On Distributed Cooperative Decision-Making in Multiarmed Bandits

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard

Abstract— We study the explore-exploit tradeoff in distributed cooperative decision-making using the context of the multiarmed bandit (MAB) problem. For the distributed cooperative MAB problem, we design the cooperative UCB algorithm that comprises two interleaved distributed processes: (i) running consensus algorithms for estimation of rewards, and (ii) upper-confidence-bound-based heuristics for selection of arms. We rigorously analyze the performance of the cooperative UCB algorithm and characterize the influence of communication graph structure on the decision-making performance of the group.

I. INTRODUCTION

Cooperative decision-making under uncertainty is ubiquitous in natural systems as well as in engineering networks. Typically in a distributed cooperative decision-making scenario, there is assimilation of information across a network followed by decision-making based on the collective information. The result is a kind of *collective intelligence*, which is of fundamental interest both in terms of understanding natural systems and designing efficient engineered systems.

A fundamental feature of decision-making under uncertainty is the *explore-exploit* tradeoff. The explore-exploit tradeoff refers to the tension between learning and optimizing: the decision-making agent needs to learn the unknown system parameters (exploration), while maximizing its decision-making objective, which depends on the unknown parameters (exploitation).

MAB problems are canonical formulations of the explore-exploit tradeoff. In a stochastic MAB problem a set of options (arms) are given. A stochastic reward with an unknown mean is associated with each option. A player can pick only one option at a time, and the player's objective is to maximize the cumulative expected reward over a sequence of choices. In an MAB problem, the player needs to balance the tradeoff between learning the mean reward at each arm (exploration), and picking the arm with maximum mean reward (exploitation).

MAB problems are pervasive across a variety of scientific communities and have found application in diverse areas including controls and robotics [25], ecology [15], [24],

P. Landgren, V. Srivastava and N. E. Leonard are with the Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA, {landgren, vaibhavs, naomi}@princeton.edu.

psychology [22], and communications [16], [1]. Despite the prevalence of the MAB problem, the research on MAB problems has primarily focused on policies for a single agent. The increasing importance of networked systems warrants the development of distributed algorithms for multiple communicating agents faced with MAB problems. In this paper, we extend a popular single-agent algorithm for the stochastic MAB problem to the distributed multiple agent setting and analyze decision-making performance as a function of the network structure.

The MAB problem has been extensively studied (see [6] for a survey). In their seminal work, Lai and Robbins [17] established a logarithmic lower bound on the expected number of times a sub-optimal arm needs to be selected by an optimal policy. In another seminal work, Auer *et al.* [3] developed the upper confidence bound (UCB) algorithm for the stochastic MAB problem that achieves the lower bound in [17] uniformly in time. Anantharam *et al.* [2] extended the results of [17] to the setting of multiple centralized players.

Recently, researchers [13], [1] have studied the MAB problem with multiple players in the decentralized setting. Primarily motivated by communication networks, these researchers assume no communication among agents and design efficient decentralized policies. Kar *et al.* [14] investigated the multiagent MAB problem in a leader-follower setting. They designed efficient policies for systems in which there is one major player that can access the rewards and the remaining minor players can only observe the sampling patterns of the major player. The MAB problem has also been used to perform empirical study of collaborative learning in social networks and the influence of network structure on decision-making performance of human agents [19].

Here, we use a running consensus algorithm [5] for assimilation of information, which is an extension of the classical DeGroot model [8] in the social networks literature. Running consensus and related models have been used to study learning [12] and decision-making [23] in social networks.

In the present paper we study the distributed cooperative MAB problem in which a set of agents are faced with a stochastic MAB problem and communicate their information with their neighbors in an undirected, connected communication graph. We use a set of running consensus algorithms for cooperative estimation of the mean reward at each arm, and we design an arm selection heuristic that leads to an order-optimal performance for the group. The major contributions of this paper are as follows.

First, we employ and rigorously analyze running consensus algorithms for distributed cooperative estimation of mean reward at each arm, and we derive bounds on several relevant

quantities.

Second, we propose and thoroughly analyze the cooperative UCB algorithm. We derive bounds on decision-making performance for the group and characterize the influence of the network structure on the performance.

Third, we introduce a novel graph centrality measure and numerically demonstrate that this measure captures the ordering of explore-exploit performance of each agent.

The remainder of the paper is organized as follows. In Section II we recall some preliminaries about the stochastic MAB problem and consensus algorithms. In Section III we present and analyze the cooperative estimation algorithm. We propose and analyze the cooperative UCB algorithm in Section IV. We illustrate our analytic results with numerical examples in Section V. We conclude in Section VI.

II. BACKGROUND

In this section we recall the standard MAB problem, the UCB algorithm, and some preliminaries on discrete-time consensus.

A. The Single Agent MAB Problem

Consider an N-armed bandit problem, i.e., an MAB problem with N arms. The reward associated with arm $i \in \{1, \ldots, N\}$ is a random variable with an unknown mean m_i. Let the agent choose arm $i(t)$ at time $t \in \{1, \ldots, T\}$ and receive a reward $r(t)$. The decision-maker's objective is to choose a sequence of arms $\{i(t)\}_{t \in \{1, \ldots, T\}}$ that maximizes the expected cumulative reward $\sum_{t=1}^{T} m_{i(t)}$, where T is the horizon length of the sequential allocation process.

For an MAB problem, the expected *regret* at time t is defined by $R(t) = m_{i^*} - m_{i(t)}$, where $m_{i^*} = \max\{m_i \mid i \in \{1, \ldots, N\}\}$. The objective of the decision-maker can be equivalently defined as minimizing the expected cumulative regret defined by $\sum_{t=1}^{T} R(t) = \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i(T)]$, where $n_i(T)$ is the cumulative number of times arm i has been chosen until time T and $\Delta_i = m_{i^*} - m_i$ is the expected regret due to picking arm i instead of arm i^*. It is known that the regret of any algorithm for an MAB problem is asymptotically lower bounded by a logarithmic function of the horizon length T [17], i.e., no algorithm can achieve an expected cumulative regret smaller than a logarithmic function of horizon length as $T \to \infty$.

