# Collaborative Learning of Stochastic Bandits over a Social Network

Ravi Kumar Kolla, Krishna Jagannathan and Aditya Gopalan

arXiv:1602.08886v2 [cs.LG] 11 Jul 2016

***Abstract*—We consider a collaborative online learning paradigm, wherein a group of agents connected through a social network are engaged in playing a stochastic multi-armed bandit game. Each time an agent takes an action, the corresponding reward is instantaneously observed by the agent, as well as its neighbours in the social network. We perform a regret analysis of various policies in this collaborative learning setting. A key finding of this paper is that natural extensions of widely-studied single agent learning policies to the network setting need not perform well in terms of regret. In particular, we identify a class of non-altruistic and individually consistent policies, and argue by deriving regret lower bounds that they are liable to suffer a large regret in the networked setting. We also show that the learning performance can be substantially improved if the agents exploit the structure of the network, and develop a simple learning algorithm based on dominating sets of the network. Specifically, we first consider a star network, which is a common motif in hierarchical social networks, and show analytically that the hub agent can be used as an information sink to expedite learning and improve the overall regret. We also derive network-wide regret bounds for the algorithm applied to general networks. We conduct numerical experiments on a variety of networks to corroborate our analytical results.**

## I. Introduction

We introduce and study a collaborative online learning paradigm, wherein a group of agents connected through a social network are engaged in learning a stochastic Multi-Armed Bandit (MAB) problem. In this setting, a set of agents are connected by a graph, representing an information-sharing network among them. At each time, each agent (a node in the social network graph) chooses an action (or *arm*) from a finite set of actions, and receives a stochastic reward corresponding to the chosen arm, from an unknown probability distribution. In addition, each agent shares the action index and the corresponding reward sample instantaneously with its neighbours in the graph. The agents are interested in maximising (minimising) their net cumulative reward (regret) over time. When there is only one learning agent, our setting is identical to the classical multi-armed bandit problem, which is a widely-studied framework for sequential learning [1], [2].

Our framework is motivated by scenarios that involve multiple decision makers acting under uncertainty towards optimising a common goal. One such example is that of a large-scale distributed recommendation system, in which a network of backend servers handles user traffic in a concurrent fashion. Each user session is routed to one of the servers running a local recommendation algorithm. Due to the high volume of recommendation requests to be served, bandwidth and computational constraints may preclude a central processor from having access to the observations from all sessions, and issuing recommendations simultaneously to them in real time. In this situation, the servers must resort to using low-rate information from their neighbours to improve their learning, which makes this a collaborative networked bandit setting.

Another application scenario is that of cooperative transportation routing with mobile applications that provide social network overlays, like Waze [3]. A user in this system is typically interested in taking the fastest or most efficient route through a city, with her app offering a choice of routes, and also recording observations from past choices. In addition, users can also add other trusted users as friends, whose observations then become available as additional information for future decision making. The social network among the users thus facilitates local information exchange, which could help users optimise their future decisions (choices of routes) faster.

In our setting, the agents use their social network to aid their learning task, by sharing their action and reward samples with their immediate neighbours in the graph. It seems reasonable that this additional statistical information can potentially help the agents to optimize their rewards faster than they would if they were completely isolated. Indeed, several interesting questions arise in this collaborative learning framework. For example, how does the structure of the social network affect the rate at which the agents can learn? Can good learning policies for the single agent setting be extended naturally to perform well in the collaborative setting? Can agents exploit their 'place' in the network to learn more efficiently? Can 'more 'privileged' agents (e.g., nodes with high degree or influence) help other agents learn faster? This work investigates and answers some of these questions analytically and experimentally.

### A. Our Contributions

We consider the collaborative bandit learning scenario, and analyse the total regret incurred by the agents (regret of the network) over a long but finite horizon $n$. Our specific contributions in this paper are as follows.

We first introduce and analyse the expected regret of the UCB-Network policy, wherein all the agents employ an extension of the celebrated UCB1 [2] policy. In this case, we derive an upper bound on the expected regret of a generic network.

R. Kolla, K. Jagannathan are with the Department of Electrical Engineering, IIT Madras, Chennai, India 600036. Email: {ee12d024, krishnaj}@ee.iitm.ac.in. A. Gopalan is with the Department of Electrical Communication Engineering, IISc, Bangalore, India 560012. Email: aditya@ece.iisc.ernet.in

The upper bound involves a graph-dependent constant, which is obtained as the solution to a combinatorial optimisation problem. We then specialize the upper bound to common network topologies such as the fully connected and the star graphs, in order to highlight the impact of the social network structure on the derived upper bound.

Second, we derive a universal lower bound on the expected regret of a generic network, for a large class of 'reasonable' policies. This lower bound is based on fundamental statistical limits on the learning rate, and is independent of the network structure. To incorporate the network structure, we derive another lower bound on the expected regret of a generic network, as a function of a graph dependent parameter. This bound holds for the class of *non-altruistic and individually consistent* (NAIC) policies, which includes appropriate extensions of well-studied single agent learning policies, such as UCB1 [2] and Thompson sampling [4] to a network setting. We then observe that the gap between the derived lower bound for the NAIC class of policies, and the upper bound of the UCB-Network policy can be quite large, even for a simple star network[1].

Third, we consider the class of star networks, and derive a refined lower bound on the expected regret of a large star network for NAIC policies. We observe that this refined lower bound matches (in an order sense) the upper bound of the UCB-Network. We thus conclude that widely-studied sequential learning policies (NAIC) which perform well in the single agent setting, may perform poorly in terms of the expected regret of the network when used in a network setting, especially when the network is highly hierarchical.

Next, motivated by the intuition built from our bounds, we seek policies which can exploit the social network structure in order to improve the learning rates. In particular, for an $m$-node star network, we propose a Follow Your Leader (FYL) policy, which exploits the centre node's role as an 'information hub'. We show that the proposed policy suffers a regret which is smaller by a factor of $m$ compared to that of any NAIC policy. In particular, the network-wide regret for the star-network under the FYL policy matches (in an order sense) the universal lower bound on regret. This serves to confirm that using the centre node's privileged role is the right information structure to exploit in a star network.

Finally, we extend the above insights to a generic network. To this end, we make a connection between the smallest *dominating set* of the network, and the achievable regret under the FYL policy. In particular, we show that the expected regret of the network is upper bounded by the product of the *domination number* and the expected regret of a single isolated agent.

In sum, our results on the collaborative bandit learning show that policies that exploit the network structure often suffer substantially lesser expected regret, compared to single-agent policies extended to a network setting.

[1]Our special interest in star graphs is motivated by the fact that social networks often posses a hub-and-spoke structure, where the star is a commonly occurring motif.

### B. Related Work

There is a substantial body of work that deals with the learning of various types of single agent MAB problems [1], [2], [5]–[7]. However, there is relatively little work on the learning of stochastic MAB problems by multiple agents. Distributed learning of a MAB problem by multiple agents has been studied in the context of a cognitive radio frame work in [8]–[10]. Unlike these models, a key novelty in our model is that it incorporates information sharing among the agents since they are connected by a network. In [11], the authors assume that each player, in each round, has access to the entire history corresponding to the actions and the rewards of all users in the network – this is a special case of our generic user network model. In [12], the authors deal with the learning of adversarial MAB problem by multiple agents connected through a network.

The primary focus in [13] is centralized learning, wherein an external agent chooses the actions for the users in the network. The learning of the stochastic MAB problem by multiple users has also been addressed from a game-theoretic perspective in [14]; the randomised algorithm proposed therein uses the parameters of the MAB problem, which are unknown to the algorithm in practice. In contrast, we propose deterministic algorithms that do not require these parameters.

In a class of MAB problems considered in [15]–[17], a sole learning agent receives side observations in each round from *other arms*, in addition to samples from the chosen arm. Another related paper is [18] – here, the model consists of a single major bandit (agent) and a set of minor bandits. While the major bandit observes its rewards, the minor bandits can only observe the actions of the major bandit. However, the bandits are allowed to exchange messages with their neighbours, to receive the reward information of the major bandit. Clearly, the models described above are rather different from the setting we consider in this work.

*Organization.* We describe the system model in Section II. Section III presents the regret analysis of the UCB-Network policy. Lower bounds on the expected regret of the network under certain classes of policies are presented in Section IV. Section V presents the regret analysis of the FYL policy. Numerical results are presented in Section VI, and Section VII concludes the paper.

## II. System Model

We first briefly outline the single agent stochastic MAB problem. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ be the set of arms available to the agent. Each arm is associated with a distribution, independent of others, say $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_K$, and let $\mu_1, \mu_2, \ldots, \mu_K$ be the corresponding means, unknown to the agent. Let $n$ be the time horizon or the total number of rounds. In each round $t$, the agent chooses an arm, for which he receives a reward, an i.i.d. sample drawn from the chosen arm's distribution. The agent can use the knowledge of the chosen arms and the corresponding rewards upto round $(t-1)$ to select an arm in round $t$. The goal of the agent is to maximize the cumulative expected reward up to round $n$.

Now, we present the model considered in this paper. We consider a set of users $V$ connected by an undirected fixed network $G = (V, E)$ [2], with $|V| = m$. Assume that each user is learning the same stochastic MAB problem i.e., faces a choice in each time from among the same set of arms $\mathcal{K}$. In the $t^{th}$ round, each user $v$ chooses an arm, denoted by $a^v(t) \in \mathcal{K}$, and receives a reward, denoted by $X^v_{a^v(t)}(t)$, an i.i.d. sample drawn from $\mathcal{P}_{a^v(t)}$. In the stochastic MAB problem set-up, for a given user $v$, the rewards from arm $i$, denoted by $\{X^v_i(t) : t = 1, 2, \ldots\}$, are i.i.d. across rounds. Moreover, the rewards from distinct arms $i$ and $j$, $X^v_i(t)$, $X^v_j(s)$, are independent. If multiple users choose the same action in a certain round, then each of them gets an independent reward sample drawn from the chosen arm's distribution. We use the subscripts $i$, $v$ and $t$ for arms, nodes and time respectively. The information structure available to each user is as follows. A user $v$ can observe the actions and the respective rewards of itself and its one hop neighbours in round $t$, before deciding the action for round $(t + 1)$.

