Stochastic Bandits with Side Observations on Networks

Swapna Buccapatnam, Atilla Eryilmaz
Department of ECE
The Ohio State University
Columbus, OH - 43210
buccapat@ece.osu.edu,
eryilmaz.2@osu.edu

Ness B. Shroff
Departments of ECE and CSE
The Ohio State University
Columbus, OH - 43210
shroff.11@osu.edu

ABSTRACT

We study the stochastic multi-armed bandit (MAB) problem in the presence of side-observations across actions. In our model, choosing an action provides additional side observations for a subset of the remaining actions. One example of this model occurs in the problem of targeting users in online social networks where users respond to their friends's activity, thus providing information about each other's preferences. Our contributions are as follows: 1) We derive an asymptotic (with respect to time) 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 two policies - a randomized policy and a policy based on the well-known upper confidence bound (UCB) policies, both of which explore each action at a rate that is a function of its network position. We show that these policies achieve the asymptotic lower bound on the regret up to a multiplicative factor independent of network structure. The upper bound guarantees on the regret of these policies are better than those of existing policies. Finally, we use numerical examples on a real-world social network to demonstrate the significant benefits obtained by our policies against other existing policies.

Categories and Subject Descriptors

H.1 [**Models and Principles**]: Miscellaneous; I.2 [**Artificial Intelligence**]: Miscellaneous

Keywords

Multiarmed bandits; Side observations; Social networks

1. INTRODUCTION

Multi-armed bandit (MAB) problems have received renewed interest over the past decade because of the emergence of content recommendation, online advertising, and social networks. In the classical MAB setting, at each time, a policy must choose an action from a set of K actions with unknown probability distributions. Choosing an action i at time t gives a random reward $X_i(t)$ drawn from the distribution of action i. The regret of any policy is defined as the difference between the total reward obtained from the action with the highest average reward and the given policy's total reward. The goal is to find policies that minimize the expected regret over a given time horizon.

In our work, we consider an important MAB setting, similar to that of [12] and [6], where choosing an action i not only generates a reward from action i, but also reveals observations for a subset of the remaining actions. An example of such a scenario is as follows: a decision maker must choose one user at each time in an online social network (FaceBook, etc.) to offer a promotion [6]. Each time the decision maker offers a promotion to a user, he also has an opportunity to survey[1] the user's neighbors in the network regarding their potential interest in a similar offer (see Figure 1). Users are found to be more responsive to such surveys using social network information compared to generic surveys [3] and this effect can be leveraged to construct side-observations. In this example, choosing an action in the multi-armed bandit problem corresponds to choosing a user in the network and side-observations across actions are captured by the links in the social network. Another example is when the actions in the MAB problem are advertisements [12] - the decision maker constructs a graph of different vacation places (Hawaii, Caribbean, Paris, etc.), where links capture similarities between different places. When a customer shows interest in one of the places, he is also asked to provide his opinion about the neighboring places in the graph.

For the setting of side-observations, the authors in [12] consider adversarial bandits, while [6] considers stochastic bandits as in our current work. In [6], the authors propose modified upper-confidence bound (UCB) based policies and show that the regret of these policies is at most $O(\bar{\chi}(G)\log(t))$, where $\bar{\chi}(G)$ is the clique partition number (see Definition 1) of the side-observation network $G(\mathcal{K})$. However, it is possible to achieve a lower regret. For example, in the star network with one central action linked to all other actions, exploring the central action yields sufficient exploration for the rest of the network. In this case, the optimal regret is at most $O(\log(t))$, while $\bar{\chi}(G)$ scales as $O(K)$ for the star network.

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.

http://dx.doi.org/10.1145/2591971.2591989.

[1]This is possible when the online network has an additional survey feature that generates side observations. For example, when user i is offered a promotion, her neighbors may be queried as follows: "User i was recently offered a promotion. Would you also be interested in the offer?"

Motivated by this observation, in our work, we aim to characterize the asymptotic lower bound on the regret for a general stochastic multi-armed bandit problem in the presence of side-observations and investigate policies that achieve this lower bound by taking the network structure into account. Our main contributions are as follows:

- We model the MAB problem in the presence of side-observations and derive an asymptotic (with respect to time) lower bound (as a function of the network structure) on the regret of any uniformly good policy which achieves the maximum long term average reward. This lower bound is presented in terms of the optimal value of a linear program (LP), namely, P_1.
- Motivated by LP P_1, we propose and investigate the performance of a randomized policy, we call ϵ_t-greedy-LP policy, as well as an upper confidence bound based policy, we call UCB-LP policy. Both of these policies *explore each action at a rate that is a function of its network position.* We show that these policies are optimal in the sense that they achieve the asymptotic lower bound on the regret up to a multiplicative constant independent of the network structure under mild assumptions.
- We also show that the upper bound on the regret of our policies scales as $O(\gamma(G)\log(t))$, where $\gamma(G)$ is the size of the minimum dominating set of the network of actions. We show that the regret performance of our policies can be strictly better than those proposed in [6] for some important network structures. Finally, we use numerical results on a social network dataset obtained from Flixster[2] to empirically compare the performance of our policies against those in [6].

The model considered in our work can be viewed as a first step in the direction of more general models of inter-dependence across actions. For this model, we show that as the number of actions becomes large, significant benefits can be obtained from policies that explicitly take network structure into account. While ϵ_t-greedy-LP policy explores actions at a rate proportional to their network position, its exploration is oblivious to the average rewards of the sub-optimal actions. On the other hand, UCB-LP policy takes into account both the upper confidence bounds on the mean rewards as well as network position of different actions at each time. The rest of the paper is organized as follows. In Section 2, we briefly discuss the related existing literature. We introduce the model in Section 3 and we present our main results in Sections 4, 5, 6. Finally, we present some numerical results in Section 7 and the proofs of our theoretical results are given in Section 8. We conclude our work in Section 9.

2. RELATED WORK

The seminal work of [10] shows that the asymptotic lower bound on the regret of any uniformly good policy scales logarithmically with time with a multiplicative constant which is a function of the distributions of actions. Further, the authors of [10] provide constructive policies called Upper Confidence Bound (UCB) policies based on the concept of optimism in the face of uncertainty that asymptotically achieve the lower bound. More recently, the authors in [1] consider the case of bounded rewards and propose simpler sample-mean based UCB policies and decreasing-ϵ_t-greedy policy that achieve logarithmic regret uniformly over time, rather than only asymptotically as in the previous works. The UCB index of any action i at time t introduced in [1] is given below:

$$\bar{x}_i(t) + \sqrt{\frac{2\log(t)}{s_i(t)}}, \tag{1}$$

where $\bar{x}_i(t)$ is the average reward and $s_i(t)$ is the total number of observations available for action i at time t.

The traditional multi-armed bandit policies incur a regret that is linear in the number of suboptimal arms. This makes them unsuitable in settings such as content recommendation, advertising, etc, where the action space is typically very large. To overcome this difficulty, richer models specifying additional information across reward distributions of different actions have been studied, such as dependent bandits [13], $\mathcal{X}$-armed bandits [5], linear bandits [14], contextual side information in bandit problems [11], etc..

More recently, [12] and [6] propose to handle the large number of actions by assuming that choosing an action reveals observations for a larger set of actions. The policies proposed in [12] achieve the best possible regret in the adversarial setting (see [4] for a survey of adversarial MABs) with side-observations and the regret bounds of these policies are in terms of the independence number of the network.

For the setting of stochastic bandits with side-observations, it can be easily shown that, except in the trivial case where all actions have a neighboring action which is optimal, the regret due to any uniformly good policy is lower bounded by $\Omega(\log(t))$. A formal proof is given in [6]. Further, the authors in [6] propose two modified UCB policies, namely, UCB-N and UCB-MaxN, and show that the regret of these policies is at most $O(\bar{\chi}(G)\log(t))$, where $\bar{\chi}(G)$ is the clique partition number (see Definition 1).

In our work, we consider the stochastic MAB problem with side-observations similar to [6] and characterize the asymptotic lower bound on the regret as a function of the network structure for the stochastic MAB problem in the presence of side-observations. Further, we propose two policies: 1) ϵ_t-greedy-LP policy, 2) UCB-LP policy, which achieve the asymptotic regret lower bound up to a multiplicative constant independent of the network structure. Further, the regret of our policies is at most $O(\gamma(G)\log(t))$, where $\gamma(G)$ is the size of the minimum dominating set of the network of actions. Since, $\gamma(G) \leq \bar{\chi}(G)$ for any network G, the regret obtained by our policies can be better than that of UCB-N and UCB-MaxN proposed in [6].

3. MODEL

In this section, we formally define the K-armed bandit problem in the presence of side observations across actions. Let $\mathcal{K} = \{1, \ldots, K\}$ denote the set of actions. A decision maker must choose an action $i \in \mathcal{K}$ at each time t. Let $X_i(t)$ denote the reward obtained by the decision maker on choosing action i at time t. The random variable $X_i(t)$ has an unknown probability distribution F_i. Let μ_i be the mean

[2]Flixster is a movie recommendation network with a social graph. Datasets from this network are made publicly available by authors of [9] at `http://www.cs.ubc.ca/~jamalim/datasets/`.

of the random variable $X_i(t)$. We assume that $\{X_i(t), t \geq 0\}$ are *i.i.d* for each i and $\{X_i(t), \forall i \in \mathcal{K}\}$ are independent for each time t. We further assume that the distribution, F_i has a bounded support in $[0, b]$ for each i. We let $b = 1$ for simplicity of exposition in our work.

