Social Imitation in Cooperative Multiarmed Bandits: Partition-based Algorithms with Strictly Local Information

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard

***Abstract*— We study distributed cooperative decision-making in a multi-agent stochastic multi-armed bandit (MAB) problem in which agents are connected through an undirected graph and observe the actions and rewards of their neighbors. We develop a novel policy based on partitions of the communication graph and propose a distributed method for selecting an arbitrary number of leaders and partitions. We analyze this new policy and evaluate its performance using Monte-Carlo simulations.**

I. Introduction

The challenge of cooperative decision-making under uncertainty is a common feature of engineered as well as natural systems. Decision-making under uncertainty hinges on the tradeoff between *exploitation*, where a decision-making agent chooses to maximize its parameter-dependent decision-making objective, and learning those system parameters though *exploration*. Multiarmed bandit (MAB) problems are canonical formulations of this *explore-exploit* tradeoff.

The stochastic MAB problem features a given set of options (arms), each with an associated stochastic reward distribution with unknown mean. An agent chooses one arm at a time, receiving a reward sampled from the associated distribution. The agent's goal is to maximize the cumulative expected reward over time. To do so the agent must balance learning the identity of the best arm (exploration) and choosing the arm with highest expected reward (exploitation).

MAB problems have a long and rich history, with applications in diverse scientific fields such as control and robotics [1], ecology [2, 3], psychology [4], and communications [5, 6]. In their seminal work, Lai and Robbins [7] established a lower bound on the expected number of times a sub-optimal arm must be selected by an optimal policy in a frequentist setting. Many algorithms have been designed that achieve the lower bound in [7] uniformly in time (see [8] for a survey). One such algorithm is the Upper Confidence Bound (UCB) policy by Auer *et al.* [9].

To date most research on the MAB problem has focused on policies for a single decision-maker. However, the rising importance of networked systems warrants the development of scalable and distributed algorithms for groups of decision-makers facing MAB problems. In this paper we build upon classical single-agent bandit policies [9] and extend them to the multi-agent setting.

The multi-agent MAB problem has been investigated in several contexts. Anantharam *et al.* [10] studied the case of multiple centralized players. Several researchers [6, 11, 12], inspired by the radio network spectrum access problem, have considered the case of decentralized multi-agent MAB with only indirect communication between agents through conflicts during arm selection. In [13–16] *cooperative* multi-agent MAB problems have been studied, where agents communicate over a network graph to maximize the cumulative reward of the group and do not interfere with one another. In our previous work [13, 14], agents communicate their estimates of mean rewards to their neighbors. In the present paper we draw inspiration from [15], in which agents communicate only decisions and rewards to neighbors rather than estimates of mean rewards. The Follow Your Leader (FYL) algorithm of [15] partitions the network into "leaders," which use the UCB1 algorithm, and "copiers," which imitate the actions of an adjacent leader. The FYL algorithm selects how many and which agents will be leaders using a dominating set of the graph; these must be computed prior to runtime. [17] studied the multi-agent MAB problem in which agents can communicate decisions and rewards with their neighbors, but they cannot imitate each other. Specifically, they considered the case of side-observations, where sampling a given arm reveals the rewards that other agents would have received had they selected the same arm.

In this paper we study the distributed cooperative MAB problem with local communication. Here agents are faced with a stochastic MAB problem and they can access the decisions and rewards of their neighbors as defined by a static communication graph. We introduce and analyze a partition-based distributed decision-making algorithm, where only one agent in each partition, a so-called leader, makes independent decisions based on its local information. The other agents in the partition, the so-called followers, imitate the decisions of the leader in the partition, either directly if the leader is a neighbor, or, otherwise, indirectly by imitating a neighbor along a path to the leader. The ability to imitate a neighbor who is itself imitating another neighbor is the key difference in the problem formulation between our work and [15]. In [15] agents can only imitate a neighbor that is a leader, and our relaxation of this constraint leads to a richer set of possible strategies and analysis.

Our problem setup is motivated by the phenomenon of social imitation, which is often encountered in natural systems [18–20]. We also define and analyze a distributed

This research has been supported in part by ONR grant N00014- 09-1-1074, ARO grant W911NG-11-1-0385, and NSF grant IIS-1734272.

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

V. Srivastava is with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA, vaibhav@egr.msu.edu

algorithm for partitioning the network and choosing leaders, where the number of leaders can be prescribed. We demonstrate that our algorithms obtain order-optimal performance for the group. We explore our results using several graphs through simulation and analytic performance bounds.

The paper is organized as follows. In Section II we introduce the cooperative MAB problem and give a lower bound on the number of times a suboptimal arm will be chosen by the network. In Section III we propose and analyze the UCB-Partition and token-passing partition generation algorithms. We analyze the UCB-Partition algorithm using Monte-Carlo simulations in Section IV and conclude in Section V.

