Sequential Learning for Optimal Monitoring of Multi-channel Wireless Networks

Pallavi Arora
Department of Computer Science
University of Houston
Houston, TX 77204, USA
E-mail: *palpal@cs.uh.edu*

Csaba Szepesvári
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada
szepesva@ualberta.ca

Rong Zheng
Department of Computer Science
University of Houston
Houston, Tx 77204, USA
rzheng@cs.uh.edu

***Abstract*—We consider the problem of optimally assigning p sniffers to K channels to monitor the transmission activities in a multi-channel wireless network. The activity of users is initially unknown to the sniffers and is to be learned along with channel assignment decisions while maximizing the benifits of this assignment, resulting in the fundamental trade-off between exploration versus exploitation. We formulate it as the linear partial monitoring problem, a super-class of multi-armed bandits. As the number of arms (sniffer-channel assignments) is exponential, novel techniques are called for, to allow efficient learning. We use the linear bandit model to capture the dependency amongst the arms and develop two policies that take advantage of this dependency. Both policies enjoy logarithmic regret bound of time-slots with a term that is sub-linear in the number of arms.**

I. Introduction

Deployment and management of wireless devices and networks are often hampered by the poor visibility of PHY and MAC characteristics, and complex interactions at various layers of the protocol stack both inside a managed network and across multiple administrative domains. The "can you hear me" Verizon wireless TV commercial is a vivid demonstration of the shortage of real-time knowledge that cellular providers have regarding the condition of operational networks. Accurate and timely estimates of network conditions and performance characteristics can yield to better performance in a number of applications, including the following:

- *Network resource management.:* Wireless service providers and network administrators need to determine the coverage of their own networks and make critical decisions such as dimensioning and allocation of network resources.
- *Wireless advisory.:* Individual devices can better adapt their operational parameters (e.g., channels, sub-carriers, hopping sequences, transmission power levels, etc.) for co-existence and better performance.
- *Trouble shooting and diagnosis.:* Availability of cross-layer information of the operational network can help network administrators to determine the root causes of service outage or performance degradation as well as identify malicious behavior and intrusion.

Passive monitoring is a technique where a dedicated set of hardware devices, called *sniffers*, are used to monitor activities in wireless networks. These devices capture transmissions of wireless devices or activities of interference sources in their vicinity, and store packet level or PHY layer information in trace files, which can be analyzed distributively or at a central location. A canonical monitoring application has three components: 1) sniffer hardware, 2) sniffer coordinator and data collector, and 3) data processor and miner.

Since most, if not all infrastructure networks utilize multiple contiguous or non-contiguous channels or bands [1], an important issue is to determine which set of frequency bands each sniffer should operate on to maximize the total amount of information gathered. This is called the *sniffer-channel assignment* problem or *channel assignment* problem for short. It is a challenging problem for two reasons. First, monitoring resources are limited, and thus it is infeasible to monitor all channels at all locations at all times. Second, intelligent channel assignment requires the knowledge of usage patterns, i.e., the likelihood of occurrence of interesting events. These are of course not known *a priori*. An interesting trade-off arises between assigning sniffers to channels known to be the busiest based on current knowledge, versus exploring channels that are undersampled. Sniffer-channel assignment with no prior knowledge of user activity is closely related to the multi-armed bandit problem (MAB) [1]. In a MAB, a gambler must decide which arm of N non-identical slot machines to play in a sequence of trials so as to maximize his payoff. In the sniffer-channel assignment problem, each of the p sniffers must be assigned to one of the K non-identical channels to monitor so as to maximize the total information gathered. The number of choices (arms) available in a round is thus $N = K^p$. In this work we assume that the payoff is proportional to the number of distinct users detected. For simplicity, we assume that a user's activity in a given channel can be described with a sequence of IID Bernoulli random variables. However, as opposed to the standard MAB problem, the observation upon a single assignment is not only the reward associated with the assignment, but also the activity patterns observed at each monitored channel. Note that the observed pattern may have correlated components, e.g. when two sniffers observe the transmission of the same set of users.

A policy for sniffer-channel assignment determines at any point in time the assignment to be chosen based on past

[1] A channel can be a single frequency band, a code in a CDMA system, or a hopping sequence in a frequency-hopping system.

information. The efficiency of different policies is measured in terms of their associated *regret*, which is defined as the difference between the expected payoff gained by a "genie" (an unattainable ideal) who always uses the optimal stationary sniffer-channel assignment, and that obtained by the given policy. The regret achieved by a policy can be evaluated in terms of its growth over time and how it scales with respect to the various problem parameters. A naive approach to the channel assignment problem would be to treat each sniffer-channel combination as an arm (action), and learn the statistics of each arm individually. With p sniffers and K channels in the network, the statistics of a total of K^p arm-payoffs needs to be learned. Direct application of known approaches to MAB (e.g., UCB[2], ϵ-greedy[3]) results in a regret bound linear in the number of arms K^p.

