Convergent Algorithms for Collaborative Filtering

(Extended Abstract)

Jon Kleinberg[*]
Department of Computer Science
Cornell University, Ithaca, NY, 14853
kleinber@cs.cornell.edu

Mark Sandler
Department of Computer Science
Cornell University, Ithaca, NY, 14853
sandler@cs.cornell.edu

ABSTRACT

A *collaborative filtering system* analyzes data on the past behavior of its users so as to make recommendations — a canonical example is the recommending of books based on prior purchases. The full potential of collaborative filtering implicitly rests on the premise that, as an increasing amount of data is collected, it should be possible to make increasingly high-quality recommendations. Despite the prevalence of this notion at an informal level, the theoretical study of such *convergent algorithms* has been quite limited.

To investigate such algorithms, we generalize a model of collaborative filtering proposed by Kumar et al., in which the recommendations made by an algorithm are compared to those of an *omniscient* algorithm that knows the hidden preferences of users. Within our generalized model, we develop a recommendation algorithm with a strong convergence property — as the amount of data increases, the quality of its recommendations approach those of the optimal omniscient algorithm. We also consider a further generalization, a *mixture model* proposed by Hofmann and Puzicha.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Clustering, Information Filtering; H.1.2 [**Information Systems**]: Models and Principles—*User/Machine systems*

General Terms

Algorithms, theory

Keywords

Mixture model, latent class models, collaborative filtering, clustering

1. INTRODUCTION

Collaborative Filtering and Convergence Properties. As individuals engage in on-line activities, they generate a large amount of usage data that — implicitly or explicitly — reflects their decisions and preferences. While the growth of this kind of data clearly raises a number of serious privacy concerns (see e.g. [8, 21]), a separate but related thread of research has argued that usage data, properly handled, can benefit individuals by allowing them to leverage the collective decisions of a large user community. A simple example, familiar to many, is the mechanism by which on-line booksellers such as Amazon.com draw on accumulated user histories to recommend related books in response to customer selections [1]. Projects such as GroupLens and EachMovie [7, 11, 17], among many others, have included a more direct approach: beyond simply observing user activities, they ask users to explicitly register their likes and dislikes, and then use the accumulation of these responses to make recommendations. The study of automated systems that draw on usage data to make targeted recommendations has grown into the field of *collaborative filtering.* (See [5, 10, 12, 17, 18, 19, 20] for early references and general surveys.)

The appealing premise underlying collaborative filtering is that, with a large enough user population, it should be possible to make high-quality recommendations in essentially a domain-independent fashion. Rather than building in specific knowledge about books, for example, a recommendation system can incorporate a notion of *similarity* among user preferences, and rely on the assumption that one should recommend to a user the things that are liked by similar users. In short, the full potential of collaborative filtering implicitly rests on a "meta-theorem" that asserts something like the following: *In the limit of infinite data, it should be possible to make arbitrarily good recommendations.*

To be able to make such a notion precise, one needs a rigorous formulation of the problem that collaborative filtering is trying to solve. Kumar et al. [14] developed a theoretical framework for studying collaborative filtering algorithms from this perspective, and they proved a number of fundamental results about the ability of such algorithms to make high-quality recommendations. Shortly afterward, Hofmann and Puzicha [13] independently formulated a much more general model, and developed a set of statistically-based algorithms within this model, without proving quantitative performance guarantees for these algorithms relative to an optimum. (For work related to Hofmann and

[*]Supported in part by a David and Lucile Packard Foundation Fellowship, an ONR Young Investigator Award, NSF ITR/IM Grant IIS-0081334, and NSF Faculty Early Career Development Award CCR-9701399.

EC'03, June 9–12, 2003, San Diego, California, USA.
Copyright 2003 ACM 1-58113-679-X/03/0006 ...$5.00.

Puzicha's model, see also [9, 16].) Both of these frameworks are based on probabilistic *generative models*, in which *hidden parameters* of the users and the items being selected give rise to the observed data through a randomized process, and an algorithm must make recommendations knowing only the observed data, not the hidden parameters. And both frameworks consider the same figure of merit in evaluating an algorithm — essentially, the probability that a user likes the item that is recommended to her, compared to the probability the user would like an item recommended by an *omniscient* algorithm that knows the hidden parameters of the system.

However, there is much we still do not understand about the following basic question. As we make the models more general, we are imposing weaker assumptions on how users behave, and so it becomes correspondingly harder to design algorithms with good performance. Thus we ask — in how general a model can one establish a precise version of the "meta-theorem" suggested above? In other words, in how general a model can we obtain a *convergent algorithm*, one whose performance converges to that of an omniscient algorithm as the amount of data increases?

This issue motivates our work here. We first propose a generalization of the model of Kumar et al., and provide a convergent algorithm for this model; the precise performance guarantee of the algorithm is described in detail below. We then explore some more general models and observe certain natural impossibility results; in particular, for the model of Hofmann and Puzicha, we show that no recommendation algorithm can achieve the type of convergent behavior we consider here.

The Basic Models. As above, we imagine a setting in which users select items — say, books — from a universe of possibilities, with each user following intrinsic preferences for certain kinds of books over others. To capture this, we define a model that consists of the following components.

- *Items, weights, clusters and users.* There is a set $\mathbf{I}$ of items, and each item i belongs to one of k disjoint *clusters* $c_1, c_2, \ldots, c_k$ (e.g. representing one of k possible genres of book). There is also a set $\mathbf{U}$ of users, and each user u has a *preference* $p(c|u) \geq 0$ for each cluster c; these are normalized so that $\sum_c p(c|u) = 1$. These preferences represent the extent to which a user u likes the items from a particular cluster (genre).

  For an item i, let $c(i)$ denote the cluster containing i. Each item i has a weight $w_i \geq 0$, intuitively reflecting its level of "appeal" to a user whose preferences are concentrated on the cluster $c(i)$. The weights are normalized so that $\sum_{i \in c} w_i = 1$.
- *User histories.* Histories of prior selections are constructed by the following generative model. Each user u selects cluster c with probability $p(c|u)$, and then selects item i from cluster c with probability w_i. This is repeated s times for each user, for a parameter s.[1]
- *Recommendations.* A recommendation algorithm is given the samples of size s selected by each user, but not any of the parameters of the system. It must then recommend one item i_u to each user u, and these recommendations yield a net *utility* of $\sum_{u \in \mathbf{U}} w_{i_u} p(c(i_u)|u)$; in other words, on each user u we get a utility equal to the probability that u chooses the item i_u. This is compared to the maximum possible utility one could accrue if knew all the parameters of the system, $\textsc{opt} = \sum_{u \in \mathbf{U}} \max_c \mathbf{w}_c p(c|u)$, where $\mathbf{w}_c = \max_{i \in c} w_i$.

We call this model the *Weighted Model*. It generalizes the *Uniform Model* of Kumar et al. [14], which is based on the same partition of items into disjoint clusters, but assumes that all items in a given cluster are equally desirable. Thus, in the terminology above, the Uniform Model assigns the same weight w_i to all the items in a given cluster, so that users select uniformly from a cluster. This special case with uniform weights introduces certain simplifying features into the recommendation problem; in particular, if we collect data from many users, and let $\#(i)$ denote the number of times that item i is selected in this dataset, then the ratio $\#(i)/\#(j)$ for items i and j in the same cluster will converge to 1 as the number of users increases. In other words, as the amount of user data increases, the hidden clusters will tend to "emerge" simply because all the items in a single cluster have the same mean.[2] But this observation appears mainly to arise from an artifact of the Uniform Model; in reality, one would not be able to identify a genre like "science fiction" from the fact that all science fiction books have an essentially identical number of sales.