II. Background and Problem Formulation

A. Cooperative MAB Problem with Local Communication

Consider an MAB problem with N arms and M agents. The reward associated to arm $i \in \{1, \dots, N\}$ is a bounded random variable in $[0, 1]$ with unknown mean m_i. The communication among agents is modeled by a connected, unweighted, undirected network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, $|\mathcal{V}| = M$. Let $\mathcal{N}(k)$ denote the set of neighbors of each agent $k \in \mathcal{V}$.

We assume that the agents can be classified into leaders and followers. We assume that every follower is connected to at least one leader through a path in $\mathcal{G}$, and it imitates one such leader, either directly or indirectly through a chain of followers. The set of leaders induces a partitioning of the graph in which every agent in a leader's partition ultimately imitates it. We assume that at each time t each leader has access to the arms chosen and rewards received by its neighbors, while each follower only has access to the arms chosen by its neighbors. We also assume that each agent k knows the degree of its neighbors: $|\mathcal{N}(j)|$ for each $j \in \mathcal{N}(k)$.

Let each agent k choose arm $i^k(t)$ at time $t \in \{1, \dots, T\}$ and receive i.i.d reward $r^k(t)$. The total number of times up to time t that agent k has selected arm i is $n_i^k(t)$ and that agent k and its neighbors have selected arm i is $\bar{n}_i^k(t) = n_i^k(t) + \sum_{j \in \mathcal{N}(k)} n_i^j(t)$. The sequence of rewards received by agent k and its neighbors from arm i is $\{r_{i,s}^k\}_{s \in \{1, \dots, \bar{n}_i^k(t)\}}$. The estimated mean of arm i at time t by agent k given its own and its neighbors' realized rewards is $\bar{\mu}_{i, \bar{n}_i^k(t)}^k = \frac{1}{\bar{n}_i^k(t)} \sum_{s=1}^{\bar{n}_i^k(t)} r_{i,s}^k$. Each leader ℓ can compute $\bar{n}_i^\ell(t)$ and $\bar{\mu}_i^\ell(t)$.

The objective of this paper is to design a distributed algorithm for partitioning the graph $\mathcal{G}$, assigning a leader to each partition such that every other agent in the partition imitates it, and determining a sequential decision-making policy for the leaders and the followers such that efficient group performance is achieved. Alternatively, a set of leaders may be assigned and the distributed algorithm should select the set of followers and, consequently, the graph partitioning.

The regret of agent k at each time t conditioned on the choice $i^k(t)$ is defined by $R^k(t) = m_{i^*} - m_{i^k(t)} \equiv \Delta_{i^k(t)}$, where $m_{i^*} = \max_{i \in \{1, \dots, N\}} m_i$. We characterize group performance in terms of the total expected cumulative regret defined by $\sum_{k=1}^M \sum_{t=1}^T \mathbb{E}[R^k(t)] = \sum_{k=1}^M \sum_{i=1}^N \Delta_i \mathbb{E}[n_i^k(T)]$, where T is the horizon length.

In this paper, we restrict our attention to policies in which the leaders follow the UCB algorithm with the estimates of the mean rewards that are computed using the rewards received by the leader and its neighbors and the followers imitate one of their neighbors.

B. Lower Bounds on Expected Cumulative Regret

It follows from [10] that the expected number of times a suboptimal arm i is selected by a fusion center with access to the reward for each agent is lower bounded by

$$\sum_{k=1}^M \mathbb{E}[n_i^k(T)] \geq \left(\frac{1}{\mathcal{D}(p_i || p_{i^*})} + o(1) \right) \ln T. \qquad (1)$$

where $\mathcal{D}(p||p_*)$ is the Kullback-Leibler divergence between the probability density p_i and p_{i^*}.

In the following we develop a policy that achieves provable performance within a constant factor of the above bound.

III. Partition Based Multi-player MAB

In this section we describe and prove upper bounds on the performance of partition-based multi-player MAB. We introduce several definitions, describe the problem and the UCB-Partition algorithm.We then establish bounds on performance of the UCB-Partition algorithm.

A. Definitions and Notation

We now introduce several definitions that formalize the leader/follower relationships inherent in the UCB-Partition algorithm. We will use these formal definitions to prove an upper bound on the cumulative expected regret of the algorithm. Fig. 1 illustrates these definitions with an example.

Let $\mathcal{G}_{\text{ldr}} = (\mathcal{V}_{\text{ldr}}, \mathcal{E}_{\text{ldr}})$ be a directed graph such that $\mathcal{V}_{\text{ldr}} = \mathcal{V}$ and

$$\mathcal{E}_{\text{ldr}} = \{(k, j) \in \mathcal{E} \mid k \ \text{ can imitate } \ j\}.$$

$\mathcal{G}_{\text{ldr}}$ encodes all possible variations of followers in the UCB-Partition algorithm: a directed edge in $\mathcal{G}_{\text{ldr}}$ indicates that the agent at the tail may follow the agent at the head. $\mathcal{G}_{\text{ldr}}$ can therefore be used to enforce operation constraints on who can or cannot follow others.

