No-regret learning for distributed social recommender systems - Online Appendix

Cem Tekin*, *Member, IEEE*, Simpson Zhang, Mihaela van der Schaar, *Fellow, IEEE*
Electrical Engineering Department, University of California, Los Angeles
Email: cmtkn@ucla.edu, mihaela@ee.ucla.edu

Abstract

In this paper, we consider decentralized sequential decision making in distributed online recommender systems, where items are recommended to users based on their search query as well as their specific background including history of bought items, gender and age, all of which comprise the context information of the user. In contrast to centralized recommender systems, in decentralized recommender systems each seller/learner only has access to the inventory of items and user information for its own products and not the products and user information of other sellers, but can get commission if it sells an item of another seller. We formulate this problem as a cooperative contextual bandit problem, analytically bound the performance of the sellers compared to the best recommendation strategy given the complete realization of user arrivals and the inventory of items, as well as the context-dependent purchase probabilities of each item, and verify our results via numerical examples on a distributed data set adapted based on Amazon data.

Index Terms

Recommender systems, social networks, online learning, cooperative learning, contextual bandits.

I. Introduction

One of the most powerful benefits of a social network is the ability for cooperation and coordination on a large scale over a wide range of different agents [1]. For example, companies can collaborate to sell products, charities can work together to raise money, and a group of workers can help each other search for jobs. Through such cooperation, agents are able to attain much greater rewards than would be possible individually. We analyze a group of agents that are connected together via a fixed network, each of whom experiences inflows of users to its page. Each time a user arrives, an agent chooses from among a set of items to offer to that user, and the user will either reject or accept each item. These items can represent a variety of things, from a good that the agent is trying to sell to a cause that the agent is trying to promote. In each application, the action of accepting or rejecting by the user will likewise have a distinct meaning. When choosing among the items to offer, the agent is uncertain about the user's acceptance probability of each item, but the agent is able to observe specific background information about the user, such as the user's gender, location, age, etc. Users with different backgrounds will have different probabilities of accepting each item, and so the agent must learn this probability over time by making different offers.

We allow for cooperation in this network by letting each agent recommend items of neighboring agents to incoming users, in addition to its own items using commissions as an incentive. When defined appropriately, this commission ensures that both sides will benefit each time a recommendation occurs and thus is able to sustain cooperation. However, since agents are decentralized, they do not directly share the information that they learn over time about user preferences for their own items. Thus agents must learn about their neighbor's acceptance probabilities through their own trial and error, unlike in other social learning papers such as [2]–[5], where agents share information directly with their neighbors.

Another key feature of our algorithm is that it is non-Bayesian unlike [2], [3]. Instead we model the learning through contextual bandits, where the context is based on the user's background. We produce a class of mechanisms that allows agents to take near-optimal actions even with decentralized learning. We prove specific bounds for the regret, which is the difference between the total expected reward of an agent using a learning algorithm and the total expected reward of the optimal policy for the agent, which is computed given perfect knowledge about acceptance probabilities for each context. We show that the regret is sublinear in time in all cases, which implies that time-averaged regret goes to 0, hence our algorithm has no-regret.

Table I provides a summary of how our work is related to other work. Of note, there are several papers that also use a similar multi-armed bandit framework for recommendations [6], [7]. Apart from these, collaborative filtering algorithms such as [8]–[16] make recommendations by predicting the user's preferences based on a similarity measure with other users. Items with the highest similarity score are then recommended to each user; for instance items may be ranked based on the number of purchases by similar users. There are numerous ways to perform the similarity groupings, such as the cluster model in [10], [13] that groups users together with a set of like-minded users and then makes recommendations based on what the users in this set choose. An important difference to keep in mind is that the recommendation systems in other works are a single centralized system, such as Amazon or Netflix. However, in this paper each agent is in effect its own separate recommendation system, since agents do not directly share information with each other. Therefore the mechanism we propose must be applied separately by every agent in the system based on that agent's history of user acceptances.

II. Problem Formulation

There are M decentralized agents/learners which are indexed by the set $\mathcal{M} := \{1, 2, \ldots, M\}$. Each agent i has an inventory of items denoted by $\mathcal{F}_i$, which it can offer to its users and the users of other agents when requested by these agents. Let $\mathcal{F} := \cup_{i \in \mathcal{M}} \mathcal{F}_i$ be the set of items of all agents. We assume that there is an unlimited supply of each type of item. This assumption holds for digital goods such as e-books, movies, videos, songs, photos, etc. An agent does not know the inventory of items of the other agents but knows an upper bound on $|\mathcal{F}_j|$[1], $j \in \mathcal{M}$ which is equal to $F_{\max}$. Let $\mathcal{K}_i = \mathcal{F}_i \cup \mathcal{M}_{-i}$ be the set of *options* of agent i. At each time step $t = 1, 2, \ldots$, a user with a specific search query indicating the type of item the user wants, or other information (price-range, age, gender etc.), arrives to agent i. We define all the properties of the arriving user known to agent i at time t as the context

[1]For a set A, $|A|$ denotes its cardinality.

	Item-based (IB), user-based (UB)	Memory-based, model-based	Uses context info.	Performance measure	Similarity distance	Centralized(C), Decentralized(D)
[17]	UB	Memory-based	No	Ranking precision	-	C
[8]	UB	Bayesian-based latent semantic model	No	MAE, RMS, 0/1 loss	Pearson correlation	C
[9]	UB	Bayesian-based Markov model	No	Precision& Recall	Pearson correlation	C
[10]	IB	Cluster model	No	-	Cosine	C
[11]	UB	Memory-based	Yes	Precision& Recall	-	C
[12]	UB	Bayesian classifier model	No	Precision& Recall	Pearson correlation	C
[13]	UB	Cluster model	No	MAE& Coverage	Pearson correlation	C
[14]	UB	MDP model	No	Recall	Self-defined similarity	C
[6]	UB	MAB model	No	Reward	Lipschitz continuous	C
[7]	UB	MAB model	Yes	Regret	Lipschitz continuous	C
Our work	UB	MAB model	Yes	Regret	Lipschitz continuous	D

TABLE I

COMPARISON WITH WORKS IN RECOMMENDER SYSTEMS.

of that user, and denote it by $x_i(t)$. We assume that the contexts of all users belong to a known space $\mathcal{X}$, which without loss of generality is taken to be $[0,1]^d$, where d is the dimension of the context space. Our results in this paper will hold without any assumptions on the context arrivals.[2] In order to incentivize the agents to recommend each other's items, they will provide commissions. These commissions are fixed at the beginning and do not change over time. The system model is shown in Fig. 1. When there is sales commission, if agent i recommends an item f_j of agent j to its user, and if that user buys the item of agent j, then agent i obtains a fixed commission which is equal to $c_{i,j} > 0$[3].

Agent i recommends N (fixed) items to its user at each time step. For example, N can be the number of recommendation slots the agent has on its website, or it can be the number of ads it can place in a magazine. An item can be chosen from the inventory of agent i, i.e., $\mathcal{F}_i$, or agent i can call another agent j and send the context

[2]Although the model we propose in this paper has synchronous arrivals, it can be easily extended to the asynchronous case where agents have different user arrival rates, and even when no user arrives in some time slots.

[3]All of our results in this paper will also hold for the case when the commission is a function of the price of the item f_j sold by agent j, i.e., $c_{i,j}(p_{f_j})$

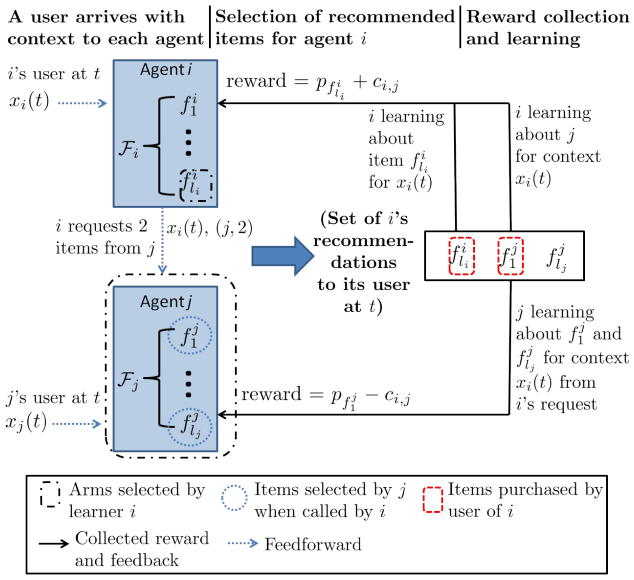

Fig. 1. Operation of the system for agent i for $N = 3$ recommendations. At each time a user arrives to agent i with context $x_i(t)$, agent i recommends a set of its own items and items from other agents.

information of the user $x_i(t)$, then agent j returns back an item f_j with price p_{f_j}[4] to be recommended to agent i based on the context information. Let $\mathcal{N}_i(t)$ be the set of items recommended by agent i to the user at time t. For simplicity, we will consider the case when the user's purchase probabilities of recommended items are independent of each other. Our framework can also be extended to the case when the purchase probabilities are dependent. Let $\mathcal{A}_N$ be the set of subsets of $\mathcal{F}$ with N items. Let $\mathcal{N} \in \mathcal{A}_N$ be a set of recommendations.