The policy $\Phi^v$ followed by a user prescribes actions at each time $t$, $\Phi^v(t) : H^v(t) \to \mathcal{K}$, where $H^v(t)$ is the information available with the user till round $t$. A policy of the network $G$, denoted by $\Phi$, comprises of the policies pertaining to all users in $G$. The performance of a policy is quantified by a real-valued random variable, called *regret*, defined as follows. The regret incurred by user $v$ for using the policy $\Phi^v$ upto round $n$ is defined as,

$$R^v_\Phi(n) = \sum_{t=1}^{n} \left(\mu^* - \mu_{a^v(t)}\right) = n\mu^* - \sum_{t=1}^{n} \mu_{a^v(t)},$$

where $a^v(t)$ is the action chosen by the policy $\Phi^v$ at time $t$, and $\mu^* = \max_{1 \le i \le K} \mu_i$. We refer to the arm with the highest expected reward as the optimal arm. The regret of the entire network $G$ under the policy $\Phi$ is denoted by $R^G_\Phi(n)$, and is defined as the sum of the regrets of all users in $G$. The expected regret of the network is given by:

$$\mathbb{E}[R^G_\Phi(n)] = \sum_{v \in V} \sum_{i=1}^{K} \Delta_i \mathbb{E}[T^v_i(n)], \tag{1}$$

where $\Delta_i = \mu^* - \mu_i$, and $T^v_i(n)$ is the number of times arm $i$ has been chosen by $\Phi^v$ upto round $n$. We omit $\Phi$ from the regret notation, whenever the policy can be understood from the context. Our goal is to devise learning policies in order to minimise the expected regret of the network.

Let $\mathcal{N}(v)$ denote the set consisting of the node $v$ and its one-hop neighbours. Let $m^v_i(t)$ be the number of times arm $i$ has been chosen by node $v$ and its one-hop neighbours till round $t$, and $\hat{\mu}_{m^v_i(t)}$ be the average of the corresponding reward samples. These are given as:

$$m^v_i(t) = \sum_{u \in \mathcal{N}(v)} T^u_i(t)$$

$$\hat{\mu}_{m^v_i(t)} = \frac{1}{m^v_i(t)} \sum_{u \in \mathcal{N}(v)} \sum_{k=1}^{t} X^u_{a^u(k)}(k) \mathbb{I}\{a^u(k) = i\},$$

where $\mathbb{I}$ denotes the indicator function. We use $m^G_i(t)$ to denote the number of times arm $i$ has been chosen by all nodes in the network till round $t$.

[2]We use the adjacency matrix $A$ to represent the network $G$. If $(i, j) \in E$ then $A(i, j) = A(j, i) = 1$, otherwise $A(i, j) = A(j, i) = 0$. We assume that $A(i, i) = 1 \; \forall i \in V$.

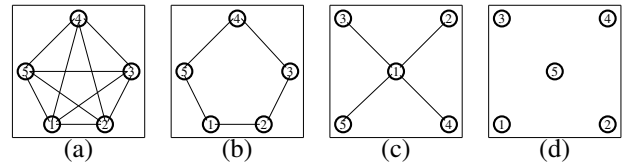


Fig. 1: Various 5-node user networks. (a) fully connected (b) circular (c) star (d) fully disconnected

## III. The UCB-Network policy

Motivated by the well-known single agent policy UCB1 [2], we propose a distributed policy called the UCB-user. This is a deterministic policy, since, for a given action and reward history, the action chosen is deterministic. When each user in the network follows the UCB-user policy, we term the network policy as UCB-Network which is outlined in Algorithm 1.

**Algorithm 1** Upper-Confidence-Bound-Network (UCB-Network)

Each user in $G$ follows UCB-user policy
**UCB-user policy for a user $v$:**
**Initialization:** For $1 \le t \le K$
- play arm $t$
**Loop:** For $K \le t \le n$
- $a^v(t+1) = \underset{j}{\operatorname{argmax}} \; \hat{\mu}_{m^v_j(t)} + \sqrt{\frac{2 \ln t}{m^v_j(t)}}$

The following theorem presents an upper bound on the expected regret of a generic network, under the UCB-Network policy.

*Theorem 1:* Assume that the network $G$ follows the UCB-Network policy to learn a stochastic MAB problem with $K$ arms. Further, assume that the rewards lie in $[0, 1]$. Then,

(i) The expected total regret of $G$ is upper bounded as:

$$\mathbb{E}\left[R^G(n)\right] \le \sum_{i:\mu_i < \mu^*} \left[C_G \frac{8 \ln n}{\Delta_i} + C_G \Delta_i\right] + b,$$

where $\Delta_i = \mu^* - \mu_i$, $\beta \in (0.25, 1)$, $b = m\left(\frac{2}{4\beta - 1} + \frac{2}{(4\beta-1)^2 \ln(1/\beta)}\right)\left(\sum_{j=1}^{K} \Delta_j\right)$, and $C_G$ is a network dependent parameter, defined as follows.

(ii) Let $\gamma_k = \min\{t \in \{1,\dots,n\} : |\{v \in V : m_i^v(t) \geq l_i = \frac{8\ln n}{\Delta_i^2}\}| \geq k\}$ denote the smallest time index when at least $k$ nodes have access to at least $l_i$ samples of arm $i$. Let $\eta_k$ be the index of the 'latest' node to acquire $l_i$ samples of arm $i$ at $\gamma_k$, such that $\eta_k \neq \eta_{k'}$ for $1 \leq k, k' \leq m$. Define $z_k = T_i(\gamma_k) := \big(T_i^1(\gamma_k),\dots,T_i^m(\gamma_k)\big)$, which contains the arm $i$ counts of all nodes at time $\gamma_k$. Then, $C_G l_i$ is the solution of the following optimisation problem:

$$\begin{aligned} &\max \quad \|z_m\|_1 \\ &\text{s.t } \exists \text{ a sequence } \{z_k\}_{k=1}^m \\ &z_j(\eta_k) = z_k(\eta_k) \;\; \forall j \geq k \\ &\langle z_k, A(\eta_k,:)\rangle \geq l_i, \;\; 1 \leq k \leq m \end{aligned} \tag{2}$$

*Proof:* Refer Appendix A. ■

*Interpretation of* (2): Under the UCB-Network policy, suppose a node has acquired at least $l_i$ samples of a sub-optimal arm $i$. As shown in the Lemma 2 in the Appendix A that such a node will not play the sub-optimal arm $i$ subsequently with high probability. Next, note that, $z_k$ is a vector of arm $i$ counts (self plays) of all nodes at time $\gamma_k$. The objective function in (2) represents the sum of arm $i$ counts of all nodes at the smallest time index, when all nodes have access to at least $l_i$ samples of arm $i$. The solution to (2) represents the maximum number of samples of arm $i$ required by the entire network such that $(a)$ Each node has access to at least $l_i$ samples of arm $i$ (the last constraint in (2)), and $(b)$ Each node stops choosing arm $i$ after it has access to $l_i$ samples of it (the penultimate constraint in (2)).

For example, the solution to (2) for an $m$-node star network (shown in Fig. 1) is $(m-1)l_i$. This corresponds to the scenario where the center node never chooses the sub-optimal arm $i$, and each leaf node chooses it $l_i$ times.

*Proof sketch:* First, we show that any node $v$ plays any sub-optimal arm $i$ in a given round $t$ with small probability after it has $l_i$ samples of it, in Lemma 2. Using Lemma 2, we then upper bound the expected regret of the network after each node has $l_i$ samples of the sub-optimal arm $i$. Next, we upper bound the maximum number of samples of the sub-optimal arm $i$ required by the entire network such that each node has access to $l_i$ samples of it, in Lemma 3. Finally, we obtain the desired upper bound by combining Lemma 2 and Lemma 3. A detailed proof, along with Lemma 2 and 3 is given in the Appendix A.

### A. Application to typical networks

Solving (2) for an arbitrary network is analytically complex. Hence, we solve the problem for a few specific networks that range from high connectivity to low connectivity; namely, the $m$-node Fully Connected (FC), circular, star and Fully Disconnected (FD) networks. For $m = 5$, these networks are shown in Fig. 1. It is easy to verify that the solution to (2) for these four networks are $l_i$, $(m-1)l_i$, $\lfloor \frac{m}{2} \rfloor l_i$ and $ml_i$, respectively. We can then evaluate the upper bounds in Theorem 1.

**Corollary 1** For an $m$-node FC network:

$$\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*} \left(\frac{8\ln n}{\Delta_i} + \Delta_i\right) + b. \tag{3}$$

**Corollary 2** For an $m$-node circular network:

$$\mathbb{E}[R^G(n)] \leq \left\lfloor \frac{m}{2} \right\rfloor \sum_{i:\mu_i<\mu^*} \left(\frac{8\ln n}{\Delta_i} + \Delta_i\right) + b. \tag{4}$$

**Corollary 3** For an $m$-node star network:

$$\mathbb{E}[R^G(n)] \leq (m-1) \sum_{i:\mu_i<\mu^*} \left(\frac{8\ln n}{\Delta_i} + \Delta_i\right) + b. \tag{5}$$

**Corollary 4** For an $m$-node FD network:

$$\mathbb{E}[R^G(n)] \leq m \sum_{i:\mu_i<\mu^*} \left(\frac{8\ln n}{\Delta_i} + \Delta_i\right) + b. \tag{6}$$

A key insight can be obtained from the above corollaries is that, the expected regret of a network decreases by a factor of $m$, 2 and $m/(m-1)$ in the cases of $m$-node FC, circular and star networks respectively, compared to FD network.

## IV. Lower bounds on the expected regret

In this section, we derive lower bounds on the expected regret of the network under various classes of policies. Our first lower bound is a universal bound which is independent of the user network, and holds for large class of 'reasonable' learning policies. Second, we derive a network-dependent lower bound for a class of *Non-Altruistic and Individually Consistent* (NAIC) policies – a class that includes network extensions of well-studied policies like UCB1 and Thompson sampling. Finally, we derive a refined lower bound for large star networks under NAIC policies.

