52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Multi-armed Bandits in the Presence of Side Observations in Social Networks

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B. Shroff

Abstract— **We consider the decision problem of an external agent choosing to execute one of M actions for each user in a social network. We assume that observing a user's actions provides valuable information for a larger set of users since each user's preferences are interrelated with those of her social peers. This falls into the well-known setting of the multi-armed bandit (MAB) problems, but with the critical new component of side observations resulting from interactions between users. Our contributions in this work are as follows: 1) We model the MAB problem in the presence of side observations and obtain an asymptotic lower bound (as a function of the network structure) on the regret (loss) of any uniformly good policy that achieves the maximum long term average reward. 2) We propose a randomized policy that explores actions for each user at a rate that is a function of her network position. We show that this policy achieves the asymptotic lower bound on regret associated with actions that are unpopular for all the users. 3) We derive an upper bound on the regret of existing Upper Confidence Bound (UCB) policies for MAB problems modified for our setting of side observations. We present case studies to show that these UCB policies are agnostic of the network structure and this causes their regret to suffer in a network setting. Our investigations in this work reveal the significant gains that can be obtained even through static network-aware policies.**

I. Introduction

The unprecedented development and adoption of on-line social networks such as Facebook, LinkedIn, MySpace, in the last few years has had a transformative impact on the scale and nature of social interaction between people. Users often endorse consumer products and engage in word-of-mouth advertising on these on-line social networks. In this paper, we investigate the learning/earning paradigm of multi-armed bandit (MAB) problems ([1], [2], [3]) in the presence of side observations resulting from the interactions on on-line social networks. We consider the setting of recommender/advertising systems embedded within an online social network that make product recommendations to users based on their individual history as well as global network information. The external agent (content provider) in these settings must choose advertisements (recommendations) from one of M advertising categories[1] to display to each user at a time. Users respond by clicking (or ignoring) these ads and we assume that the external agent uses the click-through-rate as a proxy measure for the revenue obtained from the advertisements. The external agent wishes to maximize his revenue by personalizing the display of advertisements to users based on their preferences. However, these preferences are often unknown to the external agent and learning from the click-activity by experimentation may be warranted. For each user, the external agent faces the well known multi-armed bandit problem setting, but with the additional component of side-information that each observation provides for at least a subset of the remaining users.

In this work, we model the MAB problem in the presence of side observations as follows: each time an action is chosen for a given user, the external agent receives a reward associated with her action and also observes the reward associated with that action for each of her neighbors in the social network. Such side observations are made possible in settings of on-line social networks like Facebook through mechanisms such as: 1) the past knowledge of interdependence between users' (average) preferences obtained either directly or indirectly from user behaviors such friends reposting, "liking", commenting on each other's activity on on-line social networks; 2) surveying a user's neighbors regarding their interest in the user's activity[2]. Under this model, our contributions in this work are as follows:

- We model the MAB problem in the presence of side observations and derive an asymptotic lower bound (as a function of the network structure) on the regret (loss) of any uniformly good policy that achieves the maximum long term average reward.
- We propose and investigate the performance of a randomized policy, we call ϵ-greedy-LP policy, that *explores actions for each user at a rate that is a function of her network position.* We show that this policy achieves the asymptotic lower bound on regret associated with actions that are suboptimal for all users in the network.
- We derive an upper bound on the regret of existing UCB policies for MAB problems applied to our setting of side observations. We show, using case studies, that these existing UCB policies are agnostic of the network structure, hence, could have a worse performance when compared to the ϵ-greedy-LP policy.

S. Buccapatnam and A. Eryilmaz are with the Department of Electrical and Computer Engineering (ECE) and N. B. Shroff is with the Departments of ECE and Computer Science and Engineering, The Ohio State University. Email: {buccapatnam-tiru.1, eryilmaz.2, shroff.11}@osu.edu

This work is partly supported by ARO MURI W911NF-12-1-0385, QNRF Grant NPRP 09-1168-2-455 and NSF grant CAREER-CNS-0953515.

[1] Some examples of advertising categories are sports, clothing, vacation packages, etc.

[2] This is possible when the on-line network has an additional survey feature that generates side observations. Specifically, when user i is given a recommendation for movie j, her neighbors are queried as follows: "User i was recommended movie j. Would you be interested in this movie too?"