We now define the set of all leaders by $\mathcal{L}$ and, in the following, we will denote the i-th element of $\mathcal{L}$ by ℓ_i. We also define $\mathcal{G}_{\text{ldr}}^{\text{rlz}} = (\mathcal{V}_{\text{ldr}}^{\text{rlz}}, \mathcal{E}_{\text{ldr}}^{\text{rlz}})$ such that $V_{\text{ldr}}^{\text{rlz}} = \mathcal{V}$ and

$$\mathcal{E}_{\text{ldr}}^{\text{rlz}} = \{(k, j) \subset \mathcal{E}_{\text{ldr}} \mid \nexists\, m \neq j \in \mathcal{V}_{\text{ldr}}, (k, m) \in \mathcal{E}_{\text{ldr}}^{\text{rlz}}, k \notin \mathcal{L}\}.$$

Note that $\mathcal{E}_{\text{ldr}}^{\text{rlz}}$ is defined recursively and restricts follower agent k to imitate at most one leader or follower. $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ thus encodes a possible realization of follower and leader combinations when using the UCB-Partition algorithm: a directed edge in $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ indicates that the agent at the tail will follow the agent at the head, and agents with no outgoing edges are leaders.

For a given realization $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$, we define the set of followers of leader ℓ_j as

$$\mathcal{F}_j^{\text{rlz}} = \{\ell_j\} \cup \{k \in \mathcal{V}_{\text{ldr}}^{\text{rlz}} \mid \exists \text{ directed path from } k \text{ to } \ell_j \text{ in } \mathcal{G}_{\text{ldr}}^{\text{rlz}}\}.$$

and the set of direct followers of leader ℓ_j as

$$\mathcal{F}_{j\text{-direct}}^{\text{rlz}} = \{\ell_j\} \cup \{k \in \mathcal{V}_{\text{ldr}}^{\text{rlz}} \mid (k, \ell_j) \in \mathcal{E}_{\text{ldr}}^{\text{rlz}}\}.$$

The sets $\mathcal{F}_j^{\text{rlz}}$, $j \in \{1, \ldots, |\mathcal{L}|\}$, define a partitioning of $\mathcal{G}$, where each partition contains one leader that every follower in the partition ultimately imitates, and $\mathcal{F}_{j\text{-direct}}^{\text{rlz}} \subseteq \mathcal{F}_j^{\text{rlz}}$. Fig. 1 illustrates these subgraphs for a given $\mathcal{G}$ and three example realizations $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$. We denote the length of the longest path present in $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ within the partition defined by $\mathcal{F}_j^{\text{rlz}}$ as $\text{diam}_j^{\text{rlz}}$.

Every realization of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ induces a partitioning of the graph $\mathcal{G}$. Equivalently, for any partitioning of the graph $\mathcal{G}$, we can choose a leader in each partition and construct $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$. The following analysis holds for any realization $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ and is oblivious to how $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ is constructed, i.e., whether it is induced by a given set of leaders or if it is induced by a given partitioning.

The set $\mathcal{F}_{j\text{-direct}}^{\text{rlz}}$ is used later in this paper to bound the expected cumulative regret of the UCB-Partition algorithm for both a single partition and multiple partitions of $\mathcal{G}$.

B. UCB-Network and Follow Your Leader Algorithms

The UCB-Network and Follow Your Leader (FYL) algorithms are defined in [15]. The UCB-network algorithm is equivalent to setting $\mathcal{L} = \mathcal{V}$, making every agent a leader that can access rewards of its neighbors. The UCB-Network algorithm is thus easily distributed, but it does not allow for any agent to imitate.

In the FYL algorithm the leaders $\mathcal{L}$ are defined as a dominating set[1] of $\mathcal{G}$, and the followers of ℓ_j are composed of a subset of the neighbors of ℓ_j. In the FYL algorithm the best performance is achieved when $\mathcal{L}$ is defined as the minimal dominating set. An example of leader selection corresponding to the minimal dominating set is shown in Panel C in Fig. 1.

