Collaborative Learning of Stochastic Bandits over a Social Network

Ravi Kumar Kolla, Krishna Jagannathan and Aditya Gopalan

***Abstract*—We consider a collaborative online learning paradigm, wherein a group of agents connected through a social network are engaged in learning a stochastic multi-armed bandit problem. 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.**

***Index Terms*—Online learning, multi armed bandits, regret, dominating set.**

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) 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].

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

A conference version of this paper appeared in the proceedings of the 54^{th} Annual Allerton Conference on Communication, Control and Computing, 2016.

Our framework is motivated by scenarios that involve multiple decision makers acting under uncertainty towards optimising a common goal. Potential applications of this work include cooperative distributed search by multiple robots [3], cooperative foraging [4] and multi-robot radio source localization [5], [6]. We briefly explain the setting of cooperative distributed search by multiple robots here. Consider a scenario of an under water chemical leak. Suppose that a group of robots connected by a network are tasked with locating the leak among finitely many valve locations. At each time instant, each robot goes to one of the locations and gets a sample of the chemical concentration. Due to constraints on communication costs and complexity, communication among robots is restricted only to one hop neighbours [3]. Such a setting is well modelled by using the collaborative bandit learning framework considered in this paper.

Another application scenario 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.

In our setting, the agents use the underlying 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 underlying 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 the UCB-Network policy, wherein all the agents employ an extension of the celebrated UCB1 [2] policy. Under this policy, we derive an upper bound on the expected regret of a generic network, and show that the structure of the network is captured in the upper bound through its *independence number* [7]. 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 underlying 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. Next, to incorporate the network structure, we derive another lower bound on the expected regret of a generic network as a function of the *independence number* of the graph G^2. Here, G^2 is the original graph G augmented with edges between any pair of nodes that have at least one common neighbour in G. 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 [8] 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 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 underlying 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* [7] 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* [7] 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], [9]–[11]. 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 framework in [12]–[14]. 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 [15], 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 [16], the authors study the problem of the best arm identification with fixed confidence in a MAB by multiple players connected through a complete network with communication costs. In [17], the authors consider the best arm identification with fixed budget in MAB problems by multiple agents in the context of wireless channel selection. They have studied the problem for a complete graph with communication costs. [18] deals with the regret minimisation of a non-stochastic bandit problem for multiple players with communication costs. [19] also deals with the learning of non-stochastic MAB problem by multiple agents connected through a network. Note that, these works are different from our work which deals with regret minimisation of a stochastic MAB problem by multiple players connected through a generic network.

In [20], the authors deal with the problem of the regret minimisation of stochastic bandits for peer to peer networks. The communication model considered in [20] is that, in each round, a peer can choose any other 2 peers and send messages to them. [21] deals with the regret minimisation of stochastic bandits by multiple agents. The communication model is that, in each round, an agent either chooses an arm to play or broadcasts the reward obtained in the previous round to all other agents. It is easy to see that, these communication models are different from the communication model considered in our work. [22] and [23] deal with the problem of cooperative stochastic multi-armed bandits by multiple agents for the Gaussian reward distributions and proposed policies inspired by consensus algorithms. On the other hand, we consider bounded reward distributions in our model.

The primary focus in [24] 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 been considered in [25], and the user communication model considered therein is similar to the model in our work. In [25], they address the problem from a game-theoretic perspective for a complete network. Then, they proposed a

randomised policy named ϵ-Greedy for a generic network and analysed its performance by providing an upper bound on the expected regret of the network. Note that, the ϵ-Greedy policy in [25] requires parameters c and d, as inputs, which depend on the gaps between the expected rewards of the optimal arm and the sub-optimal arms. However, these gaps are unknown to the the agent in practice, and learning these parameters is often as hard as learning the gaps between expected rewards of arms, which is as good as learning the expected values of arms. On the other hand, the policies proposed in this work are deterministic and do not any need any inputs which depend on the unknown parameters of the bandit problem. In addition, we address the problem of network regret minimisation from a collaborative learning perspective.

In a class of MAB problems considered in [26]–[28], a sole learning agent receives side observations in each round from *other arms*, in addition to samples from the chosen arm. In [29]–[32], the authors deal with the setting of regret minimisation of a non-stochastic bandit problem by a single agent with additional observations in each round. Note that, these works are different from our work because we deal with the regret minimisation of a stochastic bandit problem by multiple agents connected through a network with local information sharing. Another related paper is [33] – 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 model described above is rather different from the setting of 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 $G = (V, E)$, 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 arm in a 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 arms and the respective rewards of itself and its one-hop neighbours in round t, before deciding the action for round $(t+1)$.

We now briefly connect the applications mentioned in the Introduction with the system model. In the distributed cooperative search application, the robots correspond to the agents which are connected by a network, and the valve positions correspond to the arms of the MAB problem. In the application of the distributed recommendation systems, the servers are the agents which form the network, and products/items are the arms of the bandit problem.

The policy Φ^v followed by a user-v 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 Φ, 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 Φ^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 Φ^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 Φ is denoted by $R^G_\Phi(n)$[2], 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 Φ^v upto round n. 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: $m^v_i(t) = \sum_{u \in \mathcal{N}(v)} T^u_i(t)$, and $\hat{\mu}_{m^v_i(t)} = \frac{1}{m^v_i(t)} \sum_{u \in \mathcal{N}(v)} \sum_{k=1}^{t} X^u_{a^u(k)}(k) \mathbb{I}\{a^u(k) = i\}$, 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, G, till round t. We use the adjacency matrix A to represent the network G. If $(i, j) \in E$ then

[2]We omit Φ from the regret notation, whenever the policy can be understood from the context.

$A(i,j) = A(j,i) = 1$, otherwise $A(i,j) = A(j,i) = 0$. We assume that $A(i,i) = 1\ \forall i \in V$.

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 $1 \leq t \leq K$
- play arm t
Loop: For $K \leq t \leq n$
- $a^v(t+1) = \underset{j}{\operatorname{argmax}}\ \hat{\mu}_{m_j^v(t)} + \sqrt{\frac{2\ln t}{m_j^v(t)}}$

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 with K arms. Further, assume that the rewards lie in $[0,1]$. Then, we have

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

In the above, $\alpha(G)$ is the independence number[3] of G [7], and $b_G = (\alpha(G)d_{max} + 2.8m)\sum_{j=1}^{K}\Delta_j$, where d_{max} is the maximum degree of the network.

The following Lemma 1 and 2 are used to establish Theorem 1. Note that, Lemma 1 and 2 hold for any sub-optimal arm i. In Lemma 1, we show that the probability of playing a sub-optimal arm i by a node v in a round t is small if the node has access to at least $l_i = \left\lceil \frac{8\ln n}{\Delta_i^2} \right\rceil$ samples of arm i.

Lemma 1: For $S, t \in \mathbb{N}$, let $c_{t,S} := \sqrt{\frac{2\ln t}{S}}$. For each $v \in V$ and sub-optimal arm i, define τ_i^v as follows: $\tau_i^v := \min\{t \in \{1,2,\ldots,n\} : m_i^v(t) \geq l_i\}$. Then, for each $t > \tau_i^v, \beta \in (0.25, 1)$,

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

We make use of Hoeffding's maximal inequality [40] for proving Lemma 1. To invoke the same, we require a novel probability space construction which is given in Lemma 3 in Appendix A. A detailed proof of Lemma 1 is also given in Appendix A.