978-1-4673-5716-6/13/$31.00 ©2013 IEEE

Our work in this paper focuses on the study of N parallel M-armed bandits problem coupled by side observations. The existing UCB policies are agnostic of the network structure and this disables them to take full advantage of the side observations. Since the exploration component of ϵ-greedy-LP policy is network-aware, it is able to achieve the optimal regret associated with actions that are unpopular for all users. However, it is a static policy that explores all suboptimal actions for each user equally and could suffer from over-exploration when the actions are not unpopular for all users. Our investigations motivate the design of adaptive network-aware allocation strategies in order to take full advantage of the side-observations present in the network. We provide interesting new directions for future work along these lines in Section VI.

The rest of the paper is organized as follows. In Section II, we formulate the MAB problem in the presence of side observations. We then briefly discuss existing work in the setting of MAB problems in Section III. Our main results are presented in Section IV while we present case studies and numerical results in Section V.

II. Model

In this section, we formally define the M-armed bandit problem in the presence of side observations in a social network. Let $\mathcal{N} = \{1, \ldots, N\}$ denote the set of users (nodes) in the social network and $\mathcal{M} = \{1, \ldots, M\}$ denote the set of actions. An external agent must choose an action $a \in \mathcal{M}$ at each time t for each user i. Let $X_{ia}(t)$ denote the reward obtained by the external agent on choosing action a for user i at time t. The random variable $X_{ia}(t)$ has an unknown probability distribution F_{ia} with the univariate density function $f(x; \theta_{ia})$ and unknown parameters θ_{ia}. Let μ_{ia} be the mean of the random variable $X_{ia}(t)$. We assume that $\{X_{ia}(t), t \geq 0\}$ are *i.i.d* for each i and a and $\{X_{ia}(t), \forall i \in \mathcal{N}, \forall a \in \mathcal{M}\}$ are independent for each time t. We further assume that the distributions, F_{ia} have a bounded support of $[0, 1]$ for each i and a.

Next, we describe the side observation model captured by the network structure. The users $\mathcal{N}$ are embedded in a social network represented by the adjacency matrix $G = [g(i,j)]_{i,j \in \mathcal{N}}$ where $g(i,j) \in \{0,1\}$ and $g(i,i) = 1\ \forall i$. Let $\mathcal{N}_i$ be the set of neighbors of user i (including i), i.e, $g(j,i) = 1, \forall j \in \mathcal{N}_i$. We assume that when the external agent chooses an action a for user i, he receives a reward $X_{ia}(t)$ and also receives observations $X_{ja}(t)$, which are drawn independently from the distribution F_{ja} $\forall j \in \mathcal{N}_i$ such that $\mathbb{E}[X_{ja}(t)] = \mu_{ja}$. Such side observations are made possible in settings of on-line social networks like Facebook by surveying a user's neighbors regarding their interest in the user's activity. This is possible when the on-line network has an additional survey feature that generates side observations. Specifically, when user i is given a recommendation for movie j, her neighbors are queried as follows: "User i was recommended movie j. Would you be interested in this movie too?"[3]

An allocation strategy or policy $\boldsymbol{\phi}$ chooses the action to be played at each time for each user in the network. Formally, $\boldsymbol{\phi}$ is a sequence of random variables $\{\phi_i(t), \forall i \in \mathcal{N}, t \geq 0\}$, where $\phi_i(t) \in \mathcal{M}$ is the action chosen by policy $\boldsymbol{\phi}$ for user i at time t. Let $\mathbf{Y}_i(t)$ be the reward and side observations obtained by the policy $\boldsymbol{\phi}$ for user i at time t. Then, the event $\{\phi_i(t) = a\}$ belongs to the σ-field generated by $\{\phi_i(k), \mathbf{Y}_i(k), \forall i \in \mathcal{N}, k \leq t-1\}$. Let $T_{ia}(t)$ be the total number of times action a is chosen for user i up to time t by policy $\boldsymbol{\phi}$. Let $S_{ia}(t)$ be the total number of observations corresponding to action a available at time t for user i. For each user, rewards are only obtained for the action chosen for that user.