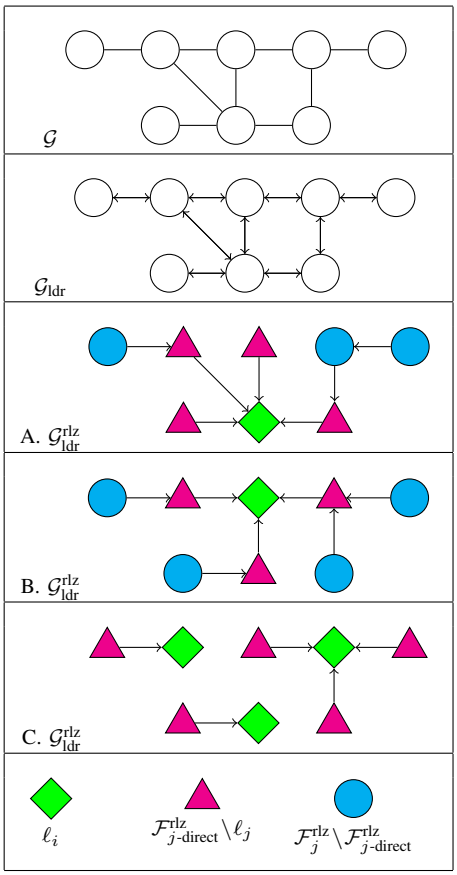

Fig. 1: Example of a communication graph $\mathcal{G}$ and a $\mathcal{G}_{\text{ldr}}$ that allows for any agent to imitate any neighbor in $\mathcal{G}$. Panels A and B show two possible realizations of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ for the case of one leader. Panel C demonstrates a realization for three leaders. The selection of each agent's role, which defines $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$, can be driven by design constraints or optimized to minimize the upper bounds on performance. Note that even if two agents are not connected in $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ they can still share sample information if connected in $\mathcal{G}$.

C. UCB-Partition Algorithm

First, we define

$$Q_i^k(t, \bar{n}_i^k(t)) = \bar{\mu}_{i,\bar{n}_i^k(t)}^k + \sqrt{\frac{2\ln(t)}{\bar{n}_i^k(t)}}. \tag{2}$$

The UCB-Partition algorithm is as follows:

(i) Initialization phase: Every leader $j \in \mathcal{L}$ chooses each arm once, and each follower $k \in \mathcal{F}_j^{\text{rlz}}$ chooses randomly for the first timestep.

(ii) Each leader $j \in \mathcal{L}$ selects arm with highest $Q_i^j(t, \bar{n}_i^j(t))$, and each follower k selects the arm selected by agent $\{m \in \mathcal{V}_{\text{ldr}}^{\text{rlz}} \mid (k,m) \in \mathcal{E}_{\text{ldr}}^{\text{rlz}}\}$ at the previous timestep.

D. Expected Cumulative Regret of UCB-Partition

Here we establish an upper bound on the cumulative expected regret of the UCB-Partition algorithm.

Theorem 1. *For the UCB-Partition algorithm with definitions given in Section III-A the following bounds hold for $i \neq i^*$ and given a $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ with $|\mathcal{L}| = 1$:*

$$\sum_{k=1}^{M} \mathbb{E}\left[n_i^k(T)\right] \leq \frac{8\ln(T)}{\Delta_i} \cdot \frac{|\mathcal{F}_1^{\text{rlz}}|}{|\mathcal{F}_{1\text{-direct}}^{\text{rlz}}|} + M^3\left(1 + \frac{\pi^2}{3}\right) + (|\mathcal{F}_1^{\text{rlz}}| - 1)\text{diam}_1^{\text{rlz}},$$

where $|\mathcal{F}_1^{\text{rlz}}| = M$ and $|\mathcal{F}_{1\text{-direct}}^{\text{rlz}}| = |\mathcal{N}(\ell_1)| + 1$.

Proof. We start by noticing that

$$\sum_{k=1}^{M} n_i^k(T) \leq |\mathcal{F}_1^{\text{rlz}}|\, n_i^{\ell_1}(T) + (|\mathcal{F}_1^{\text{rlz}}| - 1)\text{diam}_1^{\text{rlz}} \tag{3}$$

$$= |\mathcal{F}_1^{\text{rlz}}| \sum_{t=1}^{T} \mathbb{1}\left\{i^{\ell_1}(t) = i\right\} + (|\mathcal{F}_1^{\text{rlz}}| - 1)\text{diam}_1^{\text{rlz}}$$

$$\leq (|\mathcal{F}_1^{\text{rlz}}| - 1)\text{diam}_1^{\text{rlz}} + |\mathcal{F}_1^{\text{rlz}}| \left[\frac{A}{|\mathcal{F}_{1\text{-direct}}^{\text{rlz}}|} + \sum_{t=1}^{T} \mathbb{1}\left\{i^{\ell_1}(t) = i,\ n_i^{\ell_1}(t) > \frac{A}{|\mathcal{F}_{1\text{-direct}}^{\text{rlz}}|}\right\}\right], \tag{4}$$

where A is a constant that will be chosen later and the $(|\mathcal{F}_1^{\text{rlz}}|-1)\text{diam}_1^{\text{rlz}}$ term in (3) follows because every follower will not necessarily be copying their leader until the leader's

[1] A subset of nodes of a graph is called a *dominating set* if for every node not in the dominating set, there exists an adjacent node that belongs to the dominating set. The smallest dominating set is called the minimal dominating set.

choices propagate through the network. We now bound the second part of (4) using techniques from [9].