In the following result, we give an upper bound on the maximum number of samples of the sub-optimal arm i required by the entire network such that each node v has access to at least l_i and at most $l_i + |\mathcal{N}(v)| - 1$ samples of it.

Lemma 2: Let $\{\tau_i^v, v \in V, i \in \mathcal{K}\}$ be same as defined in Lemma 1. For any sub-optimal arm i, the following holds: $\sum_{v\in V} T_i^v(\tau_i^v) \leq \alpha(G)\left(l_i + d_{max}\right)$.

A detailed proof of Lemma 2 is given in Appendix A. We now prove Theorem 1 by using Lemmas 1 and 2.

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

$$\begin{aligned} B_i^v(t) &= \{\hat{\mu}_{m_j^v(t)} + c_{t,m_j^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}, \forall j \neq i\}, \\ &\subseteq \{\hat{\mu}_{m_*^v(t)} + c_{t,m_*^v(t)} \leq \hat{\mu}_{m_i^v(t)} + c_{t,m_i^v(t)}\}. \end{aligned} \tag{2}$$

Hence,

$$\begin{aligned} \mathbb{E}\left[\sum_{v=1}^{m} T_i^v(n)\right] &= \mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}[\mathbb{I}_{\{t\leq\tau_i^v, B_i^v(t)\}} + \mathbb{I}_{\{t>\tau_i^v, B_i^v(t)\}}]\right], \\ &= \underbrace{\mathbb{E}\left[\sum_{v=1}^{m} T_i^v(\tau_i^v)\right]}_{(a)} + \underbrace{\mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}\mathbb{I}_{\{t>\tau_i^v, B_i^v(t)\}}\right]}_{(b)}. \end{aligned} \tag{3}$$

Now, we upper bound (b) in (3). Let $1 \leq v \leq m$. Since, $m_i^v(t) \geq l_i$ for $t > \tau_i^v$,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}_{\{t>\tau_i^v\}}\mathbb{I}_{B_i^v(t)}\right] = \sum_{t=1}^{n}\mathbb{P}\left(B_i^v(t), \{t > \tau_i^v\}\right), \\ &\overset{(c)}{\leq} \sum_{t=1}^{\infty} 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}} \leq \int_1^{\infty} 2\left(\frac{\ln t}{\ln(1/\beta)} + 1\right)\frac{1}{t^{4\beta}}\,\mathrm{d}t \\ &= \frac{2}{4\beta - 1} + \frac{2}{(4\beta-1)^2\ln(1/\beta)} \qquad \forall \beta \in (0.25, 1) \\ &\leq 2.8 \qquad \text{(by taking infimum over } \beta), \end{aligned}$$

where (c) is due to Lemma 1. Thus, we get

$$\mathbb{E}\left[\sum_{v=1}^{m}\sum_{t=1}^{n}\mathbb{I}_{\{t>\tau_i^v\}}\mathbb{I}_{B_i^v(t)}\right] \leq 2.8m. \tag{4}$$

Using Lemma 2, we upper bound (a) in (3) as:

$$\mathbb{E}\left[\sum_{v=1}^{m} T_i^v(\tau_i^v)\right] \leq \alpha(G)\left(l_i + d_{max}\right). \tag{5}$$

Combining (3), (4) and (5) establishes the desired result. ■

A. Application to typical networks

Since evaluating the independence number of an arbitrary graph is an NP-hard problem, we now evaluate the same 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. In the following, we present the upper bounds in Theorem 1 for the same networks. Let $H(n,\Delta) = \sum_{i:\mu_i<\mu^*}\left(\frac{8\ln n}{\Delta_i} + \Delta_i\right)$.

Corollary 1 For an m-node FC network:

$$\mathbb{E}[R^G(n)] \leq H(n,\Delta) + b_G. \tag{6}$$

[3]Independence number of a graph G is defined as the cardinality of the maximum independence set of G.

Corollary 2 For an m-node circular network:

$$\mathbb{E}[R^G(n)] \leq \left\lfloor \frac{m}{2} \right\rfloor H\left(n, \Delta\right) + b_G. \tag{7}$$

Corollary 3 For an m-node star network:

$$\mathbb{E}[R^G(n)] \leq (m-1) H\left(n, \Delta\right) + b_G. \tag{8}$$

Corollary 4 For an m-node FD network:

$$\mathbb{E}[R^G(n)] \leq m H\left(n, \Delta\right) + b_G. \tag{9}$$

A key insight 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. Another key observation is that, since the UCB-Network policy's regret upper bounds for star and FD networks are almost same, it is possible that this policy may suffer large regret on star networks.

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 single-agent 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \Theta^K = \boldsymbol{\Theta}$ to denote the parameters of arms 1 to K respectively. Suppose $f(x;\theta_j)$ be the reward distribution for arm j with parameter θ_j. Let $\mu(\theta_j)$ be the mean of arm j, and $\theta^* = \arg\max_{1 \leq j \leq K} \mu(\theta_j)$. Define the parameter sets for an arm j as $\boldsymbol{\Theta_j} = \{\boldsymbol{\theta} : \mu(\theta_j) < \max_{i \neq j} \mu(\theta_i)\}$ and $\boldsymbol{\Theta_j^*} = \{\boldsymbol{\theta} : \mu(\theta_j) > \max_{i \neq j} \mu(\theta_i)\}$.

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 λ, from the distribution parametrised by β. We require the following standard assumptions on Θ and $f(\cdot;\cdot)$ [1]:

Assumption 1 *[A1]*:

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

(ii) For all $\epsilon > 0$ and β, λ such that $\mu(\lambda) > \mu(\beta)$, there 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) Θ is such that for all $\lambda \in \Theta$ and $\delta > 0$, there 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 $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ s.t. $\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)] = o(n^c)\ \forall\ c > 0$, for any sub-optimal arm j. Then, we have $\liminf_{n \to \infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta^*)}$.

Proof: Follows from Theorem 2 in [1]. ∎

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 like Bernoulli, which assigns a non-zero probability to each possible value.

Let ω be a sample path, which consists of all pairs of arms and the corresponding rewards of all nodes from rounds 1 through n: $\omega = \{(a^v(t), X^v_{a^v(t)}(t)) : v \in V, 1 \leq t \leq n\}$. Let ω_v be the sample path restricted to node v and its one-hop neighbours *i.e,* it consists of all pairs of arms and rewards of node v and its one-hop neighbours from round 1 through n. Let $\omega_{\bar{v}}$ be the sample path restricted to all nodes who are not present in the one-hop neighbourhood of node v. Mathematically,

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

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 $u \in \mathcal{N}(v) \setminus \{v\}$, $\mathbb{E}[T_i^v(n)|\omega_{\bar{v}}] = o(n^a),\ \forall\, a > 0,\ \forall\, \omega_{\bar{v}}$.

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 j, the expected number of times that the policy *plays* arm j 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.

Intuitively, a non-altruistic policy means that it plays any sub-optimal arm only a constant number of times after it has sufficient information to identify the arm as sub-optimal. It can be shown that the network extensions of UCB1 (UCB-user) and Thompson sampling [8] are NAIC policies. In particular, we show that the UCB-user policy is an NAIC policy in Appendix A of the supplementary material.