Definition 1: *(Regret)* The regret of policy $\boldsymbol{\phi}$ at time t for a fixed $\boldsymbol{\mu} = (\mu_{i1}, \ldots, \mu_{iM})_{i \in \mathcal{N}}$ is defined by

$$\begin{aligned} R_{\boldsymbol{\mu}}(t) &= \sum_{i=1}^{N} \mu_i^* t - \sum_{i=1}^{N} \sum_{a=1}^{M} \mu_{ia} \mathbb{E}[T_{ia}(t)], \\ &= \sum_{i=1}^{N} \sum_{a=1}^{M} (\mu_i^* - \mu_{ia})\, \mathbb{E}[T_{ia}(t)], \end{aligned}$$

where $\mu_i^* = \max_{a \in \mathcal{M}} \mu_{ia}$.

The objective is to find policies that minimize the rate of growth of regret with time for every fixed network G. We focus our investigation on the class of uniformly good policies defined below:

Definition 2: *(Uniformly good policies)* An allocation rule $\boldsymbol{\phi}$ is said to be uniformly good if for every fixed $\boldsymbol{\mu}$, the following condition is satisfied as $t \to \infty$:

$$R_{\boldsymbol{\mu}}(t) = o(t^b) \text{ for every } b > 0.$$

The above condition implies that uniformly good policies achieve the optimal long term average reward of $\sum_{i=1}^{N} \mu_i^*$. Next, we define two structures that will be useful later to bound the performance of allocation strategies in terms of the network structure G.

Definition 3: A *dominating set* $\mathcal{D}$ of a network G is such that every node in $\mathcal{N}$ is either in $\mathcal{D}$ or has at least one neighbor in $\mathcal{D}$. Let $\gamma(G)$ denote the size of the minimum dominating set of network G.

Definition 4: A *clique covering* $\mathcal{C}$ of a network G is a partition of nodes in $\mathcal{N}$ into sets $C \in \mathcal{C}$ such that the sub-network formed by each C is a clique. Let $\chi(G)$ be the smallest number of cliques into which the nodes of the network G can be partitioned.

III. Related Work

The seminal work of [1] shows that the asymptotic lower bound on the regret of any uniformly good policy scales logarithmically with time with a multiplicative

[3] Since, the neighbors do not have any information on whether the user i accepted the promotion, they act independently according to their own preferences in answering this survey. The network itself provides a better way for surveying and obtaining side observations.

constant, which is a function of the Kullback Leibler distance of the distributions of the actions. Further, the authors of [1] provide constructive policies called Upper Confidence Bound (UCB) policies based on the concept of optimism in the face of uncertainty, which achieve the lower bound asymptotically. Later works of [2] and [3] propose simpler sample-mean based UCB policies that achieve the logarithmic lower bound up to a multiplicative constant factor. In [3], the authors propose UCB1 and decreasing-ϵ-greedy policies that achieve logarithmic regret uniformly over time, rather than only asymptotically as in the previous works.

Recent works of [4] and [5] are related to our paper. In both [4] and [5], the authors consider a setting where the actions are embedded in a network and choosing an action reveals side observations on the neighboring actions. On the other hand, in our work, we consider N parallel M-armed bandits and assume that each time an action is chosen for a user, side observations associated with the same action are revealed for her neighbors. [4] considers an adversarial setting with no statistical assumptions on the reward distributions (see [6] for details on adversarial MABs) while [5] considers stochastic bandits. The policies proposed in [4] achieve the best possible regret in the adversarial setting with side observations and the bounds of these policies are in terms of the independence number of the network. In [5], the authors propose modified UCB1 policies and the upper bounds are in terms of $\chi(G)$.

IV. Main Results

In this section, we first obtain an asymptotic lower bound on the regret of uniformly good policies for the setting of MABs with side observations. This lower bound is expressed as the optimal value of a linear program (LP), where the constraints of the LP capture the connectivity of each user in the network.

Motivated by the LP associated with the lower bound, we propose a network-aware randomized policy called the ϵ-greedy-LP policy. Similar to the ϵ-greedy policy introduced in [3], the exploration component in our ϵ-greedy-LP policy is proportional to (1/time). However, our policy has the novel element of *exploration for each user at a rate proportional to her network position*. We provide an upper bound on the regret of this policy and show that, for actions that are suboptimal for all users, our policy achieves the asymptotic lower bound up to a constant multiplier independent of network structure. Finally, we investigate the performance of a modified UCB1 policy similar to the one proposed in [3] and provide an upper bound on its regret for our setting of side observations. We omit the full proofs due to space constraints. Interested readers can refer to [7].