Thus, if our goal is to consider algorithms that operate with a large amount of user data, it is good to have a model that can express certain additional sources of complexity that arise in collaborative filtering applications. In particular, if a science fiction book i is selected by many users, it may be because our user population contains many fans of science fiction; but it also may be because i is a very popular book in this genre, and so is reasonably likely to be selected even by users who read very little science fiction. The difficulty in resolving these two possibilities leads to much of the difficulty in estimating a user's preferences, and our generalization from the Uniform to the Weighted Model seeks to capture this issue explicitly.

Our Results. How much user data can we reasonably expect a recommendation algorithm to have access to? In formulating our results, we consider the number N of items to be fixed as the number M of users grows; this accords with the picture of a system that tries to improve its recommendations on a possibly large but given set of books, as more and more users participate. However, it is not reasonable to ask for good performance only after each user has selected a sizable fraction of all possible items; thus, while we allow the number s of selections per user to grow large relative to the number k of clusters, it should remain bounded independently of M and N.

Due to some lower bounds that we develop below, our results must depend on one further parameter. Recall that $\mathbf{w}_c = \max_{i \in c} w_i$; we define $\mathbf{w}_+ = \max_c \mathbf{w}_c$ and $\mathbf{w}_- = \min_c \mathbf{w}_c$. We then define $\mathcal{W}$ to be the ratio $\mathbf{w}_+/\mathbf{w}_-$. One can show that to obtain performance close to optimal, the number of samples per user must also depend on this quantity $\mathcal{W}$. We believe that assuming $\mathcal{W}$ to be relatively small is fairly reasonable: it simply corresponds to the assumption

[1] For now, it is not crucial whether the selection is made with or without repetition; we will return to this issue later.

[2] Note that it is possible for the items in two different clusters to all have the same mean, rendering the picture somewhat more complicated.

that each cluster contains an item of comparably high appeal to users whose preferences are concentrated on that cluster (i.e., that there is no genre consisting entirely of books that are obscure even to their fans). In this way it is much milder than assuming, for example, that each cluster contains an item whose actual weight w_i is large; and in any case, our results below show that we cannot do without $\mathcal{W}$.

We now state the main algorithmic result for the Weighted Model.

THEOREM 1.1. *There are polynomial functions p and q, and an algorithm $\mathcal{A}$, such that the following holds in the Weighted Model. If the number of users is at least $M \geq p(N, \gamma^{-1})$ and the number of selections per user is at least $s \geq q(k, \mathcal{W}, \gamma^{-1})$, then with high probability the net utility of the recommendations made by $\mathcal{A}$ is at least $(1-\gamma)$ times* OPT.

The algorithm is built from a number of intuitively natural heuristic ideas. We develop a "correlation test" that looks at the frequency with which pairs of items are selected by a common user; using this, we build a graph G by joining pairs of items that are deemed to be correlated under our definition. We show that the connected components of G correspond approximately to unions of clusters c: each component agrees with a union of clusters up to a set of items that have been selected very few times, and the clusters making up a component are "nearly indistinguishable" using the number of users and samples we have. Finally, by estimating aggregate item weights over these components, we show that the resulting recommendations are close to optimal; the crux of the analysis here is to argue that merging nearly indistinguishable clusters into a single "meta-cluster" does not greatly affect the quality of the recommendations. It is important to note that in the presentation here, we do not optimize the dependence of our bounds on the underlying parameters; this is in the interest of presenting a polynomial bound as cleanly as possible.

Although our algorithm is guaranteed to converge to near-optimal recommendations with high probability, it is not guaranteed to completely isolate all the underlying clusters c, even when given an arbitrarily large amount of data. This is a natural and inevitable consequence of the model. Indeed, consider an extreme instance in which each user u has a preference $p(c|u) = 1/k$ for each cluster c. Then the clusters are truly indistinguishable, regardless of how many users we have (since users are indifferent to the clusters); and yet it is possible to achieve near-optimality in this case by simply recommending the most heavily selected item. From this, one notices that the analysis of a recommendation algorithm must have a certain multiple-alternative flavor: either the user population is diverse enough in its preferences that the hidden clusters will be discernible, or else most users are sufficiently indifferent to the clustering that we can make good recommendations without isolating them explicitly. These considerations reflect a sense in which our goals here are quite different from what one encounters in problems with "planted structure" (e.g. [3, 4]) — while we do have an underlying generative model, it is possible to perform well even in cases where the underlying structure is provably unrecoverable.

Since the Weighted Model includes the Uniform Model, a special case of Theorem 1.1 yields a near-optimal recommendation algorithm for the Uniform Model. As we discussed above, however, if one is concerned specifically with the Uniform Model, then one can also exploit certain of its particular features to design such an algorithm directly. While Kumar et al. do not explicitly describe a convergent algorithm, it is not difficult to develop one for the Uniform Model using the analysis in Section 4 of their paper [14].[3] It is interesting that Kumar et al. restrict attention in [14] to the case of equal-sized clusters, noting that it "seems challenging" to handle unequal sizes. In the present setting, this is analogous to requiring $\mathcal{W} = 1$; thus, our results help explain quantitatively why such an assumption is important, and how to parametrize the more general cases.

To capture precisely why we cannot have the number of samples depend only on k and γ^{-1} — and hence why dependence on a parameter like $\mathcal{W}$ is necessary — we prove the following.

THEOREM 1.2. *For any functions f and g the following holds. There exists an instance of the Weighted Model with at least $f(N)$ users and at least $g(k)$ samples per user, in which no recommendation algorithm can achieve a net utility within a factor better than $O(1/k)$ times* OPT *with constant probability.*

Of course, the instances used to establish this theorem necessarily require large values of $\mathcal{W}$.

The *Mixture Model* proposed by Hofmann and Puzicha [13] can be thought of as a generalization of our Weighted Model — here, each cluster induces a probability distribution over *all* items (i.e. each item has a fractional membership in each cluster), and a user selects an item by first selecting a cluster and then selecting an item from the distribution induced by the cluster. We show that this model is essentially too strong to allow for algorithms that are convergent in our sense, even for a fixed value of $\mathcal{W}$. We discuss this topic in Section 5.

For mixture models, we also consider the problem of making recommendations when the underlying cluster structure is known, but the user preferences are not. To analyze the performance of recommendation algorithms in this setting, we introduce a further parameter Γ, which is essentially an L_1-analogue of the smallest singular value of item weights. We can show that when both $\mathcal{W}$ and Γ are bounded, it is possible to design a convergent algorithm requiring only a small amount of data from each user.

Further related work. Other recent theoretical work on collaborative filtering, specifically that of Azar et al. [2] and Drineas et al. [6], has pursued models that are not directly comparable with ours: these papers assume a latent *linear* structure (rather than probabilistic clusters), and their focus is on approximately recovering this full latent structure. Our approach, on the other hand, requires significantly less data on each user, and focuses on the task of making good recommendations regardless of whether the underlying structure can be recovered.

[3] Note that Kumar et al. explicitly give a $(1-\gamma)$-approximation for the case in which the clusters are actually known to the recommendation algorithm, rather than being hidden; this is a variant of the problem we do not consider here.