Example of a policy which is not individually consistent : Consider a 2-armed bandit problem with Bernoulli rewards with means μ_1, μ_2 *s.t.* $\mu_1 > \mu_2$. Consider the 3-node line graph with node 2 as the center. 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 Individual consistent definition. 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 provide a network-dependent lower bound for the class of NAIC policies. We need the following to introduce our next result. Let $G^2 = (V, E')$ be the original graph $G = (V, E)$ augmented with edges between any pair of nodes that have at least one common neighbour in G. Essentially, $(u, v) \in E'$ if and only if either $(u, v) \in E$ or $\mathcal{N}(u) \cap \mathcal{N}(v) \neq \phi$. Let $\alpha\left(G^2\right)$ be the independence number of the graph G^2. Note that $\alpha\left(G^2\right) \leq \alpha(G)$.

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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \boldsymbol{\Theta_J}$, then the following lower bounds hold:

$$(i) \liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)|\omega_{\bar{v}}]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta^*)}, \ \forall v \in V$$
$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta^*)}, \ \forall v \in V$$
$$(ii) \liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]}{\ln n} \geq \alpha(G^2) \cdot \frac{1}{kl(\theta_j||\theta^*)}, \tag{10}$$

where $\alpha(G^2)$ is the independence number of the graph G^2.

Proof: Refer Appendix B. ∎

Recall that evaluating the independence number of an arbitrary graph G is an NP-hard problem. Hence, we evaluate $\alpha(G^2)$ for various networks such as FC, circular, star and FD, and provide the corresponding lower bounds below. Let $\Delta_i = \mu(\theta^*) - \mu(\theta_i)$ and $J(\boldsymbol{\theta}, \Delta) = \sum_{i:\Delta_i>0} \frac{\Delta_i}{kl(\theta_i||\theta^*)}$.

Corollary 5 For an m-node FC network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq J(\boldsymbol{\theta}, \Delta). \tag{11}$$

Corollary 6 For an m-node circular network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq \left\lfloor \frac{m}{3} \right\rfloor J(\boldsymbol{\theta}, \Delta). \tag{12}$$

Corollary 7 For an m-node star network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq J(\boldsymbol{\theta}, \Delta). \tag{13}$$

Corollary 8 For an m-node FD network:

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[R^G(n)]}{\ln n} \geq mJ(\boldsymbol{\theta}, \Delta). \tag{14}$$

From Corollaries 1-8, we infer that the upper bound of the UCB-Network and the lower bound given by (10) 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 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.

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

Proof: Refer Appendix C. ∎

Theorem 4 asserts that, for a fixed large time horizon n, we can construct a star network with only $m = O(\ln n)$ nodes, whose expected regret is at least $O((m-1)\ln n)$. 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 Theorem 2 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.

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.

Remark. We can show that Theorems 1, 2, 3 and 4 of this paper hold for directed graphs too, when the adjacency matrix of the directed graph considered in the results.

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 network; for this purpose, we define the dominating set and the dominating set partition.

Definition 3. [Dominating set of a network] [7] A *dominating set* D of a network $G = (V, E)$ is a subset of V such that every node in $V \setminus D$ is connected to at least 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 network] 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 neighbours.

Note that given a dominating set for a network, 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.

Theorem 5 (FYL regret bound, star networks): Suppose the star network G with a dominating set as the center node,

Algorithm 2 Follow Your Leader (FYL) Policy

Input: Network 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 $(t-1)$

follows the FYL to learn a stochastic MAB problem with K arms. Assume that the rewards lie in $[0, 1]$. Then, we have

$$\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*} \frac{8\ln n}{\Delta_i} + (4.8m-1)\sum_{j=1}^{K}\Delta_j.$$

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 and any non-central node $u \in [2, 3 \ldots, m]$,

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

Then, we obtain the following:

$$\begin{aligned} \sum_{v=1}^{m} T_i^v(n) &= T_i^1(n) + T_i^2(n) \cdots + T_i^m(n), \\ &\leq T_i^1(n) + 1 + T_i^1(n-1) \cdots + 1 + T_i^1(n-1), \\ &\leq (m-1) + mT_i^1(n), \end{aligned} \tag{15}$$

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 τ_1 be the least time step at which $m_i^1(\tau_1)$ is at least $l_i = \lceil (8\ln n)/\Delta_i^2 \rceil$. Observe that, under FYL policy $T_i^1(\tau_1) = \lceil l_i/m \rceil$. Since the center node has chosen arm i for $\lceil l_i/m \rceil$ 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. Then,

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

By using the analysis in Theorem 1, we obtain $\mathbb{E}\Big[\sum_{t=\tau_1+1}^{n} \mathbb{I}_{B_i^1(t)}\Big] \leq 2.8$. Hence, $\mathbb{E}[T_i^1(n)] \leq \Big\lceil l_i/m \Big\rceil + 2.8$. From (15), we get that $\sum_{v=1}^{m} \mathbb{E}[T_i^v(n)] \leq (8\ln n)/\Delta_i^2 + 4.8m - 1$, where we have substituted $l_i = \lceil (8\ln n)/\Delta_i^2 \rceil$. By substituting the above in equation (1), we obtain the desired result. ∎

A key insight obtained from Theorem 5 is that an m-node star network under 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 Theorem 2. 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 we have

$$\mathbb{E}[R^G(n)] \leq \sum_{i:\mu_i<\mu^*} \frac{8|D|\ln n}{\Delta_i} + |D|(4.8m-1)\sum_{j=1}^{K}\Delta_j.$$

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. ∎

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 the tightest 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 [34]. However, fast distributed approximation algorithms for the same are well-known in the literature. For example, Algorithm 35 in [34] 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 [35]–[37].

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 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.

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. Local behaviour comparison of UCB-Network and FYL

To illustrate the local behaviour of the FYL and UCB-Network (an NAIC policy) policies, we conduct simulations in which we compare the performance of individual nodes' regret under both the policies. We consider a 10 node star network and a 10-armed bandit problem with Bernoulli arm distributions whose parameters are as $1, 0.9, 0.8 \ldots, 0.1$. We run FYL and UCB-Network policies on the aforementioned problem. Under our model, in a star network, all leaf nodes have symmetric information structure and hence we have taken an arbitrary leaf node as a representative for all leaf nodes. We have plotted expected regrets of the center and the representative of leaf nodes under the FYL and UCB-Network in Figure 1. From Figure 1, we observe that the regrets of leaf and center nodes are almost same under the FYL policy. On the other hand, there is a huge gap between the regrets of leaf and center nodes under the UCB-Network policy. Further, note that, the gap between the regrets of the center node under the FYL and UCB-Network is small and at the same time gap between the regrets of leaf node under these two policies is huge. This observation leads to the conclusion that the local behaviour of the FYL is better than the UCB-Network.