A. Lower Bound

In order to derive a lower bound on the regret, we need some mild regularity assumptions on the distributions F_{ia} that are similar to the ones in [1]. Let $D(\theta_{ia}||\theta_{ib})$ denote the Kullback Leibler (KL) distance between distributions with density functions $f(x;\theta_{ia})$ and $f(x;\theta_{ib})$ and with means μ_{ia} and μ_{ib} respectively.

Assumption *1:* We assume that $f(\cdot;\cdot)$ is such that $0 < D(\theta_{ia}||\theta_{ib}) < \infty$ whenever $\mu_{ib} > \mu_{ia}$.

Assumption *2:* For any $\epsilon > 0$ and θ_{ia}, θ_{ib} such that $\mu_{ib} > \mu_{ia}$, there exists $\Delta > 0$ for which $|D(\theta_{ia}||\theta_{ib}) - D(\theta_{ia}||\theta_{ic})| < \epsilon$ whenever $\mu_{ib} < \mu_{ic} < \mu_{ib} + \Delta$.

Assumption *3:* For each $i \in \mathcal{N}$ and $a \in \mathcal{M}$, $\theta_{ia} \in \Theta$ where the set Θ satisfies the following denseness condition: for all $\theta_{ia} \in \Theta$ and for all $\Delta > 0$, there exists $\theta_{ib} \in \Theta$ such that $\mu_{ia} < \mu_{ib} < \mu_{ia} + \Delta$.

Recall that $S_{ia}(t)$ is the total number of observations corresponding to action a available at time t for user i. The following proposition is obtained by modifying the proof of Theorem 2 in [1]. It provides an asymptotic lower bound on the total number of observations for each suboptimal action obtained by any uniformly good policy under the model described in Section II:

Proposition *1:* Suppose Assumptions 1, 2, and 3 hold. Then, under any uniformly good policy ϕ, we have that, for each user i and each action a with $\mu_{ia} < \mu_i^*$,

$$\liminf_{t\to\infty} \frac{\mathbb{E}[S_{ia}(t)]}{\log(t)} \geq \frac{1}{D(\theta_{ia}||\theta_i^*)}.$$

Proof: (Sketch) Any uniformly good policy must achieve a regret of $o(t^b)$ for all $b > 0$ for each user in the network. Also, the total number of observations of an action for a user is greater than (or equal to) the total number of times the action is chosen for that user. Using these two facts, we can modify the proof of Theorem 2 in [1] to get the above proposition for the total number of observations for each suboptimal action. The main idea in the proof of Theorem 2 in [1] is that unless we have enough (of the order $\Omega(\log(t))$) number of observations for each suboptimal action, the empirical KL-distance will not converge to the actual KL-distance and hence, the policy cannot distinguish the suboptimal action from the optimal action to obtain $o(t^b)$ regret. See [7] for a detailed proof. ■

Each time an action is chosen for a user i, we receive side observations for all her neighbors $\mathcal{N}_i$. Hence, $S_{ia}(t) = \sum_{v\in\mathcal{N}_i} T_{va}(t)$. This gives us the following corollary to Proposition 1:

Corollary *1:* Suppose Assumptions 1, 2, and 3 hold. Let $\mathcal{U}_a = \{i : \mu_{ia} < \mu_i^*\}$ be the set of users for whom action a is suboptimal. Then, under any uniformly good policy ϕ, the expected regret is asymptotically bounded below as follows:

$$\liminf_{t\to\infty} \frac{R_{\boldsymbol{\mu}}(t)}{\log(t)} \geq \sum_{a\in\mathcal{M}} \min_{i\in\mathcal{U}_a} (\mu_i^* - \mu_{ia}) c_a,$$

where c_a is the optimal value of the following linear program (LP):

$$P_1 : \min \sum_{i \in \mathcal{U}_a} w_i$$
$$\text{subject to } G_i \cdot \mathbf{w} \geq \frac{1}{D(\theta_{ia}||\theta_i^*)},\ \forall i \in \mathcal{U}_a,$$
$$\text{and } w_i \geq 0,\ \forall i \in \mathcal{N}.$$

Here, G_i is the i^{th} row of G.