Throughout this section, we assume that the distribution of each arm is parametrised by a single parameter. We use $\boldsymbol{\theta} = (\theta_1,\dots,\theta_K) \in \Theta^K = \boldsymbol{\Theta}$ to denote the parameters of arms 1 to $K$ respectively. Suppose $f(x;\theta_j)$ be the reward distribution for arm $j$ with parameter $\theta_j$. Let $\mu(\theta_j)$ be the mean of arm $j$, and $\theta^* = \underset{1\leq j\leq K}{\arg\max}\ \mu(\theta_j)$. Define the parameter sets for an arm $j$ as

$$\boldsymbol{\Theta_j} = \{\boldsymbol{\theta} : \mu(\theta_j) < \max_{i\neq j} \mu(\theta_i)\}$$
$$\boldsymbol{\Theta_j^*} = \{\boldsymbol{\theta} : \mu(\theta_j) > \max_{i\neq j} \mu(\theta_i)\}$$

Note that $\boldsymbol{\Theta_j}$ contains all parameter vectors in which the arm $j$ is a sub-optimal arm, and $\boldsymbol{\Theta_j^*}$ contains all parameter vectors in which the arm $j$ is the optimal arm. Let $kl(\beta||\lambda)$ be the KL divergence of the distribution parametrised by $\lambda$, from the distribution parametrised by $\beta$.

*[A1]* We assume that the set $\Theta$ and $kl(\beta||\lambda)$ satisfy the following [1]:

(i) $f(.;.)$ is such that $0 < kl(\beta||\lambda) < \infty$ whenever $\mu(\lambda) > \mu(\beta)$.

(ii) $\forall\epsilon > 0$ and $\forall\beta,\lambda$ such that $\mu(\lambda) > \mu(\beta), \exists\delta = \delta(\epsilon,\beta,\lambda) > 0$ for which $|kl(\beta||\lambda) - kl(\beta||\lambda')| < \epsilon$ whenever $\mu(\lambda) \leq \mu(\lambda') \leq \mu(\lambda) + \delta$.

(iii) $\Theta$ is such that $\forall \lambda \in \Theta$ and $\forall \delta > 0, \exists \lambda' \in \Theta$ such that $\mu(\lambda) < \mu(\lambda') < \mu(\lambda) + \delta$.

*Theorem 2:* Let $G$ be an $m$-node connected generic network, and suppose *[A1]* holds. Consider the set of policies for users in $G$ to learn a $K$-arm stochastic MAB problem with a parameter vector of arms as $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)] = o(n^c)\ \forall\ c > 0$, for any sub-optimal arm $j$. Then, for $\delta \in (0,1)$, the following holds.

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]}{\ln n} \geq \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta^*)}. \tag{7}$$

*Proof:* Refer Appendix B. ■

Note that the above universal lower bound is based on fundamental statistical limitations, and is independent of the network $G$. Next, we define the class of NAIC policies, and derive a network-dependent lower bound for this class. In the rest of this section, we assume that each arm is associated with a discrete reward distribution, which assigns a non-zero probability to each possible value.

Let $\omega$ be a sample path, which consists of all pairs of actions and the corresponding rewards of all nodes from rounds 1 through $n$:

$$\omega = \{(a^v(t), X^v_{a^v(t)}(t)) : v \in V, 1 \leq t \leq n\}.$$

Also, define

$$\omega_v = \{(a^u(t), X^u_{a^u(t)}(t)) : u \in \mathcal{N}(v), 1 \leq t \leq n\}$$
$$\omega_{\bar{v}} = \{(a^u(t), X^u_{a^u(t)}(t)) : u \in \mathcal{N}(v)^c, 1 \leq t \leq n\}.$$

**Definition 1** [Individually consistent policy] A policy followed by a user $v$ is said to be *individually consistent* if, for any sub-optimal arm $i$, and for any policy of a user $u \in \mathcal{N}(v) \setminus \{v\}$

$$\mathbb{E}[T_i^v(n)|\omega_{\bar{v}}] = o(n^a),\ \forall\, a > 0,\ \forall\, \omega_{\bar{v}}. \tag{8}$$

**Definition 2** [Non-altruistic policy] A policy followed by a user $v$ is said to be *non-altruistic* if there exist $a_1, a_2$, not depending on time horizon $n$, such that the following holds. For any $n$ and any sub-optimal arm $i$, the expected number of times that the policy *plays* arm $i$ after having *obtained* $a_1 \ln n$ samples of that arm is no more than $a_2$, irrespective of the policies followed by the other users in the network.

It can be shown that UCB-user and Thompson sampling [4] are NAIC policies. In particular, we show that the UCB-user policy is an NAIC policy in Lemma 4 in Appendix A.

*Example of a policy which is not individually consistent :* Consider a 2-armed stochastic bandit problem with Bernoulli rewards with means $\mu_1, \mu_2$, where $\mu_1 > \mu_2$. Consider the 3-node line graph with node 2 as the center node. Let the policy followed by node 1 be as follows: $a^1(t) = a^2(t-1)$ for $t > 1$ and $a^1(1) = 2$ (we call this policy *follow node* 2). Consider the following $\omega_{\bar{1}} = \{(a^3(t) = 2, X_2^3(t) = 0) : 1 \leq t \leq n\}$. Then, $\mathbb{E}[T_2^1(n)|\omega_{\bar{1}}] = n$ under the node 2's policy as *follow node* 3, which clearly violates the equation (8). Hence, the *follow node 2* policy for node 1 is not individually consistent.

Note that the above policy, *follow node* $u$, is in fact a non-trivial and rather well-performing policy that we will revisit in Section V. We now derive a network-dependent lower bound for the class of NAIC policies

*Theorem 3:* Let $G = (V, E)$ be a network with $m$ nodes, and suppose [A1] holds. If each node in $V$ follows an NAIC class policy to learn a $K$-arm stochastic MAB problem with a parameter vector of arms as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \boldsymbol{\Theta_j}$, and $\delta \in (0,1)$ then, the following lower bounds hold:

$$(i) \liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)|\omega_{\bar{v}}]}{\ln n} \geq \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta^*)}, \forall v \in V$$
$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta^*)}, \forall v \in V$$
$$(ii) \liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]}{\ln n} \geq L_G \cdot \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta^*)}, \tag{9}$$

where $L_G$ can be obtained from the solution to the following optimisation problem:

$$\begin{aligned} &\min \quad \|z_m\|_1 \\ &\text{s.t } \exists \text{ a sequence } \{z_k\}_{k=1}^m \\ &z_i(\eta_k) = z_k(\eta_k) \ \ \forall i \geq k \\ &\langle z_k, A(n_k, :)\rangle \geq q_j = \frac{1-\delta}{1+\delta} \cdot \frac{\ln n}{kl(\theta_j||\theta^*)}, \ \ \forall k. \end{aligned} \tag{10}$$

*Proof:* Refer Appendix C. ■

The notation used in (10) is the same as the notation in Theorem 1, except that $l_i$ is replaced with $q_j$. Further, $L_G$ is obtained by dividing the solution to (10) by $q_j$. Similar to (2), solving (10) analytically for an arbitrary network is difficult. Hence, we focus on solving (10) for the networks shown in Fig. 1, and provide the corresponding lower bounds below. Let $\Delta_i = \mu(\theta^*) - \mu(\theta_i)$.

**Corollary 5** For an $m$-node FC network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq \sum_{i:\Delta_i>0} \frac{1-\delta}{1+\delta} \cdot \frac{\Delta_i}{kl(\theta_i||\theta^*)}. \tag{11}$$

**Corollary 6** For an $m$-node circular network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq \frac{m}{3} \sum_{i:\Delta_i>0} \frac{1-\delta}{1+\delta} \cdot \frac{\Delta_i}{kl(\theta_i||\theta^*)}. \tag{12}$$

**Corollary 7** For an $m$-node star network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq \sum_{i:\Delta_i>0} \frac{1-\delta}{1+\delta} \cdot \frac{\Delta_i}{kl(\theta_i||\theta^*)}. \tag{13}$$

**Corollary 8** For an $m$-node FD network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq m \sum_{i:\Delta_i>0} \frac{1-\delta}{1+\delta} \cdot \frac{\Delta_i}{kl(\theta_i||\theta^*)}. \tag{14}$$

From corollaries 1-8, we infer that the upper bound of the UCB-Network policy and the lower bound given by (9) are of the same order, for FC $(\ln n)$, circular $(m \ln n)$ and FD $(m \ln n)$ networks. However, for star networks, there is a large gap between the UCB-Network upper bound and the lower

bound for NAIC policies in (13). Since the UCB-Network is an NAIC class policy, we proceed to ascertain if either of these bounds is too loose for star networks. Our special interest in star networks is due to the prevalence of hubs in many social networks, and as we shall see in the next section, this hierarchical structure can be exploited to enhance the learning rate.

Next, we consider a specific instance of a large star network, for which we derive a refined lower bound for the class of NAIC policies. This refined lower bound is of the same order as the regret upper bound for the UCB-Network policy, implying that the upper bound in Theorem 1 is tight in an order sense, and cannot be improved in general.

*Theorem 4:* Let $G_n = (V_n, E_n)$ be a sequence of $m_n$-node star networks learning a 2-arm stochastic MAB problem with mean rewards $\mu_a, \mu_b$ such that $\mu_a > \mu_b$. Suppose $m_n \geq 2 \cdot \frac{\ln n}{kl(\mu_b||\mu_a)}$, and that each node follows an NAIC policy. Then,

$$\liminf_{n\to\infty} \frac{\mathbb{E}[m_2^{G_n}(n)]}{(m_n - 1)\ln n} \geq \frac{1}{kl(\mu_b||\mu_a)}. \tag{15}$$

*Proof:* Refer Appendix D. ■

We now briefly explain the intuition behind Theorem 4. In a large star network, the center node learns the sub-optimal arm very quickly (in a few rounds), since it has access to a large number of samples in each round. Under an NAIC policy, once a node has enough samples to learn that an arm is sub-optimal, by definition, it stops choosing that arm with high probability. Hence, the center node stops choosing the sub-optimal arm with high probability, which in turn ensures that the leaf nodes learn the sub-optimal arm themselves, by choosing the sub-optimal arm $O(\ln n)$ times. This leads to a regret of $O((m-1)\ln n)$. Our simulation results, in Table I, also illustrates this behaviour, for the UCB-Network policy (which is NAIC) on large star networks.