$$\sum_{t=1}^{T} \mathbb{1}\left\{ i^{\ell_1}(t) = i,\ n_i^{\ell_1}(t) > \frac{A}{|\mathcal{F}^{\text{rlz}}_{\text{1-direct}}|} \right\}$$
$$\leq \sum_{t=1}^{T} \mathbb{1}\left\{ Q_{i^*}^{\ell_1}(t, \bar{n}_{i^*}^{\ell_1}(t)) < Q_i^{\ell_1}(t, \bar{n}_i^{\ell_1}(t)), n_i^{\ell_1}(t) > \frac{A}{|\mathcal{F}^{\text{rlz}}_{\text{1-direct}}|} \right\}$$
$$\leq \sum_{t=1}^{T} \mathbb{1}\left\{ Q_{i^*}^{\ell_1}(t, \bar{n}_{i^*}^{\ell_1}(t)) < Q_i^{\ell_1}(t, \bar{n}_i^{\ell_1}(t)), \bar{n}_i^{\ell_1}(t) > A - M \right\} \tag{5}$$
$$\leq \sum_{t=1}^{T} \mathbb{1}\left\{ \min_{1<a<(|\mathcal{N}(\ell_1)|+1)t} Q_{i^*}^{\ell_1}(t,a) < \max_{A-M<b<(|\mathcal{N}(\ell_1)|+1)t} Q_i^{\ell_1}(t,b) \right\}$$
$$\leq \sum_{t=1}^{T} \sum_{a=1}^{(|\mathcal{N}(\ell_1)|+1)t} \sum_{b=A-M}^{(|\mathcal{N}(\ell_1)|+1)t} \mathbb{1}\left\{ Q_{i^*}^{\ell_1}(t,a) < Q_i^{\ell_1}(t,b) \right\}$$

where (5) follows because the direct followers of the leader choose $i^{\ell_1}(t)$ at time $t+1$. In the spirit of [9], if $\mathbb{1}\left\{ Q_{i^*}^{\ell_1}(t,a) < Q_i^{\ell_1}(t,b) \right\}$ holds then at least one of the following must hold:

$$\bar{\mu}_{i^*,a} \leq m_{i^*} - \sqrt{\frac{2\ln(t)}{a}} \tag{6}$$
$$\bar{\mu}_{i,b} \geq m_i + \sqrt{\frac{2\ln(t)}{b}} \tag{7}$$
$$m_{i^*} < m_i + \sqrt{\frac{8\ln(t)}{b}} \tag{8}$$

As in [9], we bound (6) and (7) using Chernoff-Hoeffding bounds as

$$\mathbb{P}\left(\bar{\mu}_{i^*,a} \leq m_{i^*} - \sqrt{\frac{2\ln(t)}{a}} \right) \leq t^{-4}, \text{ and}$$
$$\mathbb{P}\left(\bar{\mu}_{i,b} \geq m_i + \sqrt{\frac{2\ln(t)}{b}} \right) \leq t^{-4}.$$

Setting $A = M + \frac{8\ln(t)}{\Delta_i^2}$, we see that (8) never holds. Thus,

$$|\mathcal{F}_1^{\text{rlz}}| \sum_{t=1}^{T} \mathbb{1}\left\{ i^{\ell_1}(t) = i, n_i^{\ell_1}(t) > \frac{A}{|\mathcal{F}^{\text{rlz}}_{\text{1-direct}}|} \right\}$$
$$\leq |\mathcal{F}_1^{\text{rlz}}| \sum_{t=1}^{T} \sum_{a=0}^{(|\mathcal{N}(\ell_1)|+1)t} \sum_{b=A-M}^{(|\mathcal{N}(\ell_1)|+1)t} \frac{2}{t^4}$$
$$\leq |\mathcal{F}_1^{\text{rlz}}| \sum_{t=1}^{T} \frac{2}{t^2} (|\mathcal{N}(\ell_1)|+1)^2$$
$$\leq |\mathcal{F}_1^{\text{rlz}}| (|\mathcal{N}(\ell_1)|+1)^2 (1 + \frac{\pi^2}{3}) \leq M^3 (1 + \frac{\pi^2}{3}),$$

which completes the proof. □

Corollary 1. *For the UCB-Partition algorithm with definitions given in Section III-A the following bounds hold for $i \neq i^*$ and any given $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$ with a generic set of leaders $\mathcal{L}$:*

$$\sum_{k=1}^{M} \mathbb{E}\left[n_i^k(T) \right] \leq \frac{8\ln(T)}{\Delta_i} \sum_{j \in \mathcal{L}} \frac{|\mathcal{F}_j^{\text{rlz}}|}{|\mathcal{F}^{\text{rlz}}_{j\text{-direct}}|} + M^3(1 + \frac{\pi^2}{3}) + \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}}.$$