Proof: (Sketch) The constraints in the LP P_1 are obtained by relating the total number of observations for each user's action to the total number of times an action is chosen in the user's neighborhood. ■

Next, consider the following LP:

$$P_2 : \min \sum_{i \in \mathcal{N}} z_i$$
$$\text{subject to } G_i \cdot \mathbf{z} \geq 1,\ \forall i \in \mathcal{N},$$
$$\text{and } z_i \geq 0,\ \forall i \in \mathcal{N}.$$

In the next proposition, we provide a lower bound on c_a using the optimal solution $\mathbf{z}^* = (z_i^*)_{i \in \mathcal{N}}$ of LP P_2.

PROPOSITION 2: Let $\mathcal{U}_a = \{i : \mu_{ia} < \mu_i^*\}$ be the set of users for whom action a is suboptimal. Let $\mathcal{O}_a = \{i : \mu_{ia} = \mu_i^*\}$ be the set of users for whom action a is optimal. Let $\mathcal{I}_a = \{i \in \mathcal{U}_a : \mathcal{N}_i \cap \mathcal{O}_a \neq \emptyset\}$ be the set of users in $\mathcal{U}_a$ with neighbors in $\mathcal{O}_a$. Then,

$$c_a \geq \frac{1}{\max\limits_{i \in \mathcal{U}_a} D(\theta_{ia}||\theta_i^*)} \left(\sum_{i \in \mathcal{N}} z_i^* - \gamma(G') \right), \qquad (1)$$

where G' is the sub-network of G restricted to the set of nodes $\mathcal{O}_a \cup \mathcal{I}_a$, $\gamma(G')$ is the size of the minimum dominating set of G', and $\mathbf{z}^*$ is the optimal solution of LP P_2.

Proof: (Sketch) Using the optimal solution of LP P_1, we construct a feasible solution satisfying constraints in LP P_2 for users in $\mathcal{U}_a$. In order to satisfy the constraints for users in $\mathcal{I}_a \cup \mathcal{O}_a$, we use $z_i = 1$ for all i in the minimum dominating set of $\mathcal{I}_a \cup \mathcal{O}_a$. The feasible solution constructed in this way gives an upper bound on the optimal value of LP P_2 in terms of the optimal value of LP P_1. See [7] for a detailed proof. ■

Note that the lower bound in (1) need not be asymptotically tight and might be weaker than the trivial lower bound of 0. This happens, for example, when $\gamma(G) = \gamma(G')$. However, when $\mathcal{U}_a = \mathcal{N}$, i.e., the action a is suboptimal for all users, we can see that

$$c_a \geq \frac{1}{\max\limits_{i \in \mathcal{N}} D(\theta_{ia}||\theta_i^*)} \sum_{i \in \mathcal{N}} z_i^*.$$

Hence, when $\mathcal{U}_a = \mathcal{N}$, the asymptotic lower bound on the expected regret due to action a is given by $\Omega\left(\sum_{i \in \mathcal{N}} z_i^* \log(t)\right)$, where $\mathbf{z}^*$ completely captures the dependence of the logarithmic term of the regret on the network structure G. Motivated by the LP P_2, we next propose the ϵ-greedy-LP policy.

B. ϵ-greedy-LP policy

Let $\bar{x}_{ia}(t)$ be the sample mean of observations (rewards and side observations combined) available for action a for user i up to time t. The ϵ-greedy-LP policy is described in Algorithm 1. The policy consists of two phases for each user - exploitation and exploration. For each user i, we choose $\epsilon_i(t)$ proportional to z_i^*/t, where $\mathbf{z}^*$ is the optimal solution of LP P_2. The policy explores a randomly chosen action with probability $\epsilon_i(t)$ and exploits the action with the highest sample mean with probability $1 - \epsilon_i(t)$.

Algorithm 1 : ϵ-greedy-LP

```
Input: $c > 0$, $0 < d < 1$, optimal solution $\mathbf{z}^*$ of $P_2$.
  for each time $t$ do
    for each user $i$ do
      Let $\epsilon_i(t) = z_i^* \min\left(1, \frac{cM}{d^2 t}\right)$.
      Let $a_i^* = \arg\max_{a \in \mathcal{M}} \bar{x}_{ia}(t)$,
      where $\bar{x}_{ia}(t)$ is the sample mean of observations
      available for action $a$ for user $i$ up to time $t$.
      With probability $1-\epsilon_i(t)$, pick action $\phi_i(t) = a_i^*$ and
      with probability $\epsilon_i(t)$, pick action $\phi_i(t)$ uniformly at
      random from $\mathcal{M}$.
    end for
    for each user $i$ do
      Update sample means $\bar{x}_{v\phi_i(t)}(t+1), \forall v \in \mathcal{N}_i$.
    end for
  end for
```