Assumption 1. ***Independent purchase probability:*** *For each item f offered along with the items in the set $\mathcal{N} \in \mathcal{A}_N$, a user with context x will buy the item with an unknown probability $q_f(x)$, independent of the other items in $\mathcal{N}_i(t)$, for which there exists $L > 0$, $\alpha > 0$ such that for all $x, x' \in \mathcal{X}$, we have $|q_f(x) - q_f(x')| \leq L||x - x'||^\alpha$, where $||.||$ denotes the Euclidian norm in $\mathbb{R}^d$.*

When Assumption 1 holds, the agents can estimate the purchase probability of an item by using the empirical mean of the number of times the item is purchased by users with similar context information. The goal of agent i is to maximize its total expected revenue from its own users. One-step expected revenue of agent i from recommending a set of items $\mathcal{N}_i$ to its user with context x is given by $Q_{i,\mathcal{N}_i}(x) := \sum_{f \in \mathcal{N}_i - \mathcal{F}_i} c_{i,j(f)} q_f(x) + \sum_{f \in \mathcal{N}_i - (\mathcal{N}_i - \mathcal{F}_i)} p_f q_f(x)$, where $j(f)$ is the agent who owns item f. Then, the optimal set of items for agent i is $\mathcal{N}_i^*(x) := \arg\max_{\mathcal{N} \in \mathcal{A}_N} Q_{i,\mathcal{N}_i}(x)$. Since the inventory of other agents and $q_f(x)$, $x \in \mathcal{X}$, $\mathcal{N} \in \mathcal{A}_N$ are unknown a priori to agent i, $\mathcal{N}_i^*(x)$ is unknown to agent i for all contexts $x \in \mathcal{X}$.

[4]If agent j does not want to reveal the price to agent i, then the recommendation rule can be modified as follows: Agent j's item will be recommended by agent i without a price tag, and when the user clicks to agent j's item it will be directed to agent j's website where the price will be revealed to the user.

The set of actions available to agent i at any time step is the pair (u_i, n_{u_i}), where u_i denotes the item to be recommended for $u_i \in \mathcal{F}_i$ or another agent to be recommended for $u_i \in \mathcal{M}_{-i}$, and n_{u_i} denotes whether item u_i is recommended ($n_{u_i} = 1$) or not ($n_{u_i} = 0$) for $u_i \in \mathcal{F}_i$, or how many distinct items agent u_i should recommend to agent i for $u_i \in \mathcal{M}_{-i}$. Based on this, let $\mathcal{L}_i = \{(u_i, n_{u_i}) \in \mathcal{F}_i \times \{0,1\} \text{ or } (u_i, n_{u_i}) \in \mathcal{M}_{-i} \times \{0, 1, \ldots, N\} : \sum_{u_i \in \mathcal{K}_i} n_{u_i} = N\}$, be the set of actions available to agent i. We assume that $|\mathcal{F}_j| \geq N$ for all $j \in \mathcal{M}$. Let α_i be the recommendation strategy adopted by agent i for its own users, i.e., based on its past observations and decisions, agent i chooses a vector $\alpha_i(t) \in \mathcal{L}_i$ at each time step. Let β_i be the recommendation strategy adopted by agent i when it is called by another agent to recommend its own items. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)$. Let $S^i_{\boldsymbol{\alpha},\boldsymbol{\beta}}(T)$ be the total expected reward agent i can get based only on recommendations to its own users by time T.

Agent i's goal is to maximize its total reward $S^i_{\boldsymbol{\alpha},\boldsymbol{\beta}}(T)$ from its own users for any T. Since agents are cooperative agent i also helps other agents $j \in \mathcal{M}_{-i}$ to maximize $S^j_{\boldsymbol{\alpha},\boldsymbol{\beta}}(T)$ by recommending its items to them. We assume that user arrivals to the agents are independent of each other. Therefore, agent j will also benefit from agent i if its item can be sold by agent i. In this paper, we develop distributed online learning algorithms for the agents in $\mathcal{M}$, i.e., $(\alpha_i, \beta_i)_{i \in \mathcal{M}}$ such that the expected total reward for any agent $S^i_{\boldsymbol{\alpha},\boldsymbol{\beta}}(T)$ is maximized for all $i \in \mathcal{M}$. In other words, we define the regret of agent i to be $R_i(T) := \sum_{t=1}^{T} \sum_{f \in \mathcal{N}_i^*(x_i(t)) - \mathcal{F}_i} c_{i,j} q_f(x_i(t)) + \sum_{f \in \mathcal{F}_i - (\mathcal{N}_i^*(x_i(t))) - \mathcal{F}_i)} p_f q_f(x_i(t)) - S^i_{\boldsymbol{\alpha},\boldsymbol{\beta}}(T)$, and design online learning algorithms that will minimize the regret. Note that the regret is calculated with respect to the highest expected reward agent i can obtain from its own users, but not the users of other agents. Therefore, agent i does not act strategically to attract the users of other agents, such as by cutting its own prices or paying commissions even when an item is not sold to increase its chance of being recommended by another agent. We will show that the regret of the algorithms proposed in this paper will be sublinear in time, which means that the distributed learning scheme converges to the average reward of the best recommender strategy $\mathcal{N}_i^*(x)$ for each $i \in \mathcal{M}$, $x \in \mathcal{X}$. Moreover, the regret also provides us with a bound on how fast our algorithm converges to the best recommender strategy.

III. Context Based Recommendations

We call the algorithm in this section *context based multiple recommendations* (CBMR) whose pseudocode is given in Fig. 2 and Fig. 3. The algorithm which agent i uses to recommend items to other agents when called by them is simple. Basically agent i will either explore one of its own items or exploit its item with the highest estimated purchase probability in that case. Therefore its pseudocode is not given. Basically, an agent using CBMR forms a partition of the context space $[0,1]^d$, depending on the final time T, consisting of $(m_T)^d$ sets where each set is a d-dimensional hypercube with dimensions $1/m_T \times 1/m_T \times \ldots \times 1/m_T$, and m_T is an integer that is non-decreasing in T which is an input parameter of CBMR. The sets in this partition are indexed by $\mathcal{I}_T = \{1, 2, \ldots, (m_T)^d\}$. We denote the set with index l with I_l. Agent i learns the purchase probability of the items in each set in the partition independently from the other sets in the partition based on the context information of the users that arrived to agent i and the users for which agent i is recommended by another agent. Since users with similar contexts have similar purchase probabilities, it is expected that the optimal recommendations are similar for users located in the same set

in $\mathcal{I}_T$. Since the best recommendations are learned independently for each set in $\mathcal{I}_T$, there is a tradeoff between the number of sets in $\mathcal{I}_T$ and the estimation of the best recommendations for contexts in each set in $\mathcal{I}_T$.

In order to exploit the independence of the purchase probabilities of the items, we decouple the action space $\mathcal{L}_i$ of agent i. For this, let $\mathcal{J}_{i,j} := \{1_j, 2_j, \ldots, N_j\}$ denote the set of the number of recommendations agent i can request from agent j, where we use the subscript j to denote that the recommendations are requested from agent j. Let $\tilde{\mathcal{J}}_i := \cup_{j \in \mathcal{M}_{-i}} \mathcal{J}_{i,j}$, $\mathcal{J}_i := \mathcal{F}_i \cup \tilde{\mathcal{J}}_i$ be the set of *arms* of agent i. We have $|\mathcal{J}_i| = |\mathcal{F}_i| + (M-1)N$, which is linear in $|\mathcal{F}_i|$, M, N. For an arm u, let $j(u)$ denote the agent that provides the recommendations for u and $n(u)$ denote the number of recommendations from u. CBMR has exploration and exploitation phases for each arm $u \in \mathcal{F}_i$, and exploration, exploitation and training phases for each arm $u \in \tilde{\mathcal{J}}_i$. At each time step t, CBMR forms reward estimates for each arm $u \in \mathcal{J}_i$ based on the sample mean of the observed rewards of agent i at times $t' \in \{1, \ldots, t-1\}$ agent i selected arm u in exploration and exploitation phases while $x_i(t') \in I_{l'}$, where l' is such that $x_i(t) \in I_{l'}$. When an arm u is selected in a training phase, agent i does not update the estimated reward from that arm because it believes that the items recommended by agent $j(u)$ may not be the best set of items agent $j(u)$ can offer to i. This means that agent i will form an incorrect estimate of the expected reward of arm $u \in \tilde{\mathcal{J}}_i$ if it uses observations from trainings to calculate the estimate. In contrast, agent i uses all the observations from exploration and exploitation phases to estimate the reward of an arm u. At each time step t, CBMR selects a combination of arms for agent i such that the number of recommendations from this combination is equal to N. Agent i keeps two counters for arms $u \in \tilde{\mathcal{J}}_i$. The first one, i.e., $N^i_{1,u,l}(t)$, counts the number of context arrivals to agent i in set l by time t which are also sent to agent $j(u)$ in the training phases of i. The second one, i.e., $N^i_{2,u,l}(t)$, counts the number of context arrivals to agent i in set l by time t which are used to estimate the expected reward of agent i from choosing arm u. Similarly for $u \in \mathcal{F}_i$, $N^i_{u,l}(t)$ denotes the number of context arrivals to agent i in set l by time t for which agent i recommended its item u. For notational convenience let $N^i_{u,l}(t) := N^i_{2,u,l}(t)$ for $u \in \tilde{\mathcal{J}}_i$.