Side-observation model : The actions $\mathcal{K}$ form nodes in a network $G(\mathcal{K})$, represented by the adjacency matrix $G = [g(i,j)]_{i,j\in\mathcal{K}}$, where $g(i,j) \in \{0,1\}$ and $g(i,i) = 1$ for all i. Let $\mathcal{K}_i$ be the set of neighbors of action i (including i), i.e, $g(j,i) = 1, \forall j \in \mathcal{K}_i$. While all the results in our work can be easily extended to directed networks, we assume that $G(\mathcal{K})$ is undirected for simplicity of notation.

We assume that when the decision maker chooses an action i at time t, he receives a reward $X_i(t)$ and also receives observations $X_j(t)$, for all $j \in \mathcal{K}_i$ such that $\mathbb{E}[X_j(t)] = \mu_j$. In general, not all neighboring actions are equally responsive in providing side-observations. This scenario can be modeled by assuming that each action i has a known probability p_i of providing side-observations when any of its neighbors are actually chosen. We let $p_i = 1$ for all i for the sake of clarity. Our results can be easily extended to the setting of imperfect side-observations with $p_i \leq 1$. We will discuss this extension in Remark 5 in Section 6.

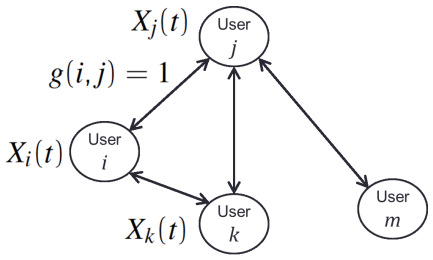

Figure 1: At time t, suppose the decision maker chooses user i to offer a promotion. He then receives a response $X_i(t)$ from user i. Using the social interconnections, he also observes responses $X_j(t)$ and $X_k(t)$ of i's neighbors j and k.

Figure 1 illustrates the side-observation model for the example of targeting users in an online social network. Such side observations are made possible in settings of online social networks like Facebook by surveying or tracking a user's neighbors reactions (likes, dislikes, no opinion, etc.) to the user's activity. This is possible when the online social network has a survey or a like/dislike indicator that generates side observations. For example, when user i is offered a promotion, her neighbors may be queried as follows: "User i was recently offered a promotion. Would you also be interested in the offer?[3]"

Objective: An allocation strategy or policy $\boldsymbol{\phi}$ chooses the action to be played at each time. Formally, $\boldsymbol{\phi}$ is a sequence of random variables $\{\phi(t), t \geq 0\}$, where $\phi(t) \in \mathcal{K}$ is the action chosen by policy $\boldsymbol{\phi}$ at time t. Let $\mathbf{X}^{\phi}(t)$ be the reward and side-observations obtained by the policy $\boldsymbol{\phi}$ at time t. Then, the event $\{\phi_i(t) = i\}$ belongs to the σ-field generated by $\{\phi(m), \mathbf{X}^{\phi}(m), m \leq t-1\}$.

[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.

Let $T_i^{\phi}(t)$ be the total number of times action i is chosen up to time t by policy $\boldsymbol{\phi}$. For each action, rewards are only obtained when the action is chosen by the policy (side-observations do not contribute to the total reward). Then, the regret of policy $\boldsymbol{\phi}$ at time t for a fixed $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ is defined by

$$R_{\boldsymbol{\mu}}^{\phi}(t) = \mu^* t - \sum_{i=1}^{K} \mu_i \mathbb{E}[T_i^{\phi}(t)] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[T_i^{\phi}(t)],$$

where $\Delta_i \triangleq \mu^* - \mu_i$ and $\mu^* \triangleq \max_{i\in\mathcal{K}} \mu_i$. Henceforth, we drop the superscript ϕ unless it is required. The objective is to find policies that minimize the rate at which the regret grows as a function of time for every fixed network $G(\mathcal{K})$. We focus our investigation on the class of uniformly good policies [10] defined below:

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 μ^*. Next, we define two structures that will be useful later to bound the performance of allocation strategies in terms of the network structure $G(\mathcal{K})$.

Definition 1. A *clique covering* $\mathcal{C}$ of a network $G(\mathcal{K})$ is a partition of all its nodes into sets $C \in \mathcal{C}$ such that the sub-network formed by each C is a clique. Let $\bar{\chi}(G)$ be the smallest number of cliques into which the nodes of the network $G(\mathcal{K})$ can be partitioned, also called the clique partition number.

Definition 2. A *dominating set* $\mathcal{D}$ of a network $G(\mathcal{K})$ is such that every node in the network 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 $G(\mathcal{K})$, also called the domination number. Note that $\gamma(G) \leq \bar{\chi}(G)$ for any network $G(\mathcal{K})$.

In the next section, we 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 action in the network.

4. REGRET LOWER BOUND

In order to derive a lower bound on the regret, we need some mild regularity assumptions (Assumptions 1, 2, and 3) on the distributions F_i that are similar to the ones in [10]. Let the probability distribution F_i have a univariate density function $f(x;\theta_i)$ with unknown parameters θ_i. Let $D(\theta||\sigma)$ denote the Kullback Leibler (KL) distance between distributions with density functions $f(x;\theta)$ and $f(x;\sigma)$ and with means $\mu(\theta)$ and $\mu(\sigma)$ respectively.

Assumption 1. *(Finiteness) We assume that $f(\cdot;\cdot)$ is such that $0 < D(\theta||\sigma) < \infty$ whenever $\mu(\sigma) > \mu(\theta)$.*

Assumption 2. *(Continuity) For any $\epsilon > 0$ and θ, σ such that $\mu(\sigma) > \mu(\theta)$, there exists $\eta > 0$ for which $|D(\theta||\sigma) - D(\theta||\rho)| < \epsilon$ whenever $\mu(\sigma) < \mu(\rho) < \mu(\sigma) + \eta$.*

ASSUMPTION 3. *(Denseness) For each $i \in \mathcal{K}$, $\theta_i \in \Theta$ where the set Θ satisfies: for all $\theta \in \Theta$ and for all $\eta > 0$, there exists $\theta' \in \Theta$ such that $\mu(\theta) < \mu(\theta') < \mu(\theta) + \eta$.*

The following proposition is obtained using Theorem 2 in [10]. It provides an asymptotic lower bound on the regret of any uniformly good policy under the model described in Section 3:

PROPOSITION 1. *Suppose Assumptions 1, 2, and 3 hold. Let $\mathcal{U} = \{i : \mu_i < \mu^*\}$ be the set of suboptimal actions. Also, let $\Delta_i = \mu^* - \mu_i$. Recall that $\mathcal{K}_i$ is the set of neighbors of i, including i, in the network $G(\mathcal{K})$. Then, under any uniformly good policy $\boldsymbol{\phi}$, the expected regret is asymptotically bounded below as follows:*

$$\liminf_{t\to\infty} \frac{R_{\boldsymbol{\mu}}(t)}{\log(t)} \geq c_{\boldsymbol{\mu}}, \tag{2}$$

where $c_{\boldsymbol{\mu}}$ is the optimal value of the following linear program (LP) P_1:

$$\begin{aligned} P_1: \ & \min \sum_{i\in\mathcal{U}} \Delta_i w_i, \\ \text{subject to: } & \sum_{j\in\mathcal{K}_i} w_j \geq \frac{1}{D(\theta_i||\theta^*)}, \ \forall i \in \mathcal{U}, \\ & w_i \geq 0, \ \forall i \in \mathcal{K}. \end{aligned}$$

PROOF. *(Sketch)* Let $S_i(t)$ be the total number of observations corresponding to action i available at time t. Then, by modifying the proof of Theorem 2 of [10], we have that, for $i \in \mathcal{U}$,

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

An observation is received for action i whenever any action in $\mathcal{K}_i$ is chosen. Hence, $S_i(t) = \sum_{j\in\mathcal{K}_i} T_j(t)$. These two facts give us the constraints in LP P_1. See Section 8 for the full proof. □

The linear program given in P_1 contains the graphical information that governs the lower bound. However, it requires the knowledge of θ_i and θ^*, which are unknown. This motivates the construction of the following linear program, LP P_2, which preserves the graphical structure while eliminating the distributional dependence on θ_i and θ^*.

$$\begin{aligned} P_2: & \min \sum_{i\in\mathcal{K}} z_i \\ \text{subject to: } & \sum_{j\in\mathcal{K}_i} z_j \geq 1, \ \forall i \in \mathcal{K}, \\ & \text{and } z_i \geq 0, \ \forall i \in \mathcal{K}. \end{aligned}$$

Let $\mathbf{z}^* = (z_i^*)_{i\in\mathcal{K}}$ be the optimal solution of LP P_2. In Sections 5 and 6, we use the above LP P_2 to modify the ϵ-greedy policy in [1] and UCB policy in [2] for the setting of side-observations. We provide regret guarantees of these modified policies in terms of the optimal value $\sum_{i\in\mathcal{K}} z_i^*$ of LP P_2. We note that the linear program P_2 is, in fact, the LP relaxation of the minimum dominating set problem on network $G(\mathcal{K})$. Since, any dominating set of network $G(\mathcal{K})$ is a feasible solution to the LP P_2, we have that the optimal value of the LP $\sum_{i\in\mathcal{K}} z_i^* \leq \gamma(G) \leq \bar{\chi}(G)$. As we will see in Remark 2, for a rich network structure, it is possible to have $\gamma(G) << \bar{\chi}(G)$.