Proof. Similar to the proof of Theorem 1, we note that

$$\sum_{k=1}^{M} n_i^k(T) \leq \sum_{j \in \mathcal{L}} |\mathcal{F}_j^{\text{rlz}}|\, n_i^{\ell_j}(T) + \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}}$$
$$= \sum_{j \in \mathcal{L}} |\mathcal{F}_j^{\text{rlz}}| \sum_{t=1}^{T} \mathbb{1}\left\{ i^{\ell_j}(t) = i \right\} + \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}}$$
$$\leq \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}} + \sum_{j \in \mathcal{L}} |\mathcal{F}_j^{\text{rlz}}| \left[\frac{A}{|\mathcal{F}^{\text{rlz}}_{j\text{-direct}}|} + \sum_{t=1}^{T} \mathbb{1}\left\{ i^{\ell_j}(t) = i,\ n_i^{\ell_j}(t) > \frac{A}{|\mathcal{F}^{\text{rlz}}_{j\text{-direct}}|} \right\} \right]$$
$$\leq \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}} + A \sum_{j \in \mathcal{L}} \frac{|\mathcal{F}_j^{\text{rlz}}|}{|\mathcal{F}^{\text{rlz}}_{j\text{-direct}}|} + M^2(1 + \frac{\pi^2}{3}) \sum_{j \in \mathcal{L}} |\mathcal{F}_j^{\text{rlz}}| \tag{9}$$
$$= \sum_{j \in \mathcal{L}} (|\mathcal{F}_j^{\text{rlz}}| - 1)\text{diam}_j^{\text{rlz}} + A \sum_{j \in \mathcal{L}} \frac{|\mathcal{F}_j^{\text{rlz}}|}{|\mathcal{F}^{\text{rlz}}_{j\text{-direct}}|} + M^3(1 + \frac{\pi^2}{3}),$$

where (9) follows from Theorem 1, completing the proof. □

E. Distributed Partition-Based Multi-agent MAB using Token Passing

The UCB-Partition algorithm and the associated bounds provide performance guarantees for a given $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$, which by definition defines $|\mathcal{L}|$ partitions of $\mathcal{G}$ and the leader-follower assignments. In this section we present a distributed method for choosing $|\mathcal{L}|$ leaders and partitions, which in turn, with follower assignments, gives $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$.

This method is comprised of two parts: leader identification and partition generation. The goal of the leader identification step is for each agent to construct, in a distributed fashion, a list of tuples of size $|\mathcal{L}|$, where each tuple contains the identify and degree of agents with top $|\mathcal{L}|$ degree. Let each agent k have a unique identity number v, and let each agent know their own degree, $|\mathcal{N}(k)|$, in $\mathcal{G}$, and identity of their neighbors. Each agent initially constructs a list of size $|\mathcal{L}|$ with only one entry: the agent's identity and degree, and the other entries are empty. Then, each agent exchanges this list with each of their neighbors and combines their own list with those received to create a new list of agents with the top $|\mathcal{L}|$ degrees in the lists (in case of ties, the agent with lower identity is selected). Each agent then repeats this process with their new list, and the procedure converges in number of timesteps equal to at most two times the diameter of graph $\mathcal{G}$ plus one.

To accomplish partition generation each agent represented in the final list identifies itself as a leader. Followers then recursively choose an agent to imitate. First, the agents that are adjacent to leader(s) commit to imitating a leader, and transmit a committed signal to their neighbors. Subsequently the uncommitted neighbors may choose to imitate one of the committed agents, until all agents are committed.

This procedure defines $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$, and the performance bounds established in Section III-D hold. In future work will we rigorously show that this strategy converges to a valid partition for a connected communication graph $\mathcal{G}$.

Fig. 2 demonstrates three examples of leader selection and follower assignment using this method for 1, 3, and 5 leaders. Note that $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ is not unique for the 3 and 5 leader cases as some followers must choose arbitrarily between two or more options. The identity number v is omitted for clarity, but it is used to break ties when choosing 5 leaders.

IV. Numerical Illustrations

In this section we compare the behavior and performance of the UCB-Partition algorithm with the algorithms in [15]. We show that the UCB-Partition algorithm performs well over a variety of graph structures and offers performance advantages over related algorithms.

All simulations are conducted with a 2-armed bandit using rewards drawn from a Bernoulli distribution with $m = [0.5,\ 0.7]$ and $T = 10^3$ or $T = 10^4$. In Figs. 3 and 4 we show cumulative regret of the group over time for different graph structures as given in Figs. 1 and 2, respectively. The cumulative regret in our simulations are computed by averaging over 8000 Monte-Carlo runs using the UCB-Partition and UCB-Network algorithms, as well as for the case with no communication between agents.

Example 1 (***Regret for Small Graphs***)**.** Fig. 3 shows group cumulative expected regret for $\mathcal{G}$ and the three versions of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ as given in Fig. 1. The UCB-Partition greatly improves performance over the UCB-Network algorithm, demonstrating the benefits of imitation. Additionally, version C of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ is a minimal dominating set partition for the FYL algorithm, so the better performance of UCB-Partition A over C here shows the advantage of the UCB-Partition algorithm over the FYL algorithm when used with suitable leaders. Finally, the the better performance of UCB-Partition A over B demonstrates the benefit of selecting agents with higher degree to be leaders.

Example 2 (***Regret for Large Graphs using Token Passing***)**.** Fig. 4 shows group cumulative expected regret for $\mathcal{G}$ and the three versions of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ corresponding to 1, 3, and 5 leaders as given in Fig. 2. Here, increasing the number of leaders results in a small increase in group cumulative regret for large T, which is also reflected in the performance bounds, a phenomenon we discuss in Example 3. As in Example 1, the UCB-Partition significantly improves performance over the UCB-Network algorithm.