2. A RECOMMENDATION ALGORITHM FOR THE WEIGHTED MODEL

We now develop the algorithm for the Weighted Model that will prove Theorem 1.1. The given collaborative filtering system, with hidden clusters $\{c\}$, preferences $\{p(c|u)\}$, and item weights $\{w_i\}$, will be referred to as the *true system.* Our $(1-\gamma)$-approximation algorithm works in three parts: we first perform a *correlation test* on pairs of items, to build a graph G; from the connected components of G we construct an *estimated system* whose parameters approximate those of the true system; and finally we make recommendations as though the parameters of the estimated system were those of the true system.

Recall that $M = |\mathbf{U}|$ and $N = |\mathbf{I}|$. We define $\#_r(i)$ to be the number of times that item i is selected, when we consider the first $r \leq s$ selections made by each user; we write $\#(i)$ for $\#_s(i)$. We define constants $\varepsilon = \gamma/6$ and $\beta = \varepsilon_1 = (\frac{\varepsilon^2}{20^2 k^6 \mathcal{W}^2})^2$.

The Correlation Test. For this part of the algorithm only, we focus on just the first two selections made by each user; this is enough to allow us to search for correlations, and it makes the analysis cleaner. We define an item i to be *light* if $\#_2(i) < \beta M/(N^2)$; we call i *heavy* otherwise. Light items are infrequent enough that it is difficult to make estimates based on them; at the same time, we do not lose much utility by ignoring them.

For each pair of heavy items, we now apply a test to estimate whether or not they belong to the same cluster. In doing this, we look at only the first two samples selected by each user. We define $\#(i,j)$ to be the number of users whose first two samples are equal to i and j; we will also refer to this as the *multiplicity* of the pair (i,j). We define the constant $\tau = \varepsilon_1^2/(32k^2)$. We declare i and j to be *correlated* if and only if the following two conditions hold:

(a) $\#(i,j) \geq \beta^2 M/N^4$, and

(b) there do not exist items ℓ_1 and ℓ_2 such that $\#(i,\ell_1)$, $\#(j,\ell_1)$, $\#(i,\ell_2)$, $\#(j,\ell_2) \geq \beta^3 M/(2N^6)$ and

$$\left|1 - \frac{\#(i,\ell_1)}{\#(j,\ell_1)} / \frac{\#(i,\ell_2)}{\#(j,\ell_2)}\right| > \tau.$$

Constructing the Estimated System. We define a graph G on the set of items $\mathbf{I}$ by joining all pairs of correlated heavy items. Let $\overline{c}_1, \ldots, \overline{c}_t$ denote the connected components of G, and assign each light item arbitrarily to one of the components.

We define a new collaborative filtering system in which the clusters are these sets $\overline{c}_1, \ldots, \overline{c}_t$; note that they form a partition of $\mathbf{I}$. We define a new weight function $\tilde{w}$ as follows: if i belongs to cluster $\overline{c}$, we set $\tilde{w}_i = \#(i)/\left(\sum_{j \in \overline{c}} \#(j)\right)$. We define the preference of u for cluster $\overline{c}$, denoted $\tilde{p}(\overline{c}|u)$, to be the fraction of items that u selected from cluster $\overline{c}$.

Making Recommendations. We now make a recommendation to each user as though the parameters of the estimated system were the true parameters. Hence, we simply recommend to each user u the item i_u that maximizes $\tilde{w}_i \tilde{p}(\overline{c}(i)|u)$.

Overview of Analysis. Our algorithm is performing optimal recommendations with respect to an estimated system that is not the true one; thus, we must show that these recommendations are not far from optimal when evaluated under the utility function of the true system. In order to do this, we construct a third collaborative filtering system, the *ideal system*, that essentially interpolates between the true and estimated systems. Its clusters are those of the estimated system, but its item weight and user preference parameters are set by aggregating the true system's parameters over the estimated clusters. Specifically, the clusters in the ideal system are $\overline{c}_1, \ldots, \overline{c}_t$, the preference of user u for cluster $\overline{c}$ is $p^*(\overline{c}|u) = \sum_{i \in \overline{c}} w_i p(c(i)|u)$, and the weight of item i is

$$w_i^* = \frac{\sum_{u \in \mathbf{U}} w_i p(c(i)|u)}{\sum_{u \in \mathbf{U}} p^*(\overline{c}(i)|u)}.$$

The analysis is then organized as follows. Using the properties of the correlation test, we show that if heavy items from different true clusters end up in the same estimated cluster $\overline{c}$, then the true clusters containing them are approximately "indistinguishable" in a sense we define below. Using this notion of indistinguishability, we show that recommendations have approximately equal utilities in the true and ideal systems. Finally, on the assumption that each user chooses a sufficiently large number of samples relative to k, $\mathcal{W}$, and γ^{-1}, we show that the optimal recommendations in the ideal and estimated systems are approximately the same. Putting this together, we see that the optimal recommendation that we make in the estimated system has near-optimal utility in the true system.

Selection with and without Repetition. We can model users as making selections from clusters either with or without repetitions. Depending on the collaborative filtering domain, one or the other possibility may make more sense — one expects users to buy books without repetition, but to visit popular Web pages (e.g. news sites or search engines) with repetition. Our analysis here focuses on the version of the model in which users make selections by sampling *with* repetition; however the results can all be carried over to the case of sampling without repetition, provided we are careful about the following issue.

Suppose each user chooses s items without repetition, and consider a cluster with the following weights. There is one item of weight $1 - \sigma_0$, there are $s - 1$ items of weight σ_1, and there are many items of weight σ_2, where $1 \gg \sigma_0 \gg \sigma_1 \gg \sigma_2$. Suppose that all users have their preferences concentrated on this cluster. Then unless the number of users is at least a function of σ_0^{-1} (which can be arbitrarily larger than the other parameters in our bounds), all users will select precisely the s heaviest items, since they are sampling without repetition, and we will have no way to distinguish the item of weight σ_0 from the items of weight σ_1.

However, if we simply assume that the maximum weight of any item is a sufficiently small constant relative to γ, then the algorithm described above still achieves a $(1-\gamma)$-approximation with high probability. Due to space limitations, we defer the details of this to the full version of the paper. Thus, under this assumption, the same algorithm works in both the model based on sampling with repetition and the model based on sampling without repetition.

3. ANALYSIS: THE CORRELATION TEST

Let us order the users arbitrarily, as $u_1, u_2, \ldots, u_M$, and define the *selection vector* of a cluster c to be the M-dimensional vector $(\mathbf{w}_c p(c|u_1), \mathbf{w}_c p(c|u_2), \ldots, \mathbf{w}_c p(c|u_M))$. We say that two clusters are *indistinguishable* if their selection vectors are parallel. Indeed, if we have two indistinguishable clusters with selection vectors x and y, there is no way to tell from the selections of users that we don't instead have a single cluster with selection vector $x + y$.

We relax the notion of indistinguishability to an approximate version, by considering the inner products of selection vectors. We say that two clusters with selection vectors x and y are *α-indistinguishable* if $(x \cdot y)/(\|x\|\|y\|) \geq 1 - \alpha$, where $x \cdot y$ denotes the inner product of x and y, and $\|x\|$ denotes the Euclidean norm of x. Note that indistinguishability corresponds to the case $\alpha = 0$.

We will need the following basic lemma about α-indistinguishability.