In the next proposition, we provide a lower bound on $c_{\boldsymbol{\mu}}$ in Equation (2) using the optimal solution $\mathbf{z}^* = (z_i^*)_{i\in\mathcal{K}}$ of LP P_2.

PROPOSITION 2. *Let $\mathcal{U} = \{i : \mu_i < \mu_i^*\}$ be the set of suboptimal actions. Let $\mathcal{O} = \{i : \mu_i = \mu^*\}$ be the set of optimal actions. Then,*

$$\frac{\max_{i\in\mathcal{U}} D(\theta_i||\theta^*)}{\min_{i\in\mathcal{U}} \Delta_i} c_{\boldsymbol{\mu}} + |\mathcal{O}| \geq \sum_{i\in\mathcal{K}} z_i^* \geq \frac{\min_{i\in\mathcal{U}} D(\theta_i||\theta^*)}{\max_{i\in\mathcal{U}} \Delta_i} c_{\boldsymbol{\mu}}. \tag{3}$$

PROOF. *(Sketch)* Let $\mathcal{I} = \{i \in \mathcal{U} : \mathcal{K}_i \cap \mathcal{O} \neq \emptyset\}$ be the set of suboptimal actions with neighbors in $\mathcal{O}$. Using the optimal solution of LP P_1, we construct a feasible solution satisfying constraints in LP P_2 for actions in $\mathcal{U} \setminus \mathcal{I}$. In order to satisfy the constraints for actions in $\mathcal{I} \cup \mathcal{O}$, we use $z_i = 1$ for all i in $\mathcal{O}$. 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. For the lower bound, any feasible solution of P_2, in particular $\mathbf{z}^*$, can be used to construct a feasible solution of P_1. See Section 8 for the full proof. □

$\sum_{i\in\mathcal{K}} z_i^* = \Theta(c_\mu)$ completely captures the time dependence of regret on network structure under the following assumption:

ASSUMPTION 4. *The quantities $|\mathcal{O}|$, $\min_{i\in\mathcal{U}} \Delta_i$, and $\min_{i\in\mathcal{U}} D(\theta_i||\theta^*)$ are constants that are independent of network size K.*

Note that the constants in the above assumption are unknown to the decision maker. In the next section, we propose the ϵ_t-greedy-LP policy which achieves the regret lower bound of $c_{\boldsymbol{\mu}} \log(t)$ up to a multiplicative constant factor that is independent of the network structure and time.

5. EPSILON-GREEDY-LP POLICY

Motivated by the LPs P_1 and P_2, we propose a *network-aware* randomized policy called the ϵ_t-greedy-LP policy. We provide an upper bound on the regret of this policy and show that it achieves the asymptotic lower bound up to a constant multiplier, independent of network structure. Let $\bar{x}_i(t)$ be the empirical average of observations (rewards and side-observations combined) available for action i up to time t. The ϵ_t-greedy-LP policy is described in Algorithm 1. The policy consists of two phases - exploitation and exploration, where the exploration probability decreases as $1/t$, similar to the ϵ_t-greedy policy proposed in [1]. However, in our policy, we choose the exploration probability for action i to be proportional to z_i^*/t, where $\mathbf{z}^*$ is the optimal solution of LP P_2, while in the original policy in [1], the exploration probability is uniform over all actions.

The following proposition provides performance guarantees on the ϵ_t-greedy-LP policy:

PROPOSITION 3. *For $0 < d < \min_{i\in\mathcal{U}} \Delta_i$, any $c > 0$, and $\alpha > 1$, the probability with which a suboptimal action i is selected by the ϵ_t-greedy-LP policy, described in Algorithm 1,*

Algorithm 1 : ϵ_t-greedy-LP

Input: $c > 0$, $0 < d < 1$, optimal solution $\mathbf{z}^*$ of LP P_2.
for each time t **do**
Let $\epsilon(t) = \min\left(1, \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2 t}\right)$ and $a^* = \arg\max_{i\in\mathcal{K}} \bar{x}_i(t)$.
With probability $1-\epsilon(t)$, pick action $\phi(t) = a^*$
With probability $\epsilon(t)$, pick action $\phi(t) = i$ with probability $\frac{z_i^*}{\sum_{i\in\mathcal{K}} z_i^*}$ for all $i \in \mathcal{K}$.
end for
Update average rewards $\bar{x}_v(t+1), \forall v \in \mathcal{K}_{\phi(t)}$.

for all $t > t' = \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2}$ *is at most*

$$\left(\frac{c}{d^2 t} z_i^*\right) + \frac{4}{d^2}\left(\frac{et'}{t}\right)^{c/2\alpha} + \frac{2\delta c}{\alpha d^2}\left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right), \tag{4}$$

where $r = \frac{3(\alpha-1)^2}{8\alpha-2}$, *and* δ *is the maximum degree in the network.*

PROOF. *(Sketch)* Since $\mathbf{z}^*$ satisfies the constraints in LP P_2, there is sufficient exploration within each suboptimal action's neighborhood. The proof is then a combination of this fact and the proof of Theorem 3 in [1]. See Section 8 for the full proof. □

In the above proposition, for large enough c, we see that second and third terms are $O(1/t^{1+\epsilon})$ for some $\epsilon > 0$ [1]. Using this fact, the following corollary bounds the expected regret of the ϵ_t-greedy-LP policy:

COROLLARY 1. *Choose parameters c and d such that,*

$$0 < d < \min_{i\in\mathcal{U}} \Delta_i, \quad \text{and} \quad c > \max(2\alpha d^2/r, 2\alpha),$$

for any $\alpha > 1$. *Then, the expected regret at time t of the* ϵ_t*-greedy-LP policy described in Algorithm 1 is at most*

$$\left(\frac{c}{d^2}\sum_{i\in\mathcal{U}} \Delta_i z_i^*\right)\log(t) + O(K), \tag{5}$$

where the $O(K)$ *term captures constants independent of time but dependent on the network structure.*

Remark 1. Under Assumption 4, we can see from Proposition 2 and Corollary 1 that, ϵ_t-greedy-LP algorithm is *order optimal* achieving the lower bound $O\left(\sum_{i\in\mathcal{U}} z_i^* \log(t)\right) = O\left(c_{\boldsymbol{\mu}} \log(t)\right)$ as the network and time scale.

While ϵ_t-greedy-LP policy is network aware, its exploration is oblivious to the average rewards of the sub-optimal actions. Further, its performance guarantees depend on the knowledge of $\min_{i\in\mathcal{U}} \Delta_i$, which is the difference between the best and the second best optimal actions. On the other hand, the UCB-LP policy proposed in the next section is network-aware taking into account the average rewards of suboptimal actions. This could lead to better performance compared to ϵ_t-greedy-LP policy in certain situations, for example, when the highly connected action is also highly suboptimal.

6. UCB-LP POLICY

In this section we propose the UCB-LP policy defined in Algorithm 2 and obtain upper bounds on its regret. The UCB-LP policy is based on the improved UCB policy proposed in [2], which can be summarized as follows: the policy estimates the values of Δ_i in each round by a value $\tilde{\Delta}_m$ which is initialized to 1 and halved in each round m. By each round m, the policy draws $n(m)$ observations for each action in the set of actions not eliminated by round m, where $n(m)$ is determined by $\tilde{\Delta}_m$. Then, it eliminates those actions whose UCB indices perform poorly. Our policy differs from the one in [2] by accounting for the presence of side-observations - this is achieved by choosing each action according to the optimal solution of LP P_2, while ensuring that $n(m)$ observations are available for each action not eliminated by round m.

Algorithm 2 : UCB-LP policy

Input: Set of actions $\mathcal{K}$, time horizon T, and optimal solution $\mathbf{z}^*$ of LP P_2.

Initialization: Let $\tilde{\Delta}_0 := 1$, $S_0 := \mathcal{K}$, and $B_0 := \mathcal{K}$
for round $m = 0, 1, 2, \ldots, \lfloor \frac{1}{2}\log_2 \frac{T}{e} \rfloor$ **do**

Action Selection: Let $n(m) := \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$

If $|B_m| == 1$: choose the single action in B_m until time T.

Else If $\sum_{i\in S_m} z_i^* \le 2|B_m|\tilde{\Delta}_m$: For each action i in S_m, choose it $z_i^*\,[n(m) - n(m-1)]$ times.

Else For each action i in B_m, choose it $[n(m) - n(m-1)]$ times.

Update the average rewards of all actions in B_m.

Action Elimination:
To get B_{m+1}, delete all actions i in B_m for which

$$\bar{x}_i(m) + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2s_i(m)}} \le \max_{a\in B_m}\left\{\bar{x}_a(m) - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2s_a(m)}}\right\},$$

where $\bar{x}_i(m)$ is the empirical average reward of action i, and $s_i(m)$ is the total number of observations for action i up to round m.