At each time t, agent i first checks which set in the partition $\mathcal{I}_T$ context $x_i(t)$ belong to. CBMR gives priority to arms that are under-explored or under-trained. An arm $u \in \mathcal{F}_i$ will be given priority to be explored if $N^i_{u,l}(t) \leq D_1(t)$. An arm $u \in \tilde{\mathcal{J}}_i$ will be given priority to be trained if $N^i_{1,u,l}(t) \leq D_{2,u}(t)$, and it will be given priority to be explored if $N^i_{1,u,l}(t) > D_{2,u}(t)$ and $N^i_{2,u,l}(t) \leq D_3(t)$, where $D_1(t), D_2(t)$ and $D_3(t)$ are monotonically non-decreasing deterministic functions of t. Let

$$\mathcal{S}_{i,l}(t) := \Big\{ u \in \mathcal{F}_i : N^i_{u,l}(t) \leq D_1(t) \text{ or } \\ u \in \tilde{\mathcal{J}}_i : N^i_{1,u,l}(t) \leq D_{2,u}(t) \text{ or } N^i_{2,u,l}(t) \leq D_3(t) \Big\}.$$

Basically, agent i exploits at time t if $\mathcal{S}_{i,l}(t) = \emptyset$. Otherwise it trains or explores. At time t if there are no other under-trained or under-explored arms, agent i chooses the remaining arms by exploitation such that the total number of recommendations at time t will be N. Basically, it chooses a set of arms $\mathcal{B}_t$ such that it is feasible, i.e., for all $u, u' \in \mathcal{B}_t - \mathcal{F}_i$, we have $j(u) \neq j(u')$ and $\sum_{u \in \mathcal{B}} n(u) = N$, and sum of the sample mean rewards of the arms in $\mathcal{B}_t$ is the maximum over all possible sets of feasible arms. The estimated reward of an arm $u_i \in \mathcal{J}_i$ can be

updated based on the received reward whenever any action $\mathcal{L}_i$ that contains arm u_i is selected by agent i. In order to analyze the regret of CBMR, we will bound the regret in exploration and training phases by showing that the number of explorations and trainings is linear in $|\mathcal{J}_i|$. Then, we will bound the regret in the exploitation phases by bounding the regret of sub-optimal and near-optimal arms selections as in the previous subsection.

Context Based Multiple Recommendations (CBMR for agent i):

Input: $D_1(t)$, $D_{2,u}(t)$, $u \in \tilde{\mathcal{J}}_i$, $D_3(t)$, T, m_T
Initialize: Partition $[0,1]^d$ into $(m_T)^d$ sets, indexed by the set $\mathcal{I}_T = \{1, 2, \ldots, (m_T)^d\}$. $N^i_{u,l} = 0, \forall u \in \mathcal{F}_i, l \in \mathcal{I}_T$, $N^i_{1,u,l} = 0, N^i_{2,u,l} = 0, \forall u \in \tilde{\mathcal{J}}_i, l \in \mathcal{I}_T$.
while $t \geq 1$ **do**
  $\mathcal{N}_i = \emptyset$, $\mathcal{N}_i^{et} = \emptyset$, $\mathcal{N}_i^e = \emptyset$, $\mathcal{N}_i^t = \emptyset$, $cnt = 0$
  while $|\mathcal{N}_i| < N$ **do**
    for $l = 1, \ldots, (m_T)^d$ **do**
      if $x_i(t) \in I_l$ **then**
        $l^* = l$
        for $u \in \mathcal{J}_i$ **do**
          if $u \in \mathcal{F}_i$ such that $N^i_{k,l} \leq D_1(t)$ **then**
            $\mathcal{N}_i = \mathcal{N}_i \cup \{u\}$, $\mathcal{N}_i^e = \mathcal{N}_i^e \cup \{u\}$, $cnt = cnt + 1$
          else if $u \in \tilde{\mathcal{J}}_i$ such that $N^i_{1,u,l} \leq D_{2,u}(t)$ and $cnt + u \leq N$ **then**
            $\mathcal{N}_i = \mathcal{N}_i \cup \{u\}$, $\mathcal{N}_i^t = \mathcal{N}_i^t \cup \{u\}$, $cnt = cnt + u$
          else if $u \in \tilde{\mathcal{J}}_i$ such that $N^i_{k,l} \leq D_3(t)$ and $cnt + u \leq N$ **then**
            $\mathcal{N}_i = \mathcal{N}_i \cup \{u\}$, $\mathcal{N}_i^e = \mathcal{N}_i^e \cup \{u\}$, $cnt = cnt + u$
          end if
        end for
      end if
    end for
  end while
  $N' = N - |\mathcal{N}_i|$
  $\mathcal{N}_i^{et} =$ **Choose**$(N', \mathcal{N}_i, (\bar{r}^i_{u,l^*})_{u \in \mathcal{J}_i})$
  $\mathcal{N}_i = \mathcal{N}_i \cup \mathcal{N}_i^{et}$
  Play$(\mathcal{N}_i, \mathcal{N}_i^e, \mathcal{N}_i^t, \mathcal{N}_i^{et}, (N^i_{u,l^*})_{u \in \mathcal{F}_i}, (N^i_{1,u,l^*})_{u \in \tilde{\mathcal{J}}_i}, (N^i_{2,u,l^*})_{u \in \tilde{\mathcal{J}}_i}, (\bar{r}^i_{u,l^*})_{u \in \mathcal{J}_i})$
  $t = t + 1$
end while

Fig. 2. Pseudocode for the CBMR algorithm.

For an arm u let $j(u)$ denote the agent which sends the recommendations when arm u is selected by agent i, and $n(u)$ denote the number of items agent $j(u)$ recommends to agent i. For an item f, let $j(f)$ denote the agent that owns the item. For simplicity, in this paper we assume that agents have different sets of items. For an item $f_i \in \mathcal{F}_i$, let $\lambda_{i,f_i}(x) := p_{f_i} q_{f_i}(x)$ be the expected reward of that item for agent i, and for an item $f \in \mathcal{F} - \mathcal{F}_i$, let $\lambda_{i,f}(x) := c_{i,j(f)} q_{f_i}(x)$ be the expected reward of that item for agent i, whent agent i's user's context is x. Recall that $\mathcal{N}_i^*(x)$ is the set of N items in $\mathcal{F}$ which maximizes agent i's expected reward for context x. For an item $f \in \mathcal{F}$ and $I_l \in \mathcal{I}_T$ let

$$\underline{\lambda}_{i,f,l} := \inf_{x \in \mathcal{I}_l} \lambda_{i,f}(x),$$

Choose(N, $\mathcal{N}$, $\boldsymbol{r}$):
1: Select arms $u \in \mathcal{J}_i - \mathcal{N}$ such that $u \notin \mathcal{J}_{i,j}$ if $\exists\ u' \in \mathcal{N} \cap \mathcal{J}_{i,j}$ for $j \in \mathcal{M}_{-i}$, $\sum_u n_u \leq N$ and $\sum_u r_u$ is maximized.
Play($\mathcal{N}$, $\mathcal{N}_1$, $\mathcal{N}_2$, $\mathcal{N}_3$ $\boldsymbol{N}$, $\boldsymbol{r}$):
1: Take action $\mathcal{N}$, get the recommendations of other agents, recommend $\mathcal{N}_i(t)$ to the user.
2: **for** $u \in \mathcal{N}$ **do**
3: **if** $u \in \mathcal{N}_i(t) \cap \mathcal{F}_i$ **then**
4: Receive reward $r_u(t) = I(u \in F_i(t))$. $r_u = \frac{N_{u,l} r_u + r_u(t)}{N_{u,l}+1}$, $N_{u,l} + +$.
5: **else if** $u \in (\mathcal{N} - \mathcal{F}_i) \cap \mathcal{N}_i^t$ **then**
6: Receive reward $r_u(t) = \sum_{f \in \mathcal{F}_j} I(f \in F_i(t))$, $N_{1,u,l} + +$
7: **else**
8: Receive reward $r_u(t) = \sum_{f \in \mathcal{F}_j} I(f \in F_i(t))$, $r_u = \frac{N_{2,u,l} r_u + r_u(t)}{N_{2,u,l}+1}$, $N_{2,u,l} + +$ $N_{2,u,l} + +$
9: **end if**
10: **end for**

Fig. 3. Pseudocode of choose and play modules.

and

$$\overline{\lambda}_{i,f,l} := \sup_{x \in \mathcal{I}_l} \lambda_{i,f}(x).$$

In order to define the set of suboptimal arms in a hypercube I_l, we will define expressions related to variation of the expected rewards of items and arms in I_l. Let $f_n(\mathcal{N}, x)$ denote the item in $\mathcal{N}$ with the nth highest expected reward for agent i for context x. The expected reward of arm u for agent i is given by

$$\mu_{i,u}(x) := \sum_{n=1}^{n(u)} \lambda_{i,f_n(\mathcal{F}_{j(u)},x)}(x).$$

For an arm u and $I_l \in \mathcal{I}_T$ let

$$\underline{\mu}_{i,u,l} := \inf_{x \in \mathcal{I}_l} \mu_{i,u}(x),$$

and

$$\overline{\mu}_{i,u,l} := \sup_{x \in \mathcal{I}_l} \mu_{i,u}(x).$$

We next define the feasible sets of arms. A set $\mathcal{B}$ of arms is feasible if for all $u, u' \in \mathcal{B} - \mathcal{F}_i$ we have $j(u) \neq j(u')$ for $u \neq u'$ and $\sum_{u \in \mathcal{B}} n(u) = N$. Let $\mathcal{C}_F$ denote the set of all sets of feasible arms. Then, for a set of arms $\mathcal{B}$, the expected reward is

$$\mu_{i,\mathcal{B}}(x) := \sum_{u \in \mathcal{B}} \mu_{i,u}(x).$$

For a set of arms $\mathcal{B}$, let $\underline{\mu}_{i,\mathcal{B},l} := \inf_{x \in I_l} \mu_{i,\mathcal{B}}(x)$, and $\bar{\mu}_{i,\mathcal{B},l} := \sup_{x \in I_l} \mu_{i,\mathcal{B}}(x)$. Let

$$\underline{\mu}^*_{i,l} := \max_{\mathcal{B} \in \mathcal{B}_F} \underline{\mu}_{i,\mathcal{B},l}.$$

When choosing an arm for a user with context in $I_l \in \mathcal{I}_T$, our learning algorithm will only use the past observations and decisions it had made for contexts belonging to I_l. Therefore it is important for us to characterize the total variation of the expected reward of an item in set I_l. For simplicity, we assume that all prices and

commisions are in the unit interval $[0,1]$. Then as a result of Assumption 1 we have for any item $f \in \mathcal{F}$ and for any $I_l \in \mathcal{I}_T$

$$\sup_{x,x' \in \mathcal{I}_l} |\mu_{i,f}(x) - \mu_{i,f}(x')| \leq L d^{\alpha/2} (m_T)^{-\alpha}. \tag{1}$$