In this paper, we focus on Gaussian rewards, i.e., the reward at arm i is sampled from a Gaussian distribution with mean m_i and variance σ^2. We assume that the variance σ^2 is known and is the same at each arm.

B. The UCB Algorithm

A popular solution to the stochastic MAB problem is the UCB algorithm proposed in [3]. The UCB algorithm initializes by sampling each arm once, and then selects an arm with maximum

$$Q_i(t) = \hat{\mu}_i(t) + C_i(t),$$

where $n_i(t)$ is the number of times arm i has been chosen up to and including time t, and $\hat{\mu}_i(t)$ and $C_i(t) = \sqrt{\frac{2 \ln(t)}{n_i(t)}}$

are the empirical mean reward of arm i and the associated measure of the uncertainty associated with that mean at time t, respectively.

The function $Q_i(t)$ is judiciously designed to balance the tradeoff between explore and exploit: the terms $\hat{\mu}_i(t)$ and $C_i(t)$ facilitate exploitation and exploration, respectively. The UCB algorithm as described above assumes that rewards have a bounded support $[0, 1]$, but this algorithm can be easily extended to distributions with unbounded support [18].

C. The Cooperative MAB Problem

The cooperative MAB problem is an extension of the single-agent MAB problem where M agents act over the same N arms. Agents maintain bidirectional communication, and the communication network can be modeled as an undirected graph $\mathcal{G}$ in which each node represents an agent and edges represent the communication between agents [7]. Let $A \in \mathbb{R}^{M \times M}$ be the adjacency matrix associated with $\mathcal{G}$ and let $L \in \mathbb{R}^{M \times M}$ be the corresponding Laplacian matrix. We assume that the graph $\mathcal{G}$ is connected, i.e., there exists a path between each pair of nodes.

In the cooperative setting, the objective of the group is defined as minimizing the expected cumulative group regret, defined by $\sum_{k=1}^{M} \sum_{t=1}^{T} R^k(t) = \sum_{k=1}^{M} \sum_{i=1}^{N} \Delta_i \mathbb{E}[n_i^k(T)]$, where $R^k(t)$ is the regret of agent k at time t and $n_i^k(T)$ is the total cumulative number of times arm i has been chosen by agent k until time T. In the cooperative setting using Gaussian rewards the lower bound [2] on the expected number of times a suboptimal arm i is selected by a fusion center that has access to reward for each agent is

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \geq \left(\frac{2\sigma^2}{\Delta_i^2} + o(1) \right) \ln T. \tag{1}$$

In the following, we will design a distributed algorithm that samples a suboptimal arm i within a constant factor of the above bound.

D. Discrete-Time Consensus

Consider a set of agents $\{1, \ldots, M\}$, each of which maintains bidirectional communication with a set of neighboring agents. The objective of the consensus algorithms is to ensure agreement among agents on a common value. In the discrete-time consensus algorithm [11], [26], agents average their opinion with their neighbors' opinions at each time. A discrete-time consensus algorithm can be expressed as

$$\mathbf{x}(t) = P \mathbf{x}(t-1), \tag{2}$$

where $\mathbf{x}(t)$ is the vector of each agent's opinion, and P is a row stochastic matrix given by

$$P = \mathcal{I}_M - \frac{\kappa}{d_{\max}} L. \tag{3}$$

$\mathcal{I}_M$ is the identity matrix of order M, $\kappa \in (0, 1]$ is a step size parameter [21], $d_{\max} = \max\{\deg(i) \mid i \in \{1, \ldots, M\}\}$, and $\deg(i)$ is the degree of node i. In the following, we assume without loss of generality that the eigenvalues of P are ordered such that $\lambda_1 = 1 > \lambda_2 \geq \ldots \geq \lambda_M > -1$.

In the context of social networks, the consensus algorithm (2) is referred to as the Degroot model [8] and has been successfully used to describe evolution of opinions [10].

One drawback of the consensus algorithm (2) is that it does not allow for incorporating new external information. This drawback can be mitigated by adding a forcing term and the resulting algorithm is called the *running consensus* [5]. Similar to (2), the running consensus updates the opinion at time t as

$$\mathbf{x}(t) = P\mathbf{x}(t-1) + P\boldsymbol{v}(t), \qquad (4)$$

where $\boldsymbol{v}(t)$ is the information received at time t. In the running consensus update (4), each agent k collects information $v_k(t)$ at time t, adds it to its current opinion, and then averages its updated opinion with the updated opinion of its neighbors.

III. COOPERATIVE ESTIMATION OF MEAN REWARDS

In this section we investigate the cooperative estimation of mean rewards at each arm. To this end, we propose two running consensus algorithms for each arm and analyze their performance.

A. Cooperative Estimation Algorithm

For distributed cooperative estimation of the mean reward at each arm i, we employ two running consensus algorithms: (i) for estimation of total reward provided at the arm, and (ii) for estimation of the total number of times the arm has been sampled.

Let $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ be agent k's estimate of the total reward provided at arm i per unit agent and the total number of times arm i has been selected until time t per unit agent, respectively. Using $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ agent k can calculate $\hat{\mu}_i^k(t)$, the estimated empirical mean of arm i at time t defined by

$$\hat{\mu}_i^k(t) = \frac{\hat{s}_i^k(t)}{\hat{n}_i^k(t)}. \qquad (5)$$

Let $i^k(t)$ be the arm sampled by agent k at time t and let $\xi_i^k(t) = \mathbb{1}(i^k(t) = i)$. $\mathbb{1}(\cdot)$ is the indicator function, here equal to 1 if $i^k(t) = i$ and 0 otherwise. For simplicity of notation we define $r_i^k(t)$ as the realized reward at arm i for agent k, which is a random variable sampled from $\mathcal{N}(m_i, \sigma^2)$, and the corresponding accumulated reward is $r^k(t) = r_i^k(t) \cdot \mathbb{1}(i^k(t) = i)$.