Theorem 4 asserts that, for a fixed, large time horizon $n$, we can construct a large star network with $m$ nodes, whose expected regret is atleast $O((m-1)\ln n)$. This lower bound matches with the upper bound for UCB-Network in Theorem 1. Thus, we conclude that the class of NAIC policies could suffer a large regret, matching the upper bound in an order sense. However, for the same star network and time horizon, the universal lower bound in (7) turns out to be $O(\ln n)$. This gap suggests the possibility that there might exist good learning policies (which are not NAIC) for a star network, with regret matching the universal lower bound. In the next section, we propose one such policy, which does not belong to the NAIC class.

## V. The Follow Your Leader (FYL) Policy

In this section, we first outline a policy called Follow Your Leader (FYL) for a generic $m$-node network. The policy is based on exploiting high-degree hubs in the graph; for this purpose, we define the dominating set and the dominating set partition.

**Definition 3** [Dominating set of a graph] [19] A *dominating set* $D$ of a graph $G = (V, E)$ is a subset of $V$ such that every node in $V \setminus D$ is adjacent to atleast one of the nodes in $D$. The cardinality of the smallest dominating set of $G$ is called as the *domination number*.

**Definition 4** [Dominating set partition of a graph] Let $D$ be a dominating set of $G$. A dominating set partition based on $D$ is obtained by partitioning $V$ into $|D|$ components such that each component contains a node in $D$ and a subset of its one hop neighbors.

Note that given a dominating set for a graph, it is easy to obtain a corresponding dominating set partition. The FYL policy for an $m$-node generic network is outlined in Algorithm 2. Under the FYL policy, all nodes in the dominating set are called *leaders* and all other nodes as *followers*; the follower nodes follow their leaders while choosing an action in a round. As we argued in Section IV, the policy deployed by a follower node in FYL is not individually consistent. The following theorem presents an upper bound on the expected regret of an $m$-node star network which employs the FYL policy.

**Algorithm 2** Follow Your Leader (FYL) Policy

**Input:** Graph $G$, a dominating set $D$ and a dominating set partition
**Leader - Each node in** $D$ :
Follows the UCB-user policy by using the samples of itself and its one-hop neighbours in the same component
**Follower - Each node in** $V \setminus D$ :
In round $t = 1$ :
- Chooses an action randomly from $\mathcal{K}$
In round $t > 1$
- Chooses the action taken by the leader in its component, in the previous round $(t-1)$

*Theorem 5 (FYL regret bound, star networks):* Suppose the star network $G$ with a dominating set as the center node, follows the FYL policy to learn a stochastic MAB problem with $K$ arms. Assume that the rewards lie in $[0, 1]$. Then,

$$\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*}^{K} \frac{8\ln n}{\Delta_i} + d,$$

where $d = \left[2m - 1 + \frac{2m}{4\beta-1}\left(1 + \frac{1}{(4\beta-1)\ln(1/\beta)}\right)\right] \sum_{j=1}^{K} \Delta_j$, $\Delta_i = \mu^* - \mu_i$ and $\beta \in (0.25, 1)$.

*Proof:* Refer Appendix E. ■

A key insight obtained from Theorem 5 is that an $m$-node star network with the FYL policy incurs an expected regret that is lower by a factor $(m-1)$, as compared to any NAIC policy. More importantly, we observe that the regret upper bound under the FYL policy meets the universal lower bound in (7). Hence, we conclude that the FYL policy is order optimal for star networks.

Finally, we present a result that asserts an upper bound on the expected regret of a generic network under the FYL policy.

*Theorem 6 (FYL regret bound, general networks):* Let $D$ be a dominating set of an $m-$node network $G = (V, E)$. Suppose $G$ with the dominating set $D$ employs the FYL policy to learn a stochastic MAB problem with $K$ arms, and the rewards lie in $[0, 1]$, then

$$\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*}^{K} \frac{8|D| \ln n}{\Delta_i} + d.$$

*Proof:* Refer Appendix F. ■

From the above theorem we infer that, the expected regret of a network scales linearly with the cardinality of a given dominating set. Hence, in order to obtain a tighter upper bound, we need to supply a smallest dominating set $D^*$ to the FYL policy. Suppose, if we provide $D^*$ as the input to the FYL policy, then we obtain an improvement of factor $m/|D^*|$ in the expected regret of an $m$-node network compared to the fully disconnected network.

It is known that, computing a smallest dominating set of a given graph is an NP-hard problem [20]. However, fast distributed approximation algorithms for the same are well-known in the literature. For example, Algorithm 35 in [20] finds a smallest dominating set with an approximation factor $\log(\text{MaxDegree}(G))$. Also, upper bounds on the domination number for specific networks such as Erdos-Renyi, power-law preferential attachment and random geometric graphs are available in [21]–[23].

## VI. Numerical Results

We now present some simulations that serve to corroborate our analysis. The simulations have been carried out using MATLAB, and are averaged over 100 sample paths. We fix the time horizon $n$ to be $10^5$.

### A. Performance of UCB-Network on various networks

We consider the following two scenarios: $(i)$ 10 node FC, circular, star and FD networks, 2 arms, Bernoulli rewards with means $0.7, 0.5$, and $(ii)$ 20 node FC, circular, star and FD networks, 10 arms, Bernoulli rewards with means $1, 0.9, 0.8, \ldots, 0.1$. We run the UCB-Network policy for these scenarios, and calculate the expected regret of the network and percentage of time the optimal arm is played by the network. The results are shown in Fig. 2 and 3. It can be observed from Fig. 2 and 3 that the expected regret of the network decreases and the percentage of time the optimal arm is chosen by the network increases, as connectivity of the network increases. This is because, an increase in the connectivity of the network increases the number of observations available to a user, in a given round.

TABLE I: Expected number of times arm 2 played by a node in star networks under UCB-Network policy, 2 armed MAB problem with Bernoulli mean rewards as 0.7 and 0.5

| Size of the network | Center Node | Leaf Node |
|---|---|---|
| 5 | 66 | 448 |
| 10 | 79 | 442 |
| 25 | 33 | 486 |
| 50 | 10 | 502 |
| 100 | 1 | 514 |
| 200 | 1 | 516 |
| 350 | 1 | 513 |

### B. Performance of UCB-Network on star networks

We consider 5, 10, 25, 50, 100, 200 and 350 node star networks, each learning a 2-armed stochastic bandit problem with Bernoulli rewards of means 0.7 and 0.5. We run the UCB-Network policy on the aforementioned networks, and summarise the results in Table I. Observe that, the expected number of times the center node chooses arm 2 (sub-optimal arm) decreases as the network size increases. This forces each leaf node to choose arm 2 on its own in order to learn. Therefore, as the star network size increases, the expected regret of the network can be approximated as the product of the network size and the expected regret of an isolated node.

### C. Comparison of UCB-Network and FYL policies

We consider 25, 100 and 350 node star networks learning a 2-arm stochastic bandit problem with Bernoulli rewards of means 0.7 and 0.5. We run both UCB-Network and FYL policies on the above-mentioned networks. It can be observed from Fig. 4 that the star networks incur much smaller expected regret under the FYL policy, as compared to UCB-Network, and learn the optimal arm much faster.

## VII. Concluding Remarks

We studied the collaborative learning of a stochastic MAB problem by a group of users connected through a social network. We analysed the regret performance of widely-studied single-agent learning policies, extended to a network setting. Specifically, we showed that the class of NAIC policies (such as UCB-Network) could suffer a large expected regret in the network setting. We then proposed and analysed the FYL policy, and demonstrated that exploiting the structure of the network leads to a substantially lower expected regret. In particular, the FYL policy's upper bound on the expected regret matches the universal lower bound, for star networks, proving that the FYL policy is order optimal. This also suggests that using the center node as an information hub is the right information structure to exploit.

In terms of future research directions, we plan to study this model for other flavours of MAB problems such as linear stochastic [24] and contextual bandits [25]. Even in the basic stochastic bandit model considered here, several fundamental questions remain unanswered. For a given network structure, what is the least regret achievable by *any* local information-constrained learning strategy? Is it possible in a general

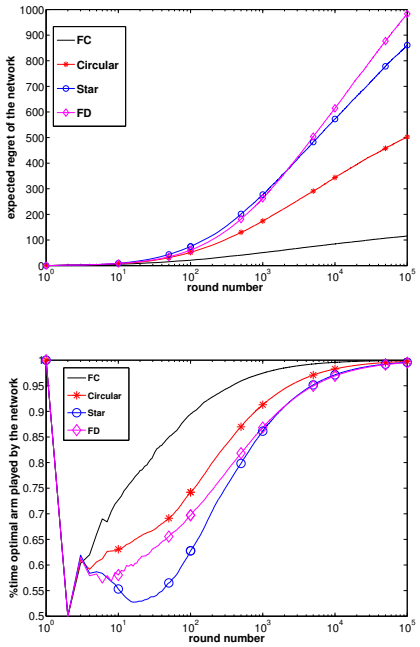


Fig. 2: Performance comparison of UCB-Network policy on various 10 node networks: 2 arms, Bernoulli rewards with means 0.7 and 0.5

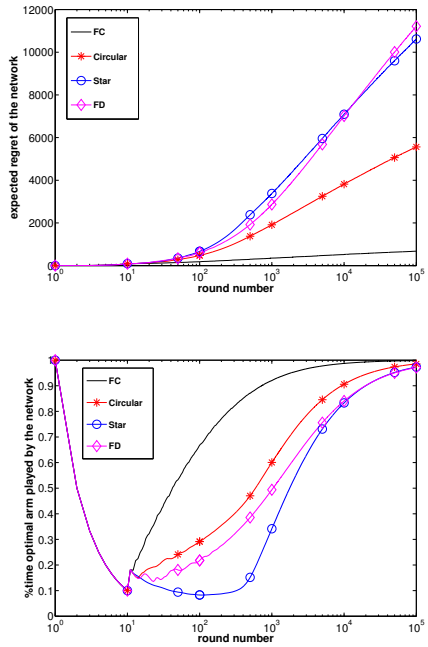


Fig. 3: Performance comparison of UCB-Network policy on various 20 node networks: 10 arms, Bernoulli rewards with means $1, 0.9, \ldots, 0.1$