LEMMA 3.1. *Let x and y be α-indistinguishable selection vectors. Let $v = (x/\|x\|) - (y/\|y\|)$. Then $\|v\| \leq \sqrt{2\alpha}$, and the L_1 norm of v (the sum of the absolute values of its coordinates) is at most $\sqrt{2\alpha M}$.*

We also define the notion of an *essential* item, which is closely related to the notion of an item being *heavy*. Let E_i denote the expected number of times item i is selected when we consider a single selection by each user; that is, $E_i = \sum_{u\in\mathbf{U}} w_i p(c(i)|u)$. We say that item i is *essential* if $E_i \geq \beta M/N^2$. Thus the expected value of $\#_2(i)$ for any essential item i is at least $2\beta M/N^2$; since this quantity is a sum of independent 0-1 random variables, standard tail inequalities imply that for M sufficiently large relative to N and β, no essential item will be selected fewer than $\beta M/N^2$ times with high probability, and so all essential items will be considered heavy by our algorithm.

LEMMA 3.2. *Define $\alpha = 2\tau$, where τ is the constant in the correlation test. Then with high probability, the following holds for all pairs of items i and j, provided that M is sufficiently large relative to N and γ.*

(i) If i and j are essential items that belong to the same cluster in the true system, then they are joined by an edge in G.

(ii) If i and j are joined by an edge in G, then they belong to α-indistinguishable clusters.

Proof. Let E_{ij} denote the expected value of $\#(i,j)$, if we get two samples from each user. If i and j come from the same cluster c, then $E_{ij} = 2w_iw_j \sum_{u\in\mathbf{U}} p(c|u)^2$. If $i \in c$ and $j \in c'$, where the clusters c and c' are distinct, then $E_{ij} = 2w_iw_j \sum_{u\in\mathbf{U}} p(c|u)p(c'|u)$. For those pairs (i,j) with $E_{ij} \geq \beta^3 M/(4N^6)$, let $\mathcal{F}_{ij}$ denote the event that E_{ij} and $\#(i,j)$ differ by a factor of at most $(1 \pm \tau/64)$. For those pairs (i,j) with $E_{ij} < \beta^3 M/(4N^6)$, let $\mathcal{F}_{ij}$ denote the event that $\#(i,j) < \beta^3 M/(2N^6)$. We write $\mathcal{F} = \cap_{i,j}\mathcal{F}_{ij}$. Our condition on M is that it be large enough so that the probability of each event $\mathcal{F}_{ij}$ is at least $1 - N^{-3}$; applying the Union Bound, it follows that $\mathcal{F}$ has probability at least $1 - N^{-1}$. In particular, if $\mathcal{F}$ occurs, then all pairs (i,j) considered by the correlation test will have the property that E_{ij} and $\#(i,j)$ differ by a factor of at most $(1 \pm \tau/64)$.

Consequently, if i and j are essential items from the same cluster c, then

$$\begin{aligned} E_{ij} &= 2w_iw_j \textstyle\sum_{u\in\mathbf{U}} p(c|u)^2 \geq \frac{2}{M} w_iw_j \left(\sum_{u\in\mathbf{U}} p(c|u)\right)^2 \\ &= \tfrac{2}{M}(E_i)(E_j) \geq \tfrac{2\beta^2 M}{N^4}. \end{aligned}$$

Thus, given that $\mathcal{F}$ occurs, we have $\#(i,j) \geq \beta^2 M/(N^4)$, as required by the correlation test. Moreover, if ℓ is any other item, then $E_{i\ell}/E_{j\ell} = w_i/w_j$. Hence if the pairs $(i,\ell_1), (j,\ell_1), (i,\ell_2), (j,\ell_2)$ all have sufficiently large multiplicity, then given $\mathcal{F}$ all are within a factor of $(1 \pm \tau/64)$ of their expectations, and so we have

$$\left|1 - \frac{\#(i,\ell_1)}{\#(j,\ell_1)} / \frac{\#(i,\ell_2)}{\#(j,\ell_2)}\right| \leq \tau.$$

It follows that the pair (i,j) passes both parts of the correlation test. This proves part (i).

For part (ii), suppose we have two heavy items $i \in c$ and $j \in c'$ such that c and c' are not α-indistinguishable. We want to show that, given $\mathcal{F}$, they will not be joined by an edge in G. Clearly, if $\#(i,j) < \beta^2 M/N^4$, then this will be the case. Otherwise, consider some other item i_1 in c and $j_1 \in c'$. By the argument from part (i), we must have $\#(i,i_1) \geq \beta^2 M/(N^4)$, since i and i_1 belong to the same cluster. Moreover, since i_1 is essential, we have $E_{i_1 j}/E_{ij} = w_{i_1}/w_j \geq \beta/N^2$, and so given $\mathcal{F}$ we have

$$\#(j, i_1) \geq \beta^3 M/(2N^6).$$

A symmetric argument applies to $\#(j,j_1)$ and $\#(j,i_1)$; thus, given $\mathcal{F}$, all four of these pairs have sufficient multiplicity to be considered in part (b) of the correlation test for (i,j).

Now,

$$E_{ii_1}/E_{ji_1} = (w_i \sum_{u\in\mathbf{U}} p(c|u)^2)/(w_j \sum_{u\in\mathbf{U}} p(c|u)p(c'|u))$$

and

$$E_{ij_1}/E_{jj_1} = (w_i \sum_{u\in\mathbf{U}} p(c|u)p(c'|u))/(w_j \sum_{u\in\mathbf{U}} p(c'|u)^2).$$

Thus, if we let x denote the selection vector c and x' denote the selection vector of c', we have

$$\begin{aligned} \frac{E_{ii_1}}{E_{ji_1}} / \frac{E_{ij_1}}{E_{jj_1}} &= \frac{\left(\sum_{u\in\mathbf{U}} p(c|u)p(c'|u)\right)^2}{\left(\sum_{u\in\mathbf{U}} p(c|u)^2\right)\left(\sum_{u\in\mathbf{U}} p(c'|u)^2\right)} = \\ &= \frac{(x \cdot x')^2}{\|x\|^2\|x'\|^2} \leq (1-\alpha)^2 < (1-\alpha), \end{aligned}$$

using the fact that c and c' are not α-indistinguishable. Since all these pairs have sufficiently large multiplicity, all their values $\#(\cdot,\cdot)$ will be within a factor of $(1 \pm \tau/64)$ of their expectations, given $\mathcal{F}$, and hence we will have

$$\left|1 - \frac{\#(i,i_1)}{\#(j,i_1)} / \frac{\#(i,j_1)}{\#(j,j_1)}\right| > \frac{\alpha}{2} = \tau.$$

It follows that the pair (i,j) will not pass the correlation test. ■

Now, when we consider the connected components of G, we see that all the essential items from any true cluster belong to a single component. Other pairs of items joined by an edge come from α-indistinguishable clusters; but since α-indistinguishability is not transitive, we cannot immediately draw a similar conclusion for essential items that belong to the same component. For this, we need the following lemma.

LEMMA 3.3. *Suppose we have r clusters $c_1 \dots c_r$, and cluster c_i is ε_i^2 indistinguishable from c_{i+1}, where $\varepsilon_i \le 1/r$. Then c_1 is $8\left(\sum \varepsilon_i\right)^2$-indistinguishable from c_r.*

If essential items i and j belong to the same component of G, then there is an i-j path in G such that the items on the path change cluster membership at most $k-1$ times. This, together with Lemmas 3.2 and 3.3, gives us the second statement in the following theorem.