The estimates $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ are updated using running consensus as follows

$$\hat{\mathbf{n}}_i(t) = P\hat{\mathbf{n}}_i(t-1) + P\boldsymbol{\xi}_i(t), \qquad (6)$$
$$\text{and} \quad \hat{\mathbf{s}}_i(t) = P\hat{\mathbf{s}}_i(t-1) + P(\mathbf{r}_i(t) \circ \boldsymbol{\xi}_i(t)), \qquad (7)$$

where $\hat{\mathbf{n}}_i(t)$, $\hat{\mathbf{s}}_i(t)$, $\boldsymbol{\xi}_i(t)$, and $\mathbf{r}_i(t)$ are vectors of $\hat{n}_i^k(t)$, $\hat{s}_i^k(t)$, $\xi_i^k(t)$, and $r_i^k(t)$, $k \in \{1, \ldots, M\}$, respectively, and $\circ$ denotes element-wise multiplication (Hadamard product).

B. Analysis of the Cooperative Estimation Algorithm

We now analyze the performance of the estimation algorithm defined by (5), (6) and (7). Let $n_i^{\text{cent}}(t) \equiv \frac{1}{M} \sum_{\tau=1}^{t} \mathbf{1}_M^\top \boldsymbol{\xi}_i(\tau)$ be the total number of times arm i has been selected per unit agent up to and including time t, and let $s_i^{\text{cent}}(t) \equiv \frac{1}{M} \sum_{\tau=1}^{t} \boldsymbol{\xi}_i^\top(t) \mathbf{r}_i(t)$ be the total reward provided at arm i per unit agent up to and including time t. Also, let λ_i denote the i-th largest eigenvalue of P, $\mathbf{u}_i$ the eigenvector corresponding to λ_i, u_i^d the d-th entry of $\mathbf{u}_i$, and

$$\epsilon_n = \sqrt{M} \sum_{p=2}^{M} \frac{|\lambda_p|}{1 - |\lambda_p|}. \qquad (8)$$

Note that $\lambda_1 = 1$ and $\mathbf{u}_1 = \mathbf{1}_M / \sqrt{M}$. Let us define

$$\nu_{pj}^{\text{+sum}} = \sum_{d=1}^{M} u_p^d u_j^d \mathbb{1}(u_p^k u_j^k \geq 0)$$

$$\text{and} \quad \nu_{pj}^{\text{-sum}} = \sum_{d=1}^{M} u_p^d u_j^d \mathbb{1}(u_p^k u_j^k \leq 0).$$

We also define

$$a_{pj}(k) = \begin{cases} \nu_{pj}^{\text{+sum}} u_p^k u_j^k, & \text{if } \lambda_p \lambda_j \geq 0 \ \& \ u_p^k u_j^k \geq 0, \\ \nu_{pj}^{\text{-sum}} u_p^k u_j^k, & \text{if } \lambda_p \lambda_j \geq 0 \ \& \ u_p^k u_j^k \leq 0, \\ \nu_{pj}^{\max} |u_p^k u_j^k|, & \text{if } \lambda_p \lambda_j < 0, \end{cases} \qquad (9)$$

where $\nu_{pj}^{\max} = \max\{|\nu_{pj}^{\text{-sum}}|, \nu_{pj}^{\text{+sum}}\}$. Furthermore, let

$$\epsilon_c^k = M \sum_{p=1}^{M} \sum_{j=2}^{M} \frac{|\lambda_p \lambda_j|}{1 - |\lambda_p \lambda_j|} a_{pj}(k). \qquad (10)$$

We note that both ϵ_n and ϵ_c^k depend only on the topology of the communication graph. These are measures of distributed cooperative estimation performance.

*Proposition 1 (**Performance of cooperative estimation**):* For the distributed estimation algorithm defined in (5), (6) and (7), and a doubly stochastic matrix P defined in (3), the following statements hold

(i) the estimate $\hat{n}_i^k(t)$ satisfies

$$n_i^{\text{cent}}(t) - \epsilon_n \leq \hat{n}_i^k(t) \leq n_i^{\text{cent}}(t) + \epsilon_n;$$

(ii) the following inequality holds for the estimate $\hat{n}_i^k(t)$ and the sequence $\{\xi_i^j(\tau)\}_{\tau \in \{1,\ldots,t\}}, j \in \{1, \ldots, M\}$

$$\sum_{\tau=1}^{t} \sum_{j=1}^{M} \left(\sum_{p=1}^{M} \lambda_p^{t-\tau+1} u_p^k u_p^j \right)^2 \xi_i^j(\tau) \leq \frac{\hat{n}_i^k(t) + \epsilon_c^k}{M}.$$

Proof: We begin with the first statement. From (6) it follows that

$$\hat{\mathbf{n}}_i(t) = P^t \hat{\mathbf{n}}_i(0) + \sum_{\tau=1}^{t} P^{t-\tau+1} \boldsymbol{\xi}_i(\tau)$$

$$= \sum_{\tau=0}^{t} \left[\frac{1}{M} \mathbf{1}_M \mathbf{1}_M^\top \boldsymbol{\xi}_i(\tau) + \sum_{p=2}^{M} \lambda_p^{t-\tau+1} \mathbf{u}_p \mathbf{u}_p^\top \boldsymbol{\xi}_i(\tau) \right]$$

$$= n_i^{\text{cent}}(t) \mathbf{1}_M + \sum_{\tau=1}^{t} \sum_{p=2}^{M} \lambda_p^{t-\tau+1} \mathbf{u}_p \mathbf{u}_p^\top \boldsymbol{\xi}_i(\tau). \qquad (11)$$

We now bound the k-th entry of the second term on the right hand side of (11):

$$\sum_{\tau=1}^{t}\sum_{p=2}^{M}\lambda_p^{t-\tau+1}\left(\mathbf{u}_p\mathbf{u}_p^\top\boldsymbol{\xi}_i(\tau)\right)_k \le \sum_{\tau=1}^{t}\sum_{p=2}^{M}|\lambda_p^{t-\tau+1}|\|\mathbf{u}_p\|_2^2\|\boldsymbol{\xi}_i(\tau)\|_2$$

$$\le \sqrt{M}\sum_{\tau=1}^{t}\sum_{p=2}^{M}|\lambda_p^{t-\tau+1}| \le \epsilon_n.$$

This establishes the first statement.