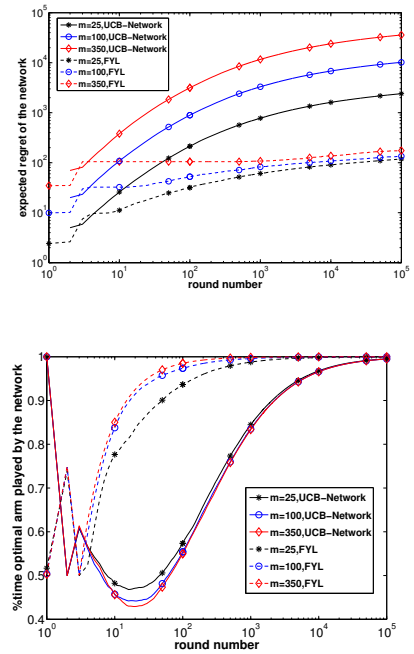


Fig. 4: Performance comparison of UCB-Network and FYL policies on various star networks: 2 arms, Bernoulli rewards with means 0.7 and 0.5

network to outperform 'good single-agent' policies (i.e., those that work well individually, like UCB) run independently throughout the network? If so, what kind of information sharing/exchange might an optimal strategy perform? It is conceivable that there could be sophisticated distributed bandit strategies that could signal within the network using their action/reward sequences, which in turns begs for an approach relying on information-theoretic tools.

## 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] 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, 2002.

[3] W. Mobile, "Waze," *Retrieved January*, vol. 3, p. 2013, 2012.

[4] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," *arXiv preprint arXiv:1111.1797*, 2011.

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

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[7] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

[8] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *Signal Processing, IEEE Transactions on*, vol. 58, no. 11, pp. 5667–5681, 2010.

[9] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: learning under competition," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[10] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multi-player multi-armed bandits," *arXiv preprint arXiv:1505.00553*, 2015.

[11] S. Liu, C. Chen, and Z. Zhang, "Distributed multi-armed bandits: Regret vs. communication," *arXiv preprint arXiv:1504.03509*, 2015.

[12] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," *arXiv preprint arXiv:1602.04741*, 2016.

[13] S. Buccapatnam, A. Eryilmaz, and N. B. Shroff, "Multi-armed bandits in the presence of side observations in social networks," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 2013, pp. 7309–7314.

[14] S. Buccapatnam, J. Tan, and L. Zhang, "Information sharing in distributed stochastic bandits," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*. IEEE, 2015, pp. 2605–2613.

[15] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, "Leveraging side observations in stochastic bandits," *arXiv preprint arXiv:1210.4839*, 2012.

[16] S. Mannor and O. Shamir, "From bandits to experts: On the value of side-observations," in *Advances in Neural Information Processing Systems*, 2011, pp. 684–692.

[17] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren, "Online learning with feedback graphs: Beyond bandits," *arXiv preprint arXiv:1502.07617*, 2015.

[18] S. Kar, H. V. Poor, and S. Cui, "Bandit problems in networks: Asymptotically efficient distributed allocation rules," in *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*. IEEE, 2011, pp. 1771–1778.

[19] T. W. Haynes, S. Hedetniemi, and P. Slater, *Fundamentals of domination in graphs*. CRC Press, 1998.

[20] F. Kuhn, *Lecture notes on Network Algorithms*. `http://bit.ly/1SjmOTt`, Summer term, 2013.

[21] B. Wieland and A. P. Godbole, "On the domination number of a random graph," *The electronic journal of combinatorics*, vol. 8, no. 1, p. R37, 2001.

[22] F. Molnár Jr, N. Derzsy, É. Czabarka, L. Székely, B. K. Szymanski, and G. Korniss, "Dominating scale-free networks using generalized probabilistic methods," *Scientific reports*, vol. 4, 2014.

[23] A. Bonato, M. Lozier, D. Mitsche, X. Pérez-Giménez, and P. Prałat, "The domination number of on-line social networks and random geometric graphs," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2015, pp. 150–163.
[24] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.
[25] 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*. ACM, 2010, pp. 661–670.
[26] S. Bubeck, "Jeux de bandits et fondations du clustering," Ph.D. dissertation, Citeseer, 2010.

## Appendix A

We require the following Lemma 1, 2, 3 and inequality to prove Theorem 1.

*Hoeffding's Maximal Inequality* [26]: Let $X_1, X_2, \dots$ be centered i.i.d random variables lying in $[0,1]$. Then, for any $x > 0$ and $t \geq 1$,

$$\mathbb{P}\left(\exists s \in \{1,\dots,t\}, \sum_{i=1}^{s} X_i > x\right) \leq \exp\left(-\frac{2x^2}{t}\right).$$

In order to introduce Lemma 1, we need the following.

Consider a new probability space with probability measure $\tilde{\mathbb{P}}$, for the rewards corresponding to all arms. First, for a fixed node $v \in V$, for each action $i \in \mathcal{K}$, we consider a sequence of i.i.d. random variables $\{Y_i(k)\}_{k=1}^{\infty}$ with arm $i$'s distribution. If a node $v$ or its neighbours choose an arm $i$, then they receive the rewards from the sequence $\{Y_i(k)\}_{k=1}^{\infty}$. Next, for each $u \in V \setminus \mathcal{N}(v)$, for each action $i \in \mathcal{K}$, we consider a sequence of i.i.d. random variables $\{X_i^u(k)\}_{k=1}^{\infty}$ with arm $i$'s distribution. If a node $u \in V \setminus \mathcal{N}(v)$ chooses an arm $i$, then it receives a reward from the sequence $\{X_i^u(k)\}_{k=1}^{\infty}$. Recall that, in the setting described in Section II, if a user $v$ chooses arm $i$, then it receives a reward from the sequence $\{X_i^u(k)\}_{k=1}^{\infty}$. In this probability space, we considered the probability measure to be $\mathbb{P}$.

We prove that the probability of a sample path of the network in both probability spaces are equal, in the following lemma. Hence, this allows us to equivalently work in the new probability space, as and when appropriate.

*Lemma 1:* Consider an $m$-node undirected user graph. Let $A(t)$ and $Z(t)$ be the random variables which indicate the actions chosen by all nodes and the corresponding rewards, in round $t$. Let $E(k) = (A(k), Z(k), \dots, A(1), Z(1))$. Then, $\forall t \geq 1$,

$$\mathbb{P}[E(t) = (\bar{a}_{1:t}, \bar{z}_{1:t})] = \tilde{\mathbb{P}}[E(t) = (\bar{a}_{1:t}, \bar{z}_{1:t})],$$

where $\bar{a}_{1:t} = (\bar{a}_1, \dots, \bar{a}_t)$, $\bar{z}_{1:t} = (\bar{z}_1, \dots, \bar{z}_t)$ with $\bar{a}_k \in \mathcal{K}^m$ and $\bar{z}_k \in [0,1]^m$ for any $k \geq 1$.

*Proof:* We establish the result using induction on $t$. The result trivially holds for $t = 1$, since a policy does not possess any information in the very first round itself. Assume that it is true for $t = k$. Then,

$$\mathbb{P}[E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})] = \tilde{\mathbb{P}}[E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})].$$

Now, we prove that the result holds for $t = k+1$.

$$\begin{aligned}
&\mathbb{P}[E(k+1) = (\bar{a}_{1:k+1}, \bar{z}_{1:k+1})] \\
&= \mathbb{P}[A(k+1) = \bar{a}_{k+1}, Z(k+1) = \bar{z}_{k+1}, E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})], \\
&= \mathbb{P}[A(k+1) = \bar{a}_{k+1}, Z(k+1) = \bar{z}_{k+1} | E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})] \\
&\qquad \cdot \mathbb{P}[E(k) = (a_{1:k}, \bar{z}_{1:k})], \\
&= \mathbb{P}[A(k+1) = \bar{a}_{k+1}, Z(k+1) = \bar{z}_{k+1} | E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})] \\
&\qquad \cdot \tilde{\mathbb{P}}[E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})], \qquad (16)
\end{aligned}$$

since we assumed that the result is true for $t = k$. Note that, in our model, the actions taken by a policy in round $(k+1)$ for a given $E(k)$, are independent of the probability space from which the rewards are generated. Further, the reward distributions of arms are identical in both probability spaces $\mathbb{P}$ and $\tilde{\mathbb{P}}$. Therefore,

$$\begin{aligned}
&\mathbb{P}[A(k+1) = \bar{a}_{k+1}, Z(k+1) = \bar{z}_{k+1} | E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})] \\
&= \tilde{\mathbb{P}}[A(k+1) = \bar{a}_{k+1}, Z(k+1) = \bar{z}_{k+1} | E(k) = (\bar{a}_{1:k}, \bar{z}_{1:k})]. \qquad (17)
\end{aligned}$$

By substituting (17) in (16), we obtain

$$\mathbb{P}\left[E(k+1) = (\bar{a}_{1:k+1}, \bar{z}_{1:k+1})\right] = \tilde{\mathbb{P}}\left[E(k+1) = (\bar{a}_{1:k+1}, \bar{z}_{1:k+1})\right],$$

which completes the proof. ■

*Lemma 2:* Let $c_{t,S} = \sqrt{\frac{2\ln t}{S}}$, $\beta \in (0,1)$. For each $v \in V$ and sub-optimal arm $i$, define $\tau_i^v$ as follows:

$$\tau_i^v = \min\{t \in [n] : m_i^v(t) \geq l_i = \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil\}.$$

Then, for each $t > \tau_i^v$,

$$\mathbb{P}\left(\{\hat{\mu}_{m_*^v(t)} + c_{t,m_*^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}\}\right) \leq 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}.$$

*Proof:* For convenience, we denote

$$A_i^v(t) = \{\hat{\mu}_{m_*^v(t)} + c_{t,m_*^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}\}.$$

Note that,

$$\mathbb{P}\left(A_i^v(t) \cap \{t > \tau_i^v\}\right) = \mathbb{P}\left(A_i^v(t) \cap \{m_i^v(t) \geq l_i\}\right). \qquad (18)$$