Using (1), for any arm u and for any $I_l \in \mathcal{I}_T$, we have

$$\sup_{x,x' \in \mathcal{I}_l} |\mu_{i,u}(x) - \mu_{i,u}(x')| \leq n(u) L d^{\alpha/2} (m_T)^{-\alpha}. \tag{2}$$

This implies that for any $I_l \in \mathcal{I}_T$ and $x \in I_l$

$$\begin{aligned} \bar{\mu}_{i,\mathcal{B},l} &= \sup_{x \in I_l} \left(\sum_{u \in \mathcal{B}} \mu_{i,u}(x) \right) \\ &\geq \sum_{u \in \mathcal{B}} \left(\bar{\mu}_{i,u,l} - n(u) L d^{\alpha/2} (m_T)^{-\alpha} \right) \\ &= \sum_{u \in \mathcal{B}} \bar{\mu}_{i,u,l} - N L d^{\alpha/2} (m_T)^{-\alpha}, \end{aligned} \tag{3}$$

and

$$\begin{aligned} \underline{\mu}_{i,\mathcal{B},l} &= \inf_{x \in I_l} \left(\sum_{u \in \mathcal{B}} \mu_{i,u}(x) \right) \\ &\leq \sum_{u \in \mathcal{B}} \left(\underline{\mu}_{i,u,l} + n(u) L d^{\alpha/2} (m_T)^{-\alpha} \right) \\ &= \sum_{u \in \mathcal{B}} \underline{\mu}_{i,u,l} + N L d^{\alpha/2} (m_T)^{-\alpha}. \end{aligned} \tag{4}$$

Using the results of (3) and (4) we get

$$\sum_{u \in \mathcal{B}} \bar{\mu}_{i,u,l} - N L d^{\alpha/2} (m_T)^{-\alpha} \leq \bar{\mu}_{i,\mathcal{B},l} \leq \sum_{u \in \mathcal{B}} \bar{\mu}_{i,u,l}, \tag{5}$$

and

$$\sum_{u \in \mathcal{B}} \underline{\mu}_{i,u,l} \leq \underline{\mu}_{i,\mathcal{B},l} \leq \sum_{u \in \mathcal{B}} \underline{\mu}_{i,u,l} + N L d^{\alpha/2} (m_T)^{-\alpha}. \tag{6}$$

For the set I_l of the partition $\mathcal{I}_T$, the set of suboptimal feasible sets of arms for agent i at time t is given by

$$\mathcal{U}_l^i(t) := \left\{ \mathcal{B} \in \mathcal{C}_F \text{ such that } \underline{\mu}_{i,l}^* - \bar{\mu}_{i,\mathcal{B},l} \geq a_1 t^{\theta} \right\}, \tag{7}$$

where we will optimize over a_1 and θ[5]. We divide the regret into three parts: $R_i^e(T)$, $R_i^s(T)$ and $R_i^n(T)$, where $R_i^e(T)$ is the regret due to trainings and explorations by time T, $R_i^s(T)$ is the regret due to suboptimal action selections by time T, and $R_i^n(T)$ is the regret due to near optimal arm selection by time T. We bound each of these terms seperately. In the following lemma, we bound the regret of CBMR due to explorations and trainings. Let Y_R be the difference between expected rewards of the best N items for agent i and the worst N items for agent

[5]This optimization is done to derive the regret bound. The actual performance of our algorithm does not depend on what is a_1 and θ. By optimizing over them, we get tighter performance bounds for our algorithm.

i. When the prices and commissions are in the unit interval, an upper bound on Y_R is N. However, depending on the prices, commission and the purchase probabilities Y_R can be much smaller than N.

Lemma 1. *When CBMR is run by agent i with parameters $D_1(t) = t^z \log t$, $D_{2,u}(t) = \binom{F_{\max}}{n(u)} t^z \log t$, $u \in \tilde{\mathcal{J}}_i$, $D_3(t) = t^z \log t$ and $m_T = \lceil T^\gamma \rceil$, where $0 < z < 1$ and $0 < \gamma < 1/d$, we have*

$$E[R_i^e(T)] \leq Y_R 2^d (|\mathcal{J}_i| + (M-1)N) T^{\gamma d}$$
$$+ Y_R 2^d \left(|\mathcal{J}_i| + (M-1) \sum_{a=1}^{N} \binom{F_{\max}}{a} \right) T^{z+\gamma d} \log T.$$

Proof: For a set $I_l \in \mathcal{I}_T$, the number of exploration steps of agent i is bounded by $|\mathcal{J}_i| \lceil T^z \log T \rceil$. Agent i spends at most $\sum_{z=1}^{N} \binom{F_{\max}}{z} \lceil T^z \log T \rceil$ time steps to train agent j. Note that this is the worst-case number of trainings for which agent j does not learn about the purchase probabilities of its items in set I_l from its own users, and from the users of agents other than agent i. The worst case expected regret at each training or exploration step is Y_R. The result follows from summing over all sets in $\mathcal{I}_T$. ■

In the next lemma, we bound $E[R_i^s(T)]$.

Lemma 2. *When CBMR is run with parameters $D_1(t) = t^z \log t$, $D_{2,u}(t) = \binom{F_{\max}}{n(u)} t^z \log t$, $u \in \tilde{\mathcal{J}}_i$, $D_3(t) = t^z \log t$ and $m_T = \left\lceil T^{z/2\alpha} \right\rceil$, where $0 < z < 1$, we have*

$$E[R_i^s(T)] \leq Y_R(2N\beta_2 + (M-1)N^2 F_{\max}\beta_2) + T^{z/2}(2Y_R(M-1)N^2 F_{\max}\beta_2/z).$$

Proof: Let Ω denote the space of all possible outcomes, and let w be a sample path. Let $l_i(t)$ denote the set in $\mathcal{I}_T$ which includes context $x_i(t)$. When clear from the context of presentation, to simplfy the notation, we will use l instead of $l_i(t)$ to denote the set that includes $x_i(t)$. Let

$$\mathcal{W}^i(t) := \{w \in \Omega : S_{i,l_i(t)}(t) = \emptyset\}$$

denote the event that CBMR is in the exploitation phase at time t. The idea is to bound the probability that agent i selects a suboptimal set of arms in an exploitation phase, and then using this to bound the expected number of times a suboptimal set of arms is selected by agent i. For an arm u, $\bar{r}^i_{u,l}(t)$ denotes the sample mean of the rewards collected from explorations and trainings of arm u in set I_l by learner i by time t. Similarly for a set of arms $\mathcal{B} \in \mathcal{C}_F$, $\bar{r}^i_{\mathcal{B},l}(t)$ denotes the sum of the sample mean rewards of arms in $\mathcal{B}$, i.e,

$$\bar{r}^i_{\mathcal{B},l}(t) = \sum_{u \in \mathcal{B}} \bar{r}^i_{u,l}(t).$$

When the agent we refer to is clear from the context we will drop the superscript in the notation of the sample mean rewards. The set of suboptimal arms for agent i at time t is given by

$$\mathcal{C}^i_S(t) = \mathcal{C}_F - \mathcal{U}^i_{l_i(t)}(t).$$

Let

$$\mathcal{B}^*_S(t) := \arg\max_{\mathcal{B} \in \mathcal{C}^i_S(t)} \bar{r}^i_{u,l_i(t)}(t)$$

denote the best (or one of the best if there are multiple) suboptimal set of arms at time t, i.e., the suboptimal set of arms whose sample mean reward is highest among all feasible suboptimal set of arms.

Let $\mathcal{V}^i(t)$ be the event that a suboptimal set of arms in $\mathcal{C}_S^i(t)$ is chosen by agent i at time t. Since $R_s(T)$ is a random variable we have

$$R_i^s(T) \leq Y_R \sum_{t=1}^{T} I(\mathcal{V}^i(t), \mathcal{W}^i(t)),$$

with probability one, where $I(A)$ is the indicator function of event A which is equal to 1 if event A happened and 0 otherwise. Taking the expectation with respect to the randomness of the rewards, we get

$$E[R_i^s(T)] \leq Y_R \sum_{t=1}^{T} P(\mathcal{V}^i(t), \mathcal{W}^i(t)). \tag{8}$$

In the next part of the proof, we will bound $P(\mathcal{V}^i(t), \mathcal{W}^i(t))$ with a decaying function of t.