Reset:
The set S_{m+1} of actions with neighbors in B_{m+1} is given as $S_{m+1} = \{i : \mathcal{K}_i \cap B_{m+1} \neq \emptyset\}$. Note that $B_{m+1} \subseteq S_{m+1}$.
Let $\tilde{\Delta}_{m+1} = \frac{\tilde{\Delta}_m}{2}$.

end for

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

PROPOSITION 4. *For action i, define round* m_i *as follows:*

$$m_i := \min\left\{m : \tilde{\Delta}_m < \frac{\Delta_i}{2}\right\}.$$

Define $\bar{m} = \min\left\{m : \sum_{i\in\mathcal{K}} z_i^* > \sum_{i:m_i>m} 2^{-m+1}\right\}$ *and the set* $B = \{i \in \mathcal{U} : m_i > \bar{m}\}$.

Then, the expected regret due to the UCB-LP policy described in Algorithm 2 is at most

$$\sum_{i\in\mathcal{U}\setminus B} \Delta_i z_i^* \frac{32\log(T\hat{\Delta}_i^2)}{\hat{\Delta}_i^2} + \sum_{i\in B} \frac{32\log(T\Delta_i^2)}{\Delta_i} + O(K\delta), \quad (6)$$

where $\hat{\Delta}_i = \max\{2^{-\bar{m}+2}, \min_{\mathcal{K}_i}\{\Delta_j\}\}$ *and* (z_i^*) *is the solution of LP* P_2. δ *is the maximum degree in the network. The* $O(K\delta)$ *term captures constants independent of time. Further, under Assumption 4, the regret is also at most*

$$O\left(\sum_{i\in\mathcal{K}} z_i^* \log(T)\right) + O(K\delta), \quad (7)$$

where (z_i^*) *entirely captures the time dependence on network structure.*

PROOF. *(Sketch)* The $\log(T)$ term in the regret follows from the fact that, with high probability, each suboptimal action i is eliminated (from the set B_m) on or before the first round m such that $\tilde{\Delta}_m < \Delta_i/2$. See Section 8 for the full proof. □

Next, we briefly describe the policies UCB-N and UCB-MaxN proposed in [6]. In UCB-N policy, at each time, the action with the highest UCB index (see (1)) is chosen similar to UCB1 policy in [1]. In UCB-MaxN policy, at each time t, the action i with the highest UCB index (1) is identified and its neighboring action $j \in \mathcal{K}_i$ with the highest empirical average reward at time t is chosen.

Remark 2. The regret upper bound of UCB-N policy is

$$\inf_{\mathcal{C}} \sum_{C\in\mathcal{C}} \frac{8\max_{i\in C}\Delta_i}{\min_{i\in C}\Delta_i^2}\log(T) + O(K),$$

where $\mathcal{C}$ is a clique covering of the sub-network of suboptimal actions. The regret upper bound for UCB-MaxN is the same as that for UCB-N with an $O(|\mathcal{C}|)$ term instead of the time-invariant $O(K)$ term. We show a better regret performance for UCB-LP policy and ϵ_t-greedy-LP policies with respect to the $\log(T)$ term because $\sum_{i\in\mathcal{K}} z_i^* \leq \gamma(G) \leq \bar{\chi}(G)$. However, the time-invariant term in our policies is $O(K)$ and $O(K\delta)$, which can be worse than the time-invariant term $O(|\mathcal{C}|)$ in UCB-MaxN.

To see the possible gap between $\gamma(G)$ and $\bar{\chi}(G)$, consider an Erdos-Renyi random graph $G(K,p)$. As noted in [12], as $K \to \infty$, for any fixed $p > 0$, $\gamma(G)$ is at most $O(\log(K))$ while $\bar{\chi}(G)$ is at least $O(K/\log(K))$. On the other hand, it has been shown [7] that for power law graphs, both $\gamma(G)$ and $\bar{\chi}(G)$ scale linearly with N, although $\gamma(G)$ has a lower slope. We note that the social network of interest may or may not display a power law behavior. We find that the subgraphs of the Flixster network have a degree distribution that is a straight line on a log-log plot indicating a power law distribution display while the authors in [15] show that the degree distribution of the global Facebook network is not a straight line on log-log plot. Our numerical results in Section 7 show that our policies outperform existing policies even for the Flixster network.

Remark 3. All uniformly good policies that ignore side-observations incur a regret that is at least $\Omega(|\mathcal{U}|\log(t))$ [10], where $|\mathcal{U}|$ is the number of suboptimal actions. This could be significantly higher than the guarantees on the regret of both ϵ_t-greedy-LP policy and UCB-LP policy for a rich network structure as discussed in Remark 2.

Remark 4. While ϵ_t-greedy-LP does not require knowledge of the time horizon T, UCB-LP policy requires the knowledge of T. UCB-LP policy can be extended to the case of an unknown time horizon similar to the suggestion in [2]. Start with $T_0 = 2$ and at end of each T_l, set $T_{l+1} = T_l^2$. The regret bound for this case is expected to be similar to the one in Proposition 4.

Remark 5. In our work, we assumed that the side observations are always available. However, in reality, side observations may only be obtained sporadically. Suppose that when action i is chosen, side-observations are obtained for each neighboring action j with a known probability p_j. In this case, Proposition 1 holds with the replacement of LP P_1 with LP P_1' as follows:

$$P_1': \ \min \sum_{i\in\mathcal{U}} \Delta_i w_i,$$
$$\text{subject to: } w_i + p_i \sum_{j\in\mathcal{K}_i\setminus\{i\}} w_j \geq \frac{1}{D(\theta_i||\theta^*)}, \ \forall i \in \mathcal{U},$$
$$w_i \geq 0, \ \forall i \in \mathcal{K}.$$

Both of our policies work for this setting by changing the LP P_2 to P_2' as follows:

$$P_2': \ \min \sum_{i\in\mathcal{K}} z_i$$
$$\text{subject to: } z_i + p_i \sum_{j\in\mathcal{K}_i\setminus\{i\}} z_j \geq 1, \ \forall i \in \mathcal{K},$$
$$\text{and } z_i \geq 0, \ \forall i \in \mathcal{K}.$$

The regret bounds of our policies will now depend on the optimal solution of LP P_2'.

7. NUMERICAL RESULTS

We consider the Flixster network dataset for the numerical evaluation of our algorithms. The authors in [9] collected this social network data, which contains about 1 million users and 14 million links. We use graph clustering [8] to identify two strongly clustered sub-networks of sizes 1000 and 2000 nodes. Both these sub-networks have a degree distribution that is a straight line on a log-log plot indicating a power law distribution commonly observed in social networks.

Our empirical setup is as follows. Each user in the network is offered a promotion at each time, and accepts the promotion with probability $\mu_i \in [0.3, 0.9]$. The decision maker receives a random reward of 1 if a user accepts the promotion or 0 reward otherwise. μ_i is chosen uniformly at random from $[0.3, 0.8]$ and there are 50 randomly chosen users with optimal $\mu_i = 0.9$. Figures 2 and 3 show the regret performance as a function of time for the two sub-networks of sizes 1000 and 2000 respectively. For the ϵ_t-greedy-LP policy, we let $c = 5$ and $d = 0.2$. For both networks, we see that our policies outperform the UCB-N and UCB-MaxN policies.

We also observe that the improvement obtained by UCB-N policy over the baseline UCB1 policy is marginal.

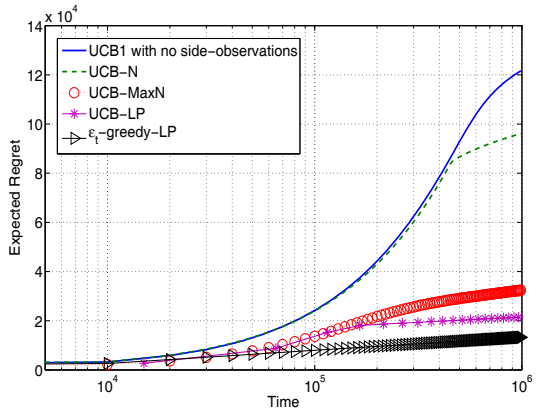

Figure 2: Regret of all the policies for a network of size 1000.

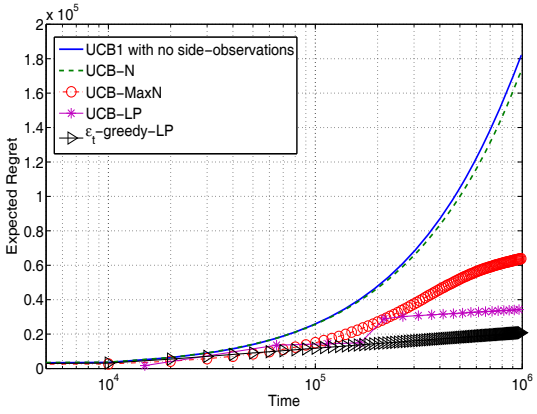

Figure 3: Regret of all the policies for a network of size 2000.

8. PROOFS

In what follows, we give the proofs of all propositions stated in the earlier sections. These proofs make use of Lemmas 1, 2, and 3, and Proposition 5 given in the Appendix.