The following proposition provides performance guarantees on the expected regret due to ϵ-greedy-LP policy:

PROPOSITION 3: Let $\Delta_{ia} = \mu_i^* - \mu_{ia}$. For $0 < d < \min\limits_{a \in \mathcal{M}}\{\min\limits_{i \in \mathcal{U}_a} \Delta_{ia}\}$, and $c > \max(4\alpha(4\alpha-1)d^2/3(\alpha-1)^2, 2\alpha)$ for any $\alpha > 1$, the expected regret of ϵ-greedy-LP policy described in Algorithm 1 due to each action a is at most

$$\left(\frac{c}{d^2} \max_{i \in \mathcal{U}_a} \Delta_{ia} \sum_{i \in \mathcal{U}_a} z_i^* \right) \log(t) + O(1). \qquad (2)$$

Proof: (Sketch) Since $\mathbf{z}^*$ satisfies the constraints in LP P_2, there is sufficient exploration of suboptimal actions in each user's neighborhood. This ensures that the regret from the exploitation phase is finite. Further, the regret from the exploration phase is bounded logarithmically due to the decreasing ϵ. See [7] for a detailed proof. ■

COROLLARY 2: When $\mathcal{U}_a = \mathcal{N}$, i.e., action a is suboptimal for all users in G, the ϵ-greedy-LP policy achieves $O\left(\sum_{i \in \mathcal{N}} z_i^* \log(t)\right)$ regret, where $\mathbf{z}^*$ completely captures the time dependence of the regret on the network structure G. Also, $\sum_{i \in \mathcal{N}} z_i^* \leq \gamma(G)$, where $\gamma(G)$ is the size of the minimum dominating set of the network G. Hence, the regret due to an action a with $\mathcal{U}_a = \mathcal{N}$ is $O\left(\gamma(G) \log(t)\right)$.

While the above corollary holds true for actions that are unpopular for all users, we believe that the policy is nearoptimal when different actions are optimal within well-separated clusters in the network, i.e, the sets, $\mathcal{U}_a$ partition the network into well-separated clusters.

C. UCB1-SO policy

Next, consider the UCB1-SO policy in Algorithm 2 that is a modified version of the UCB1 algorithm described in [3]. UCB1-SO has been modified to take into account both rewards and side observations while computing the UCB index of each action.

Algorithm 2 : UCB1-SO

```
for each time $t$ do
  for each user $i$ do
    Pick action $\phi_i'(t)$ such that,
        $\phi_i'(t) = \arg\max_{a \in \mathcal{M}} \bar{x}_{ia}(t) + \sqrt{\frac{2\log(t)}{s_{ia}(t)}},$
    where $s_{ia}(t)$ is the number of observations available
    for action $a$ for user $i$ at time $t$.
  end for
  for each user $i$ do
    Update sample means $\bar{x}_{v\phi_i'(t)}(t+1), \forall v \in \mathcal{N}_i$.
    $s_{v\phi_i'(t)}(t+1) = s_{v\phi_i'(t)} + 1, \forall v \in \mathcal{N}_i$.
  end for
end for
```

The following proposition provides performance guarantees on the expected regret due to UCB1-SO policy:

PROPOSITION 4: Let $\mathcal{I}_a = \{i \in \mathcal{U}_a : \mathcal{N}_i \cap \mathcal{O}_a \neq \emptyset\}$ be the set of users in $\mathcal{U}_a$ with at least one neighbor for whom action a is optimal. Then, under UCB1-SO policy described in Algorithm 2,

$$\limsup_{t\to\infty} \mathbb{E}[T_{va}(t)] < \infty, \forall v \in \mathcal{I}_a. \qquad (3)$$

Let $\Delta_{ia} = \mu_i^* - \mu_{ia}$. Then, the expected regret of the UCB1-SO policy due to each action a is at most

$$\inf_{\mathcal{C}_a} \left(\sum_{C \in \mathcal{C}_a} \frac{\max_{i \in C} \Delta_{ia}}{\min_{i \in C} \Delta_{ia}^2} \right) 8\log(t) + O(1), \qquad (4)$$

where $\mathcal{C}_a$ is the clique covering of all users in $\mathcal{U}_a \setminus \mathcal{I}_a$.