Let $\mathcal{E}^i_{u,l_i(t)}(t)$ denote the set of rewards collected by agent i from arm u as a result of recommendations to agent i's users whose context are in the set $I_{l_i(t)}$ by time t. Let $\mathcal{H}^i_{l_i(t)}(t)$ be the event that for all arms $u \in \tilde{\mathcal{J}}_i$ at most t^ϕ samples in $\mathcal{E}^i_{u,l_i(t)}(t)$ come from recommendations made by agent $j(u)$ such that at least one of the $n(u)$ items recommended by agent $j(u)$ is a suboptimal item for that particular recommendation. To be more precise, let $\sigma_j^i = (\sigma_j^i(1), \sigma_j^i(2), \ldots, \sigma_j^i(|\mathcal{F}_j|))$ be an ordering of items of agent j in terms of their expected rewards for agent i such that $\sigma_j^i(k) \geq \sigma_j^i(k+1)$ for all $k \in \{1, \ldots, |\mathcal{F}_j| - 1\}$. When clear from the context, we will drop the subscript an superscript denoting the agents from the notation. For arm u and $I_l \in \mathcal{I}_T$, an item $f \in \mathcal{F}_{j(u)}$ is suboptimal if

$$\underline{\lambda}_{i,\sigma^i_{j(u)}(n(u)),l} - \overline{\lambda}_{i,f,l} \geq a_2 t^\theta. \tag{9}$$

Next, we define three events which are going to be used to bound the probability that a suboptimal set of arms is chosen at exploitation steps.

$$\mathcal{O}_1(t) := \left\{ \bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) \geq \overline{\mu}_{i,\mathcal{B}^*_S(t),l_i(t)} + H_t, \mathcal{H}^i_{l_i(t)}(t), \mathcal{W}^i(t) \right\}$$

$$\mathcal{O}_2(t) := \left\{ \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t) \leq \underline{\mu}_{i,\mathcal{B}^*(l_i(t)),l_i(t)} - H_t, \mathcal{H}^i_{l_i(t)}(t), \mathcal{W}^i(t) \right\}$$

$$\mathcal{O}_3(t) := \Big\{ \bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) \geq \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t), \bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) < \overline{\mu}_{i,\mathcal{B}^*_S(t),l_i(t)} + H_t, \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t) > \underline{\mu}_{i,\mathcal{B}^*(l_i(t)),l_i(t)} - H_t,$$
$$\mathcal{H}^i_{l_i(t)}(t), \mathcal{W}^i(t) \Big\}.$$

We have

$$\begin{aligned}
&\{\mathcal{V}^i(t), \mathcal{W}^i(t)\} \\
&\subset \{\bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) \geq \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t), \mathcal{W}^i(t)\} \\
&\subset \{\bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) \geq \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t), \mathcal{W}^i(t), \mathcal{H}^i_{l_i(t)}(t)\} \cup \{\bar{r}_{\mathcal{B}^*_S(t),l_i(t)}(t) \geq \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t), \mathcal{W}^i(t), (\mathcal{H}^i_{l_i(t)}(t))^C\} \\
&\subset \mathcal{O}_1(t) \cup \mathcal{O}_2(t) \cup \mathcal{O}_3(t) \cup \left\{ (\mathcal{H}^i_{l_i(t)}(t))^C, \mathcal{W}^i(t) \right\},
\end{aligned} \tag{10}$$

for some $H_t > 0$, where for an event $\mathcal{A}$, $\mathcal{A}^C$ denotes the complement of that event. This implies that

$$P\left(\mathcal{V}^i(t), \mathcal{W}^i(t)\right) \leq P(\mathcal{O}_1(t)) + P(\mathcal{O}_2(t)) + P(\mathcal{O}_3(t)) + P(\mathcal{H}^i_{l_i(t)}(t))^C, \mathcal{W}^i(t)) \tag{11}$$

Next, we prove that the following condition (**C1**)

$$\textbf{C1}: \qquad \left(2NLd^{\alpha/2}(m_T)^{-\alpha} + 2H_t - a_1 t^{\theta} \leq 0\right),$$

implies that $P(\mathcal{O}_3(t)) = 0$.

Step 1: Under **C1** since $-a_1 t^{\theta} \geq \overline{\mu}_{i,\mathcal{B}_S^*(t),l_i(t)} - \underline{\mu}_{i,\mathcal{B}^*(l_i(t)),l_i(t)}$ we have

$$\textbf{C1} \Rightarrow \overline{\mu}_{i,\mathcal{B}_S^*(t),l_i(t)} + NLd^{\alpha/2}(m_T)^{-\alpha} - \left(\underline{\mu}_{i,\mathcal{B}^*(l_i(t)),l_i(t)} - NLd^{\alpha/2}(m_T)^{-\alpha}\right) + 2H_t \leq 0. \tag{12}$$

Step 2: Using the relations between $\mu_{i,\mathcal{B},l}$ and $\mu_{i,u,l}$, $u \in \mathcal{B}$, $\mathcal{B} \in \mathcal{C}_F$, $l \in \mathcal{I}_T$ given by (5) and (6) together with (12), we get

$$(12) \Rightarrow \sum_{u \in \mathcal{B}_S^*(t)} \overline{\mu}_{i,u,l_i(t)} - \sum_{u \in \mathcal{B}^*(l_i(t))} \underline{\mu}_{i,u,l_i(t)} + 2H_t \leq 0.$$

Step 3: In $\mathcal{O}_3(t)$, we have

$$\bar{r}_{\mathcal{B}_S^*(t),l_i(t)}(t) < \overline{\mu}_{i,\mathcal{B}_S^*(t),l_i(t)} + H_t \leq \sum_{u \in \mathcal{B}_S^*(t)} \overline{\mu}_{i,u,l_i(t)} + H_t,$$

and

$$-\bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t) < -\underline{\mu}_{\mathcal{B}^*(l_i(t)),l_i(t)} + H_t \leq -\sum_{u \in \mathcal{B}^*(l_i(t))} \underline{\mu}_{i,u,l_i(t)} + H_t,$$

hence

$$\bar{r}_{\mathcal{B}_S^*(t),l_i(t)}(t) - \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t) < \sum_{u \in \mathcal{B}_S^*(t)} \overline{\mu}_{i,u,l_i(t)} - \sum_{u \in \mathcal{B}^*(l_i(t))} \underline{\mu}_{i,u,l_i(t)} + 2H_t. \tag{13}$$

Step 4: Equations (12) and (13) together imply that

$$\bar{r}_{\mathcal{B}_S^*(t),l_i(t)}(t) < \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t).$$

However on event $\mathcal{O}_3(t)$ we must have $\bar{r}_{\mathcal{B}_S^*(t),l_i(t)}(t) \geq \bar{r}_{\mathcal{B}^*(l_i(t)),l_i(t)}(t)$. Therefore **C1** implies $P(\mathcal{O}_3(t)) = 0$.

Next, we will bound $P(\mathcal{O}_1(t))$. Recall that $\mathcal{B}_S^*(t)$ is a random variable since rewards are random, and $C_S^i(t)$ is a deterministic set.

Step 1: By law of total probability we have

$$\begin{aligned}
P(\mathcal{O}_1(t)) &= \sum_{\mathcal{B} \in C_S^i(t)} P(\mathcal{O}_1(t) | \mathcal{B}_S^*(t) = \mathcal{B}) P(\mathcal{B}_S^*(t) = \mathcal{B}) \\
&\leq \sum_{\mathcal{B} \in C_S^i(t)} P(\mathcal{B}_S^*(t) = \mathcal{B}) \max_{\mathcal{B} \in C_S^i(t)} P(\mathcal{O}_1(t) | \mathcal{B}_S^*(t) = \mathcal{B}) \\
&= \max_{\mathcal{B} \in C_S^i(t)} P(\mathcal{O}_1(t) | \mathcal{B}_S^*(t) = \mathcal{B}).
\end{aligned}$$

Step 2: For any $\mathcal{B} \in C_S^i(t)$,

$$\begin{aligned}
P(\mathcal{O}_1(t)|\mathcal{B}_S^*(t) = \mathcal{B}) &= P\left(\bar{r}_{\mathcal{B},l_i(t)}(t) \geq \overline{\mu}_{i,\mathcal{B},l_i(t)} + H_t, \mathcal{H}_{l_i(t)}^i(t), \mathcal{W}^i(t)\right) \\
&= P\left(\sum_{u\in\mathcal{B}} \bar{r}_{u,l_i(t)}(t) \geq \overline{\mu}_{i,\mathcal{B},l_i(t)} + H_t, \mathcal{H}_{l_i(t)}^i(t), \mathcal{W}^i(t)\right) \\
&\leq P\left(\sum_{u\in\mathcal{B}} \bar{r}_{u,l_i(t)}(t) \geq \sum_{u\in\mathcal{B}} \overline{\mu}_{i,u,l_i(t)} - NLd^{\alpha/2}(m_T)^{-\alpha} + H_t, \mathcal{H}_{l_i(t)}^i(t), \mathcal{W}^i(t)\right) \\
&\leq P\left(\sum_{u\in\mathcal{B}} \bar{r}_{u,l_i(t)}^{\text{best}}(|\mathcal{E}_{u,l_i(t)}^i(t)|) \geq \sum_{u\in\mathcal{B}} \overline{\mu}_{i,u,l_i(t)} - NLd^{\alpha/2}(m_T)^{-\alpha} + H_t\right)
\end{aligned}$$

For any $u \in \mathcal{J}_i$, $\mathcal{B} \in \mathcal{C}_F$ and $I_l \in \mathcal{I}_T$ let

$$\mathcal{O}_{u,\mathcal{B},l}^{\text{best}}(t) := \left\{\bar{r}_{u,l}^{\text{best}}(|\mathcal{E}_{u,l}^i(t)|) \geq \overline{\mu}_{i,u,l} - \frac{N}{|\mathcal{B}|}Ld^{\alpha/2}(m_T)^{-\alpha} + \frac{H_t}{|\mathcal{B}|}\right\},$$

and

$$\mathcal{Z}_{\mathcal{B},l}^{\text{best}}(t) := \left\{\sum_{u\in\mathcal{B}} \bar{r}_{u,l}^{\text{best}}(|\mathcal{E}_{u,l}^i(t)|) \geq \sum_{u\in\mathcal{B}} \overline{\mu}_{i,u,l} - NLd^{\alpha/2}(m_T)^{-\alpha} + H_t\right\}.$$

We have

$$\bigcap_{u\in\mathcal{B}} (\mathcal{O}_{u,\mathcal{B},l}^{\text{best}}(t))^C \subset (\mathcal{Z}_{\mathcal{B},l}^{\text{best}}(t))^C \Rightarrow \mathcal{Z}_{\mathcal{B},l}^{\text{best}}(t) \subset \bigcup_{u\in\mathcal{B}} \mathcal{O}_{u,\mathcal{B},l}^{\text{best}}(t).$$

Hence

$$P\left(\mathcal{Z}_{\mathcal{B},l_i(t)}^{\text{best}}(t)\right) \leq \sum_{u\in\mathcal{B}} P\left(\mathcal{O}_{u,\mathcal{B},l_i(t)}^{\text{best}}(t)\right). \tag{14}$$

Therefore

$$\begin{aligned}
P(\mathcal{O}_1(t)) &\leq \sum_{u\in\mathcal{B}} P\left(\mathcal{O}_{u,\mathcal{B},l_i(t)}^{\text{best}}(t)\right) \\
&\leq Ne^{-2t^z \log t\left(\frac{H_t}{|\mathcal{B}|} - \frac{N}{|\mathcal{B}|}Ld^{\alpha/2}(m_T)^{-\alpha}\right)^2} \\
&\leq Ne^{-2t^z \log t\left(\frac{H_t}{N} - Ld^{\alpha/2}(m_T)^{-\alpha}\right)^2}.
\end{aligned} \tag{15}$$

Next, we will bound $P(\mathcal{O}_2(t))$. Steps are similar to bounding $P(\mathcal{O}_1(t))$. However, $\mathcal{B}_{l_i(t)}^*(t)$ is deterministic.