Proof of Proposition 1

Let $\mathcal{U} = \{i : \mu_i < \mu^*\}$ be the set of suboptimal actions. Also, let $\Delta_i = \mu^* - \mu_i$. Recall that $\mathcal{K}_i$ is the set of neighbors of i, including i, in the network $G(\mathcal{K})$. Also, $T_i(t)$ is the total number of times action i is chosen up to time t by policy $\boldsymbol{\phi}$. Let $S_i(t)$ be the total number of observations corresponding to action i available at time t. From Proposition 5 given in the Appendix, we have,

$$\liminf_{t\to\infty} \frac{\mathbb{E}[S_i(t)]}{\log(t)} \geq \frac{1}{D(\theta_i||\theta^*)}, \ \forall i \in \mathcal{U}. \tag{8}$$

An observation is received for action i whenever any action in $\mathcal{K}_i$ is chosen. Hence,

$$S_i(t) = \sum_{j\in\mathcal{K}_i} T_j(t). \tag{9}$$

Now, from Equations (8) and (9), for each $i \in \mathcal{U}$,

$$\liminf_{t\to\infty} \frac{\sum_{j\in\mathcal{K}_i} \mathbb{E}[T_j(t)]}{\log(t)} \geq \frac{1}{D(\theta_i||\theta^*)}. \tag{10}$$

Using (10), we get the constraints of LP P_1. Further, we have from definition of regret that,

$$\liminf_{t\to\infty} \frac{R_\mu(t)}{\log(t)} = \liminf_{t\to\infty} \sum_{i\in\mathcal{U}} \Delta_i \frac{\mathbb{E}[T_i(t)]}{\log(t)}.$$

The above equation along with the constraints of the LP P_1 obtained from (10) gives us the required lower bound on regret.

Proof of Proposition 2

Let $\mathcal{I} = \{i \in \mathcal{U} : \mathcal{K}_i \cap \mathcal{O} \neq \emptyset\}$ be the set of suboptimal actions with neighbors in $\mathcal{O}$. Let $(z_i^*)_{i\in\mathcal{K}}$ be the optimal solution of LP P_2.

We will first prove the upper bound in Equation 3. Using the optimal solution $(w_i^*)_{i\in\mathcal{K}}$ of LP P_1, we construct a feasible solution satisfying constraints in LP P_2 in the following way: For actions $i \in \mathcal{U}$, let $z_i = \left(\max_{i\in\mathcal{U}} D(\theta_i||\theta^*)\right) w_i^*$. Then $(z_i)_{i\in\mathcal{U}}$ satisfy constraints for all actions $i \in \mathcal{U} \setminus \mathcal{I}$ because w_i^* satisfy constraints of LP P_1.

In order to satisfy the constraints for actions in $\mathcal{I} \cup \mathcal{O}$, we use $z_i = 1$ for all i in $\mathcal{O}$. The feasible solution constructed in this way gives an upper bound on the optimal value of LP P_2. Hence,

$$\begin{aligned}
\sum_{i\in\mathcal{K}} z_i^* &\leq \sum_{i\in\mathcal{U}} z_i + |\mathcal{O}| \\
&\leq \sum_{i\in\mathcal{U}} \left(\max_{i\in\mathcal{U}} D(\theta_i||\theta^*)\right) w_i^* + |\mathcal{O}| \\
&\leq \frac{\max_{i\in\mathcal{U}} D(\theta_i||\theta^*)}{\min_{i\in\mathcal{U}} \Delta_i} \sum_{i\in\mathcal{U}} \Delta_i w_i^* + |\mathcal{O}| \\
&\leq \frac{\max_{i\in\mathcal{U}} D(\theta_i||\theta^*)}{\min_{i\in\mathcal{U}} \Delta_i} c_{\boldsymbol{\mu}} + |\mathcal{O}|
\end{aligned}$$

For the lower bound, any feasible solution of P_2, in particular $\mathbf{z}^*$, can be used to construct a feasible solution of P_1. For actions $i \in \mathcal{K}$, let $w_i = \dfrac{z_i^*}{\min_{i\in\mathcal{U}} D(\theta_i||\theta^*)}$. Then $(w_i)_{i\in\mathcal{K}}$ satisfies the constraints of LP P_1 and hence gives an upper bound on its optimal value. Therefore, we have

$$\begin{aligned}
c_{\boldsymbol{\mu}} &= \sum_{i\in\mathcal{U}} \Delta_i w_i^*, \\
&\leq \sum_{i\in\mathcal{K}} \frac{\Delta_i z_i^*}{\min_{i\in\mathcal{U}} D(\theta_i||\theta^*)} \\
&\leq \sum_{i\in\mathcal{K}} \frac{\max_{i\in\mathcal{U}} \Delta_i z_i^*}{\min_{i\in\mathcal{U}} D(\theta_i||\theta^*)}
\end{aligned}$$

which gives us the required lower bound.

Proof of Proposition 3

Since $\mathbf{z}^*$ satisfies the constraints in LP P_2, there is sufficient exploration within each suboptimal action's neighborhood. The proof is then a combination of this fact and the proof of Theorem 3 in [1]. Let $\bar{X}_i(t)$ be the random variable denoting the sample mean of all observations available for

action i at time t. Let $\bar{X}^*(t)$ be the random variable denoting the sample mean of all observations available for an optimal action at time t. Let $X_{i,m}$ denote the sample mean of m random variables drawn from the distribution F_i. Fix a suboptimal action i. For some $\alpha > 1$, define m_i as follows,

$$m_i = \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=1}^{t} \epsilon(m)$$

Let $\phi(t)$ be the action chosen by ϵ_t-greedy-LP policy at time t. Then,

$$\mathbb{P}[\phi(t) = i] \leq \frac{\epsilon(t) z_i^*}{\sum_{i\in\mathcal{K}} z_i^*} + (1 - \epsilon(t))\mathbb{P}[\bar{X}_i(t) \geq \bar{X}^*(t)]$$

We also have that,

$$\mathbb{P}[\bar{X}_i(t) \geq \bar{X}^*(t)] \leq \mathbb{P}\left[\bar{X}_i(t) \geq \mu_i + \frac{\Delta_i}{2}\right] + \mathbb{P}\left[\bar{X}^*(t) \leq \mu^* - \frac{\Delta_i}{2}\right].$$

The analysis of both the terms in the right hand side of the above expression is similar. Let $S_i^{(R)}(t)$ be the total number of observations available for action i from the exploration phase of the policy up to time t. Let $S_i(t)$ be the total number of observations available for action i up to time t. Hence, we have,

$$\begin{aligned}
&\mathbb{P}\left[\bar{X}_i(t) \geq \mu_i + \frac{\Delta_i}{2}\right] = \sum_{m=1}^{t} \mathbb{P}\left[S_i(t) = m; \bar{X}_i(t) \geq \mu_i + \frac{\Delta_i}{2}\right] \\
&= \sum_{m=1}^{t} \mathbb{P}\left[S_i(t) = m | \bar{X}_{i,m} \geq \mu_i + \frac{\Delta_i}{2}\right] \mathbb{P}\left[\bar{X}_{i,m} \geq \mu_i + \frac{\Delta_i}{2}\right] \\
&\leq \sum_{m=1}^{\lfloor t \rfloor} \mathbb{P}\left[S_i(t) = m | \bar{X}_{i,m} \geq \mu_i + \frac{\Delta_i}{2}\right] e^{\frac{-\Delta_i^2 m}{2}} \\
&\qquad \text{(follows from Chernoff-Hoeffding bound in Lemma 1)} \\
&\leq \sum_{m=1}^{\lfloor m_i \rfloor} \mathbb{P}\left[S_i(t) = m | \bar{X}_{i,m} \geq \mu_i + \frac{\Delta_i}{2}\right] + \frac{2}{\Delta_i^2} e^{\frac{-\Delta_i^2 m_i}{2}} \\
&\qquad \left(\text{since } \sum_{m+1}^{\infty} e^{-ku} = \frac{1}{k} e^{-km}\right) \\
&\leq \sum_{m=1}^{\lfloor m_i \rfloor} \mathbb{P}\left[S_i^{(R)}(t) \leq m | \bar{X}_{i,m} \geq \mu_i + \frac{\Delta_i}{2}\right] + \frac{2}{\Delta_i^2} e^{\frac{-\Delta_i^2 m_i}{2}} \\
&\leq m_i \mathbb{P}\left[S_i^{(R)}(t) \leq m_i\right] + \frac{2}{d^2} e^{\frac{-d^2 m_i}{2}}
\end{aligned}$$

In the above, the last equation follows from the fact that $S_i^{(R)}(t)$, which is the total number of observations for action i available from the exploration phase of the policy up to time t, is independent of the sample means of all actions. Now,

$$\begin{aligned}
\mathbb{E}\left[S_i^{(R)}(t)\right] &= \sum_{m=1}^{t} \epsilon(m) \sum_{j\in\mathcal{K}_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \\
&= \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=1}^{t} \epsilon(m) = \alpha m_i
\end{aligned}$$

$$\begin{aligned}
var\left[S_i^{(R)}(t)\right] &= \sum_{m=1}^{t} \left[\epsilon(m) \sum_{j\in\mathcal{K}_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \right. \\
&\qquad \left. - \left(\epsilon(m) \sum_{j\in\mathcal{K}_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_i^*}\right)^2\right] \\
&\leq \sum_{m=1}^{t} \epsilon(m) \sum_{j\in\mathcal{K}_i} \frac{z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} = \mathbb{E}\left[S_i^{(R)}(t)\right]
\end{aligned}$$