To prove the second statement, we note that

$$\sum_{\tau=1}^{t}\sum_{j=1}^{M}\left(\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^k u_p^j\right)^2\xi_i^j(\tau)$$

$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=1}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^k u_w^k\sum_{j=1}^{M}u_p^j u_w^j\xi_i^j(\tau)$$

$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^k u_w^k\nu_{pwi}(\tau)$$

$$+\frac{1}{M}\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{j=1}^{M}\lambda_p^{t-\tau+1}u_p^k u_p^j\xi_i^j(\tau)$$

$$=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^k u_w^k\nu_{pwi}(\tau)+\frac{1}{M}\hat{n}_i^k(t),$$

$$(12)$$

where $\nu_{pwi}(\tau)=\sum_{j=1}^{M}u_p^j u_w^j\xi_i^j(\tau)$.

We now analyze the first term of (12):

$$\sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}(\lambda_p\lambda_w)^{t-\tau+1}u_p^k u_w^k\nu_{pwi}(\tau)$$

$$\le \sum_{\tau=1}^{t}\sum_{p=1}^{M}\sum_{w=2}^{M}|(\lambda_p\lambda_w)^{t-\tau+1}||u_p^k u_w^k\nu_{pwi}(\tau)|$$

$$\le \sum_{\tau=0}^{t-1}\sum_{p=1}^{M}\sum_{w=2}^{M}|\lambda_p\lambda_w|^{t-\tau+1}a_{pw}(k)$$

$$\le \sum_{p=1}^{M}\sum_{w=2}^{M}\frac{|\lambda_p\lambda_w|}{1-|\lambda_p\lambda_w|}a_{pw}(k). \quad (13)$$

Bounds in (13) establish the second statement. ∎

We now derive bounds on the deviation of the estimated mean when using the cooperative estimation algorithm. We use techniques from [9]. Recall that for $i\in\{1,\dots,N\}$ and $k\in\{1,\dots,M\}$ we let $\{r_i^k(t)\}_{t\in\mathbb{N}}$ be the sequence of i.i.d. Gaussian random variables with mean $m_i\in\mathbb{R}$. Let $\mathcal{F}_t$ be the filtration defined by the sigma-algebra of all the measurements until time t. Let $\{\xi_i^k(t)\}_{t\in\mathbb{N}}$ be a sequence of Bernoulli variables such that $\xi_i^k(t)$ is deterministically known given $\mathcal{F}_{t-1}$, i.e., $\xi_i^k(t)$ is pre-visible w.r.t. $\mathcal{F}_{t-1}$. Additionally, let $\phi_i(\beta)=\ln\left(\mathbb{E}[\exp\left(\beta r_i^k(t)\right)]\right)$ denote the cumulant generating function of $r_i^k(t)$.

*Theorem 1 (**Estimator Deviation Bounds**):* For the estimates $\hat{s}_i^k(t)$ and $\hat{n}_i^k(t)$ obtained using equations (6) and (7),

the following concentration inequality holds

$$\mathbb{P}\left(\frac{\hat{s}_i^k(t)-m_i\hat{n}_i^k(t)}{(\frac{1}{M}(\hat{n}_i^k(t)+\epsilon_c^k))^{1/2}}>\delta\right)<\left\lceil\frac{\ln(t+\epsilon_n)}{\ln(1+\eta)}\right\rceil\exp\left(\frac{-\delta^2}{2\sigma^2}G(\eta)\right),$$

$$(14)$$

where $\delta>0$, $\eta>0$, $G(\eta)=(1-\frac{\eta^2}{16})$, and ϵ_c^k and ϵ_n are defined in (10) and (8), respectively.

Proof: We begin by noting that $\hat{s}_i^k(t)$ can be decomposed as

$$\hat{s}_i^k(t)=\sum_{\tau=1}^{t}\sum_{p=1}^{M}\lambda_p^{t-\tau+1}\sum_{j=1}^{M}u_p^k u_p^j r_i^j(\tau)\xi_i^j(\tau). \quad (15)$$

Let $\hat{s}_i^{kp}(t)=\sum_{\tau=1}^{t}\lambda_p^{t-\tau+1}\sum_{j=1}^{M}u_p^k u_p^j r_i^j(\tau)\xi_i^j(\tau)$. Then,

$$\sum_{p=1}^{M}\hat{s}_i^{kp}(t)=\sum_{p=1}^{M}\sum_{j=1}^{M}\lambda_p u_p^k u_p^j r_i^j(t)\xi_i^j(t)+\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1).$$

$$(16)$$

It follows from (15) and (16) that for any $\Theta>0$

$$\mathbb{E}\left[\exp\left(\Theta\hat{s}_i^k(t)\right)\big|\mathcal{F}_{t-1}\right]=\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\hat{s}_i^{kp}(t)\right)\bigg|\mathcal{F}_{t-1}\right]$$

$$=\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\sum_{j=1}^{M}u_p^k u_p^j r_i^j(t)\xi_i^j(t)\right)\bigg|\mathcal{F}_{t-1}\right]$$

$$\times\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1)\right)$$

$$=\exp\left(\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_p u_p^k u_p^j r_i^j(t)\right)\xi_i^j(t)\right)$$

$$\times\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1)\right),$$

where the last line follows using the fact that conditioned on $\mathcal{F}_{t-1}$, $\xi_i^j(t)$ is a deterministic variable and $r_i^j(t)$ are i.i.d. for each $j\in\{1,\dots,M\}$. Therefore, it follows that

$$\mathbb{E}\left[\exp\left(\Theta\sum_{p=1}^{M}\hat{s}_i^{kp}(t)-\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_p u_p^k u_p^j r_i^j(t)\right)\right.\right.$$

$$\left.\left.\times\xi_i^j(t)\right)\bigg|\mathcal{F}_{t-1}\right]=\exp\left(\Theta\sum_{p=1}^{M}\lambda_p\hat{s}_i^{kp}(t-1)\right).$$