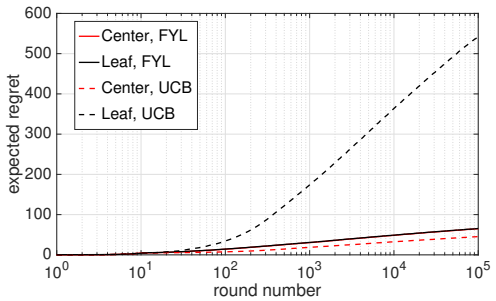


Fig. 1: Illustration of local behaviour of FYL and UCB-Network policies on 10 node star networks; 10-armed Bernoulli bandit problem with means are as $1, 0.9 \ldots, 0.1$.

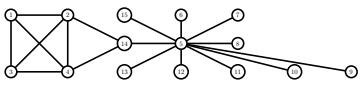


Fig. 2: N/W #1

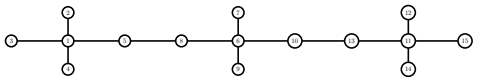


Fig. 3: N/W #2

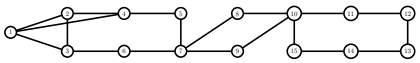


Fig. 4: N/W #3

C. Comparison of FYL policy and a centralised policy

We now proceed to compare the FYL policy with a centralised policy on various networks. We first briefly introduce the centralized policy considered in the simulations. Under centralized control setting, we assume that there is an external agent who has access to information of all the nodes and then he suggests actions to all nodes in each round. For a given network, we devised the centralized policy as follows. The external agent chooses a node in the network and suggests that node to play the UCB policy by taking information from all nodes into account. He also suggests all other nodes to follow the node which is using the UCB policy. It is easy to see that, this policy is same as FYL policy. Since FYL policy achieves the universal regret lower bound on star networks, we can treat this policy as a centralized policy. Note that, we can have this policy on any network under the assumption that centralized control is possible.

Figures 2-5 show the various networks considered in the simulations. To maintain consistency in the simulations, we have fixed the number of nodes in all the networks are as 15. We call these networks as "NW #1", "NW #2", "NW #3", "NW #4". Note that, the cardinality of the smallest dominating set of these networks are 2, 3, 4 and 5 respectively. We considered a 10-armed Bernoulli bandit problem with means drawn uniformly random from $[0, 1]$, whose details are mentioned in the captions of Figure 6. We have run the FYL and the aforementioned centralised policy on these various networks. In Figure 6, we have plotted the expected regrets of FYL policy and the centralized policy on these four networks. From Figure 6, we observe that the FYL policy's expected regret increases as the cardinality of the smallest dominating set increases.

D. Comparison of FYL and ϵ-Greedy policy in [25]

Since the model considered in [25] is similar to ours, we now compare the performance of the proposed FYL policy with the ϵ-Greedy policy proposed in [25] on various networks. Note that, the ϵ-Greedy policy requires c and d, as input parameters, which require the knowledge of difference between the expected values of the optimal and sub-optimal arms of the bandit problem.

We first compare the FYL and the ϵ-Greedy policies on star networks. We consider 10-armed bandit problem consists of Bernoulli rewards with mean values are as $1, 0.9, 0.8 \ldots, 0.1$,

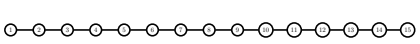


Fig. 5: N/W #4

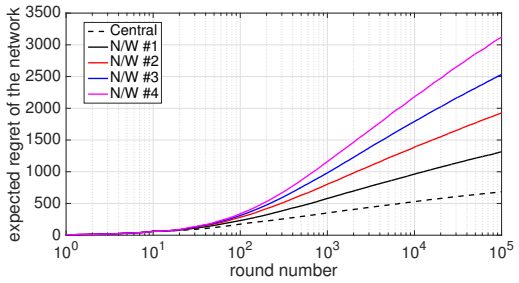
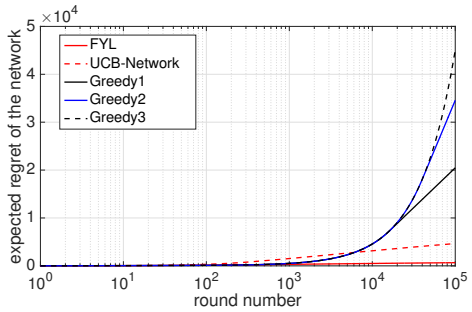
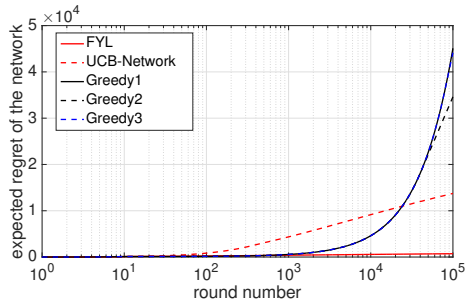


Fig. 6: FYL policy on various networks; μ = [0.65 0.54 0.45 0.33 0.29 0.17 0.14 0.08 0.06 0.01]



(a) 10 node and the following parameters are for ϵ-greedy policies: c_1 = 4, d_1 = 0.05, c_2 = 10, d_2 = 0.05, c_3 = 20, d_3 = 0.01, z = (1 0 0 . . . 0).



(b) 25 node and the following parameters are for ϵ-greedy policies: c_1 = 4, d_1 = 0.01, c_2 = 10, d_2 = 0.05, c_3 = 20, d_3 = 0.05, z = (1 0 0 . . . 0).

Fig. 7: Comparison of our FYL, UCB-Network policies, and the ϵ-Greedy policy [25] for star networks on an 10-armed Bernoulli bandit problem with means are as 1, 0.9, 0.8, . . . , 0.1.

and 10 and 25 node star networks. We ran our FYL, UCB-Network policies and ϵ-Greedy policy in [25] with various c and d parameters on the aforementioned problem. We consider three sets of c and d parameters for the ϵ-Greedy policy and named these policies as "Greedy-1", "Greedy-2", "Greedy-3" in the simulations. The parameter z which is mentioned in the figure captions is a solution to (5) in [25] which is used in the ϵ-Greedy policy. From Figure 7, we observe that the FYL policy outperforms the ϵ-Greedy policy for all the various c and d parameters considered.

We now compare FYL and ϵ-Greedy policies on the networks shown in Figure 3 and 4. We consider a 10-armed Bernoulli bandit problem with means are drawn uniformly at random from $[0, 1]$. We have run both policies on these networks and the results are shown in Figure 8 and 9. Figures captions contain the details of arm means of the bandit problem, and the parameters of the ϵ-Greedy policy. From Figures 8-9, we observe that the FYL policy outperforms the ϵ-Greedy policy in [25] on all the considered networks.

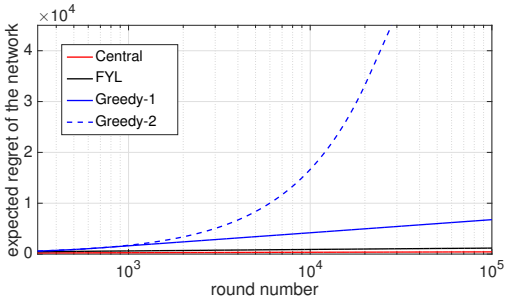


Fig. 8: FYL vs ϵ-Greedy on N/W #2; Parameters of the bandit problem and the policies: μ = [0.99 0.65 0.63 0.60 0.42 0.39 0.35 0.22 0.14 0.02], z = [1 0 0 0 0 1 0 0 0 0 1 0 0 0 0], d_1 = 0.1723, c_1 = 2, d_2 = 0.0689, c_2 = 10.