Step 1: We have

$$\begin{aligned}
P(\mathcal{O}_2(t)) &\leq P\left(\sum_{u\in\mathcal{B}_{l_i(t)}^*(t)} \bar{r}_{u,l_i(t)}(t) \leq \sum_{u\in\mathcal{B}_{l_i(t)}^*(t)} \underline{\mu}_{i,u,l_i(t)} + NLd^{\alpha/2}(m_T)^{-\alpha} - H_t, \mathcal{H}_{l_i(t)}^i(t), \mathcal{W}^i(t)\right) \\
&\leq P\left(\sum_{u\in\mathcal{B}_{l_i(t)}^*(t)} \bar{r}_{u,l_i(t)}^{\text{worst}}(|\mathcal{E}_{u,l}^i(t)|) \leq \sum_{u\in\mathcal{B}_{l_i(t)}^*(t)} \underline{\mu}_{i,u,l_i(t)} + NLd^{\alpha/2}(m_T)^{-\alpha} + Nt^{\phi-1} - H_t\right),
\end{aligned}$$

where the last inequality follows from the fact that there even when all arms in $\mathcal{B}_{l_i(t)}^*(t)$ had all their recommendations from suboptimal items in t^ϕ time slots, the sum of the sample means of all arms in $\mathcal{B}_{l_i(t)}^*(t)$ will be at

most Nt^{ϕ} lower than the sample mean when all the recommendation in those t^{ϕ} time slots came from the best items. This is true because the commissions and prices are in unit interval. This argument can be easily extended to the case when commissions and prices are in a bounded interval by multiplying this factor by the maximum commission or price. For any $u \in \mathcal{J}_i$, $\mathcal{B} \in \mathcal{C}_F$ and $I_l \in \mathcal{I}_T$ let

$$\mathcal{O}^{\text{worst}}_{u,\mathcal{B},l}(t) := \left\{ \bar{r}^{\text{worst}}_{u,l}(|\mathcal{E}^i_{u,l}(t)|) \leq \underline{\mu}_{i,u,l} + \frac{N}{|\mathcal{B}|} L d^{\alpha/2} (m_T)^{-\alpha} + \frac{N}{|\mathcal{B}|} t^{\phi-1} - \frac{H_t}{|\mathcal{B}|} \right\}$$

and

$$\mathcal{Z}^{\text{worst}}_{\mathcal{B},l}(t) := \left\{ \sum_{u\in\mathcal{B}} \bar{r}^{\text{worst}}_{u,l}(|\mathcal{E}^i_{u,l}(t)|) \leq \sum_{u\in\mathcal{B}} \underline{\mu}_{i,u,l} + N L d^{\alpha/2} (m_T)^{-\alpha} + N t^{\phi-1} - H_t \right\}$$

We have

$$\bigcap_{u\in\mathcal{B}} (\mathcal{O}^{\text{worst}}_{u,\mathcal{B},l}(t))^C \subset (\mathcal{Z}^{\text{worst}}_{\mathcal{B},l}(t))^C \Rightarrow \mathcal{Z}^{\text{worst}}_{\mathcal{B},l}(t) \subset \bigcup_{u\in\mathcal{B}} \mathcal{O}^{\text{worst}}_{u,\mathcal{B},l}(t).$$

Hence

$$P\left(\mathcal{Z}^{\text{worst}}_{\mathcal{B}^*_{l_i(t)}(t),l_i(t)}(t)\right) \leq \sum_{u\in\mathcal{B}^*_{l_i(t)}(t)} P\left(\mathcal{O}^{\text{worst}}_{u,\mathcal{B}^*_{l_i(t)}(t),l_i(t)}(t)\right). \tag{16}$$

Therefore

$$\begin{aligned} P(\mathcal{O}_2(t)) &\leq \sum_{u\in\mathcal{B}^*_{l_i(t)}(t)} P\left(\mathcal{O}^{\text{worst}}_{u,\mathcal{B}^*_{l_i(t)}(t),l_i(t)}(t)\right) \\ &\leq N e^{-2t^z \log t \left(\frac{H_t}{|\mathcal{B}|} - \frac{N}{|\mathcal{B}|} t^{\phi-1} - \frac{N}{|\mathcal{B}|} L d^{\alpha/2} (m_T)^{-\alpha}\right)^2} \\ &\leq N e^{-2t^z \log t \left(\frac{H_t}{N} - t^{\phi-1} - L d^{\alpha/2} (m_T)^{-\alpha}\right)^2}. \end{aligned} \tag{17}$$

Finally, we bound $P((\mathcal{H}^i_{l_i(t)}(t))^C, \mathcal{W}^i(t))$. For an arm $u \in \tilde{\mathcal{J}}_i$ and $l \in \mathcal{I}_T$, let $X^i_{u,l}(t)$ denote the random variable which is the number of times at least one suboptimal item of agent $j(u)$ is recommended to agent i in exploitation steps in set I_l of agent i by time t. We have

$$P((\mathcal{H}^i_{l_i(t)}(t))^C, \mathcal{W}^i(t)) \leq \sum_{u\in\tilde{\mathcal{J}}_i} P(X^i_{u,l_i(t)}(t) > t^{\phi}) \leq \sum_{u\in\tilde{\mathcal{J}}_i} E[X^i_{u,l_i(t)}(t)]/t^{\phi}.$$

Let $\Xi^i_{u,l}(t)$ be the event that a suboptimal item of agent $j(u)$ is recommended to agent i, agent i is in exploitation step at time t and the context of the user of agent i is in set I_l. Let $\mathcal{S}^j_{i,u,l}(t)$ denote the set of suboptimal items of agent j for agent i at time t for arm u for set I_l. We have $\mathcal{S}^j_{i,u,l}(t) \subset \{\sigma(n(u)) + 1, \ldots, \sigma(|\mathcal{F}_{j(u)}|)\}$. Then

$$E[X^i_{u,l_i(t)}(t)] = \sum_{t'=1}^{t} P(\Xi^i_{u,l_i(t)}(t)).$$

We have

$$P\left(\Xi_{u,l}^{i}(t)\right) \leq \sum_{a=1}^{n(u)} \sum_{b\in\mathcal{S}_{i,u,l}^{j}(t)} P(\bar{s}_{i,\sigma(b),l}^{j}(t) \geq \bar{s}_{i,\sigma(a),l}^{j}(t))$$

$$\leq \sum_{a=1}^{n(u)} \sum_{b\in\mathcal{S}_{i,u,l}^{j}(t)} \left(P\left(\bar{s}_{i,\sigma(b),l}^{j}(t) \geq \overline{\lambda}_{i,\sigma(b),l} + H_t, \mathcal{W}^i(t)\right)\right.$$

$$+P\left(\bar{s}_{i,\sigma(a),l}^{j}(t) \leq \underline{\lambda}_{i,\sigma(a),l} - H_t, \mathcal{W}^i(t)\right) + P\left(\bar{s}_{i,\sigma(b),l}^{j}(t) \geq \bar{s}_{i,\sigma(a),l}^{j}(t),\right.$$

$$\left.\left.\bar{s}_{i,\sigma(b),l}^{j}(t) < \overline{\lambda}_{i,\sigma(b),l} + H_t, \bar{s}_{i,\sigma(a),l}^{j}(t) > \underline{\lambda}_{i,\sigma(a),l} - H_t, \mathcal{W}^i(t)\right)\right),$$

For a suboptimal item of agent j, $\sigma(b)$ and an $n(u)$-best item $\sigma(a)$, using (9), we have $\overline{\lambda}_{i,\sigma(b),l} - \underline{\lambda}_{i,\sigma(a),l} \leq -a_2 t^{\theta}$. This together with the second and third events in the last probability above imply that

$$\bar{s}_{i,\sigma(b),l}^{j}(t) - \bar{s}_{i,\sigma(a),l}^{j}(t) \leq 2H_t - a_2 t^{\theta}.$$

This implies that when $2H_t - a_2 t^{\theta} \leq 0$, we have the third probability equal to 0. Since during an exploitation step of agent i, at least $t^z \log t$ recommendations are made for each item of each agent j in set $I_{l_i(t)}$, we have

$$P\left(\Xi_{u,l}^{i}(t)\right) \leq 2n(u)(F_{\max} - n(u))e^{-2(H_t)^2 t^z \log t}$$

$$\leq 2NF_{\max}e^{-2(H_t)^2 t^z \log t}. \tag{18}$$

Bounding $\mathbf{E}[\mathbf{R_i^s(T)}]$:

Step 1: In order to bound $E[R_i^s(T)]$ given in (8), we need to use the bound for $P(\mathcal{V}^i(t), \mathcal{W}^i(t))$ given in (11). This equation can be further bounded by using our bounds in (15), (17) and (18), all of which holds when conditions **C1**: $2NLd^{\alpha/2}(m_T)^{-\alpha} + 2H_t - a_1 t^{\theta} \leq 0$ and **C2**: $2H_t - a_2 t^{\theta} \leq 0$ hold. The exponential terms of the bounds in (15), (17) and (18) are $e^{-2t^z \log t\left[H_t/N - Ld^{\alpha/2}(m_T)^{-\alpha}\right]^2}$, $e^{-2t^z \log t\left[H_t/N - t^{\phi-1} - Ld^{\alpha/2}(m_T)^{-\alpha}\right]^2}$ and $e^{-2t^z \log t(H_t)^2}$ respectively. If **C3**: $H_t/N - t^{\phi-1} - Ld^{\alpha/2}(m_T)^{-\alpha} > 0$, then the largest of these three terms will be $e^{-2t^z \log t\left[H_t/N - t^{\phi-1} - Ld^{\alpha/2}(m_T)^{-\alpha}\right]^2}$, hence an upper bound on this will be an upper bound on all three terms. Since $m_T = \lceil T \rceil^{\gamma}$, for $t \leq T$, we have $t^{-\gamma} \geq (m_T)^{-1}$, hence

$$2NLd^{\alpha/2}(m_T)^{-\alpha} + 2H_t - a_1 t^{\theta} \leq 2NLd^{\alpha/2}t^{-\gamma\alpha} + 2H_t - a_1 t^{\theta}.$$

Thus **C1** holds if **C4**: $2NLd^{\alpha/2}t^{-\gamma\alpha} + 2H_t - a_1 t^{\theta} \leq 0$ holds. Let $H_t = A_1 t^{\phi-1}$ for some $A_1 > 0$ which we will set later. In the next lemma, we will show that time order of the regret bound due to selecting near optimal sets of arms increases exponentially with θ. Therefore we should set θ as small as possible to minimize the regret bound due to near optimal selections. Similarly from the result of Lemma 1, we need to choose the slicing exponent γ as small as possible in the algorithm since the number of trainings and explorations, hence the time order of the regret bound for trainings and explorations depend exponentially on γ. Since condition **C4** hols if

$$\frac{2NLd^{\alpha/2}}{t^{\gamma\alpha}} + 2A_1 t^{\phi-1} \leq a_1 t^{\theta},$$

and since $1/t^{\gamma\alpha}$ decreases in γ for $t > 1$, the smallest γ, θ and a_1 values for which **C4** holds are $\theta = \phi - 1$, $\gamma = (1-\phi)/\alpha$ and $a_1 = 2NLd^{\alpha/2} + 2A_1$. We also need **C3** to hold, moreover, we need the exponential term to decay fast enough such that when we can bound the sum of the exponential terms from $t = 1$ to T by a small number much less than T that does not depend on T. Again, since $t^{-\gamma} \geq (m_T)^{-1}$, for $H_t/N - t^{\phi-1} - Ld^{\alpha/2}t^{-\gamma\alpha} > 0$ all three exponential terms are upper bounded by

$$e^{-2t^z \log t\left[H_t/N - t^{\phi-1} - Ld^{\alpha/2}t^{-\gamma\alpha}\right]^2}.$$

Let $H_t/N - t^{\phi-1} - Ld^{\alpha/2}t^{-\gamma\alpha} = t^{\phi-1}$. Thus, we should have $A_1 = N(2 + Ld^{\alpha/2})$. Finally when $a_1 = a_2$, **C2** will hold when **C1** holds.

Step2: Given $H_t = N(2 + Ld^{\alpha/2})t^{\phi-1}$, $\gamma = (1-\phi)/\alpha$, $\theta = \phi - 1$, $a_1 = a_2 = 4N(1 + Ld^{\alpha/2})$, all of the conditions **C1**, **C2** and **C3** holds and we have

$$\begin{aligned}
P(\mathcal{O}_1(t)) &\leq Ne^{-2t^{z+2\phi-2}\log t}, \\
P(\mathcal{O}_2(t)) &\leq Ne^{-2t^{z+2\phi-2}\log t}, \\
P(\mathcal{O}_3(t)) &= 0, \\
P(\Xi^i_{u,l}(t)) &\leq 2NF_{\max}e^{-2t^{z+2\phi-2}\log t},
\end{aligned}$$

for all $u \in \tilde{\mathcal{J}}_i$, $l \in \mathcal{I}_T$. Let $\phi = 1 - z/2$. Then we have $P(\mathcal{O}_1(t)) \leq N/t^2$, $P(\mathcal{O}_2(t)) \leq N/t^2$, $P(\Xi^i_{u,l}(t)) \leq 2NF_{\max}/t^2$. Thus $E[X^i_{u,l_i(t)}] \leq 2NF_{\max}\beta_2$ for all $u \in \tilde{\mathcal{J}}_i$, and

$$\begin{aligned}
P((\mathcal{H}^i_{l_i(t)}(t))^C, \mathcal{W}^i(t)) &\leq \sum_{u\in\tilde{\mathcal{J}}_i} E[X^i_{u,l_i(t)}(t)]/t^\phi \\
&\leq (M-1)N^2F_{\max}\beta_2/t^\phi.
\end{aligned}$$

Hence by using (11)

$$P(\mathcal{V}^i(t), \mathcal{W}^i(t)) \leq \frac{2N}{t^2} + \frac{(M-1)N^2F_{\max}\beta_2}{t^{1-z/2}}.$$

Then,

$$\begin{aligned}
E[R^s_i(T)] &\leq Y_R \sum_{t=1}^{T} P(\mathcal{V}^i(t), \mathcal{W}^i(t)) && (19)\\
&\leq Y_R \left(2N\beta_2 + (M-1)N^2F_{\max}\beta_2\left(1 + \frac{2T^{z/2}}{z}\right)\right) && (20)\\
&= Y_R(2N\beta_2 + (M-1)N^2F_{\max}\beta_2) + T^{z/2}(2Y_R(M-1)N^2F_{\max}\beta_2/z), && (21)
\end{aligned}$$

where the inequality follows from the result of Appendix A.

■

Note that $E[R^s_i(T)]$ is linear in $\mathcal{J}_i$ instead of $\mathcal{L}_i$. In the next lemma, we bound the regret due to near-optimal arm selections by agent i by time T, i.e., $E[R^n_i(T)]$.

Lemma 3. *When CBMR is run with parameters* $D_1(t) = t^z \log t$, $D_{2,u}(t) = \binom{F_{\max}}{n(u)} t^z \log t$, $u \in \tilde{\mathcal{J}}_i$, $D_3(t) = t^z \log t$ *and* $m_T = \lceil T^\gamma \rceil$, *where* $0 < z < 1$ *and* $0 < \gamma < 1/d$, *given that* $2LY_R(\sqrt{d})^\alpha t^{-\gamma\alpha} + 2(Y_R+2)t^{-z/2} \leq a_1 t^\theta$ *we have*

$$E[R_i^n(T)] \leq 4N(N+1)(1+Ld^{\alpha/2})\frac{T^{1-z/2}}{1-z/2} + 2Y_R N^2 F_{\max}\beta_2.$$

Proof: Consider the case that agent i chooses a near optimal set of arms given in (7), with $a_1 = 4N(1+Ld^{\alpha/2})$ and $\theta = \phi - 1 = -z/2$ as given in the proof of Lemma 2. At all the time steps t in which agent i had chosen a near optimal set of arms $\mathcal{B}_t$, when all agents $j(u)$ will recommend their $n(u)$ near optimal items to agent i for $u \in \mathcal{B}_t \cap \tilde{\mathcal{J}}_i$, agent i's one step expected regret at time t will be $(N+1)a_1 t^\theta$. However, there can be some time steps which are exploitation steps for agent i, but when agent i selects a near optimal set of arms $\mathcal{B}_t$, an agent $j(u)$ may recommend one of its suboptimal items to agent i. This event is given by $\Xi^i_{u,l_i(t)}(t)$ in the proof of Lemma 2, and it is bounded by $2NF_{\max}/t^2$ for all $u \in \tilde{\mathcal{J}}_i$. Since at most N arms can be chosen at each time step, the expected regret due to such events by time T is upper bounded by $Y_R N(2NF_{\max})\beta_2 = 2Y_R N^2 F_{\max}\beta_2$. Thus we have

$$\begin{aligned} E[R_i^n(T)] &\leq \sum_{t=1}^{T}(N+1)4N(1+Ld^{\alpha/2})t^\theta + 2Y_R N^2 F_{\max}\beta_2 \\ &\leq 4N(N+1)(1+Ld^{\alpha/2})\frac{\theta + T^{1+\theta}}{1+\theta} + 2Y_R N^2 F_{\max}\beta_2, \end{aligned}$$

where last inequality follows from the bound in Appendix A. ■

Combining the above lemmas, we obtain the finite time, uniform regret bound for agents using CBMR given in the following theorem.

Theorem 1. *Let CBMR run with control functions* $D_1(t) = D_3(t) = t^{2\alpha/(3\alpha+d)} \log t$, $D_{2,u}(t) = \binom{F_{\max}}{n(u)} t^{2\alpha/(3\alpha+d)} \log t$, $u \in \tilde{\mathcal{J}}_i$, *and* $m_T = \lfloor T^{1/(3\alpha+d)} \rfloor$. *Then,*

$$\begin{aligned} R_i(t) \leq\ & T^{\frac{2\alpha+d}{3\alpha+d}} \times \left(Y_R 2^d Z_i \log T + 4N(N+1)(1+Ld^{\alpha/2})\frac{3\alpha+d}{2\alpha+d} \right) \\ & + T^{\frac{d}{3\alpha+d}} \times \left(Y_R 2^d (|\mathcal{J}_i| + (M-1)N) \right) \\ & + T^{\frac{\alpha}{3\alpha+d}} \times \left(Y_R (M-1) M^2 F_{\max}\beta_2 \frac{3\alpha+d}{\alpha} \right) \\ & + 2Y_R N\beta_2 + Y_R N^2 F_{\max}\beta_2 (M+1), \end{aligned}$$

where $Z_i = |\mathcal{J}_i| + (M-1)\sum_{a=1}^{N}\binom{F_{\max}}{a}$. *Concisely we have* $R_i(T) = O\left(|\mathcal{J}_i| T^{\frac{2\alpha+d}{3\alpha+d}}\right)$.