Using the above argument recursively with the fact that $s_i^k(0)=0$, we obtain

$$\mathbb{E}\left[\exp\left(\Theta\hat{s}_i^k(t)\right.\right.$$

$$\left.\left.-\sum_{\tau=1}^{t}\sum_{j=1}^{M}\phi_i\left(\Theta\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^k u_p^j r_i^j(\tau)\right)\xi_i^j(\tau)\right)\right]=1.$$

Since for Gaussian random variables $\phi_i(\beta) = \beta m_i + \frac{1}{2}\sigma^2\beta^2$, we have

$$
\begin{aligned}
1 = \mathbb{E}\bigg[&\exp\bigg(\Theta\big(\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)\big) \\
&- \frac{\sigma^2}{2}\sum_{\tau=1}^{t}\sum_{j=1}^{M}\bigg(\Theta\sum_{p=1}^{M}\lambda_p^{t-\tau+1}u_p^k u_p^j\bigg)^2 \xi_i^j(\tau)\bigg)\bigg] \\
\geq \mathbb{E}\bigg[&\exp\bigg(\Theta\big(\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)\big) - \frac{\sigma^2\Theta^2}{2M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)\bigg],
\end{aligned}
$$

where the last inequality follows from the second statement of Proposition 1. Now using the Markov Inequality, we obtain

$$
\begin{aligned}
e^{-a} &\geq \mathbb{P}\bigg(\exp\bigg(\Theta\big(\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)\big) - \frac{\sigma^2\Theta^2}{2M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg) \geq e^a\bigg) \\
&= \mathbb{P}\bigg(\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} \geq \frac{a}{\Theta}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{-\frac{1}{2}} \\
&\qquad + \frac{\sigma^2\Theta}{2}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{\frac{1}{2}}\bigg). \quad (17)
\end{aligned}
$$

The right hand side of the above inequality contains a random variable $\hat{n}_i^k(t)$ which is dependent on the random variable on the left hand side. Therefore, we use union bounds on $\hat{n}_i^k(t)$ to obtain the desired concentration inequality. Towards this end, we consider an exponentially increasing sequence of time indices $\{(1+\eta)^{g-1} \mid g \in \{1, \ldots, D\}\}$, where $D = \left\lceil \frac{\ln(t+\epsilon_n)}{\ln(1+\eta)} \right\rceil$ and $\eta > 0$. For every $g \in \{1, \ldots, D\}$, define

$$
\Theta_g = \frac{1}{\sigma}\sqrt{\frac{2aM}{(1+\eta)^{g-\frac{1}{2}} + \epsilon_c^k}}. \quad (18)
$$

Thus, if $(1+\eta)^{g-1} \leq \hat{n}_i^k(t) \leq (1+\eta)^g$, then

$$
\begin{aligned}
&\frac{a}{\Theta_g}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{-\frac{1}{2}} + \frac{\sigma^2\Theta_g}{2}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{\frac{1}{2}} \\
&= \sigma\sqrt{\frac{a}{2}}\bigg(\bigg(\frac{(1+\eta)^{g-\frac{1}{2}} + \epsilon_c^k}{\hat{n}_i^k(t) + \epsilon_c^k}\bigg)^{\frac{1}{2}} + \bigg(\frac{\hat{n}_i^k(t) + \epsilon_c^k}{(1+\eta)^{g-\frac{1}{2}} + \epsilon_c^k}\bigg)^{\frac{1}{2}}\bigg) \\
&\leq \sigma\sqrt{\frac{a}{2}}\bigg(\bigg(\frac{(1+\eta)^{g-\frac{1}{2}}}{\hat{n}_i^k(t)}\bigg)^{\frac{1}{2}} + \bigg(\frac{\hat{n}_i^k(t)}{(1+\eta)^{g-\frac{1}{2}}}\bigg)^{\frac{1}{2}}\bigg) \\
&\leq \sigma\sqrt{\frac{a}{2}}\bigg((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\bigg), \quad (19)
\end{aligned}
$$

where the second-to-last inequality follows from the fact that for $a, b > 0$, the function $\epsilon \mapsto \sqrt{\frac{a+\epsilon}{b+\epsilon}} + \sqrt{\frac{b+\epsilon}{a+\epsilon}}$ with domain $\mathbb{R}_{\geq 0}$ is monotonically non-increasing, and the last inequality follows from the fact the for $\eta > 0$, the function $x \mapsto \sqrt{\frac{(1+\eta)^{g-\frac{1}{2}}}{x}} + \sqrt{\frac{x}{(1+\eta)^{g-\frac{1}{2}}}}$ with domain $[(1+\eta)^{g-1}, (1+\eta)^g]$ achieves its maximum at either of the boundaries.

It follows from (17) that

$$
\begin{aligned}
e^{-a} \geq \mathbb{P}\bigg(&\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} > \frac{a}{\Theta_g}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{-\frac{1}{2}} \\
&+ \frac{\sigma^2\Theta_g}{2}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{\frac{1}{2}} \\
&\& \; (1+\eta)^{g-1} \leq \hat{n}_i^k(t) + \epsilon_c^k < (1+\eta)^g\bigg),
\end{aligned}
$$

for any $g \in \{1, \ldots, D\}$. Therefore,

$$
\begin{aligned}
De^{-a} \geq \sum_{g=1}^{D} \mathbb{P}\bigg(&\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} > \frac{a}{\Theta_g}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{-\frac{1}{2}} \\
&+ \frac{\sigma^2\Theta_g}{2}\bigg(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\bigg)^{\frac{1}{2}} \\
&\& \; (1+\eta)^{g-1} \leq \hat{n}_i^k(t) + \epsilon_c^k < (1+\eta)^g\bigg) \\
\geq \mathbb{P}\bigg(&\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} > \sigma\sqrt{\frac{a}{2}}\bigg((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\bigg)\bigg),
\end{aligned}
$$

where the last inequality follows from inequality (19). Setting $\sigma\sqrt{\frac{a}{2}}\big((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\big) = \delta$, this yields

$$
\begin{aligned}
&\mathbb{P}\bigg(\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} > \delta\bigg) \\
&\qquad \leq D\exp\bigg(\frac{-2\delta^2}{\sigma^2\big((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\big)^2}\bigg).
\end{aligned}
$$