THEOREM 3.4. *If two essential items belong to the same cluster, then they belong to the same component of G. If two essential items belong to the same component of G, then they belong to clusters that are $(8k^2\alpha)$-indistinguishable.*

4. ANALYSIS: THE IDEAL SYSTEM

In relating the quality of our recommendation to the optimal utility in the true system, there are a number of sources of error to bound — both in the fact that the components of G do not really correspond to the clusters, and in the fact that we only have a bounded number of samples for each user. To make this process more tractable, we now make use of the *ideal system* defined at the end of Section 2.

A recommendation r to the full population of users can be viewed as a vector of items, one item r_u corresponding to the individual recommendation for each user u. Let $\Lambda(r)$ denote the utility of the recommendation vector r, evaluated according to the parameters of the true system. Let $\Lambda^*(r)$ denote the utility of r evaluated according to the parameters of the ideal system. The following theorem is the second main step in the analysis, showing that we do not lose much through the merging of clusters implicit in the construction of G. It is crucial that this theorem applies not just to optimal recommendation vectors, as we will be using it with r equal to our algorithm's recommendation, which is not necessarily optimal for either system.

THEOREM 4.1. *For every recommendation vector r, we have*

$$|\Lambda(r) - \Lambda^*(r)| \le \varepsilon \cdot \text{OPT}. \qquad (1)$$

Proof. First we outline the proof and then will go into details.

Recall that E_i denotes the expected number of times item i is selected in the true system when we consider a single selection by each user; that is, $E_i = \sum_{u\in\mathbf{U}} w_i p(c(i)|u)$. We define the corresponding quantity $E_i^* = \sum_{u\in\mathbf{U}} w_i^* p^*(\overline{c}(i)|u)$, and observe that by our construction of the ideal system, we have

$$\begin{aligned} E_i^* &= w_i^* \sum_{u\in\mathbf{U}} p^*(\overline{c}(i)|u) \\ &= \frac{\sum_{u\in\mathbf{U}} w_i p(c(i)|u)}{\sum_{u\in\mathbf{U}} p^*(\overline{c}(i)|u)} \times \sum_{u\in\mathbf{U}} p^*(\overline{c}(i)|u) \\ &= \sum_{u\in\mathbf{U}} w_i p(c(i)|u) = E_i \end{aligned}$$

for every item i. Thus all items have the same expectation in both systems.

For a recommendation vector r, let r_u denote the item recommended to user u. We show that

$$\sum_{u\in\mathbf{U}} |w_{r_u} p(c(r_u)|u) - w_{r_u}^* p^*(c(r_u)|u)| \le \varepsilon\text{OPT}, \qquad (2)$$

which is sufficient to establish the theorem. The left-hand side of (2) is a sum of M terms, and to bound this we divide the users into three sets: those users for whom r_u is an inessential item; those users for whom r_u belongs to a cluster c in the true system with very low total preference (satisfying $\sum_{u\in\mathbf{U}} p(c|u) \le \frac{1}{8\mathcal{W}k^2}$); and all other users. The contribution of the terms associated with the first two kinds of users can be bounded using the fact that the corresponding items do not represent very much utility.

Now, for each cluster c in the true system, we know by Theorem 3.4 that there is a cluster $\overline{c}$ in the ideal system containing all the essential items of c; we denote this by $c \sqsubseteq \overline{c}$. Consider a user u of the third kind, who was recommended an essential item i belonging to cluster c in the true system, such that $c \sqsubseteq \overline{c}$ for a cluster $\overline{c}$ in the ideal system. Even if we are just interested in this single user, we need to estimate the sum $\sum_u p^*(\overline{c}|u)$ since it appears in the definition of the item weight w_i^*. We do this by considering all true clusters $c_1, \dots, c_b \sqsubseteq \overline{c}$; using Lemma 3.1 and the fact that all these clusters are approximately indistinguishable (by Theorem 3.4) we show how to approximate $\sum_u p^*(\overline{c}|u)$ by the sum of $p(c_a|u)$ over these clusters ($a = 1, 2, \dots, b$), scaled by the lengths of their selection vectors. This will not be possible for all users; but we can show that the set of users to which our approximation does not apply also represents a small contribution to the left-hand side of (2).

Now let us go into details of the proof. Again, we say that a cluster c in the true system is *essentially* a subset of a cluster $\overline{c}$ in the ideal system if every essential element of c belongs to $\overline{c}$, and we denote this by $c \sqsubseteq \overline{c}$. By Theorem 3.4, every cluster c in the true system is essentially a subset of some $\overline{c}$.

By the definition of utility we can rewrite (1) as

$$|\sum_{u\in\mathbf{U}} w_{r_u} p(c(r_u)|u) - w_{r_u}^* p^*(c(r_u)|u)| \le \varepsilon\text{OPT}.$$

It is sufficient to prove the following:

$$\sum_{u\in\mathbf{U}} |w_{r_u} p(c(r_u)|u) - w_{r_u}^* p^*(c(r_u)|u)| \le \varepsilon\text{OPT}. \qquad (3)$$

First we consider users who were recommended inessential items. Let for item i, let X_i denote the set of users u for whom $r_u = i$, and let X denote the union of X_i over all inessential items i. If i is an inessential item, then we have

$$\begin{aligned} &\sum_{u\in X_i} |w_i p(c(i)|u) - w_i^* p^*(c(i)|u)| \\ &\qquad \le \sum_{u\in X_i} |w_i p(c(i)|u)| + \sum_{u\in X_i} |w_i^* p^*(c(i)|u)| \\ &\qquad \le 2\sum_{u\in\mathbf{U}} |w_i p(c(i)|u)| \le 2\frac{\varepsilon_1}{N^2} M \end{aligned}$$

and since there are at most N inessential items, it follows that users who were recommended inessential items can contribute to the left-hand side of (3) at most

$$2N\frac{\varepsilon_1}{N^2} M \le \frac{\varepsilon}{4}\text{OPT}.$$

Now we focus on users who were recommended essential items. We partition these users into two types: those who were recommended an item from a cluster c satisfying

$$\sum_{u\in\mathbf{U}} p(c|u) \le \frac{1}{8\mathcal{W}k^2},$$

and all others. Let X' denote the set of users of the first type. Users in X' can be handled by means analogous to what we used above for users who were recommended inessential items. Indeed, let U_c denote the set of users who were recommended an item from cluster c, where c is essentially a subset of $\overline{c}$. Then we have

$$\sum_{u \in U_c} w^*_{r_u} p^*(\overline{c}|u) \leq \sum_{u \in U_c} w^*_l p^*(\overline{c}|u)$$

where l is the heaviest item in cluster $\overline{c}$. Thus

$$\begin{array}{rcl} \sum_{u \in U_c} w^*_l p^*(\overline{c}|u) & \leq & \sum_{u \in U} w^*_l p^*(\overline{c}|u) = \sum_{u \in U} w_l p(c|u) \\ & \leq & \mathbf{w}_+ \sum_{u \in U} p(c|u) \leq \dfrac{\varepsilon \mathbf{w}_+}{8k^2 \mathcal{W}} M \\ & \leq & \frac{\varepsilon}{8k} \text{OPT}, \end{array}$$

and by similar reasoning we also have $\sum_{u \in U_c} w_{r_u} p(c|u) \leq \frac{\varepsilon}{8k}\text{OPT}$. Since there are k clusters, users of this type can contribute to the left-hand side of (3) at most $2\frac{\varepsilon}{8}\text{OPT} = \frac{\varepsilon}{4}\text{OPT}$.