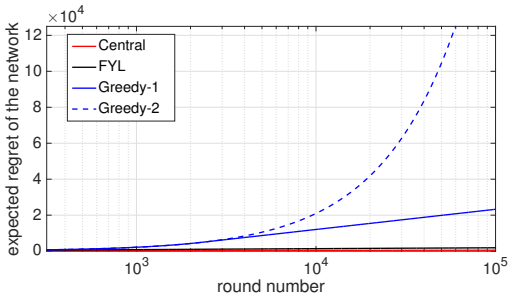


Fig. 9: FYL vs ϵ-Greedy on N/W #3;Parameters of the bandit problem and the policies: μ = [0.96 0.95 0.91 0.90 0.81 0.63 0.54 0.29 0.13 0.09], z = [1 0 0 0 0 0 1 0 0 0 1 0 0 1 0], d_1 = 0.0037, c_1 = 2, d_2 = 0.0015, c_2 = 10.

VII. Concluding Remarks

We studied the collaborative learning of a stochastic MAB problem by a group of users connected through a 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 [38] and contextual bandits [39]. 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

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. Another interesting line of future work is to consider the strategic behaviour among the agents and study when each agent tries to individually optimize a notion of utility. In this regard, an important open question would be – what is the utility function for each player such that the appropriate equilibrium or locally optimal response is to somehow coordinate their efforts in this manner? This would help in understanding incentive mechanisms for the nodes to cooperate, and is a subject of independent, follow-up research.

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] S. Li, R. Kong, and Y. Guo, "Cooperative distributed source seeking by multiple robots: Algorithms and experiments," *IEEE/ASME Transactions on mechatronics*, vol. 19, no. 6, pp. 1810–1820, 2014.

[4] K. Sugawara, T. Kazama, and T. Watanabe, "Foraging behavior of interacting robots with virtual pheromone," in *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, vol. 3. IEEE, 2004, pp. 3074–3079.

[5] D. Song, C.-Y. Kim, and J. Yi, "Simultaneous localization of multiple unknown and transient radio sources using a mobile robot," *IEEE Transactions on Robotics*, vol. 28, no. 3, pp. 668–680, 2012.

[6] C.-Y. Kim, D. Song, Y. Xu, and J. Yi, "Localization of multiple unknown transient radio sources using multiple paired mobile robots with limited sensing ranges," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5167–5172.

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

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] E. Hillel, Z. S. Karnin, T. Koren, R. Lempel, and O. Somekh, "Distributed exploration in multi-armed bandits," in *Advances in Neural Information Processing Systems*, 2013, pp. 854–862.

[17] Y. Xue, P. Zhou, T. Jiang, S. Mao, and X. Huang, "Distributed learning for multi-channel selection in wireless network monitoring," in *Sensing, Communication, and Networking (SECON), 2016 13th Annual IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[18] V. Kanade, Z. Liu, and B. Radunovic, "Distributed non-stochastic experts," in *Advances in Neural Information Processing Systems*, 2012, pp. 260–268.

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

[20] B. Szorenyi, R. Busa-Fekete, I. Hegedus, R. Ormándi, M. Jelasity, and B. Kégl, "Gossip-based distributed stochastic bandit algorithms," in *International Conference on Machine Learning*, 2013, pp. 19–27.

[21] M. Chakraborty, K. Y. P. Chua, S. Das, and B. Juba, "Coordinated versus decentralized exploration in multi-agent multi-armed bandits," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 164–170.

[22] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *Control Conference (ECC), 2016 European*. IEEE, 2016, pp. 243–248.

[23] ——, "Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms," in *Decision and Control (CDC), 2016 IEEE 55th Conference on*. IEEE, 2016, pp. 167–172.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] T. Kocák, G. Neu, M. Valko, and R. Munos, "Efficient learning by implicit exploration in bandit problems with side observations," in *Advances in Neural Information Processing Systems*, 2014, pp. 613–621.

[30] K. Amin, S. Kale, G. Tesauro, and D. S. Turaga, "Budgeted prediction with expert advice." in *AAAI*, 2015, pp. 2490–2496.

[31] S. Kale, "Multiarmed bandits with limited expert advice," in *Conference on Learning Theory*, 2014, pp. 107–122.

[32] Y. Seldin, P. Bartlett, K. Crammer, and Y. Abbasi-Yadkori, "Prediction with limited advice and multiarmed bandits with paid observations," in *International Conference on Machine Learning*, 2014, pp. 280–287.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] S. Bubeck, "Jeux de bandits et fondations du clustering," Ph.D. dissertation, Citeseer, 2010.

APPENDIX A

We require the following concentration inequality and Lemma 3 to prove Lemmas 1 and 2.

Hoeffding's Maximal Inequality [40]: Let $X_1, X_2, \dots$ be centered i.i.d. r.v.s lying in $[0,1]$. Then, for any $x > 0$ and $t \geq 1$, we have $\mathbb{P}\left(\exists s \in \{1,\dots,t\} \text{s.t.} \sum_{i=1}^{s} X_i > x\right) \leq \exp\left(-\frac{2x^2}{t}\right)$.

In order to introduce Lemma 3, we need the following. Consider a new probability space with the probability measure $\tilde{\mathbb{P}}$, for the following 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 $\{X_i^v(k)\}_{k=1}^{\infty}$. In this probability space, we considered the probability measure to be $\mathbb{P}$.

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

Lemma 3: Consider an m-node undirected network. 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 $k \geq 1$.
A detailed proof of Lemma 3 is given in Appendix A of the Supplementary material.

Proof of Lemma 1.

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,

$$\mathbb{P}\left(A_i^v(t) \cap \{t > \tau_i^v\}\right) = \mathbb{P}\left(A_i^v(t) \cap \{m_i^v(t) \geq l_i\}\right). \quad (16)$$

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

$$\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\}, \quad (17)$$
$$\{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\}, \quad (18)$$
$$\{\mu^* < \mu_i + 2c_{t,m_i^v(t)}\}. \quad (19)$$

Note that, the event given in (19) does not occur when the event $\{m_i^v(t) \geq l_i\}$ occurs. Hence,

$$\begin{aligned}
&\mathbb{P}\left(A_i^v(t) \cap \{m_i^v(t) \geq l_i\}\right) \leq \\
&\mathbb{P}(\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\} \cup \{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\} \\
&\qquad\qquad \cap \{m_i^v(t) \geq l_i\}), \\
&\leq \mathbb{P}\left(\{\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\}\right) \\
&\qquad + \mathbb{P}\left(\{\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\}\right). \quad (20)
\end{aligned}$$