Now, using Bernstein's inequality given in Lemma 2, we have

$$\begin{aligned}
\mathbb{P}\left[S_i^{(R)}(t) \leq m_i\right] &= \mathbb{P}\left[S_i^{(R)}(t) \leq \mathbb{E}\left[S_i^{(R)}(t)\right] - (\alpha - 1)m_i\right] \\
&\leq \exp\left(-r m_i\right),
\end{aligned}$$

where $r = \frac{3(\alpha-1)^2}{8\alpha-2}$. Now, we will obtain upper and lower bounds on m_i. For the upper bound, for any $t > t' = \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2}$,

$$\begin{aligned}
m_i &= \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=1}^{t} \epsilon(m) \\
&= \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} t' + \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=t'+1}^{t} \frac{c\sum_{j\in\mathcal{K}} z_j^*}{d^2 t} \\
&\leq \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \frac{c\sum_{i\in\mathcal{K}} z_i^*}{d^2} + \frac{\delta c}{\alpha d^2} \sum_{m=t'+1}^{t} \frac{1}{t} \\
&\leq \frac{\delta c}{\alpha d^2} \log\left(\frac{e^2 t}{t'}\right),
\end{aligned}$$

where δ is the maximum degree in the network. In the above, $\sum_{i\in\mathcal{K}_i} z_i^* \leq \delta$ because $z_i^* \leq 1$, which is due to the fact that $(z_i^*)_{i\in\mathcal{K}}$ is the optimal solution of LP P_2. Next, for the lower bound, we use the fact that $\sum_{j\in\mathcal{K}_i} z_i^* \geq 1$ for all i because $(z_i^*)_{i\in\mathcal{K}}$ satisfies the constraints of LP P_2. Thus

$$\begin{aligned}
m_i &= \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=1}^{t} \epsilon(m) \\
&\geq \frac{1}{\alpha} \frac{\sum_{j\in\mathcal{K}_i} z_j^*}{\sum_{j\in\mathcal{K}} z_i^*} \sum_{m=t'+1}^{t} \frac{c\sum_{j\in\mathcal{K}} z_j^*}{d^2 t} \\
&\geq \frac{c}{\alpha d^2} \sum_{m=t'+1}^{t} \frac{1}{t} \geq \frac{c}{\alpha d^2} \log\left(\frac{t}{et'}\right).
\end{aligned}$$

Hence, combining the inequalities above,

$$\begin{aligned}
&\mathbb{P}\left[\bar{X}_i(t) \geq \mu_i + \frac{\Delta_i}{2}\right] \leq m_i \mathbb{P}\left[S_i^{(R)}(t) \leq m_i\right] + \frac{2}{d^2} e^{\frac{-d^2 m_i}{2}} \\
&\leq m_i \exp\left(-r m_i\right) + \frac{2}{d^2} e^{\frac{-d^2 m_i}{2}} \\
&\leq \frac{\delta c}{\alpha d^2} \left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right) + \frac{2}{d^2} \left(\frac{et'}{t}\right)^{c/2\alpha}.
\end{aligned}$$

Now, similarly for the optimal action, we have, for all $t > t'$

$$\begin{aligned}
\mathbb{P}\left[\bar{X}^*(t) \leq \mu^* - \frac{\Delta_i}{2}\right] &\leq \frac{\delta c}{\alpha d^2} \left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right) \\
&\quad + \frac{2}{d^2} \left(\frac{et'}{t}\right)^{c/2\alpha}.
\end{aligned}$$

Combining everything, we have for any suboptimal action i, for all $t > t'$

$$\mathbb{P}[\phi(t) = i] \leq \left(\frac{c}{d^2 t} z_i^*\right) + \frac{4}{d^2}\left(\frac{et'}{t}\right)^{c/2\alpha} + \frac{2\delta c}{\alpha d^2}\left(\frac{et'}{t}\right)^{cr/\alpha d^2} \log\left(\frac{e^2 t}{t'}\right).$$

Proof of Proposition 4

The proof technique in similar to that in [2]. We will analyze the regret by conditioning on two disjoint events. The first event is that each suboptimal action a is eliminated by an optimal action on or before the first round m such that $\tilde{\Delta}_m < \Delta_a/2$. This happens with high probability and leads to logarithmic regret. The compliment of the first event yields linear regret in time but occurs with probability proportional to $1/T$. The main difference from the proof in [2] is that on the first event, the number of times we choose each action is proportional to $z_i^* \log(T)$ in the exploration phase of the policy. This gives us the required upper bound in terms of optimal solution $\mathbf{z}^*$ of LP P_2.

Let $*$ denote any optimal action. Let m^* denote the round in which the last optimal action $*$ is eliminated. For each suboptimal action, define round $m_i := \min\{m : \tilde{\Delta}_m < \frac{\Delta_i}{2}\}$. For an optimal action i, $m_i = \infty$ by convention. Then, by the definition of m_i, for all rounds $m < m_i$, $\Delta_i \leq 2\tilde{\Delta}_m$, and

$$\frac{2}{\Delta_i} < 2^{m_i} = \frac{1}{\tilde{\Delta}_{m_i}} \leq \frac{4}{\Delta_i} < \frac{1}{\Delta_{m_i+1}} = 2^{m_i+1}. \qquad (11)$$

From Lemma 3 in the Appendix, the probability that action i is not eliminated in round m_i by $*$ is at most $\frac{2}{T\tilde{\Delta}^2_{m_i}}$.

Let $\mathcal{U}_i = \mathcal{K}_i \cap \mathcal{U}$ be the set of suboptimal neighbors of action i. Let $I(t)$ be the action chosen at time t by the UCB-LP policy.

Let E_{m^*} be the event that all suboptimal actions with $m_i \leq m^*$ are eliminated by $*$ on or before their respective m_i. Then, the complement of E_{m^*}, denoted as $E^c_{m^*}$, is the event that there exists some suboptimal action i with $m_i \leq m^*$, which is not eliminated in round m_i. Let E_i^c be the event that action i is not eliminated by round m_i by $*$. Let $m_f = \lfloor \frac{1}{2}\log_2 \frac{T}{e} \rfloor$ and $I(t)$ denote the action chosen at time t by the policy. Recall that regret is denoted by $R_{\boldsymbol{\mu}}(T)$. Let $\mathbb{P}[m^* = m]$ be denoted by p_m. Hence, $\sum_{m=0}^{m_f} p_m = 1$.

$$\begin{aligned}
\mathbb{E}\left[R_{\boldsymbol{\mu}}(T)\right] &= \sum_{m=0}^{m_f} \mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^* = m\}\right]\mathbb{P}[m^* = m] \\
&= \sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}} \Delta_j \mathbb{P}\left[I(t) = j|\{m^* = m\}\right] p_m \\
&= \sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E_{m^*}|\{m^* = m\}\right] p_m \\
&\quad + \sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E^c_{m^*}|\{m^* = m\}\right] p_m \\
&= (i) + (ii)
\end{aligned}$$

Next we will show that term (i) leads to logarithmic regret while term (ii) leads to a constant regret with time.

First, consider the term (ii) of the regret expression. Recall that $\mathcal{U}_j = \mathcal{K}_j \cap \mathcal{U}$ is the set of suboptimal neighbors of j. For each $j \in \mathcal{U}$, we have,

$$\begin{aligned}
&\sum_{m=0}^{m_f}\sum_{t=1}^{T} \mathbb{P}\left[\{I(t) = j\} \cap E^c_{m^*}|\{m^* = m\}\right]\mathbb{P}[m^* = m] \\
&\leq \sum_{m=0}^{m_f}\sum_{t=1}^{T} \mathbb{P}\left[\{I(t) = j\} \cap \left(\cup_{i\in\mathcal{U}:m_i\leq m^*} E_i^c\right)|\{m^* = m\}\right] p_m \\
&= \sum_{m=0}^{m_f}\sum_{t=1}^{T} \mathbb{P}\left[\{I(t) = j\} \cap \left(\cup_{i\in\mathcal{U}_j:m_i\leq m^*} E_i^c\right)|\{m^* = m\}\right] p_m \\
&\qquad \text{(because } \{I((t) = j\} \text{ depends only on neighbors of } j.) \\
&\leq \sum_{m=0}^{m_f}\sum_{t=1}^{T} \Big(\mathbb{P}\left[\{I(t) = j\}|\left(\cup_{i\in\mathcal{U}_j:m_i\leq m^*} E_i^c\right), \{m^* = m\}\right] \\
&\qquad\qquad \mathbb{P}\left[\cup_{i\in\mathcal{U}_j:m_i\leq m^*} E_i^c|\{m^* = m\}\right] p_m\Big) \\
&\leq T\mathbb{P}\left[\cup_{i\in\mathcal{U}_j} E_i^c|\{m^* = m_f\}\right] \sum_{m=0}^{m_f} p_m \\
&\leq T\sum_{i\in\mathcal{U}_j} \frac{2}{T\tilde{\Delta}^2_{m_i}}, \\
&\quad \left(\text{using Lemma 3, } P\left[E_i^c|\{m^* = m_f\}\right] \leq \frac{2}{T\tilde{\Delta}^2_{m_i}}\right) \\
&\leq \sum_{i\in\mathcal{U}_j} \frac{32}{\Delta_i^2},
\end{aligned}$$

where the last inequality follows from Equation (11). Hence, the term (ii) of regret is

$$\sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E^c_{m^*}|\{m^* = m\}\right] p_m \leq \sum_{j\in\mathcal{U}} \Delta_j \sum_{i\in\mathcal{U}_j} \frac{32}{\Delta_i^2} = O(K\delta), \qquad (12)$$

where δ is the maximum degree in the network.