Proof: (Sketch) We use the following proof technique from [4] and [5] in our analysis. We bound the regret of UCB1-SO policy for the whole network above with the regret for the clique covering. The latter can then be bounded above using the analysis of UCB1 policy from [3]. See [7] for a detailed proof. ■

For the case of actions a with $\mathcal{U}_a = \mathcal{N}$, the upper bound in (4) is $O(\chi(G)\log(t))$, where $\chi(G)$ is the size of the minimum clique covering of all nodes in the network G. For any network G, we have $\chi(G) \geq \gamma(G)$. Hence, we have, from Corollary 2, that the upper bound in (4) is greater than or equal to regret of the ϵ-greedy-LP policy for actions a with $\mathcal{U}_a = \mathcal{N}$.

In the next section, we present case studies and numerical results to better understand the working on UCB1-SO policy and to compare the performance of the two policies described in this section.

V. CASE STUDIES AND NUMERICAL RESULTS

We see from (3) that UCB1-SO policy does not explore when there are enough side observations in the system. Despite this adaptive nature of UCB1-SO policy, the UCB1-SO policy remains unaware of the network structure. Next, we present two case studies that demonstrate this. Consider the star network G_{star} in Figure 1 with center node c such that $g(c,j) = g(j,c) = 1$ for all j and $g(i,j) = 0$ otherwise.

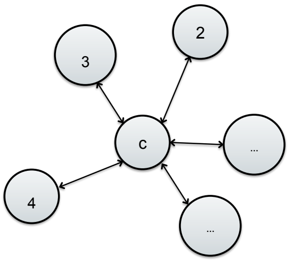

Fig. 1: A star network.

Case Study 1*:* Consider an action a such that a is optimal for user 2 and suboptimal for all other users. Then, under UCB1-SO policy, owing to Proposition 4, $\limsup_{t\to\infty} \mathbb{E}[T_{ca}(t)]$ is a finite constant. Hence, none of the other users in the network receive substantial side observations for action a. Now, UCB1-SO policy is a uniformly good policy. Hence, for all $i \neq c$, there exists a constant k such that, $\mathbb{E}[T_{ia}(t)] \gtrsim k\log(t)$ as $t \to \infty$,[4] owing to the lower bound in Proposition 1. Hence, the regret of UCB1-SO policy for action a can be bounded above as:

$$R_a(t) \gtrsim k'(N-2)\log(t) \quad \text{as } t \to \infty,$$

where k' is a constant independent of network structure and time.

Case Study 2*:* Next, consider an action a that is suboptimal for all users. Then, as we will verify soon numerically (see Figure 3) that, under UCB1-SO policy, $\mathbb{E}[T_{ca}(t)] \leq \mathbb{E}[T_{ia}(t)]$ for all i. Hence, we have

$$\mathbb{E}[S_{ia}(t)] = \mathbb{E}[T_{ca}(t)] + \mathbb{E}[T_{ia}(t)] \leq 2\mathbb{E}[T_{ia}(t)],$$

for all $i \neq c$. Once again, combining the above fact with the lower bound in Proposition 1, we have that, for some constant k, $\mathbb{E}[T_{ia}(t)] \gtrsim k\log(t)$ as $t \to \infty$ for all $i \neq c$. This gives us the following bound on the expected regret of UCB1-SO policy for action a :

$$R_a(t) \gtrsim k'(N-1)\log(t) \quad \text{as } t \to \infty,$$

where k' is a constant independent of network size and time. On the other hand, in both the case studies above, the optimal solution of the LP P_2 for the star network is $z_c^* = 1$ and

[4] We say that $f(t) \gtrsim g(t)$ as $t \to \infty$ if $\liminf_{t\to\infty} f(t)/g(t) \geq 1$.

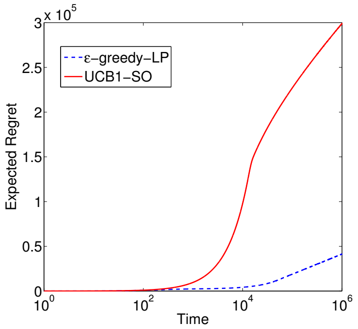

Fig. 2: Comparing ϵ-greedy-LP policy and UCB1-SO policy when all users have same action profiles in a star network.