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, $\mathbb{P}\left(\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\right) \leq \mathbb{P}\left(\exists s_* \in \{|\mathcal{N}(v)|, \dots, |\mathcal{N}(v)|t\} : \hat{\mu}_{s_*} \leq \mu^* - c_{t,s_*}\right)$, Now, we use the peeling argument [40] on a geometric grid over $[a, at]$ which is given here. For any $\beta \in (0.25, 1)$, $a \geq 1$, if $s \in \{a, \dots, at\}$ then there exists $j \in \{0, \dots, \frac{\ln t}{\ln(1/\beta)}\}$ such that $a\beta^{j+1}t < s \leq a\beta^j t$. Hence,

$$\begin{aligned}
&\mathbb{P}\left(\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\right) \\
&\leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \mathbb{P}\Big(\exists s_* : |\mathcal{N}(v)|\beta^{j+1}t < s_* \leq |\mathcal{N}(v)|\beta^j t, \\
&\qquad\qquad s_*\hat{\mu}_{s_*} \leq s_*\mu^* - \sqrt{2s_* \ln t}\Big), \quad (21) \\
&\leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \mathbb{P}\Big(\exists s_* : |\mathcal{N}(v)|\beta^{j+1}t < s_* \leq |\mathcal{N}(v)|\beta^j t, \\
&\qquad\qquad s_*\hat{\mu}_{s_*} \leq s_*\mu_* - \sqrt{2|\mathcal{N}(v)|\beta^{j+1}t \ln t}\Big). \quad (22)
\end{aligned}$$

Now, we proceed to bound the RHS in the above equation by using Hoeffding's maximal inequality. Note that, the random variables present in the RHS of (22) are drawn from various i.i.d. sequences. Since the Hoeffding's maximal inequality requires i.i.d. random variables drawn from a single sequence, we invoke the same in the new probability space with the measure $\tilde{\mathbb{P}}$ due to Lemma 3. Thus,

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{m_*^v(t)} \leq \mu^* - c_{t,m_*^v(t)}\right) &\leq \sum_{j=0}^{\frac{\ln t}{\ln(1/\beta)}} \exp\left(-4\beta \ln t\right), \\
&\leq \left(\frac{\ln t}{\ln(1/\beta)} + 1\right) \frac{1}{t^{4\beta}}. \quad (23)
\end{aligned}$$

Similarly, we can show that

$$\mathbb{P}\left(\hat{\mu}_{m_i^v(t)} \geq \mu_i + c_{t,m_i^v(t)}\right) \leq \left(\frac{\ln t}{\ln(1/\beta)} + 1\right) \frac{1}{t^{4\beta}}. \quad (24)$$

Substituting (23) and (24) in (20) gives the desired result. ■

In order to present the proof of Lemma 2, we need the following notation. Let γ_k denote the smallest time index when at least k nodes have access to at least l_i samples of sub-optimal arm i *i.e.,* $\gamma_k = \min\{t \in \{1,\dots,n\} : |\{v \in V : m_i^v(t) \geq l_i\}| \geq k\}$. Let η_k be the index of the node to acquire l_i samples of sub-optimal arm i at γ_k, such that $\eta_k \neq \eta_{k'}$ for all $1 \leq k' < k$. For instance $\eta_5 = 3$, it means that node-3 is the 5^{th} node to acquire l_i samples of arm-i. Let $z_k = (z_k(1), z_k(2), \dots, z_k(m)) = T_i(\gamma_k) := \left(T_i^1(\gamma_k), \dots, T_i^m(\gamma_k)\right)$, which contains the sub-optimal arm i counts of all nodes at time γ_k.

Consider the following optimisation problem that is also required in the proof of Lemma 2.

$$\begin{aligned}
&\max \ \ \|z_m\|_1 \\
&\text{s.t } \exists \text{ a sequence } \{z_k\}_{k=1}^m \text{ and} \\
&\exists \{\eta_k\}_{k=1}^m \text{ a permutation of } \{1,2,\dots,m\} \\
&z_j(\eta_k) = z_k(\eta_k) \ \ \forall j \geq k, \forall k \\
&l_i \leq \langle z_k, A(\eta_k, :)\rangle < l_i + |\mathcal{N}(\eta_k)|, \ \ 1 \leq k \leq m \\
&z_k \in \{0,1,2,\cdots,l_i\}^m, \ \ 1 \leq k \leq m.
\end{aligned} \quad (25)$$

Interpretation of (25): Under the UCB-Network policy, suppose a node has acquired at least l_i samples of a sub-optimal arm i. As shown in Lemma 1, 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 γ_k. The objective function in (25) 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 (25) 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 (25)), and
(b) Each node stops choosing arm i after it has access to l_i samples of it (the penultimate constraint in (25)).

With the above notation, we are ready to prove Lemma 2.

Proof of Lemma 2.

Proof: We establish this lemma in two steps. In the first step, we construct an optimisation problem then argue that $\sum_{v=1}^{m} T_i^v(\tau_i^v)$ is upper bounded by the solution to the optimisation problem. In the second step, we show that the solution to this optimisation problem is upper bounded by $\alpha(G)\left(l_i + d_{max}\right)$.

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 sub routine gives the value of the above mentioned random variable for a realization. Consider an m length column vector of zeros, say y. Basically, it contains the arm i counts (self plays) of all nodes.

Sub routine:

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 at least l_i and at most $(l_i + \|A(I,:)\|_1 - 1)$. 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.

Since we seek for an upper bound on $\sum_{v=1}^{m} T_i^v(\tau_i^v)$, we can cease the growth of any node v (coordinate) after it has access to at least l_i and at most $(l_i + |\mathcal{N}(v)| - 1)$ samples of arm i including its one-hop neighbours. The aforementioned is ensured by step 4. Observe that $\|y\|_1$, where y is the vector returned by the above sub routine, yields an upper bound on 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 (25) captures the above. The 5^{th} in (25) ensures that the node η_k has l_i samples of sub-optimal arm i at time instance γ_k. Recall that, γ_k is a random variable which tracks the *least* time at which at least k nodes have more than l_i samples of arm i. The 4^{th} line in (25) ensures that sub-optimal arm i count of node η_k does not increase(or stop playing arm i) after time instance γ_k. Hence, a feasible point in the optimisation problem in (25) is a sequence $\{z_k\}_{k=1}^{m}$ which satisfies the aforementioned two constraints. Therefore, for a given realization, z_m in the optimisation problem in (25) corresponds to a y returned by the procedure mentioned above. 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. This establishes that, $\sum_{v=1}^{m} T_i^v(\tau_i^v)$ is upper bounded by the solution to (25).

We now show that the solution to (25) is upper bounded by $\alpha(G)\left(l_i + d_{max}\right)$. We now try to upper bound the objective function value of (25) for each of its feasible points. Let $\{z_k\}_{k=1}^{m}$ and $\{\eta_k\}_{k=1}^{m}$ be a feasible point pair of (25). We now construct a *maximal independence set* of the graph G based on $\{\eta_k\}_{k=1}^{m}$. Let $m_1 = m$ and define $m_d = \max\{j : \eta_j \notin \cup_{k=1}^{d-1}\mathcal{N}(\eta_{m_k})\}$ for $d > 1$. Define the above m_d's until $\cup_{k=1}^{d-1}\mathcal{N}(\eta_{m_k})$ becomes $\{1, 2, \cdots, m\}$. Assume that the above process results $m_1, m_2, \cdots m_p$ with $p \leq m$ and $m_p < m_{p-1} < \cdots < m_2 < m_1 = m$. Define $C := \{\eta_{m_1}, \eta_{m_2}, \cdots, \eta_{m_p}\}$. It is easy to see that C is a maximal independence set.