Example 3 (***Time Dependency of Optimal Leader Selection***)**.** Fig. 5 compares the relative performance of the UCB-

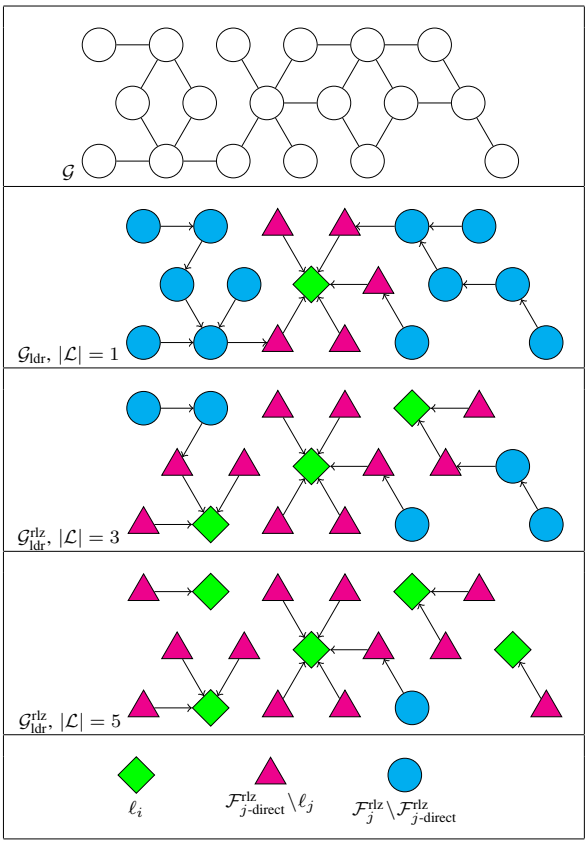

Fig. 2: Example of a large communication graph $\mathcal{G}$ and a $\mathcal{G}_{\text{ldr}}$ (not shown) that allows for any agent to follow any neighbor in $\mathcal{G}$. Three panels show three possible realizations of $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ with 1, 3, and 5 leaders, respectively, where the leaders and followers are selected using the token passing method described in III-E.

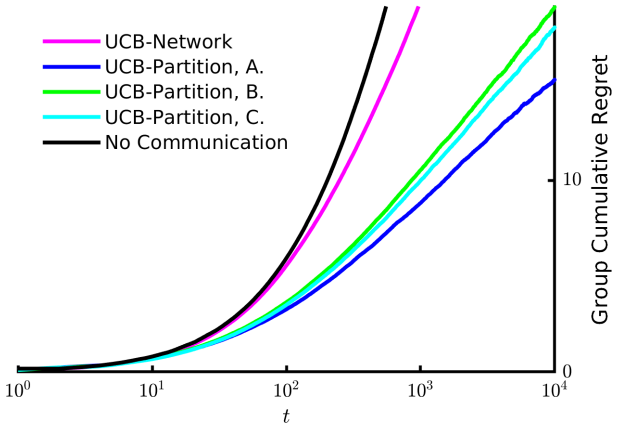

Fig. 3: Simulation results of expected cumulative regret for the UCB-Network and UCB-Partition algorithms using $\mathcal{G}$ and $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ as given in Fig. 1.

Partition algorithm using the 1, 3, and 5 leader realizations $\mathcal{G}_{\text{ldr}}^{\text{rlz}}$ in Fig. 2 at each timestep t for $T = 10^3$. Early on, the 3 leader network outperforms the 1 leader network, but as t grows the 1 leader network begins to perform the best, a trend which can be seen continuing in Fig. 4.

This is expected as bounds expressed in Theorem 1 and Corollary 1 indicate that as $T \to \infty$ the lowest regret will be obtained when the agent or agents with the highest overall degree in $\mathcal{G}$ are the only leaders. This result is indicated by the domination of the logarithmic term in the bound and is

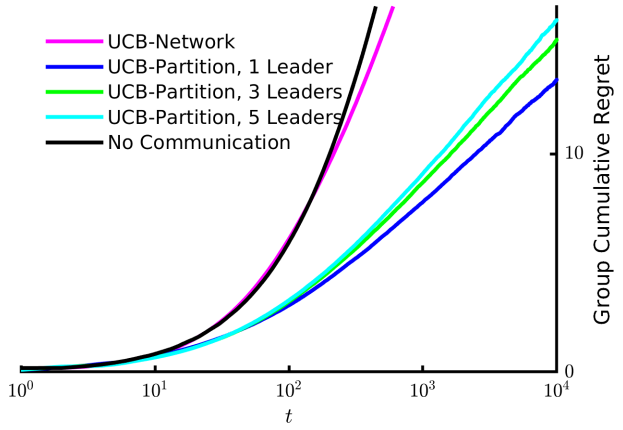

Fig. 4: Simulation results of expected cumulative regret for the UCB-Network and UCB-Partition algorithms using $\mathcal{G}$ and $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$ as given in Fig. 2.