It can be verified that the first three terms in the Taylor series for $\frac{4}{\big((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\big)^2}$ provide a lower bound, i.e.,

$$
\frac{4}{\big((1+\eta)^{\frac{1}{4}} + (1+\eta)^{-\frac{1}{4}}\big)^2} \geq 1 - \frac{\eta^2}{16}.
$$

Therefore, it holds that

$$
\begin{aligned}
&\mathbb{P}\bigg(\frac{\hat{s}_i^k(t) - m_i\hat{n}_i^k(t)}{\big(\frac{1}{M}\big(\hat{n}_i^k(t) + \epsilon_c^k\big)\big)^{\frac{1}{2}}} > \delta\bigg) \leq D\exp\bigg(\frac{-\delta^2}{2\sigma^2}\bigg(1 - \frac{\eta^2}{16}\bigg)\bigg) \\
&\qquad = \bigg\lceil \frac{\ln(t+\epsilon_n)}{\ln(1+\eta)} \bigg\rceil \exp\bigg(\frac{-\delta^2}{2\sigma^2}\bigg(1 - \frac{\eta^2}{16}\bigg)\bigg).
\end{aligned}
$$

$\blacksquare$

IV. Cooperative Decision-Making

In this section, we extend the UCB algorithm [3] to the distributed cooperative setting in which multiple agents can communicate with each other according to a given graph topology. Intuitively, compared to the single agent setting, in the cooperative setting each agent will be able to perform better due to communication with neighbors. However, the extent of an agent's performance advantage depends on the

network structure. We compute bounds on the performance of the group in terms of the expected group cumulative regret. We also propose a metric that orders the contribution of agents to the cumulative group regret in terms of their location in the graph.

A. Cooperative UCB Algorithm

The cooperative UCB algorithm is analogous to the UCB algorithm, and uses a modified decision-making heuristic that captures the effect of the additional information an agent receives through communication with other agents as well as the rate of information propagation through the network.

The cooperative UCB algorithm is initialized by each agent sampling each arm once and proceeds as follows. At each time t each agent k selects the arm with maximum $Q_i^k(t-1) = \hat{\mu}_i^k(t-1) + C_i^k(t-1)$, where

$$C_i^k(t-1) = \sigma \sqrt{ \frac{2\gamma}{G(\eta)} \cdot \frac{\hat{n}_i^k(t-1) + \epsilon_c^k}{M\hat{n}_i^k(t-1)} \cdot \frac{\ln(t-1)}{\hat{n}_i^k(t-1)} }, \quad (20)$$

and receives realized reward $r_i^k(t)$, where $\gamma > 1$, $G(\eta) = (1 - \eta^2/16)$, and $\eta \in (0,4)$. Each agent k updates its cooperative estimate of the mean reward at each arm using the distributed cooperative estimation algorithm described in (5), (6), and (7). Note that the heuristic Q_i^k requires the agent k to know ϵ_c^k, which depends on the global graph structure. This requirement can be relaxed by replacing ϵ_c^k with an increasing sub-logarithmic function of time. We leave rigorous analysis of the alternative policy for future investigation.

B. Regret Analysis of the Cooperative UCB Algorithm

We now derive a bound on the expected cumulative group regret using the distributed cooperative UCB algorithm. This bound recovers the upper bound given in (1) within a constant factor. The contribution of each agent to the group regret is a function of its location in the network.

*Theorem 2 (**Regret of the Cooperative UCB Algorithm**):* For the cooperative UCB algorithm and the Gaussian multiarmed bandit problem the number of times a suboptimal arm i is selected by all agents until time T satisfies

$$\sum_{k=1}^{M} \mathbb{E}[n_i^k(T)] \leq \max\left\{ M, \left\lceil M\epsilon_n + \sum_{k=1}^{M} \frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(T)}{M\Delta_i^2} \right\rceil \right\}$$
$$+ \frac{2M}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2} + \frac{\ln((1+\epsilon_n)(1+\eta))}{\gamma-1} + 2 \right)$$

where $\eta > 0$ and $\gamma > 1$.

Proof: We proceed similarly to [3]. The number of times a suboptimal arm i is selected by all agents until time

T is

$$\sum_{k=1}^{M} n_i^k(T) = \sum_{k=1}^{M}\sum_{t=1}^{T} \mathbb{1}(i^k(t) = i^k)$$
$$\leq \sum_{k=1}^{M}\sum_{t=1}^{T} \mathbb{1}(Q_i^k(t-1) \geq Q_{i^*}^k(t-1))$$
$$\leq A + \sum_{k=1}^{M}\sum_{t=1}^{T} \mathbb{1}(Q_i^k(t-1) \geq Q_{i^*}^k(t-1),$$
$$Mn_i^{\text{cent}}(t-1) \geq A), \quad (21)$$

where $A > 0$ is a constant that will be chosen later.