Define $Q_d = \{\eta_j : \eta_j \in \mathcal{N}(\eta_{m_d}) \text{ and } \eta_j \notin \cup_{k=1}^{d-1}\mathcal{N}(\eta_{m_k})\}$ for $1 \leq d \leq p$. Essentially, Q_d contains the node η_{m_d} and its neighbours which are not connected to any of the nodes $\eta_{m_1}, \eta_{m_2}, \cdots, \eta_{m_{d-1}}$. Furthermore, Q_d's are disjoint and $\cup_{d=1}^{p} Q_d = \{1, 2, \cdots, m\}$. Hence, we write $\sum_{d=1}^{m} z_m(d) = \sum_{d=1}^{p}\sum_{\eta_j \in Q_d} z_m(\eta_j)$.

Claim: For any $\eta_j \in Q_d$, we have $z_{m_d}(\eta_j) = z_m(\eta_j)$.

We now argue that no node in Q_d has satisfied the 5^{th} line in (25) strictly after node η_{m_d}. Suppose a node $\eta_a \in Q_d$ has satisfied 5^{th} line in (25) strictly after node η_{m_d}. It implies that $m_d < a \leq m$. Since $\eta_a \notin \cup_{k=1}^{d-1}\mathcal{N}(\eta_{m_k})$ and $a > m_d$, it gives that $\eta_a \in C$. However, $\eta_a \notin C$, because $\eta_a \in \mathcal{N}(\eta_{m_d})$ and $\eta_{m_d} \in C$. It is a contradiction. Therefore, no node in Q_d has satisfied the 5^{th} line in (25) strictly after node η_{m_d}. It implies that $z_{m_d}(\eta_j) = z_m(\eta_j)$, $\forall \eta_j \in Q_d$ due to the 4^{th} line in (25). It completes the proof of the claim. From the above claim, we write that $\sum_{\eta_j \in Q_d} z_m(\eta_j) = \sum_{\eta_j \in Q_d} z_{m_d}(\eta_j) \leq \sum_{\eta_j \in \mathcal{N}(\eta_{m_d})} z_{m_d}(\eta_j) \leq l_i + |\mathcal{N}(\eta_{m_d})| - 1$ where the last inequality is due to the 5^{th} line in (25). Therefore, $\sum_{d=1}^{m} z_m(d) = \sum_{d=1}^{p}\sum_{\eta_j \in Q_d} z_m(\eta_j) \leq \sum_{d=1}^{p}(l_i + |\mathcal{N}(\eta_{m_d})| - 1) \leq pl_i + \sum_{d=1}^{p} d_{max} \leq \alpha(G)\left(l_i + d_{max}\right)$, where the last inequality holds due to the fact that p is the size of a maximal independence set. Since the above inequality is true for any feasible point pair of (25), it is true that solution of (25) is also upper bounded by $\alpha(G)\left(l_i + d_{max}\right)$. ∎

APPENDIX B

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 4: 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:

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

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta^*)}.$$

Proof of Lemma 4: 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 $\mu(\lambda) > \mu(\theta^*)$. 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 arm 2's distribution. 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

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

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

$$\mathbb{P}_{\boldsymbol{\gamma}}\{m_1^v(n) = n_1, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n|\omega_{\bar{v}}\} = \mathbb{E}_\theta\left[\mathbb{I}_{\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\hat{kl}_{n_2}\leq h_n\}} \prod_{u\in\mathcal{N}(v)} \prod_{t=1}^{T_2^u(n)=q_u} \frac{f(X_{2,t}^u;\lambda)}{f(X_{2,t}^u;\theta_2)}\right]. \quad (26)$$

However, $\prod_{u\in\mathcal{N}(v)} \prod_{t=1}^{q_u} \frac{f(X_{2,t}^u;\lambda)}{f(X_{2,t}^u;\theta_2)} = \exp(-\hat{kl}_{n_2})$. Therefore, $\mathbb{P}_{\boldsymbol{\gamma}}\{m_1^v(n) = n_1, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n|\omega_{\bar{v}}\} = \mathbb{E}_\theta\left[\mathbb{I}_{\{m_1^v(n)=n_1,\ldots,m_K^v(n)=n_K,\hat{kl}_{n_2}\leq h_n\}} \exp(-\hat{kl}_{n_2})\right]$. Note that, $\exp(-\hat{kl}_{n_2}) \geq n^{-(1-a)}$, since $\hat{kl}_{n_2} \leq h_n$ in the region of integration. Therefore,

$$\mathbb{P}_{\boldsymbol{\gamma}}\{m_1^v(n) = n_1, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n|\omega_{\bar{v}}\} \geq n^{a-1}\mathbb{P}_{\boldsymbol{\theta}}\{m_1^v(n) = n_1, \ldots, m_K^v(n) = n_K, \hat{kl}_{n_2} \leq h_n|\omega_{\bar{v}}\}.$$

Hence, $\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) \geq n^{-(1-a)}\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}})$. Now, we upper bound $\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}})$ in order to upper bound $\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}})$, as follows: $\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) \leq \mathbb{P}_{\boldsymbol{\gamma}}\left(m_2^v(n) < g_n|\omega_{\bar{v}}\right)$. Since $\{m_2^v(n) < g_n\} \subseteq \{T_2^v(n) < g_n\}$, we have $\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) \leq \mathbb{P}_{\boldsymbol{\gamma}}\left(T_2^v(n) < g_n|\omega_{\bar{v}}\right) = \mathbb{P}_{\boldsymbol{\gamma}}\left(n - T_2^v(n) > n - g_n|\omega_{\bar{v}}\right)$. 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

$$\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) \leq \frac{\mathbb{E}_{\boldsymbol{\gamma}}[n - T_2^v(n)|\omega_{\bar{v}}]}{n - g_n} = \frac{\sum_{i=1,i\neq 2}^{K} \mathbb{E}_{\boldsymbol{\gamma}}[T_i^v(n)|\omega_{\bar{v}}]}{n - g_n}.$$

Due to Individual Consistent property of the policy, we write $\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) = \frac{(K-1)o(n^a)}{n-O(\ln n)}$, for $0 < a < \delta$, since arm 2 is the unique optimal arm under γ. Hence,

$$\mathbb{P}_{\boldsymbol{\theta}}(C_n^v|\omega_{\bar{v}}) \leq n^{(1-a)}\mathbb{P}_{\boldsymbol{\gamma}}(C_n^v|\omega_{\bar{v}}) = o(1). \quad (27)$$

Observe that,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(C_n^v|\omega_{\bar{v}}\right) \geq \mathbb{P}_{\boldsymbol{\theta}}\Big(m_2^v(n) < g_n, \frac{1}{g_n}\max_{i\leq g_n} \hat{kl}_i \leq \frac{kl(\theta_2||\lambda)(1-a)}{(1-\delta)}\Big|\omega_{\bar{v}}\Big), \quad (28)$$

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\frac{1}{g_n}\max_{i\leq g_n} \hat{kl}_i \leq \frac{kl(\theta_2||\lambda)(1-a)}{(1-\delta)}\right) \to 1, \quad (29)$$

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 [11]: Let $\{X_t\}$ be a sequence of independent real-valued r.vs 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 (27), (28) and (29), we obtain

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

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 $\mu(\theta_1) \leq \mu(\lambda) \leq \mu(\theta_1) + \delta$ for some δ. Let ϵ be $\delta kl(\theta_2||\theta_1)$. Hence, we get the following:

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

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

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

By taking $\delta \to 0$, we obtain $\mathbb{E}_{\boldsymbol{\theta}}[m_2^v(n)|\omega_{\bar{v}}] \geq \frac{\ln n}{kl(\theta_2||\theta_1)}(1 - o(1))$. Hence, we have proved that for any $v \in V$, $\omega_{\bar{v}}$ and any sub-optimal arm j, $\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)|\omega_{\bar{v}}]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta_1)}$ and $\liminf_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta_1)}$, 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 4 implies that for each $v \in V$, $\exists\, n_v \in \mathbb{N}$ such that

$$\frac{\mathbb{E}_{\boldsymbol{\theta}}[m_j^v(n)]}{\ln n} \geq \frac{1}{kl(\theta_j||\theta_1)}, \ \forall n \geq n_v. \quad (30)$$

Let $n' = \max(n_v : v \in V)$. Using (30) for each $v \in V$, and for $n \geq n'$, we determine a lower bound for $\mathbb{E}_{\boldsymbol{\theta}}[m_j^G(n)]$. Due to Non Altruistic property of the policy, 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 $n \geq n'$.

$$\begin{aligned} &\text{minimize} \ \ \|z_m\|_1 \\ &\text{s.t } \exists \text{ a sequence } \{z_k\}_{k=1}^m \text{ and} \\ &\exists \{\eta_k\}_{k=1}^m \text{which is a permutation of} \{1,2,\ldots,m\} \\ &z_i(\eta_k) = z_k(\eta_k) \ \ \forall i \geq k, \forall k \\ &\langle z_k, A(\eta_k,:)\rangle \geq q_j = \left\lceil \frac{\ln n}{kl(\theta_j||\theta_1)} \right\rceil \ \ \forall k \\ &z_k \in \{0,1,2,\cdots,q_j\}^m, \ \ \forall k. \end{aligned} \tag{31}$$

Note that, the notation in (31) is same as used in (25). In order to lower bound the solution to (31), we now lower bound the objective function value in (31) at each feasible point of it.

Let $\{z_k\}_{k=1}^m$ and $\{\eta_k\}_{k=1}^m$ be a feasible point pair of (31). Let $S = \{s_1, s_2, \cdots, s_{\alpha(G^2)}\}$ be a maximum independence set of the graph G^2. Here, $G^2 = (V, E')$ is the graph $G = (V, E)$ augmented with all edges between any pair of nodes if they have at least one common neighbour. Essentially, $(u,v) \in E'$ if and only if either $(u,v) \in E$ or $\mathcal{N}(u) \cap \mathcal{N}(v) \neq \phi$. It implies that, for any two distinct nodes $s_i, s_j \in S$, we have $\mathcal{N}(s_i) \cap \mathcal{N}(s_j) = \phi$. Then, $\sum_{k=1}^m z_m(k) \geq \sum_{i=1}^{\alpha(G^2)} \sum_{d \in \mathcal{N}(s_i)} z_m(d) \overset{(*)}{\geq} \sum_{i=1}^{\alpha(G^2)} q_j = \alpha\left(G^2\right) q_j$, where $(*)$ is due to the fact that, in the vector z_m, every node has access to at least q_j samples of sub-optimal arm-j. Since the inequality above is true for any feasible point, it is true that solution to (31) is also lower bounded by $\alpha\left(G^2\right) q_j$. Hence, we get that $\mathbb{E}[m_j^G(n)] \geq \alpha(G^2) \cdot \frac{\ln n}{kl(\theta_j||\theta_1)}, \forall n \geq n'$. It implies that $\liminf_{n\to\infty} \frac{\mathbb{E}[m_j^G(n)]}{\ln n} \geq \alpha(G^2) \cdot \frac{1}{kl(\theta_j||\theta_1)}$, which establishes the desired result. ■

APPENDIX C

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 μ_b in the first round as follows. $\mathbb{E}[m_b^{G_n}(1)] = \sum_{v \in V} \left(\frac{p_2}{2} + \frac{p_1}{2}\right) = \frac{m_n}{2} \geq \frac{\ln n}{kl(\mu_b||\mu_a)}$, since MAB is (μ_a, μ_b) with probability $\frac{1}{2}$, and is (μ_b, μ_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, $\mathbb{E}[m_2^{G_n,1}(1)] \geq \frac{\ln n}{kl(\mu_2||\mu_1)}, \ \ \forall n \in \mathbb{N}$. From Theorem 3, it follows that

$$\liminf_{n\to\infty} \frac{\mathbb{E}[m_2^{G_n,v}(n)]}{\ln n} \geq \frac{1}{kl(\mu_2||\mu_1)}, \ \ \forall \ v \in V_n. \tag{32}$$

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 (32), 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, $\mathbb{E}[m_2^{G_n}(n)] \geq (m_n - 1)\left(\frac{\ln n}{kl(\mu_2||\mu_1)} - O(1) - 1\right) \forall n \geq n'$. It implies that $\liminf_{n\to\infty} \frac{\mathbb{E}[m_2^{G_n}(n)]}{(m_n-1)\ln n} \geq \frac{1}{kl(\mu_2||\mu_1)}$, which completes the proof. ■

ACKNOWLEDGEMENTS

This work is undertaken as a part of an Information Technology Research Academy (ITRA), Media Labs Asia, project titled "De-congesting India's transportation networks". Also, this work is partially supported by a DST INSPIRE Faculty grant [IFA13-ENG-69] awarded to Aditya Gopalan.

Ravi Kumar Kolla is a Ph.D. student in the Department of Electrical Engineering at the Indian Institute of Technology Madras, Chennai, India. He received the M.Tech degree from NIT Calicut, India, and the B.Tech degree from Sri Venkateswara University, Tirupati, India. His research interests include online learning, multi-armed bandit problems and statistical inference.

Krishna Jagannathan obtained his B.Tech. in Electrical Engineering from IIT Madras in 2004, and the S.M. and Ph.D. degrees in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology (MIT) in 2006 and 2010 respectively. During 2010-2011, he was a visiting post-doctoral scholar in Computing and Mathematical Sciences at Caltech, and an off-campus post-doctoral fellow at MIT. Since November 2011, he has been an assistant professor in the Department of Electrical Engineering, IIT Madras. He worked as a consultant at the Mathematical Sciences Research Center, Bell Laboratories, Murray Hill, NJ in 2005, and as an engineering intern at Qualcomm, Campbell, CA in 2007. His research interests lie in the stochastic modeling and analysis of communication networks, transportation networks, network control, and queuing theory. Dr. Jagannathan serves on the editorial boards of the journals *IEEE/ACM Transactions on Networking* and *Performance Evaluation*. He is the recipient of a best paper award at WiOpt 2013, and the Young Faculty Recognition Award for excellence in Teaching and Research at IIT Madras (2014).

Aditya Gopalan is an Assistant Professor and INSPIRE Faculty Fellow at the Indian Institute of Science, Electrical Communication Engineering. He received the Ph.D. degree in electrical engineering from the University of Texas at Austin, and the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology Madras. He was an Andrew and Erna Viterbi Post-Doctoral Fellow at the Technion-Israel Institute of Technology. His research interests include machine learning and statistical inference, control, and algorithms for resource allocation in communication networks.