Observe that, the event $A_i^v(t)$ occurs only if atleast one of the following events occur.

$$\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\}, \qquad (19)$$

$$\{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\}, \qquad (20)$$

$$\{\mu^* < \mu_i + 2c_{t,m_i^v(t)}\}. \qquad (21)$$

Note that, the event given by (21) does not occur when the event $\{m_i^v(t) \geq l_i\}$ occurs. Hence,

$$\begin{aligned}
&\mathbb{P}\left(A_i^v(t) \cap \{m_i^v(t) \geq l_i\}\right) \leq \\
&\mathbb{P}(\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\} \cup \{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\} \\
&\qquad \cap \{m_i^v(t) \geq l_i\}), \\
&\leq \mathbb{P}\left(\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\}\right) \\
&\qquad + \mathbb{P}\left(\{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\}\right). \qquad (22)
\end{aligned}$$

For each node $v \in V$ and each arm $i$, the initialization phase of the UCB-user policy implies that $|\mathcal{N}(v)| \leq m_i^v(t) \leq |\mathcal{N}(v)|t$. Therefore,

$$\mathbb{P}\left(\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\right) \leq \mathbb{P}\left(\exists s_* \in \{|\mathcal{N}(v)|, \ldots, |\mathcal{N}(v)|t\} : \hat{\mu}_{s_*} \leq \mu^* - c_{t,s_*}\right),$$

$$\leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \mathbb{P}\Big(\exists s_* : |\mathcal{N}(v)|\beta^{j+1}t < s_* \leq |\mathcal{N}(v)|\beta^j t, \; s_*\hat{\mu}_{s_*} \leq s_*\mu^* - \sqrt{2s_* \ln t}\Big), \tag{23}$$

$$\leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \mathbb{P}\Big(\exists s_* : |\mathcal{N}(v)|\beta^{j+1}t < s_* \leq |\mathcal{N}(v)|\beta^j t, \; s_*\hat{\mu}_{s_*} \leq s_*\mu_* - \sqrt{2|\mathcal{N}(v)|\beta^{j+1}t \ln t}\Big). \tag{24}$$

Here, (23) is due to the peeling argument on geometric grid over $[|\mathcal{N}(v)|, |\mathcal{N}(v)|t]$. This implies that, for $\beta \in (0,1)$, $a \geq 1$, if $s \in \{a, \ldots, at\}$ then there exists $j \in \{0, \ldots, \frac{\ln t}{\ln(1/\beta)}\}$ such that $a\beta^{j+1}t < s \leq a\beta^j t$. Now, we proceed to bound the probability of the event given by (24) using Hoeffding's maximal inequality and Lemma 1. Hence,

$$\mathbb{P}\left(\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\right) \leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \exp\left(-4\beta \ln t\right), \leq \left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}. \tag{25}$$

Similarly, we can show that

$$\mathbb{P}\left(\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\right) \leq \left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}. \tag{26}$$

Substituting (25) and (26) in (22) gives the desired result. ■

*Lemma 3:* Let $\tau_i^v$ $\forall v \in V$, and $l_i, \forall 1 \leq i \leq K$ be as defined in the Lemma 2. Assume that a node $v$ stops playing the sub-optimal arm $i$ at time $\tau_i^v$. Then, for an arm $i$, $\sum_{v \in V} T_i^v(\tau_i^v) \leq C_G l_i$, where $C_G l_i$ is the solution to the optimisation problem in (2).

*Proof:* We first evaluate the value of the random variable $\sum_{v=1}^{m} T_i^v(\tau_i^v)$ for all realizations. Then, we determine the maximum value of the random variable over all realizations. The following algorithm gives the value of the above mentioned random variable for a realization. Consider an $m$ length column vector of zeros, say $y$.

*Algorithm*:

Step 1: Select an integer $I$ from $B = \{1, 2, \ldots, m\}$.

Step 2: Increase $y(I)$ by 1, *i.e.,* $y(I) = y(I) + 1$.

Step 3: Find the indices (say $C$) corresponding to elements in $Ay$ which are atleast $l_i$. Here, $A$ is the adjacency matrix of the graph $G$.

Step 4: Update $B = B \setminus C$ and $A$ by removing rows corresponding to $C$ in $A$

Step 5: Go to step 1, if $B$ is non-empty else stop by returning $y$.

Here, step 4 ensures that nodes having $l_i$ samples of arm $i$ stops playing arm $i$ further. Observe that $\|y\|_1$, where $y$ is the vector returned by the above algorithm, yields the value of the random variable $\sum_{v=1}^{m} T_i^v(\tau_i^v)$ for a realization. Therefore, it suffices to maximize $\|y\|_1$ over all realizations.

The optimisation problem in (2) captures the above. The final constraint in (2) ensures that the node $\eta_k$ has $l_i$ samples of sub-optimal arm $i$ at time instance $\gamma_k$. Recall that, $\gamma_k$ is a random variable which tracks the *least* time at which atleast $k$ nodes have more than $l_i$ samples of arm $i$. The penultimate constraint ensures that sub-optimal arm $i$ count of node $\eta_k$ does not increase(or stop playing arm $i$) after time instance $\gamma_k$. Hence, a feasible point in the above optimisation problem is a sequence $\{z_k\}_{k=1}^m$ which satisfies the aforementioned two constraints. Then, $\|z_m\|_1$ corresponds to the value of the random variable $\sum_{v=1}^{m} T_i^v(\tau_i^v)$ for a realization. ■

By using the above lemmas, we now prove Theorem 1.

*Proof:* From (1), we need to upper bound $\mathbb{E}[T_i^v(n)]$ for all $v \in V$ in order to upper bound the expected regret of $G$. Let $B_i^v(t)$ be the event that node-$v$ plays sub-optimal action-$i$ in round $t$:

$$B_i^v(t) = \{\hat{\mu}_{m_j^v(t)} + c_{t,m_j^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}, \forall j \neq i\}, \subseteq \{\hat{\mu}_{m_*^v(t)} + c_{t,m_*^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}\}. \tag{27}$$

Hence,

$$\mathbb{E}\left[\sum_{v=1}^{m} T_i^v(n)\right] = \mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}[\mathbb{I}_{\{t \leq \tau_i^v, B_i^v(t)\}} + \mathbb{I}_{\{t > \tau_i^v, B_i^v(t)\}}]\right], = \underbrace{\mathbb{E}\left[\sum_{v=1}^{m} T_i^v(\tau_i^v)\right]}_{(a)} + \underbrace{\mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}\mathbb{I}_{\{t > \tau_i^v, B_i^v(t)\}}\right]}_{(b)}. \tag{28}$$

Now, we upper bound $(b)$ in (28). Let $1 \leq v \leq m$. Since, $m_i^v(t) \geq l_i$ for $t > \tau_i^v$,

$$\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}_{\{t > \tau_i^v\}}\mathbb{I}_{B_i^v(t)}\right] = \sum_{t=1}^{n}\mathbb{P}\left(B_i^v(t), \{t > \tau_i^v\}\right), \overset{(c)}{\leq} \sum_{t=1}^{\infty} 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}, \leq \int_1^{\infty} 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}\,\mathrm{d}t, = \frac{2}{4\beta - 1} + \frac{2}{(4\beta-1)^2\ln(1/\beta)},$$

where $(c)$ is due to Lemma 2. Thus, $(b)$ in (28) upper bounded as

$$\mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}\mathbb{I}_{\{t > \tau_i^v\}}\mathbb{I}_{B_i^v(t)}\right] \leq m\left(\frac{2}{4\beta - 1} + \frac{2}{(4\beta-1)^2\ln(1/\beta)}\right). \tag{29}$$

Now, we upper bound the random variable in $(a)$ in (28) for all realizations. Consider a new system in which each node $v$ stops playing sub-optimal arm $i$ for $t > \tau_i^v$. By using Lemma 3, we can calculate an upper bound on $\sum_{v=1}^{m} T_i^v(\tau_i^v)$. It is easy to see that the same upper bound also holds for $(a)$ in (28). Hence,

$$\mathbb{E}\left[\sum_{v=1}^{m} T_i^v(\tau_i^v)\right] \leq C_G l_i. \tag{30}$$

Combining (28), (29) and (30) establishes the desired result. ■

*Lemma 4:* Consider a network $G = (V, E)$ learning a $K$-arm stochastic MAB problem with mean rewards $\mu_1 \geq \mu_2 \geq \ldots \mu_K$. Assume that, each arm distribution is discrete and it assigns a non-zero probability to each possible value. Then, the UCB-user policy followed by any user $v$ in $G$ to learn the above MAB problem is non-altruistic and individually consistent (NAIC) policy.

*Proof:* First, we prove the non-altruistic part. Lemma 2 gives an upper bound on the probability that a node $v$ following the UCB-user policy plays any sub-optimal arm $i$ in round $t$, after it has obtained $l_i = \frac{8 \ln n}{\Delta_i^2}$ samples of the arm $i$, where $n$ is the time horizon. We can treat $\frac{8}{\Delta_i^2}$ as $a$ in the definition of non-altruistic policy. Observe that, in (29), we upper bounded the expected number of times a node $v$ chooses any sub-optimal arm $i$, after it has access to $l_i$ samples of arm $i$, till $n$. Note that, this upper bound is a constant. Hence, the UCB-user policy satisfies the non-altruistic property. Now, we prove the individually consistent part. Recall that, $\omega_{\bar{v}}$ contains actions and the corresponding rewards of the nodes outside the neighbourhood of node $v$, from round 1 to $n$. Note that, the event $A_i^v(t)$ defined in the proof of Lemma 2 is independent of any $\omega_{\bar{v}}$, given the event $\{m_i^v(t) = a, m_*^v(t) = b\}$. Hence, on the lines of Lemma 2, for $\beta \in (0, 1)$, $t > \tau_i^v$ (same as defined in Lemma 2),

$$\mathbb{P}\left(A_i^v(t)|\omega_{\bar{v}}\right) \leq 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}. \tag{31}$$

Thus,

$$\begin{aligned}\mathbb{E}[T_i^v(n)|\omega_{\bar{v}}] &\leq l_i + \sum_{t=\tau_i^v+1} \mathbb{P}\left(A_i^v(t)|\omega_{\bar{v}}\right), \\ &\leq \frac{8 \ln n}{\Delta_i^2} + \sum_{t=1}^{\infty} \mathbb{P}\left(A_i^v(t)|\omega_{\bar{v}}\right), \\ &= \left(\frac{8 \ln n}{\Delta_i^2} + O(1)\right) \in o(n^c), \text{ for any } c > 0.\end{aligned}$$

Therefore, the UCB-user policy followed by a node $v$ satisfy individually consistent property, which completes the proof. ■

## Appendix B

*Proof of Theorem 2.*

*Proof:* Follows from Theorem 2 in [1], by considering $m_i^G(n)$ instead of $T_i(n)$ in the event $C_n$ defined in the respective proof. ■

## Appendix C

*Proof of Theorem 3.*

We now prove $(i)$ in Theorem 3, in the following lemma. With the aid of this lemma, we then prove the second part of the theorem.