Proof: The highest orders of regret come from explorations, trainings and near optimal arms, which are $O(T^{(2\alpha+1)z/(2\alpha)})$ (Lemma 1) and $O(T^{1-z/2})$ (Lemma 3) respectively. Note that the first one is increasing in z, while the second one is decreasing in z. Thus the best value of z is when they are equal to each other which is given by $z = 2\alpha/(3\alpha+d)$. We have

$$R_i(T) = E[R_i^e(T)] + E[R_i^s(T)] + E[R_i^n(T)].$$

Thus summing the results of Lemmas 1, 2 and 3, we get

$$\begin{aligned} R_i(t) \leq\ & T^{\frac{2\alpha+d}{3\alpha+d}} \times \left(Y_R 2^d Z_i \log T + 4N(N+1)(1+Ld^{\alpha/2}) \frac{3\alpha+d}{2\alpha+d} \right) \\ & + T^{\frac{d}{3\alpha+d}} \times \left(Y_R 2^d (|\mathcal{J}_i| + (M-1)N) \right) \\ & + T^{\frac{\alpha}{3\alpha+d}} \times \left(Y_R (M-1) M^2 F_{\max} \beta_2 \frac{3\alpha+d}{\alpha} \right) \\ & + 2Y_R N \beta_2 + Y_R N^2 F_{\max} \beta_2 (M+1). \end{aligned}$$

■

The result of Theorem 1 indicates that the regret of CBMR is sublinear in time and linear in $|\mathcal{J}_i|$. This proves that CBMR's performance converges to the best distributed recommendation strategy given the purchase probabilities are exactly known by the sellers. The regret increases with the dimension of the context space d.

IV. Numerical Results

We simulate CBMR using a distributed data set adapted based on Amazon data [18]. The Amazon product co-purchasing network data set includes product IDs, sales ranks of the products, and for each product the IDs of products which are frequently purchased with that product. This data is collected by crawling the Amazon website [18] and contains $410,236$ products and $3,356,824$ edges between products that are frequently co-purchased together. We simulate CBMR using the following distributed data set adapted based on Amazon data. For a set of N_1 chosen products, we take that product and the F_1 products that are frequently co-purchased with that product.

The set of products that are taken in the first step of the above procedure is denoted by $\mathcal{C}_h$. The set of all products $\mathcal{F}$ contains these N_1 products and the products co-purchased frequently with them, which we denote by set $\mathcal{C}_f$. We assume that each item has a unit price of 1, but have different purchase probabilities for different types of users. Since user information is not present in the data set, we generate it by assuming that a user searches for a specific item. This search query will then be the context information of the user. The context space is discrete, thus we set $\mathcal{I}_T = \mathcal{C}_h$. Based on this, the agent that the user arrives to recommends N items to the user. The agent's goal is to maximize the total number of items sold to the users.

We generate the purchase probabilities in the following way: When a product recommended for context x is in the set of frequently co-purchased products, the purchase probability of that product will be g_c. When it is not, the purchase probability of that product will be g_{nc}, for which we have $g_c > g_{nc}$. We assume that there are 3 agents and evaluate the performance of agent 1 based on the number of users arriving to agent 1 with a specific context x^*, which we take as the first item in set $\mathcal{C}_h$. We assume that $T = 100,000$, which means that $100,000$ users with context x^* arrive to agent 1. Since the arrival rate of context x^* can be different for the agents, we assume arrivals with context x^* to other agents are drawn from a random process. We take $N_1 = 20$, $F_1 = 2$ and $N = 2$. As a result, we get 30 distinct items in $\mathcal{F}$ which are distributed among the agents such that $|\mathcal{F}_i| = 10$.

A. Effect of commission on the performance

We assume that agent 1 has one of the frequently co-purchased items for context x^*, while agent 3 has the other frequently co-purchased item. The total reward of agent 1 as a function of the commissions $c_{1,2} = c_{1,3} = c$ is given

in Table II. We note that there is no increase in the total reward when the commission is increased to 0.1, because this amount is not enough to incentivize agent 1 to recommend other agent's items. However, for commissions greater than 0.1, the optimal policy recommends the two frequently co-purchased items together, hence agent 1 learns that it should get recommendations from other agents. Therefore, when commission is greater than 0.1, the total reward of the agent is increasing in the commission. Selecting commissions adaptively over time is a future research topic.

B. Effect of the set of items of each agent on the performance

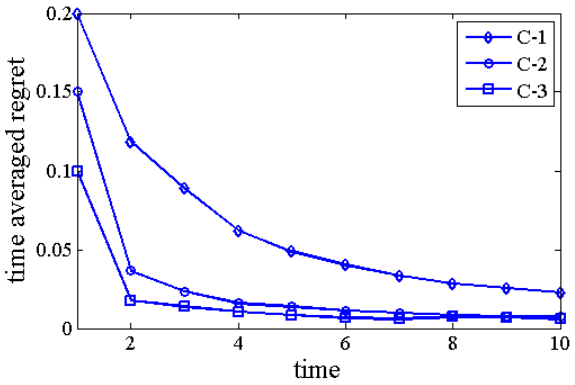

Fig. 4. Time averaged regret of CBMR for independent purchase probabilities when agent 1 has both frequently co-purchased items (C-1), only one of the frequently co-purchased items (C-2) and none of the frequently co-purchased items (C-3).

In C-1 agent 1 has both items that are frequently co-purchased in context x^*, in C-2 it has one of the items that is frequently co-purchased in context x^*, and in C-3 it has none of the items that are frequently co-purchased in context x^*. The total reward of agent 1 for these cases is 17744, 14249 and 9402 respectively, while the total expected reward of the optimal policy is 20000, 15000 and 10000 respectively. Note that the total reward for C-3 is almost half of the total reward for C-1 since the commission agent 1 gets for a frequently co-purchased item is 0.5. The time averaged regret of CBMR for all these cases is given in Figure 4. We see that the convergence rate for C-1 is slower than C-2 and C-3. This is due to the fact that in all of the trainings step in C-1 a suboptimal set of items is recommended, while for C-2 and C-3 in some of the training steps the optimal set of items is recommended.

APPENDIX A
A BOUND ON DIVERGENT SERIES

For $p > 0$, $p \neq 1$, $\sum_{t=1}^{T} 1/(t^p) \leq 1 + (T^{1-p} - 1)/(1-p)$.

Proof: See [19]. ■

Commission c	0	0.1	0.2	0.3	0.4	0.5
Reward (CBMR)	10471	10422	11476	12393	13340	14249

TABLE II

THE TOTAL REWARD OF AGENT 1 AS A FUNCTION OF THE COMMISSION IT CHARGES TO OTHER AGENTS.

REFERENCES

[1] Kwang-Cheng Chen, Mung Chiang, and H.Vincent Poor, "From technological networks to social networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 9, pp. 548–572, 2013.

[2] Vikram Krishnamurthy, "Quickest detection pomdps with social learning: Interaction of local and global decision makers," *Information Theory, IEEE Transactions on*, vol. 58, no. 8, pp. 5563–5587, 2012.

[3] Vikram Krishnamurthy and H Vincent Poor, "Social learning and bayesian games in multiagent signal processing: How do local and global decision makers interact?," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 43–57, 2013.

[4] Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi, "Non-bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[5] Angelia Nedic and Asuman Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *Automatic Control, IEEE Transactions on*, vol. 54, no. 1, pp. 48–61, 2009.

[6] Yash Deshpande and Andrea Montanari, "Linear bandits in high dimension and recommendation systems," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 1750–1754.

[7] Pushmeet Kohli, Mahyar Salek, and Greg Stoddard, "A fast bandit algorithm for recommendations to users with heterogeneous tastes," 2013.

[8] Thomas Hofmann, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 89–115, 2004.

[9] Nachiketa Sahoo, Param Vir Singh, and Tridas Mukhopadhyay, "A hidden markov model for collaborative filtering," *MIS Quarterly*, vol. 36, no. 4, pp. 1329–1356, 2012.

[10] Greg Linden, Brent Smith, and Jeremy York, "Amazon. com recommendations: Item-to-item collaborative filtering," *Internet Computing, IEEE*, vol. 7, no. 1, pp. 76–80, 2003.

[11] Umberto Panniello, Alexander Tuzhilin, and Michele Gorgoglione, "Comparing context-aware recommender systems in terms of accuracy and diversity," *User Modeling and User-Adapted Interaction*, pp. 1–31, 2012.

[12] Koji Miyahara and Michael J Pazzani, "Collaborative filtering with the simple bayesian classifier," in *PRICAI 2000 Topics in Artificial Intelligence*. 2000, pp. 679–689, Springer.

[13] Mark OConnor and Jon Herlocker, "Clustering items for collaborative filtering," in *Proceedings of the ACM SIGIR workshop on recommender systems*. UC Berkeley, 1999, vol. 128.

[14] Guy Shani, Ronen I Brafman, and David Heckerman, "An mdp-based recommender system," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 453–460.

[15] Zhenlei Yan and Jie Zhou, "User recommendation with tensor factorization in social networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3853–3856.

[16] Shang Shang, Sanjeev R Kulkarni, Paul W Cuff, and Pan Hui, "A randomwalk based model incorporating social information for recommendations," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.

[17] Marko Balabanović and Yoav Shoham, "Fab: content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[18] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, pp. 5, 2007.

[19] Edward Chlebus, "An approximate formula for a partial sum of the divergent p-series," *Applied Mathematics Letters*, vol. 22, no. 5, pp. 732–737, 2009.