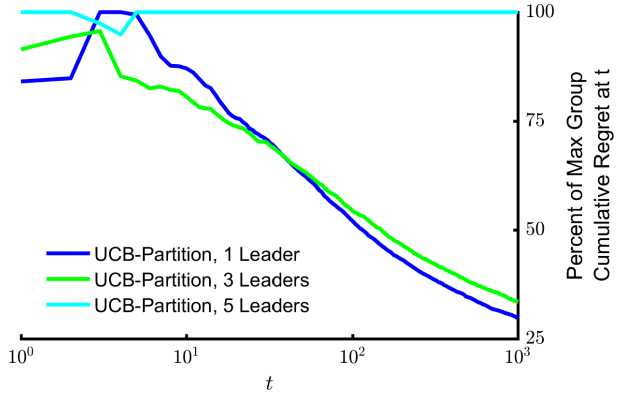

Fig. 5: Simulation results of the expected cumulative group regret of the 1, 3, and 5 leader $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$'s in Fig. 2 as a percentage of the algorithm with highest regret at each time t. Lower percentage values indicate lower regret.

intuitive, as over large timescales it is beneficial to wait and imitate, through one's neighbors, a leader with the highest possible number of available samples.

However, for $T < \infty$ the $\mathcal{G}^{\text{rlz}}_{\text{ldr}}$-dependent constant terms in Theorem 1 and Corollary 1 can be significant relative to the logarithmic term, and having more leaders may be advantageous as this tends to reduce $|\mathcal{F}^{\text{rlz}}_j|$ and $\text{diam}^{\text{rlz}}_j$. Additionally, this factor would be particularly important for non-stationary MAB problems in social settings, where the mean rewards from the arms can change in time.

These results suggest that selecting the optimal leaders or optimal number of leaders is a function not only of $\mathcal{G}$ but also of the time horizon T. We intend to explore this trade-off in future work.

V. Final Remarks

In this paper we investigated cooperative decision-making in networks using the cooperative multi-agent MAB problem. We developed the UCB-Partition algorithm and proved bounds on its performance. Additionally, we developed a distributed policy that utilizes token-passing, does not require knowledge of the full communication graph, and can select an arbitrary number of leaders for use with the UCB-Partition algorithm. We demonstrated the utility of the UCB-Partition using several different examples of communication graphs and explored the time dependency of selecting the optimal number of leaders.

Future research directions include tightening the performance bounds of the UCB-Partition algorithm and constructing algorithms for leader selection as a function of time. Additionally, alternate metrics for choosing when to imitate or lead may offer performance benefits, and a tight lower bound on expected regret as a function of the local communication graph remains an open problem. It would also be interesting to use our results in studies of human or animal networks facing problems described by the MAB.

References

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

[2] J. R. Krebs, A. Kacelnik, and P. Taylor, "Test of optimal sampling by foraging great tits," *Nature*, vol. 275, no. 5675, pp. 27–31, 1978.

[3] V. Srivastava, P. Reverdy, and N. E. Leonard, "On optimal foraging and multi-armed bandits," in *Allerton Conference on Communication, Control, and Computing*, Oct. 2013, pp. 494–499.

[4] P. B. Reverdy, V. Srivastava, and N. E. Leonard, "Modeling human decision making in generalized Gaussian multiarmed bandits," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 544–571, 2014.

[5] 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*. Pacific Grove, CA: IEEE, Oct. 2008, pp. 98–102.

[6] 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*, vol. 29, no. 4, pp. 731–745, 2011.

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

[8] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

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

[10] 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*, vol. 32, no. 11, pp. 968–976, Nov 1987.

[11] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[12] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Transactions on Signal Processing*, vol. 62, no. 23, pp. 6184–6193, 2014.

[13] P. Landgren, V. Srivastava, and N. E. Leonard, "On distributed cooperative decision-making in multiarmed bandits," in *European Control Conference*, Aalborg, Denmark, Jun. 2016, pp. 243–248.

[14] ——, "Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms," in *IEEE Conference on Decision and Control*, Las Vegas, NV, USA, Dec. 2016, pp. 167–172.

[15] R. K. Kolla, K. Jagannathan, and A. Gopalan, "Collaborative learning of stochastic bandits over a social network," in *Allerton Conference on Communication, Control, and Computing*, Sept 2016, pp. 1228–1235.

[16] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

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

[18] L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland, "Why copy others? insights from the social learning strategies tournament," *Science*, vol. 328, no. 5975, pp. 208–213, 2010.

[19] K. H. Schlag, "Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits," *Journal of Economic Theory*, vol. 78, no. 1, pp. 130–156, 1998.

[20] U. Toelch, M. J. Bruce, M. T. H. Meeus, and S. M. Reader, "Humans copy rapidly increasing choices in a multiarmed bandit problem," *Evolution and Human Behavior*, vol. 31, no. 5, pp. 326–333, 2010.