Let Y denote the set of users of the second type; this consists of all users not considered so far. We show that for all but a tiny fraction of them, the corresponding terms in the left-hand side of (3) are less then $\frac{\varepsilon}{4M}\text{OPT}$; and the tiny fraction for which this fails to hold is small enough that it cannot contribute much.

Consider a user $u \in Y$ for whom $i = r_u \in c_1 \sqsubseteq \overline{c}$. Let x denote the selection vector of c_1. The idea of the remainder of the proof is to express $p^*(\overline{c}|u)w_i$ in terms of $p(c|u)w_i$, or to show that this user is from the tiny fraction mentioned above. We divide the analysis into three parts.

1. First, we want to express $p^*(\overline{c}|u)$ in terms of the original preferences. By definition we have:

$$p^*(\overline{c}|u) = \sum_{i \in \overline{c}} p(c|u) w_i$$

Since each cluster $\overline{c}$ is just a union of true clusters, up to inessential items, we have

$$\sum_{c \sqsubseteq \overline{c}} p(c|u) - \varepsilon_1/N \leq p^*(\overline{c}|u) \leq \sum_{c \sqsubseteq \overline{c}} p(c|u) + \varepsilon_1/N, \qquad (4)$$

and thus

$$\sum_{u \in \mathbf{U}} \sum_{c \sqsubseteq \overline{c}} p(c|u) - \frac{\varepsilon_1 M}{N} \leq \sum_{u \in \mathbf{U}} p^*(\overline{c}|u) \leq \sum_{u \in \mathbf{U}} \sum_{c \sqsubseteq \overline{c}} p(c|u) + \frac{\varepsilon_1 M}{N} \qquad (5)$$

Since the remainder of the proof involves a number of inequalities similar to the above, we introduce a more compact notation. For real numbers x, a, and ε, we write $x \in a \pm \varepsilon$ to denote the pair of inequalities $a - \varepsilon \leq x \leq a + \varepsilon$ More generally, we use this notation with more than one $\pm$ to indicate that the left-hand side lies between the smallest and largest one can obtain by substituting $+$ or $-$ for each occurrence of $\pm$ on the right-hand side.

2. We now want to express $p^*(\overline{c}|u)$ in terms of $p(c_1|u)$ and some global characteristics of other subclusters of $\overline{c}$. This will not be possible for all users, but we will show that it is possible for almost all.

All clusters in $\overline{c}$ are $\frac{1}{2}\varepsilon_1^2$-indistinguishable; therefore by Lemma 3.1, any cluster $c' \sqsubseteq \overline{c}$ with selection vector v satisfies

$$\sum |v_i - \frac{|v|}{|x|} x_i| \leq \sqrt{\varepsilon_1^2 M} |v|. \qquad (6)$$

Let $|c| = \sqrt{\sum_{u \in \mathbf{U}} [p(c|u)]^2}$ and $q_c = \sum_{u \in \mathbf{U}} p(c|u)$. Now, by the definition of the selection vector, we have

$$|q_{c'} - \frac{|c'|}{|c_1|} q_{c_1}| \leq \sqrt{\varepsilon_1^2 M} |c'|$$

and therefore

$$q_{c'} \;\in\; \frac{|c'|}{|c_1|} q_c \pm \varepsilon_1 \sqrt{M} |c'|.$$

Substituting this into (5) we have

$$\sum_{u \in \mathbf{U}} p^*(\overline{c}|u) \;\in\; \sum_{c' \sqsubseteq \overline{c}} q_{c_1} \left[\frac{|c'|}{|c_1|} \pm \frac{\varepsilon_1 \sqrt{M} |c'|}{q_{c_1}} \right] \pm \frac{\varepsilon M}{N} \qquad (7)$$

By using (6) again for any cluster c' in $\overline{c}$ we have that at most $\sqrt{\varepsilon_1} M$ users do not satisfy the following constraint:

$$|p(c'|u) - \frac{|c'|}{|c_1|} p(c_1|u)| \leq \sqrt{\varepsilon_1}. \qquad (8)$$

Thus the total number of users who do not satisfy (8) for *some* cluster in $\overline{c}$ is at most $k\sqrt{\varepsilon_1} M$, and hence there are at most $k^2 \sqrt{\varepsilon_1} M$ users in (3) for whom (8) is not satisfied for their cluster c_1 and some other cluster c'. Clearly the maximum amount of utility these users can contribute is at most

$$k^2 \sqrt{\varepsilon_1} M \mathbf{w}_+ \leq \frac{\varepsilon}{4k} \mathbf{w}_- M \leq \frac{\varepsilon}{4} \text{OPT}. \qquad (9)$$

For the rest of the users, the constraint (8) applies for all essential subclusters of their new cluster, and thus

$$\frac{|c'|}{|c_1|} p(c_1|u) - \sqrt{\varepsilon_1} \leq p(c'|u) \leq \frac{|c'|}{|c_1|} p(c_1|u) + \sqrt{\varepsilon_1}.$$

Combining this with (4) we have

$$p(\overline{c}|u) \;\in\; p(c_1|u) [\sum_{c' \sqsubseteq \overline{c}} \frac{|c'|}{|c_1|}] \pm k\sqrt{\varepsilon_1} \pm \frac{\varepsilon_1}{N}, \qquad (10)$$

3. Now we are ready to estimate $w^*_i p^*(\overline{c}|u)$. We have

$$w^*_i p^*(\overline{c}|u) = p^*(\overline{c}|u) \frac{w_i q_{c_1}}{\sum_{u \in \mathbf{U}} p^*(\overline{c}|u)}$$

and substituting here (10) and (7) we have

$$w^*_i p^*(\overline{c}|u) \;\in\; p(c_1|u) \frac{\left[\sum_{c' \sqsubseteq \overline{c}} \dfrac{|c'|}{|c_1|} \pm \dfrac{k\sqrt{\varepsilon_1} + \frac{\varepsilon_1}{N}}{p(c_1|u)} \right] w_i q_{c_1}}{q_{c_1} \left[\sum_{c' \sqsubseteq \overline{c}} (\dfrac{|c'|}{|c_1|} \pm \dfrac{\varepsilon_1 \sqrt{M} |c'|}{q_{c_1}}) \pm \dfrac{\varepsilon_1 M}{N q_{c_1}} \right]}.$$

Using here the fact that $q_c \geq \frac{\varepsilon}{8\mathcal{W}k^3} M$ and $|c_1| < \sqrt{M}$, and $\varepsilon_1 \leq \frac{\varepsilon^2}{64\mathcal{W}^2 k^4}$ we have

$$w^*_i p^*(\overline{c}|u) \;\in\; \frac{w_i p(c_1|u) \left[\sum_{c' \sqsubseteq \overline{c}} \dfrac{|c'|}{|c_1|} \pm \dfrac{\varepsilon}{32k\mathcal{W} p(c_1|u)} \right]}{\left[\sum_{c' \sqsubseteq \overline{c}} \dfrac{|c'|}{|c_1|} (1 \pm \dfrac{\varepsilon}{8\mathcal{W}k}) \pm \dfrac{\varepsilon}{8\mathcal{W}k} \right]}$$

and by introducing $\gamma = \sum_{c' \sqsubseteq \overline{c}} \frac{|c'|}{|c_1|} \geq 1$, we can rewrite this as

$$w_i^* p^*(\overline{c}|u) \;\in\; w_i p(c_1|u) \frac{\left[1 \pm \dfrac{\varepsilon}{32\gamma k \mathcal{W} p(c_1|u)}\right]}{\left[1 \pm \dfrac{\varepsilon}{8\mathcal{W}k} \pm \dfrac{\varepsilon}{8\gamma\mathcal{W}k}\right]}.$$