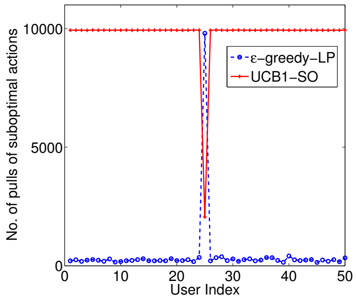

Fig. 3: Center user explores the most in ϵ-greedy-LP, while end users explore the most in the UCB1-SO in a star network.

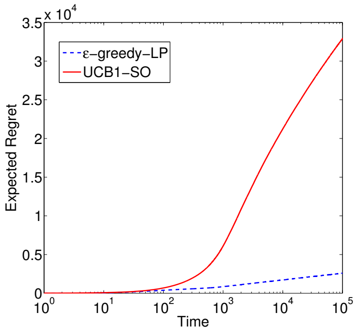

Fig. 4: ϵ-greedy-LP policy has better regret than the UCB1-SO policy even when different clusters of users have different best actions.

$z_i^* = 0$ for all other i. Hence, in both cases, the regret of ϵ-greedy-LP policy is at most $k'' \log(t)$, where k'' is a constant independent of network size and time.

Next, we present some numerical evaluation to compare the performance of ϵ-greedy-LP and UCB1-SO policies described earlier. First, we consider a star network with 50 users (user 25 is the center node) and 50 actions. For each user i, $\mu_{i2} = 0.9$ for action 2 and $\mu_{ia} = 0.7$ for all other actions. For the ϵ-greedy-LP policy, we let $c = 10$ and $d = 0.1$ and the optimal solution of the LP P_2 for the star network is $z_1^* = 1$ and $z_i^* = 0$ for all other i. In Figure 2, we see that the ϵ-greedy-LP outperforms UCB1-SO as expected from Case Study 2. From Figure 3, we see that the center node explores most in the ϵ-greedy-LP policy while the center node explores the least in UCB1-SO policy verifying our claim in Case Study 2. Next, we consider the network formed by connecting two star networks, G_1 and G_2 of size 20 each. We assume that action 2 is optimal for users in G_1 and action 3 is optimal for users in G_2. We then add 100 random links between nodes of G_1 and G_2. The addition of these links makes z_i^* non-zero for non-central users in both G_1 and G_2. All optimal actions have a mean reward 0.9 and suboptimal actions have a mean reward of 0.7. For the ϵ-greedy-LP policy, we let $c = 2$ and $d = 0.1$. In Figure 4, we see that the ϵ-greedy-LP still has better performance than UCB1-SO policy.

VI. Conclusion and Future Work

In this work, we modeled and investigated the stochastic bandit problem in the presence of side observations. The network-aware exploration of ϵ-greedy-LP policy leads to optimal performance in terms of network structure when the action is suboptimal for all users. However, this policy does not adapt to the presence of side observations and this could lead to performance loss when there is more diversity in the optimal actions in the network. On the other hand, we showed, using case studies in Section V that, UCB1-SO policy is agnostic of the network structure. Our investigations motivate the search for adaptive network-aware allocation strategies. Next, we present two promising directions for future work in the development of such strategies.

Making UCB1 network-aware: The first idea is to change the exploration term in the index of UCB1-SO policy to reflect the network structure. More concretely, we propose to investigate changes in the index of the form $\bar{x}_{ia}(t) + \sqrt{\frac{h(i)\log(t)}{s_{ia}(t)}}$, where $h(i)$ captures the network position of user i.

Making ϵ-greedy-LP adaptive: The second idea is to make the exploration component of ϵ-greedy adaptive. As we have more observations in the system with time, we can infer which users belong to the sets, $\mathcal{U}_a$, $\mathcal{I}_a$, and $\mathcal{O}_a$, respectively. Using this inference, one could modify the constraints sets and objective function in LP P_2 hoping to converge to the solution of LP P_1 for each action a.

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Mathematics*, vol. 27, pp. 1054–1078, 1995.

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, May 2002.

[4] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *NIPS*, 2011, pp. 684–692.

[5] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, "Leveraging side observations in stochastic bandits," in *Proc. of Uncertainty in Artificial Intelligence (UAI), Catalina Island, USA*, August 2012.

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 7, no. 68, 2000.

[7] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Multi-armed bandits in the presence of side observations in social networks," The Ohio State University, Tech. Rep., March 2013, "http://www2.ece.ohio-state.edu/~buccapat/mabReport.pdf".