At a given time $t+1$ an individual agent k will choose a suboptimal arm only if $Q_i^k(t) \geq Q_{i^*}^k(t)$. For this condition to be true at least one of the following three conditions must hold:

$$\hat{\mu}_{i^*}^k(t) \leq m_{i^*} - C_{i^*}^k(t) \quad (22)$$
$$\hat{\mu}_i^k(t) \geq m_i + C_i^k(t) \quad (23)$$
$$m_{i^*} < m_i + 2C_i^k(t). \quad (24)$$

We now bound the probability that (23) holds. Applying Theorem 1, it follows for $t \geq N$ (i.e., after initialization) that

$$\mathbb{P}\left(\hat{\mu}_i^k(t) \geq m_i + C_i^k(t)\right) = \mathbb{P}\left(\frac{\hat{s}_i^k - m_i\hat{n}_i^k}{\sqrt{\frac{1}{M}\left(\hat{n}_i^k(t)+\epsilon_c^k\right)}} \geq \sigma\sqrt{\frac{2\gamma\ln(t)}{G(\eta)}} \right)$$
$$\leq \left\lceil \frac{\ln(t+\epsilon_n)}{\ln(1+\eta)} \right\rceil \exp(-\gamma\ln(t))$$
$$\leq \left(\frac{\ln(t+\epsilon_n)}{\ln(1+\eta)} + 1 \right) \exp(-\gamma\ln(t))$$
$$= \left(\frac{\ln\left(t^{\frac{t+\epsilon_n}{t}}\right)}{\ln(1+\eta)} + 1 \right) \frac{1}{t^\gamma}$$
$$\leq \left(\frac{\ln(t(1+\epsilon_n))}{\ln(1+\eta)} + 1 \right) \frac{1}{t^\gamma}$$
$$= \left(\frac{\ln(t)}{\ln(1+\eta)} + \frac{\ln(1+\epsilon_n)}{\ln(1+\eta)} + 1 \right) \frac{1}{t^\gamma}.$$

It follows analogously with a slight modification to Theorem 1 that

$$\mathbb{P}((22)\text{ holds}) \leq \left(\frac{\ln(t)}{\ln(1+\eta)} + \frac{\ln(1+\epsilon_n)}{\ln(1+\eta)} + 1 \right) \frac{1}{t^\gamma}.$$

Finally, we examine the probability that (24) holds. It follows that

$$m_{i^*} < m_i + 2C_i^k(t)$$
$$\implies n_i^{\text{cent}}(t) < \left\lceil \epsilon_n + \frac{8\sigma^2\gamma(\hat{n}_i^k(t)+\epsilon_c^k)\ln(t)}{M\Delta_i^2(\hat{n}_i^k(t))^2} \right\rceil$$
$$\leq \left\lceil \epsilon_n + \frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(t)}{M\Delta_i^2} \right\rceil.$$

From monotonicity of $\ln(t)$, it follows that (24) does not hold if $n_i^{\text{cent}}(t) \geq \left\lceil \epsilon_n + \frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(T)}{M\Delta_i^2} \right\rceil$. Now, let $A =$

$\max\left\{M, \lceil M\epsilon_n + \sum_{k=1}^{M} \frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(T)}{M\Delta_i^2}\rceil\right\}$, where the first element of the set corresponds to the selection of each arm i once by each player during the initialization and the second element ensures that (24) does not hold. Then, it follows from (21) that

$$\sum_{k=1}^{M}\mathbb{E}[n_i^k(T)] \leq \max\left\{M, \left\lceil M\epsilon_n + \sum_{k=1}^{M}\frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(T)}{M\Delta_i^2}\right\rceil\right\}$$
$$+ 2\sum_{k=1}^{M}\sum_{t=N}^{T-1}\left(\frac{\ln(t)}{\ln(1+\eta)} + \frac{\ln(1+\epsilon_n)}{\ln(1+\eta)} + 1\right)\frac{1}{t^\gamma}$$
$$= \max\left\{M, \left\lceil M\epsilon_n + \sum_{k=1}^{M}\frac{8\sigma^2\gamma(1+\epsilon_c^k)\ln(T)}{M\Delta_i^2}\right\rceil\right\}$$
$$+ \frac{2M}{\ln(1+\eta)}\left(\sum_{t=1}^{T}\frac{\ln((1+\epsilon_n)(1+\eta))}{t^\gamma} + \sum_{t=1}^{T}\frac{\ln(t)}{t^\gamma}\right)$$
$$\leq \left\lceil M\epsilon_n + \sum_{k=1}^{M}\frac{8\sigma^2\gamma(1+\epsilon_c^k)}{M\Delta_i^2}\ln(T)\right\rceil$$
$$+ \frac{2M}{\ln(1+\eta)}\left(\frac{1}{(\gamma-1)^2} + \frac{\gamma\ln((1+\epsilon_n)(1+\eta))}{\gamma-1} + 1\right).$$

This establishes the proof. ∎

*Remark 1 (**Towards Explore-Exploit Centrality**):*
Theorem 2 provides bounds on the performance of the group as a function of the graph structure. However, the bound is dependent on the values of ϵ_c^k for each individual agent. In this sense, $\varsigma^k \equiv 1/\epsilon_c^k$ can be thought of as a measure of node certainty in the context of explore-exploit problems. For $\epsilon_c^k = 0$, the agent behaves like a centralized agent. Higher values of ϵ_c^k reflect behavior of an agent with sparser connectivity. Rigorous connections between ϵ_c^k and standard notions of network centralities [20] is an interesting open problem that we leave for future investigation. □

V. NUMERICAL ILLUSTRATIONS

In this section, we elucidate our theoretical analyses from the previous sections with numerical examples. We first demonstrate that the ordering on the performance of nodes obtained through numerical simulations is identical to the ordering by our upper bounds. We then investigate the effect of connectivity on the performance of agents in random graphs.

For all simulations below we consider a 10-armed bandit problem with mean rewards as in Table I, $\sigma = 30$, $\eta = 0$, and $\gamma = 1$.

TABLE I: Rewards at each arm i for simulations

i	1	2	3	4	5	6	7	8	9	10
m_i	40	50	50	60	70	70	80	90	92	95

*Example 1 (**Regret on Fixed Graphs**):* Consider the set of agents communicating according to the graph in Fig. 1 and using the cooperative UCB algorithm to handle the explore-exploit tradeoff in the distributed cooperative MAB problem.
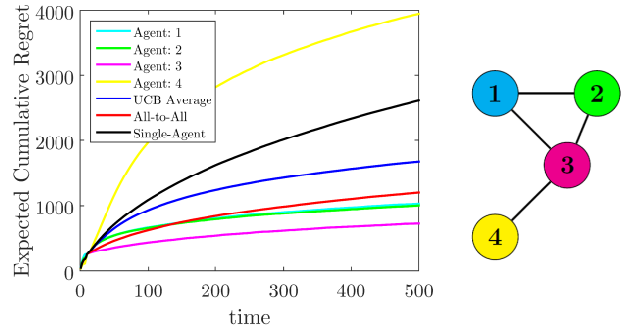

Fig. 1: Simulation results comparing expected cumulative regret for different agents in the fixed network shown, using P as in (3) and $\kappa = \frac{d_{\max}}{d_{\max}-1}$. Note that agents 1 and 2 have nearly identical expected regret.