This interval can be bounded (assuming $\varepsilon << 1$) by

$$w_i p(c_1|u)(1 \pm \frac{\varepsilon}{6\mathcal{W}k}) \pm \frac{\varepsilon w_i}{32k\mathcal{W}}(1 \pm \frac{\varepsilon}{8\mathcal{W}k})$$

and bounding w_i by $\mathbf{w}_+$ and $p(c_1|u)$ by 1, we have

$$w_i^* p^*(\overline{c}|u) \;\in\; w_i p(c_1|u) \pm \frac{\varepsilon \mathbf{w}_+}{4\mathcal{W}k}.$$

Using the fact that $OPT \geq \frac{M\mathbf{w}_-}{k} = \frac{M\mathbf{w}_-}{\mathcal{W}k}$, we immediately have

$$w_i^* p^*(\overline{c}|u) \;\in\; w_i p(c_1|u) \pm \frac{\text{OPT}}{4M},$$

and combining this with (9) we have

$$\sum_{u \in Y} |p^*(\overline{c}(r_u)|u) w^*_{r_u} - p(c(r_u)|u) w_{r_u}| \leq \frac{\varepsilon}{2}\text{OPT}.$$

Combining the contributions from users in the sets X, X', and Y, we see that the left-hand side of (3) is at most $(\frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2})\text{OPT} = \varepsilon\text{OPT}$. ■

Now, for a recommendation vector r, let $\tilde{\Lambda}(r)$ denote the utility of r evaluated according to the parameters of the estimated system. (Recall that the recommendation returned by our algorithm is optimal for the estimated system.) Relating $\Lambda^*(r)$ and $\tilde{\Lambda}(r)$ is much more straightforward, since both involve the same clusters; one must simply argue that with enough samples per user, we can determine the item i maximizing $w_i^* p^*(c(i)|u)$ to within a close approximation, for all users except a small fraction. The analysis here is similar to Theorem 4 of Kumar et al. [14], adapted to the case of the Weighted Model.

LEMMA 4.2. *Suppose there are at least s selections per user, where $s = \kappa k \mathcal{W} \varphi^{-2} \log(k\varphi^{-1})$ for a constant κ and for $\varphi = \varepsilon/(4k\mathcal{W})$. If r denotes an optimal recommendation vector for either the ideal or estimated system, then with high probability we have $(1-\varepsilon)\Lambda^*(r) \leq \tilde{\Lambda}(r) \leq (1+\varepsilon)\Lambda^*(r)$.*

We can now put all these bounds together to prove the main result.

Proof of Theorem 1.1. Let r, r^*, and $\tilde{r}$ be recommendation vectors such that r is optimal for the true system, r^* is optimal for the ideal system, and $\tilde{r}$ is optimal for the estimated system. Then by Theorem 4.1 and the optimality of r^*, we have $\Lambda(r) \leq \Lambda^*(r) + \varepsilon \cdot \text{OPT} \leq \Lambda^*(r^*) + \varepsilon \cdot \text{OPT}$. Applying Lemma 4.2 and the optimality of $\tilde{r}$, we have $\Lambda^*(r^*) \leq (1{+}\varepsilon)\tilde{\Lambda}(r^*) \leq (1{+}\varepsilon)\tilde{\Lambda}(\tilde{r}) \leq (1{+}\varepsilon)^2\Lambda^*(\tilde{r})$. Finally, Theorem 4.1 again shows that $\Lambda^*(\tilde{r}) \leq \Lambda(\tilde{r}) + \varepsilon \cdot \text{OPT}$.

Combining these bounds, we have $\Lambda(r) \leq (1+\varepsilon)^2\Lambda(\tilde{r}) + 2\varepsilon{\cdot}\text{OPT}$. Since $\text{OPT} = \Lambda(r)$, we have $\Lambda(\tilde{r}) \geq \frac{(1-2\varepsilon)}{(1+\varepsilon)^2}\Lambda(r)$; since ε is small enough so that $\frac{(1-2\varepsilon)}{(1+\varepsilon)^2} \geq 1 - \gamma$, this completes the proof. ■

5. PARAMETRIZATIONS FOR DIFFERENT MODELS

The parameter $\mathcal{W}$. We now give a family of examples that establishes Theorem 1.2, and shows why the parameter $\mathcal{W}$ is needed in our sample bounds. The examples will actually be instances in which all items in a given cluster have the same weight, and thus they help explain why restricting to clusters of equal size in the Uniform Model (as in [14]) can be crucial.

We choose large quantities b and x; their relation to the other parameters of the system will be established below. Let $c_1, c_2, \ldots, c_k$ be the k clusters; let c_a have size x, for $a \leq k-1$, and let c_k have size x^3. All the items in each cluster have equal weight. We partition the set of users into k *groups* $\mathbf{U}_1, \ldots, \mathbf{U}_k$, where $\mathbf{U}_a$ has size bx^2 for $i \leq k-1$, and $\mathbf{U}_k$ has size bx^3. Users $u \in \mathbf{U}_a$, for $a = 1, 2, \ldots, k-1$ have preferences that place probability mass x^{-1} on items in cluster c_a, and probability mass $1 - x^{-1}$ on items in cluster c_k. Users $u \in \mathbf{U}_k$ have preferences that place probability mass 1 on items in cluster c_k. We can make b as large as we want, and hence make the number of users arbitrarily larger than the number of items.

Now, if we were to recommend an item from cluster c_a to each user from group $\mathbf{U}_a$, we obtain a utility of $(k-1)bx^2(x^{-2}) + bx^3(x^{-3}) = kb$. But if each user selects only $g(k)$ items, and if x is large enough relative to $g(k)$, then with high probability at most $O(kbx)$ users will select any item from the set $c_1 \cup \cdots \cup c_{k-1}$, and we can obtain a utility of at most $O(kbxx^{-2}) = o(b)$ from them. For the remainder, we see only samples from c_k, and it is easy to show that no algorithm will achieve a utility better than $b + o(b)$ on this set with high probability. Hence, no algorithm can perform better than $O(1/k)$ times OPT with constant probability.

We observe that this example is tight, in view of the following result.

PROPOSITION 5.1. *If the number of users is sufficiently large relative to the number of items, then the algorithm that simply recommends to all users the item that has been selected the most times achieves a utility that is $\Omega(1/k)$ times* OPT *with high probability.*

The mixture model. Hofmann and Puzicha [13] proposed a *Mixture Model* for collaborative filtering that can be viewed as the following generalization of our Weighted Model. There are clusters $c_1, \ldots, c_k$, and user preferences over clusters, as before. But instead of items being partitioned into clusters, each cluster now induces a distribution over the set of all items. Let $w(i|c)$ denote the probability mass placed on item i by cluster c. To select an item, a user u first chooses a cluster c with probability $p(c|u)$, and then chooses an item i with probability $w(i|c)$. As before, the utility of recommending i to u is the probability that u would select i; but since u can now select i through any cluster, this probability is computed as $\sum_c w(i|c)p(c|u)$.