Next, we consider the term (i). Recall that, in this term, we consider the case that all suboptimal actions i with $m_i \leq m^*$ are eliminated by $*$ on or before m_i.

$$\begin{aligned}
(i) &= \sum_{m=0}^{m_f}\sum_{t=1}^{T}\sum_{j\in\mathcal{U}} \Delta_j \mathbb{P}\left[\{I(t) = j\} \cap E_{m^*}|\{m^* = m\}\right] p_m \\
&= \sum_{m=0}^{m_f} \mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^* = m\}, E_{m^*}\right]\mathbb{P}[E_{m^*}|\{m^* = m\}]p_m \\
&\leq \sum_{m=0}^{m_f} \Big(\mathbb{E}\left[\text{Regret from } \{i : m_i \leq m^*\}|\{m^* = m\}, E_{m^*}\right] \\
&\qquad + \mathbb{E}\left[\text{Regret from } \{i : m_i > m^*\}|\{m^* = m\}, E_{m^*}\right]\Big)p_m \\
&\leq \sum_{m=0}^{m_f} \Big(\mathbb{E}\left[\text{Regret from } \{i : m_i \leq m_f\}|\{m^* = m_f\}, E_{m_f}\right] \\
&\qquad + \mathbb{E}\left[\text{Regret from } \{i : m_i > m^*\}|\{m^* = m\}, E_{m^*}\right]\Big)p_m
\end{aligned}$$

$$\leq \mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^*=m_f\},E_{m_f}\right]\sum_{m=0}^{m_f}p_m$$
$$+\sum_{m=0}^{m_f}\mathbb{E}\left[\text{Regret from } \{i:m_i>m^*\}|\{m^*=m\},E_{m^*}\right]p_m$$
$$=(ia)+(ib)$$

Once again, we will consider the above two terms separately. For the term (ia), under the event E_{m_f}, each suboptimal action i is eliminated by $*$ by round m_i. Define round $\bar{m}$ and the set B as follows:

$$\bar{m}=\min\{m:\sum_{i\in\mathcal{K}}z_i^*>\sum_{i:m_i>m}2^{-m+1}\},$$
$$B=\{i\in\mathcal{U}:m_i>\bar{m}\}.$$

After round $\bar{m}$, Algorithm 2 chooses only those actions with $m_i>\bar{m}$. Also, by the definition of the **Reset** phase of Algorithm 2, we have that any suboptimal action $i\notin B$ is chosen (i.e. appears in the set S_m at round m) only until all actions in its neighborhood are eliminated or until $\bar{m}$, whichever happens first. Define $n_i=\min\{\bar{m},\max_{\mathcal{K}_i}\{m_j\}\}$ for each suboptimal action i. Then any suboptimal action $i\notin B$ is chosen for at most n_i rounds.

$$(ia)=\mathbb{E}\left[R_{\boldsymbol{\mu}}(T)|\{m^*=m_f\},E_{m_f}\right]$$
$$\leq\sum_{i\in\mathcal{U}\setminus B}\Delta_i z_i^*\frac{2\log(T\tilde{\Delta}_{n_i}^2)}{\tilde{\Delta}_{n_i}^2}+\sum_{i\in B}\Delta_i\frac{2\log(T\tilde{\Delta}_{m_i}^2)}{\tilde{\Delta}_{m_i}^2}$$
$$\leq\sum_{i\in\mathcal{U}\setminus B}\Delta_i z_i^*\frac{32\log(T\hat{\Delta}_i^2)}{\hat{\Delta}_i^2}+\sum_{i\in B}\Delta_i\frac{32\log(T\Delta_i^2)}{\Delta_i^2},\quad(13)$$

where $\hat{\Delta}_i=\max\{2^{-\bar{m}+2},\min_{\mathcal{K}_i}\{\Delta_j\}\}$ and (z_i^*) is the solution of LP P_2.

Finally, we consider the term (ib). An optimal action $*$ is not eliminated in round m^* if (16) holds for $m=m^*$. Hence, using (17) and (18), the probability p_m that $*$ is eliminated by a suboptimal action in any round m^* is at most $\frac{2}{T\tilde{\Delta}_{m^*}^2}$. Hence, term (ib) is given as:

$$\sum_{m=0}^{m_f}\mathbb{E}\left[\text{Regret from } \{i:m_i>m^*\}|\{m^*=m\},E_{m^*}\right]p_m$$
$$\leq\sum_{m=0}^{m_f}\sum_{i\in\mathcal{U}:m_i\geq m}\frac{2}{T\tilde{\Delta}_m^2}.T\max_{\mathcal{U}}\Delta_j$$
$$\leq\max_{\mathcal{U}}\Delta_j\sum_{m=0}^{m_f}\sum_{i\in\mathcal{U}:m_i\geq m}\frac{2}{\tilde{\Delta}_m^2}.$$
$$\leq\sum_{i\in\mathcal{U}}\sum_{m=0}^{m_i}\frac{2}{\tilde{\Delta}_m^2}$$
$$=\sum_{i\in\mathcal{U}}2.2^{2m_i+2}\leq\sum_{i\in\mathcal{U}}\frac{32}{\Delta_i^2}=O(K).\quad(14)$$

Now we get the result (6) by combining the bounds in (12), (13), and (14).

Further, the definition of set B ensures that we have

$$\sum_{i\in B}\Delta_i\leq\sum_{i\in\mathcal{K}}z_i^*.$$

Also, using the Assumption 4, $\frac{32\Delta_i\log(T\hat{\Delta}_i^2)}{\hat{\Delta}_i^2}$, $\frac{32\log(T\Delta_i^2)}{\Delta_i^2}$ are bounded by $C\log(T)$, where C is a constant independent of network structure. Hence, (13) can be bounded as:

$$\sum_{i\in\mathcal{U}\setminus B}\Delta_i z_i^*\frac{32\log(T\hat{\Delta}_i^2)}{\hat{\Delta}_i^2}+\sum_{i\in B}\Delta_i\frac{32\log(T\Delta_i^2)}{\Delta_i^2}$$
$$\leq\sum_{i\in\mathcal{U}\setminus B}z_i^*C\log(T)+\sum_{i\in B}\Delta_i C\log(T)$$
$$\leq\sum_{i\in\mathcal{U}\setminus B}z_i^*C\log(T)+\sum_{i\in B}2^{-\bar{m}+1}C\log(T)$$
$$\leq 2\sum_{i\in\mathcal{K}}z_i^*C\log(T).\quad(15)$$

Hence, we get (7) from (15), (12), and (14).

9. CONCLUSION

In this work, we studied the stochastic multi-armed bandit problem in the presence of side-observations across actions that are embedded in a network. We obtained an asymptotic (with respect to time) lower bound as a function of the network structure on the regret of any uniformly good policy. Further, we proposed two policies: 1) the ϵ_t-greedy-LP policy, and 2) the UCB-LP policy, both of which are optimal in the sense that they achieve the asymptotic lower bound on the regret, up to a multiplicative constant that is independent of the network structure. These policies can have a better regret performance than existing policies for some important network structures. The ϵ_t-greedy-LP policy is a network-aware any-time policy, but its exploration is oblivious to the average rewards of the suboptimal actions. On the other hand, UCB-LP considers both the network structure and the average rewards of actions. Finally, using numerical examples on the Flixster network dataset, we confirmed the significant benefits obtained by our policies against other existing policies.

Acknowledgments

This work is supported by the NSF grants: CAREER-CNS-0953515, CCF-0916664, CNS-1012700, CNS-1065136, and by grants from the Army Research Office: W911NF-08-1-0238 and W911NF-12-1-0385. Also, the work of A. Eryilmaz was supported in part by the QNRF grant number NPRP 09-1168-2-455.

10. REFERENCES

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, May 2002.

[2] P. Auer and R. Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

[3] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. I. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489:295–298, 2012.

[4] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[5] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. *X*-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695, 2011.

[6] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat. Leveraging side observations in stochastic bandits. In *UAI*, pages 142–151. AUAI Press, 2012.

[7] C. Cooper, R. Klasing, and M. Zito. Lower bounds and algorithms for dominating sets in web graphs. *Internet Mathematics*, 2:275–300, 2005.

[8] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

[9] M. Jamali and M. Ester. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 135–142. ACM, 2010.

[10] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[11] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670. ACM, 2010.

[12] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *NIPS*, pages 684–692, 2011.

[13] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 721–728, New York, NY, USA, 2007. ACM.

[14] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

[15] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

APPENDIX

Notation: $S_n = \frac{1}{n}\sum_{j=1}^{n} X_j$ is called the sample mean of the random variables $X_1, \ldots, X_n$. The first two lemmas below state the Chernoff-Hoeffding inequality and Bernstein's inequality.