The values of ϵ_c^k for nodes $1, 2, 3$, and 4 are $2.31, 2.31, 0$, and 5.43, respectively. As predicted by Remark 1, agent 3 should have the lowest regret, agents 1 and 2 should have equal and intermediate regret, and agent 4 should have the highest regret. These predictions are validated in our simulations shown in Fig. 1. The expected cumulative regret in our simulations is computed using 500 Monte-Carlo runs.

We now explore the effect of ϵ_c^k on the performance of an agent in an Erdös-Réyni random (ER) graph. ER graphs are a widely used class of random graphs where any two agents are connected with a given probability ρ [4].

*Example 2 (**Regret on Random Graphs**):* Consider a set of 10 agents communicating according to an ER graph and using the cooperative UCB algorithm to handle the explore-exploit tradeoff in the aforementioned MAB problem. In our simulations, we consider 100 connected ER graphs, and for each ER graph we compute the expected cumulative regret of agents using 30 Monte-Carlo simulations. We show the behavior of the expected cumulative regret of each agent as a function of ς^k in Fig. 2.

It is evident that increased ς^k results in a sharp increase in performance. Conversely, low ς^k is indicative of very poor performance. This strong disparity is due to agents with lower ς^k doing a disproportionately high amount of exploration over the network, allowing other agents to exploit. The disparity is also seen in Fig. 1, as agent 4 does considerably worse than the others. Additionally, as shown in Fig. 1 the expected cumulative regret averaged over all agents is higher than the centralized (all-to-all) case.

VI. FINAL REMARKS

Here we used the distributed multi-agent MAB problem to explore cooperative decision-making in networks. We designed the cooperative UCB algorithm that achieves logarithmic regret for the group. Additionally, we investigated the performance of individual agents in the network as a function of the graph topology. We derived a node certainty measure ς^k that predicts the relative performance of the agents.

Several directions of future research are of interest. First, the arm selection heuristic designed in this paper requires some knowledge of global parameters. Relaxation of this constraint will be addressed in a future publication. Another

interesting direction is to study the tradeoff between the communication and the performance. Specifically, if agents do not communicate at each time but only intermittently, then it is of interest to characterize performance as a function of the communication frequency.
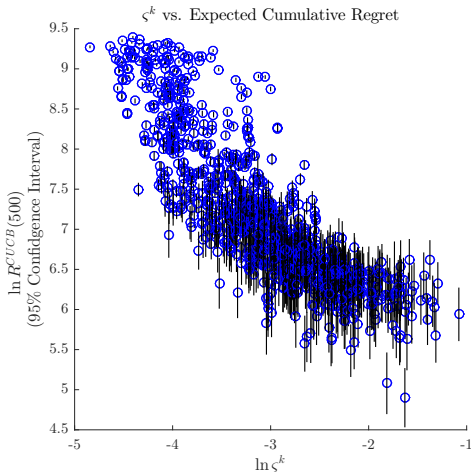

Fig. 2: Simulation results of expected cumulative regret as a function of ς^k for nodes in ER graphs with $\rho = \frac{\ln(10)}{10}$, P as in (3), and $\kappa = \frac{d_{\max}}{d_{\max}-1}$.

REFERENCES

[1] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

[2] V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: I.I.D. rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, Nov 1987.

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

[4] B. Bollobás. *Random Graphs*. Springer, 1998.

[5] P. Braca, S. Marano, and V. Matta. Enforcing consensus while monitoring the environment in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(7):3375–3380, 2008.

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

[7] F. Bullo, J. Cortés, and S. Martínez. *Distributed Control of Robotic Networks*. Princeton University Press, 2009.

[8] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[9] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

[10] B. Golub and M. O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–149, 2010.

[11] A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003.

[12] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi. Non-Bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

[13] D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[14] S. Kar, H. V. Poor, and S. Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. In *IEEE Conference on Decision and Control and European Control Conference*, pages 1771–1778, Orlando, FL, December 2011.

[15] J. R. Krebs, A. Kacelnik, and P. Taylor. Test of optimal sampling by foraging great tits. *Nature*, 275(5675):27–31, 1978.

[16] L. Lai, H. Jiang, and H. V. Poor. Medium access in cognitive radio networks: A competitive multi-armed bandit framework. In *Asilomar Conference on Signals, Systems and Computers*, pages 98–102, Pacific Grove, CA, October 2008. IEEE.

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

[18] K. Liu and Q. Zhao. Extended UCB policy for multi-armed bandit with light-tailed reward distributions. *arXiv:1112.1768*, December 2011.

[19] W. Mason and D. J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.

[20] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[21] R. Olfati-Saber and R. M. Murray. Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9):1520–1533, 2004.

[22] P. B. Reverdy, V. Srivastava, and N. E. Leonard. Modeling human decision making in generalized Gaussian multiarmed bandits. *Proceedings of the IEEE*, 102(4):544–571, 2014.

[23] V. Srivastava and N. E. Leonard. Collective decision-making in ideal networks: The speed-accuracy trade-off. *IEEE Transactions on Control of Network Systems*, 1(1):121–132, 2014.

[24] V. Srivastava, P. Reverdy, and N. E. Leonard. On optimal foraging and multi-armed bandits. In *Allerton Conference on Communication, Control, and Computing*, pages 494–499, Monticello, IL, USA, October 2013.

[25] V. Srivastava, P. Reverdy, and N. E. Leonard. Surveillance in an abruptly changing world via multiarmed bandits. In *IEEE Conference on Decision and Control*, pages 692–697, Los Angeles, CA, December 2014.

[26] J. N. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, November 1984.