This model has the appealing feature that items can belong to multiple genres. However it is general enough so that parameterization based on k and $\mathcal{W}$ is not sufficient. We can show that there are instances of this model, where natural analogue of the $\mathcal{W}$ parameter is equal to 1; and yet no convergent algorithm is possible. One can naturally define $\mathcal{W}$

for this model to be the maximum ratio of the utility one can obtain from a user whose preferences are completely concentrated on cluster i to the utility one can obtain from a user whose preferences are completely concentrated on cluster j, over all i and j.

THEOREM 5.2. *For any functions f and g the following holds. There exists an instance of the Mixture Model with at least $f(N)$ users and at least $g(k)$ samples per user, in which no recommendation algorithm can achieve a net utility within a factor better than $O(1/k)$ times* OPT *with constant probability. This holds even when it is possible to obtain the same utility from every user, regardless of her preference vector (and hence the analogue of $\mathcal{W}$ is equal to 1).*

Proof. Suppose that the set of items consists of k *special items* $i_1, \ldots, i_k$, and a very large number x of *standard items.* For a very small number $\sigma > 0$, cluster c_a places probability mass σ on special item i_a, probability mass 0 on each other special item, and probability mass σ^2 on each standard item. The users are divided into k groups of equal size; the users in group a place all their preference weight on cluster c_a.

Now, if we recommend the special item i_a to each user of group a, we obtain a utility of σ on each user. However, if each user selects $g(k) \geq k$ items, and if σ is much smaller than $1/g(k)$, then a very small fraction of all users will select any special item. For the rest, we see only a sample of standard items; and it is easy to show that on these users, any algorithm will obtain utility at most $O(\sigma/k)$ per user with high probability. (Indeed, the strategy of recommending a random special item to each of these users achieves (σ/k) per user, and the strategy of recommending a standard item is worse since $\sigma^2 < \sigma/k$.) ■

One interesting and important point about this construction is that the impossibility holds even for algorithms that know the underlying cluster structure $w(i|c)$ of the mixture model, but not the user preferences.

Now, the following natural question arises. Suppose we know $w(i|c)$ but not the user preferences; can we find a parametrization of the general Mixture Model under which we can construct a convergent algorithm? In other words, if the value of such a characterizing parameter remain fixed as the amount of data per user grows, is there an algorithm whose performance is $(1-\gamma)$ times optimal, for γ converging to 0? We have recently been able to show that such a parametrization is indeed possible, as follows.

We use two parameters: $\mathcal{W}$ as defined above, and a new parameter Γ, which is defined as follows. Let $\mathbf{W}$ denote the matrix of mixture model parameters, with (i, c) entry equal to the weight value $w(i|c)$, and set

$$\Gamma = \min_{\|x\|_1 \neq 0} \frac{\|\mathbf{W}x\|_1}{\|x\|_1}$$

Thus, Γ can be viewed as an L_1-analogue of the smallest singular value of $\mathbf{W}$ (since the standard singular value σ_{min} would be obtained by replacing the L_1 norm with the L_2 norm).

For $\Gamma > 0$, we can show that there is a polynomial function f and an algorithm that, provided with at least $f(\gamma, \Gamma, \mathcal{W}, k, \delta)$ samples per user yields recommendations with net utility at least $(1-\gamma)$OPT with probability at least $1-\delta$. Further, we can show that some lower bound $\Gamma > 0$ is in a sense necessary for making good recommendations; indeed, there are families of instances of the 2-cluster Mixture Model for which $\Gamma \to 0$, and no algorithm can produce better than a $(\frac{1}{2} - o(1))$-approximation with constant probability.[4]

We note that this parametrization in terms of Γ is much stronger than the corresponding analysis in terms of the standard L_2 minimum singular value σ_{min}. Indeed, for making good recommendations, it is sufficient to have a lower bound on σ_{min}; however, we note that $\Gamma = 1$ for any instance of the Weighted Model, while there are instances of the Weighted Model in which σ_{min} is arbitrarily small.

Whether it is possible to obtain convergent algorithms when the cluster parameters of the Mixture Model are unknown, subject to a lower bound on Γ, remains an interesting open question.

6. REFERENCES

[1] Amazon.com, *Amazon.com Services: Recommendations*, at http://www.amazon.com/exec/obidos/tg/browse/-/help/508506/ref=pd_ir_how_nav/.

[2] Y. Azar, A. Fiat, A. Karlin, F. McSherry, J. Saia "Spectral analysis of data" *Proc. ACM STOC*, 2000

[3] A. Blum, J. Spencer, "Coloring random and semi-random k-colorable graphs." *J. Algorithms*, 19(1995).

[4] R. Boppana, "Eigenvalues and graph bisection: An average-case analysis," *Proc. IEEE Symp. on Foundations of Computer Science*, 1987.

[5] J.S. Breese, D. Heckerman, C. Kadie. "Empirical analysis of predictive algorithms for collaborative filtering," *Proc. Conf. on Uncertainty in Artificial Intelligence*, 1998.

[6] P. Drineas, I. Kerendis, P. Raghavan "Competive recommendation systems", *Proc. ACM STOC*, 2002

[7] EachMovie collaborative filtering data set, at http://www.research.compaq.com/SRC/eachmovie/

[8] Federal Trade Commission, *Privacy Online: A Report to Congress*, 1998.

[9] L. Getoor, M. Sahami, "Using Probabilistic Relational Models for Collaborative Filtering," *WebKDD'99*, 1999.

[10] D. Goldberg, D. Nichols, B. Oki, D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM* 35(1992).

[11] GroupLens home page, http://www.cs.umn.edu/Research/GroupLens/index.html

[12] W. Hill, L. Stead, M. Rosenstein, G. Furnas, "Recommending And Evaluating Choices In A Virtual Community Of Use," *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems*, 1995.

[13] T. Hofmann, J. Puzicha, "Latent Class Models for Collaborative Filtering," *Proc. International Joint Conference in Artificial Intelligence*, 1999.

[14] S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Recommendation systems: A probabilistic analysis," *Proc. IEEE FOCS*, 1998.

[4] In fact this factor of $\frac{1}{2}$ is a tight bound for $k = 2$, since a simple greedy algorithm yields a $\frac{1}{k}$-approximation for arbitrary instances of the k-cluster Mixture Model.

[15] R. Motwani, P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[16] A. Popescul, L. Ungar, D. Pennock, S. Lawrence, “Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments,” *Proc. Conf. on Uncertainty in Artificial Intelligence*, 2001.

[17] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, “GroupLens: An open architecture for collaborative filtering of Netnews,” *Proc. ACM Conf. on CSCW*, 1994.

[18] P. Resnick, H. Varian, “Recommender systems,” *Communications of the ACM*, 40(1997) 56-58. Introduction to special issue on recommender systems.

[19] Upendra Shardanand, Pattie Maes, Social Information Filtering: Algorithms for Automating “Word of Mouth,” *Proc. ACM SIGCHI Conf. on Human Factors in Computing Systems*, 1995, 210-217.

[20] ACM SIGGROUP, Resources on Collaborative Filtering, at http://www.acm.org/siggroup/collab.html.

[21] *Stanford Legal Review, 52* (July 2000), special issue on privacy.