*Lemma 5:* Consider a node $v$ in a network $G$. Assume that node $v$ follows an NAIC policy, and suppose [A1] holds. Further, assume that each arm is associated with a discrete distribution such that it assigns a non-zero positive probability to each possible value. Then, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta_j}$, and for any $\omega_{\bar{v}}$, the following holds:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)|\omega_{\bar{v}}]}{\ln n} \geq \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta_1)},$$
$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1-\delta}{1+\delta} \cdot \frac{1}{kl(\theta_j||\theta_1)}.$$

*Proof:* Without loss of generality, assume that $\theta_1 = \theta^*$ and $j = 2 \Rightarrow \boldsymbol{\theta} \in \boldsymbol{\Theta_2}$. Consider a new parameter vector $\boldsymbol{\gamma} = (\theta_1, \lambda, \theta_3, \ldots, \theta_K)$ such that $\mu(\lambda) > \mu(\theta^*)$, $j \neq 1$. Note that, arm 1 is optimal under parameter vector $\boldsymbol{\theta}$, while arm 2 is optimal under parameter vector $\boldsymbol{\gamma}$. Let $X_{2,1}, \ldots, X_{2,n}$ be $n$ i.i.d samples generated from the sub-optimal arm 2's distribution with parameter vector $\boldsymbol{\theta}$. Define

$$\hat{kl}_s = \sum_{t=1}^{s} \ln\left(\frac{f(X_{2,t};\theta_2)}{f(X_{2,t};\lambda)}\right).$$

For any $v \in V$ and any sub-optimal arm $j$, and $0 < a < \delta$, we define

$$C_n^v = \{m_2^v(n) < \frac{(1-\delta)\ln n}{kl(\theta_2||\lambda)} \text{ and } \hat{kl}_{m_2^v(n)} \leq (1-a)\ln n\}, \tag{32}$$

where $\hat{kl}_{m_2^v(n)} = \sum_{u\in\mathcal{N}(v)} \sum_{t=1}^{T_2^u(n)} \ln\left(\frac{f(X_{2,t}^u;\theta_2)}{f(X_{2,t}^u;\lambda)}\right)$, since $\{X_{2,t}^u\}_{u\in\mathcal{N}(v)}$ are i.i.d. For convenience, let $g_n = \frac{(1-\delta)\ln n}{kl(\theta_2||\lambda)}$ and $h_n = (1-a)\ln n$. For a given $\omega_{\bar{v}}$, observe that $C_n^v$ is a disjoint union of events of the form $\{m_1^v(n) = n_1, m_2^v(n) = n_2, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n\}$ with $n_1 + n_2 \cdots + n_K = n|\mathcal{N}(V)|$ and $n_2 \leq g_n$. Further, $\{m_2^v(n) = n_2\}$ is also a disjoint union of the events of the form $\{\cap_{u\in\mathcal{N}(v)} T_2^u(n) = q_u\}$ with $\sum_{u\in\mathcal{N}(v)} q_u = n_2$. Since $\boldsymbol{\gamma} = (\theta_1, \lambda, \theta_3, \ldots, \theta_K)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \ldots, \theta_K)$, we write

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\gamma}}\{m_1^v(n) = n_1, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n|\omega_{\bar{v}}\} = \\ &\mathbb{E}_{\theta}\Bigg[\mathbb{I}_{\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\hat{kl}_{n_2}\leq h_n\}} \\ &\prod_{u\in\mathcal{N}(v)} \prod_{t=1}^{T_2^u(n)=q_u} \frac{f(X_{2,t}^u;\lambda)}{f(X_{2,t}^u;\theta_2)}\Bigg].\end{aligned} \tag{33}$$

However, $\prod_{u\in\mathcal{N}(v)}\prod_{t=1}^{q_u}\frac{f(X^u_{2,t};\lambda)}{f(X^u_{2,t};\theta_2)}=\exp(-\hat{kl}_{n_2})$. Therefore,

$$\mathbb{P}_{\boldsymbol{\gamma}}\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\hat{kl}_{n_2}\le h_n|\omega_{\bar{v}}\}= \mathbb{E}_\theta\left[\mathbb{I}_{\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\hat{kl}_{n_2}\le h_n\}}\exp(-\hat{kl}_{n_2})\right].$$

Note that, $\exp(-\hat{kl}_{n_2})\ge n^{-(1-a)}$, since $\hat{kl}_{n_2}\le h_n$ in the region of integration. Therefore,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\gamma}}\{&m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\\ &\hat{kl}_{n_2}\le h_n|\omega_{\bar{v}}\}\\ &\ge n^{-(1-a)}\mathbb{P}_{\boldsymbol{\theta}}\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\\ &\hat{kl}_{n_2}\le h_n|\omega_{\bar{v}}\}. \quad (34)\end{aligned}$$

Hence,

$$\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})\ge n^{-(1-a)}\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}}). \quad (35)$$

Now, we bound $\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}})$ as follows:

$$\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})\le\mathbb{P}_{\boldsymbol{\gamma}}\left(m_2^v(n)<g_n|\omega_{\bar{v}}\right).$$

Since $T_2^v(n)\le m_2^v(n)$,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})&\le\mathbb{P}_{\boldsymbol{\gamma}}\left(T_2^v(n)<g_n|\omega_{\bar{v}}\right),\\ &=\mathbb{P}_{\boldsymbol{\gamma}}\left(n-T_2^v(n)>n-g_n|\omega_{\bar{v}}\right).\end{aligned}$$

Note that, $n|\mathcal{N}(v)|-m_2^v(n)$ is a non-negative random variable and $kl(\theta_2||\lambda)>0$. Therefore, applying Markov's inequality to the right-hand side in the above equation, we obtain

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})&\le\frac{\mathbb{E}_{\boldsymbol{\gamma}}[n-T_2^v(n)|\omega_{\bar{v}}]}{n-g_n},\\ &=\frac{\sum_{i=1,i\ne2}^{K}\mathbb{E}_{\boldsymbol{\gamma}}[T_i^v(n)|\omega_{\bar{v}}]}{n-g_n}=\frac{(K-1)o(n^a)}{n-O(\ln n)},\end{aligned}$$

for $0<a<\delta$, since arm 2 is the unique optimal arm under $\gamma$. Hence,

$$\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}})\le n^{(1-a)}\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})=o(1). \quad (36)$$

Observe that,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\left(C_n^v|\omega_{\bar{v}}\right)\ge\mathbb{P}_{\boldsymbol{\theta}}\Big(&m_2^v(n)<g_n,\\ &\frac{1}{g_n}\max_{i\le g_n}\hat{kl}_i\le\frac{kl(\theta_2||\lambda)(1-a)}{(1-\delta)}\Big|\omega_{\bar{v}}\Big), \quad (37)\end{aligned}$$

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\frac{1}{g_n}\max_{i\le g_n}\hat{kl}_i\le\frac{kl(\theta_2||\lambda)(1-a)}{(1-\delta)}\right)\to1, \quad (38)$$

due to $\frac{1-a}{1-\delta}>1$ and the maximal version of the Strong Law of Large Numbers which is given below.

*Maximal version of SLLN* [7]: Let $\{X_t\}$ be a sequence of independent real-valued random variables with positive mean $\mu>0$. Then,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}X_t=\mu\ a.s.\Rightarrow\lim_{n\to\infty}\frac{1}{n}\max_{s=1,\ldots,n}\sum_{t=1}^{s}X_t=\mu\ a.s.$$

From (36), (37) and (38), we obtain

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)<g_n|\omega_{\bar{v}}\right)=o(1),\ \forall\omega_{\bar{v}},\\ &\Rightarrow\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)<g_n\right)=o(1).\end{aligned}$$

Part (iii) of assumption, $[A1]$, guarantees the existence of a $\lambda\in\Theta$ such that $\mu(\theta_1)<\mu(\lambda)<\mu(\theta_1)+\delta$ holds. Combining $\mu(\theta_1)>\mu(\theta_2)$ with the part (i) of $[A1]$, we obtain $0<kl(\theta_2||\theta_1)<\infty$. From part (ii) of $[A1]$, we deduce that $|kl(\theta_2||\theta_1)-kl(\theta_2||\lambda)|<\epsilon$, since $\mu(\theta_1)\le\mu(\lambda)\le\mu(\theta_1)+\delta$ for some $\delta$. Let $\epsilon$ be $\delta kl(\theta_2||\theta_1)$. Hence, we write the following:

$$|kl(\theta_2||\lambda)-kl(\theta_2||\theta_1)|<\delta kl(\theta_2||\theta_1),\qquad\text{for }0<\delta<1.$$

Hence,

$$\begin{aligned}&\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)<\frac{1-\delta}{1+\delta}\cdot\frac{\ln n}{kl(\theta_2||\theta_1)}\Big|\omega_{\bar{v}}\right)=o(1),\\ &\Rightarrow\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)<\frac{1-\delta}{1+\delta}\cdot\frac{\ln n}{kl(\theta_2||\theta_1)}\right)=o(1).\end{aligned}$$

Furthermore,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[m_2^v(n)|\omega_{\bar{v}}]=\sum_i i\cdot\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)=i|\omega_{\bar{v}}\right),\\ &\ge\left(\frac{1-\delta}{1+\delta}\right)\frac{\ln n}{kl(\theta_2||\theta_1)}\mathbb{P}_{\boldsymbol{\theta}}\left(m_2^v(n)>\frac{1-\delta}{1+\delta}\cdot\frac{\ln n}{kl(\theta_2||\theta_1)}\Big|\omega_{\bar{v}}\right),\\ &=\left(\frac{1-\delta}{1+\delta}\right)\frac{\ln n}{kl(\theta_2||\theta_1)}(1-o(1)).\end{aligned}$$

Hence, we have proved that for any $v\in V$, $\omega_{\bar{v}}$ and any sub-optimal arm $j$,

$$\begin{aligned}\liminf_{n\to\infty}\frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)|\omega_{\bar{v}}]}{\ln n}&\ge\frac{1-\delta}{1+\delta}\cdot\frac{1}{kl(\theta_j||\theta_1)},\\ \liminf_{n\to\infty}\frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n}&\ge\frac{1-\delta}{1+\delta}\cdot\frac{1}{kl(\theta_j||\theta_1)},\end{aligned}$$

which completes the proof of this lemma, and establishes $(i)$ in Theorem 3. ■

With the help of this, we now prove the second part of Theorem 3.