LEMMA 1. *Let $X_1, \ldots, X_n$ be a sequence of random variables with support $[0,1]$ and $\mathbb{E}[X_t] = \mu$ for all $t \leq n$. Let $S_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. Then, for all $\epsilon > 0$, we have,*

$$\mathbb{P}[S_n \geq \mu + \epsilon] \leq e^{-2n\epsilon^2}$$
$$\mathbb{P}[S_n \leq \mu - \epsilon] \leq e^{-2n\epsilon^2}.$$

LEMMA 2. *Let $X_1, \ldots, X_n$ be a sequence of random variables with support $[0,1]$ and $\sum_{k=1}^{t} var[X_k | X_1, \ldots, X_{k-1}] \leq \sigma^2$ for all $t \leq n$. Let $S_n = \sum_{j=1}^{n} X_j$. Then, for all $\epsilon > 0$, we have,*

$$\mathbb{P}[S_n \geq \mathbb{E}[S_n] + \epsilon] \leq \exp\left\{-\frac{\epsilon^2}{2\sigma^2 + \frac{2}{3}\epsilon}\right\}$$

$$\mathbb{P}[S_n \leq \mathbb{E}[S_n] - \epsilon] \leq \exp\left\{-\frac{\epsilon^2}{2\sigma^2 + \frac{2}{3}\epsilon}\right\}.$$

The next lemma is used in the proof of Proposition 4.

LEMMA 3. *The probability that action i is not eliminated in round m_i by $*$ is at most $\frac{2}{T\tilde{\Delta}^2_{m_i}}$.*

PROOF. Let $\bar{X}_i(m)$ be the sample mean of all observations for action i available in round m. Let $\bar{X}^*(m)$ be the sample mean of the optimal action. The constraints of LP P_2 ensure that at the end of each round m, for all actions in B_m, we have $n(m) := \left\lceil \frac{2\log(T\tilde{\Delta}^2_m)}{\tilde{\Delta}^2_m} \right\rceil$ observations. Now, for $m = m_i$, if we have,

$$\bar{X}_i(m) \leq \mu_i + \sqrt{\frac{\log(T\tilde{\Delta}^2_m)}{2n(m)}} \text{ and } \bar{X}^*(m) \geq \mu^* - \sqrt{\frac{\log(T\tilde{\Delta}^2_m)}{2n(m)}}, \tag{16}$$

then, action i is eliminated by $*$ in round m_i. In fact, in round m_i, we have

$$\sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}} \leq \frac{\tilde{\Delta}_{m_i}}{2} < \frac{\Delta_i}{4}.$$

Hence, in the elimination phase of the UCB-LP policy, if (16) holds for action i in round m_i, we have,

$$\begin{aligned}
\bar{X}_i(m_i) + \sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}} &\leq \mu_i + 2\sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}} \\
&< \mu_i + \Delta_i - 2\sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}} \\
&= \mu^* - 2\sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}} \\
&\leq \bar{X}^*(m_i) - \sqrt{\frac{\log(T\tilde{\Delta}^2_{m_i})}{2n(m_i)}},
\end{aligned}$$

and action i is eliminated. Hence, the probability that action i is not eliminated in round m_i is the probability that either one of the inequalities in (16) do not hold. Using Chernoff-Hoeffding bound (Lemma 1), we can bound this as follows,

$$\mathbb{P}\left[\bar{X}_i(m) > \mu_i + \sqrt{\frac{\log(T\tilde{\Delta}^2_m)}{2n(m)}}\right] \leq \frac{1}{T\tilde{\Delta}^2_m} \tag{17}$$

$$\mathbb{P}\left[\bar{X}^*(m) < \mu^* - \sqrt{\frac{\log(T\tilde{\Delta}^2_m)}{2n(m)}}\right] \leq \frac{1}{T\tilde{\Delta}^2_m}. \tag{18}$$

Summing the above two inequalities for $m = m_i$ gives us that the probability that action i is not eliminated in round m_i by $*$ is at most $\frac{2}{T\tilde{\Delta}^2_{m_i}}$. □

The next proposition is a modified version of Theorem 2 in [10]. We use it to obtain the regret lower bound in Proposition 1.

PROPOSITION 5. *Suppose Assumptions 1, 2, and 3 hold. Then, under any uniformly good policy $\boldsymbol{\phi}$, we have that, for each action i with $\mu_i < \mu^*$,*

$$\liminf_{t\to\infty} \frac{\mathbb{E}[S_i(t)]}{\log(t)} \geq \frac{1}{D(\theta_i || \theta^*)}. \tag{19}$$

PROOF. This proof follows from the proof of Theorem 2 in [10]. To fix ideas, suppose $i = 1$ is a suboptimal action and suppose action 2 is optimal. Let the parameters of the reward distributions be $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ and the associated means be $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. Then, for any $0 < \delta < 1$, due to Assumptions 1, 2, 3, we have that there exists a parameter λ and mean μ_λ associated with the density function $f(\cdot, \lambda)$ such that

$$\mu_\lambda > \mu_2 \text{ and } |D(\theta_1||\theta_2) - D(\theta_1||\lambda)| \leq \delta D(\theta_1||\theta_2). \quad (20)$$

Now, consider the new sets of parameters $\boldsymbol{\eta} = (\lambda, \ldots, \theta_K)$, where the mean rewards are changed to $(\mu_\lambda, \mu_2, \ldots, \mu_K)$. For this set of parameters, action 1 is the unique optimal. Then, for any uniformly good policy, for $0 < b < \delta$,

$$\mathbb{E}_{\boldsymbol{\eta}}[t - T_1(t)] = o(t^b)$$

and therefore,

$$\mathbb{P}_{\boldsymbol{\eta}}\left[T_1(t) < (1-\delta)\log(t)/D(\theta_1||\lambda)\right] = o(t^{b-1}),$$

similar to the asymptotic lower bound proof in [10]. Now, using the fact that $S_1(t) \geq T_1(t)$, we have

$$\mathbb{P}_{\boldsymbol{\eta}}\left[S_1(t) < (1-\delta)\log(t)/D(\theta_{11}||\lambda)\right] = o(t^{b-1}).$$

Now the rest of the proof of Theorem 2 in [10] applies directly to $S_1(t)$. We will repeat it below for completeness. Let $(Y_i(k))_{k\geq 1}$ be the observations drawn from distribution F_i and define

$$L_m = \sum_{k=1}^{m} \log\left(\frac{f(Y_1(k);\theta_1)}{f(Y_1(k);\lambda)}\right).$$

Now, we have that $\mathbb{P}_{\boldsymbol{\eta}}[C_t] = o(t^{b-1})$ where $C_t = \{S_1(t) < (1-\delta)\log(t)/D(\theta_1||\lambda)$ and $L_{S_1(t)} \leq (1-b)\log(t)\}$.

Now,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\eta}}[T_1(t) = t_1, \ldots, T_K(t) = t_K, L_{s_1} \leq (1-b)\log(t)] \\
&= \mathbb{E}_{\boldsymbol{\eta}}\left[\mathbb{P}_{\boldsymbol{\eta}}\left[T_1(t) = t_1, \ldots, T_K(t) = t_K,\right.\right. \\
&\qquad \left.\left. L_{s_1} \leq (1-b)\log(t) | (Y_i(k))_{\{i\in\mathcal{K}, k=1,\ldots,t}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\prod_{k=1}^{s_1} \frac{f(Y_1(k);\lambda)}{f(Y_1(k);\theta_1)}\right. \\
&\qquad \mathbb{P}_{\boldsymbol{\theta}}\left[T_1(t) = t_1, \ldots, T_K(t) = t_K,\right. \\
&\qquad \left.\left.\left. L_{s_1} \leq (1-b)\log(t) | (Y_i(k))_{\{i\in\mathcal{K}, k=1,\ldots,s_i}\right]\right] \\
&\geq \exp(-(1-b)\log(t))\mathbb{P}_{\boldsymbol{\theta}}\left[T_1(t) = t_1, \ldots,\right. \\
&\qquad \left. T_K(t) = t_K, L_{s_1} \leq (1-b)\log(t)\right], \quad (21)
\end{aligned}$$

In the above, we used the definition $s_1 = \sum_{j\in\mathcal{K}_1} t_j$.

Also, C_t is a disjoint union of events of the form $\{T_1(t) = t_1, \ldots, T_K(t) = t_K, L_{s_1} \leq (1-b)\log(t)\}$ with $t_1 + \ldots + t_M = t$ and $s_1 \leq (1-\delta)\log(t)/D(\theta_1||\lambda)$. Hence using (21),

$$P_{\boldsymbol{\theta}}[C_t] \leq t^{(1-b)} P_{\boldsymbol{\eta}}[C_t] \to 0. \quad (22)$$

By strong law of large numbers $L_m/m \to D(\theta_1||\lambda)$ as $m \to \infty$ and $\max_{k\leq m} L_k/m \to D(\theta_1||\lambda)$ almost surely. Now, since $1 - b > 1 - \delta$, it follows that as $t \to \infty$,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}\left[L_k \geq (1-b)\log(t) \text{ for some}\right. \\
&\qquad \left. k < (1-\delta)\log(t)/D(\theta_1||\lambda)\right] \to 0. \quad (23)
\end{aligned}$$

Hence, from (22) and (23),

$$\mathbb{P}_{\boldsymbol{\theta}}\left[S_1(t) < (1-\delta)\log(t)/D(\theta_1||\lambda)\right] = 0.$$

Now using the above equation with (20) gives us the asymptotic lower bound in (19). □