*Proof:* Lemma 5 implies that for each $v\in V$, $\exists\,n_v\in\mathbb{N}$ such that

$$\frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n}\ge\frac{1-\delta}{1+\delta}\cdot\frac{1}{kl(\theta_j||\theta_1)},\ \forall n\ge n_v. \quad (39)$$

Let $n'=\max(n_v:v\in V)$. Using (39) for each $v\in V$, and for $n\ge n'$, we determine a lower bound for $\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]$. It is easy to see that the solution to the following optimisation problem is a valid lower bound for $\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]$ for $n\ge n'$.

$$\begin{aligned}&\text{minimize}\ \ \|z_m\|_1\\ &\text{s.t}\ \exists\text{ a sequence }\{z_k\}_{k=1}^m\\ &z_i(\eta_k)=z_k(\eta_k)\ \ \forall i\ge k,\\ &\langle z_k,A(n_k,:)\rangle\ge\frac{1-\delta}{1+\delta}\cdot\frac{1}{kl(\theta_j,\theta_1)}\ln n\ \ \forall k.\end{aligned} \quad (40)$$

Note that, the notation in (40) is same as used in Theorem 1, Lemma 3. Let $L_G\left(\frac{1-\delta}{1+\delta}\right)\frac{\log n}{kl(\theta_j||\theta_1)}$ be the solution of (40). Thus,

$$\mathbb{E}[m_j^G(n)] \geq L_G\left(\frac{1-\delta}{1+\delta}\right)\frac{\ln n}{kl(\theta_j||\theta_1)}, \ \ \forall n \geq n',$$
$$\Rightarrow \liminf_{n\to\infty}\frac{\mathbb{E}[m_j^G(n)]}{\ln n} \geq L_G\left(\frac{1-\delta}{1+\delta}\right)\frac{1}{kl(\theta_j||\theta_1)},$$

which establishes the desired result. ■

## APPENDIX D

*Proof of Theorem 4.*

*Proof:* Without loss of generality we consider that node 1 is the center node and node 2 through $m_n$ are leaf nodes. Since a policy does not possess any information in the first round, it chooses arm 1 with probability $p_1$ and arm 2 with probability $p_2$, such that $0 \leq p_1, p_2 \leq 1$ and $p_1 + p_2 = 1$. Now, we find the expected number of nodes that chose the arm with parameter $\mu_b$ in the first round as follows:

$$\mathbb{E}[m_b^{G_n}(1)] = \sum_{v\in V}\left(\frac{1}{2}p_2 + \frac{1}{2}p_1\right) = \frac{m_n}{2} \geq \frac{\ln n}{kl(\mu_b,\mu_a)}, \tag{41}$$

since MAB is $(\mu_a, \mu_b)$ with probability $\frac{1}{2}$, and is $(\mu_b, \mu_a)$ with probability $\frac{1}{2}$. Henceforth, for convenience, we replace $a$ with 1 and $b$ with 2. Let $m_i^{G,v}(t)$ be a random variable indicating the total number of times arm $i$ has been chosen by node $v$ and its one hop neighbours till round $t$, in the network $G$. Note that, $m_2^{G_n}(1)$ is equals to $m_2^{G_n,1}(1)$, since the network in consideration is a star network with node 1 as the center node. Therefore,

$$\mathbb{E}[m_2^{G_n,1}(1)] \geq \frac{\ln n}{kl(\mu_2,\mu_1)}, \ \ \forall n \in \mathbb{N}, \tag{42}$$

From Theorem 3, it follows that

$$\liminf_{n\to\infty}\frac{\mathbb{E}[m_2^{G_n,v}(n)]}{\ln n} \geq \frac{1}{kl(\mu_2,\mu_1)}, \ \ \forall\ v \in V_n. \tag{43}$$

The above inequalities imply that, for any $v \in V_n$, $\exists\, n_v \in \mathbb{N}$ such that $\frac{\mathbb{E}[m_2^{G_n,v}(n)]}{\ln n} \geq \frac{1}{kl(\mu_2,\mu_1)}\forall n \geq n_v$. Let $n' = \max(n_v : v \in V_n)$.

For all $n \in \mathbb{N}$, since the center node has obtained $\frac{\ln n}{kl(\mu_2,\mu_1)}$ samples of arm 2 in the very first round, and the policy is non-altruistic, it chooses arm 2 at most $O(1)$ number of times further. For all $n \geq n'$, in order to satisfy all the inequalities in (43), each leaf node has to choose the arm 2 at least $\left(\frac{\ln n'}{kl(\mu_2,\mu_1)} - O(1)\right)$ times. Hence,

$$\mathbb{E}[m_2^{G_n}(n)] \geq (m_n - 1)\left(\frac{\ln n}{kl(\mu_2,\mu_1)} - O(1) - 1\right)\ \forall n \geq n',$$
$$\Rightarrow \liminf_{n\to\infty}\frac{\mathbb{E}[m_2^{G_n}(n)]}{(m_n-1)\ln n} \geq \frac{1}{kl(\mu_2,\mu_1)},$$

which completes the proof. ■

## APPENDIX E

*Proof of Theorem 5.*

*Proof:* Without loss of generality, we assume that node 1 is the center node in the star network. Under FYL policy, for $2 \leq u \leq m$, $a^u(t) = a^1(t-1)$ for $t > 1$. Hence, for any sub-optimal arm $i$,

$$\begin{aligned}T_i^u(n) =& \mathbb{I}_{\{a^u(1)=i\}} + \mathbb{I}_{\{a^u(2)=i\}}\cdots + \mathbb{I}_{\{a^u(n)=i\}},\\ =& \mathbb{I}_{\{a^u(1)=i\}} + \mathbb{I}_{\{a^1(1)=i\}}\cdots + \mathbb{I}_{\{a^1(n-1)=i\}},\\ \leq& 1 + T_i^1(n-1).\end{aligned}$$

Therefore, we obtain the following:

$$\begin{aligned}\sum_{v=1}^{m} T_i^v(n) &= T_i^1(n) + T_i^2(n)\cdots + T_i^m(n),\\ &\leq T_i^1(n) + 1 + T_i^1(n-1)\cdots + 1 + T_i^1(n-1),\\ &\leq (m-1) + mT_i^1(n),\end{aligned} \tag{44}$$

since $T_i^1(n-1) \leq T_i^1(n)$. Now, we find an upper bound on $T_i^1(n)$ under FYL policy. Let $\tau_1$ be the least time step at which $m_i^1(\tau_1)$ is atleast $l_i = \frac{8\ln n}{\Delta_i^2}$. Observe that, under FYL policy $T_i^1(\tau_1) = \lceil\frac{l_i}{m}\rceil$. Since, the center node has chosen arm $i$ for $\lceil\frac{l_i}{m}\rceil$ times, $(m-1)$ leaf nodes must have also selected arm $i$ for the same number of times. This leads to $m_i^1(\tau_1) = l_i$. Let $B_i^1(t)$ be the event that node-1 chooses arm $i$ in round $t$. Hence,

$$T_i^1(n) = T_i^1(\tau_1) + \sum_{t=\tau_1+1}^{n}\mathbb{I}_{B_i^1(t)} = \left\lceil\frac{l_i}{m}\right\rceil + \sum_{t=\tau_1+1}^{n}\mathbb{I}_{B_i^1(t)}.$$

By using the analysis in Theorem 1, we obtain

$$\mathbb{E}\Big[\sum_{t=\tau_1+1}^{n}\mathbb{I}_{B_i^1(t)}\Big] \leq \frac{2}{4\beta-1} + \frac{2}{(4\beta-1)^2\ln(1/\beta)}.$$

Hence,

$$\mathbb{E}[T_i^1(n)] \leq \left\lceil\frac{l_i}{m}\right\rceil + \frac{2}{4\beta-1} + \frac{2}{(4\beta-1)^2\ln(1/\beta)}.$$

From (44),

$$\begin{aligned}\sum_{v=1}^{m}\mathbb{E}[T_i^v(n)] \leq& \frac{8\ln n}{\Delta_i^2} + 2m - 1 +\\ &\frac{2m}{4\beta-1}\left(1 + \frac{1}{(4\beta-1)\ln(1/\beta)}\right),\end{aligned}$$

where we have substituted $l_i = \frac{8\ln n}{\Delta_i^2}$. Therefore, the expected regret of the FYL policy on an $m$-node star network upto $n$ number of rounds is upper bounded as:

$$\begin{aligned}\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*}^{K}\frac{8\ln n}{\Delta_i} + \Big[2m - 1 + \frac{2m}{4\beta-1}\cdot\\ \left(1 + \frac{1}{(4\beta-1)\ln(1/\beta)}\right)\Big]\sum_{j=1}^{K}\Delta_j,\end{aligned}$$

which completes the proof. ■

## APPENDIX F

*Proof of Theorem 6.*

*Proof:* Since the leader node (a node in the given dominating set) in a particular component uses samples only from its neighbours in the same component, we can upper bound the expected regret of each component using Theorem 5. We get the desired result by adding the expected regrets of all the components. ■