In this paper, we formulate the optimal channel assignment problem as a multi-agent multi-arm partial information problem with linearly parameterized payoff. Our proposed policies are centralized and slotted in nature, namely, a fusion center collects the information from each sniffer in each slot and make decision regarding the channel assignment for the next slot. Utilizing the dependency among arms, we reduce the unknown parameter space to $K \cdot 2^p$. We devise two order-optimal policies both having a regret bound that grows logarithmically with respect to time with an associated constant that grows sub-linearly in the number of arms. The key improvement compared to the naive approach comes from the concept of *spanner arms*, i.e, a small collection of arms which provide information about all parameters. The policies and regret bounds are derived for general correlation structures among the sniffers, and remain valid for special cases where the sniffer's observations are identical or independent. In both cases methods exist to identify the best arm to play in each slot which are linear in the number of arms and exponential in the number of sniffers.

The rest of the paper is organized as follows. In Section II, related work on wireless monitoring and sequential learning is summarized. We present the problem formulation in Section III. Details and analysis of the two policies are provided in Section IV and Section V, respectively. Simulation results are presented in Section VI, followed by conclusion and a list of future work in Section VII.

II. Related Work

Wireless monitoring is an active area of research that has received much attention from several perspectives. There has been much work done on wireless monitoring from a *system-level* viewpoint, in an attempt to design complete systems, and address the interactions among the components of such systems [4], [5], [6], [8], [9]. The authors of these works have argued both qualitatively and quantitatively the need for monitoring on the wireless side.

To determine the optimal allocation of monitoring resources to maximize captured information remains, in [10], where Shin and Bagchi consider the selection of monitoring nodes and their associated channels for monitoring wireless mesh networks. The optimal monitoring is formulated as a maximum coverage problem with group budget constraints, which was previously studied by Chekuri and Kumar in [11]. In [12], we introduced a quality of monitoring (QoM) metric defined by the expected number of active users monitored, and investigated the problem of maximizing QoM by judiciously assigning sniffers to channels based on knowledge of user activities in a multi-channel wireless network. Two capture models are considered. The first one, called the *user-centric model* assumes frame-level capturing capability of sniffers such that the activities of different users can be distinguished. The second one, called the *sniffer-centric model* utilizes binary channel information only(active or not) at a sniffer.

The above works assume that certain statistics regarding the users' activity are given [10], [11] or can be inferred [12]. When such statistics are not known a priori, sequential learning is needed. Sequential decision making in presence of uncertainty, faces the fundamental tradeoff between *exploration* and *exploitation*. On one hand, it is desirable to put sniffers to the channels where most activities have been observed and thus more information is likely to be gathered (exploitation). On the other hand, exploring the channels that are under-sampled helps to reduce uncertainty and thus avoid being misled by imprecise information. Such tradeoffs are vividly illustrated by the famous multi-armed bandit problem (MAB). A large volume of work has been been devoted to designing good strategies for variations of the MAB problem and to the understanding of the theoretical limits of such procedures, among which, just to name a few, Lai and Robbin [2] established logarithmic upper and lower bounds for dent stochastic arms with parametric payoff distributions; Agrawal [13] considered a class of sample-mean based policies for the same setting; Auer *et al* analyzed upper confidence bound (UCB) based and ϵ-greedy policies for non-parametric stochastic bandit problems [3]. Recently, bandit problems with linear parameterized payoff are considered in [15], [27]. Regret minimization under partial monitoring is investigated in [16], where the player in a repeated game, instead of observing the action chosen by the opponent in each game round, receives a feedback generated by the combined choice of the two players.

Recognizing the connection between the MAB and spectrum access in cognitive radio networks, Lai *et al.* applied the UCB1 algorithm [3] to single user-channel selection in [17], and later extended it to consider Markovian payoffs and for the case of multiple users in [18]. Liu and Zhao [19] formulated the problem of secondary user channel selection as a decentralized multi-armed bandit problem, and presented a policy that achieves asymptotically logarithmic regret in time. Anandkumar [20] proposed two policies for distributed learning and access with order-optimal cognitive system throughput under self play. In addition to learning the channel availability, the second users also learn the other users' strategies and the number of total users in the system through channel feedback. Existing work applying MAB in the cognitive radio context assumes identical channel view with the exception of Gai *et al* [21]. However, the model considered in this work, in fact makes the implicit assumption that all secondary users are co-located (“if there are multiple users on the channel, then we assume that, due to interference, at most one of the conflicting users gets reward”). Since co-located secondary users likely observe identical primary user activities, a contradiction arises to the claim of “allowing the reward process on the same channel to be different” [21].

In contrast to existing work, we consider a model where sniffers are in general configuration and may observe different sets of users in the same channel. This encompasses models when either sniffers are co-located or when they are sufficiently far apart. The algorithms and analytical bounds devised are directly applicable to the specialized cases. Admittedly, due to its generality, the model suffers from a higher computation and storage complexity. Unfortunately, this unavoidable as a result of the NP-hardness of the nominal resource allocation problem when all statistics are known as shown in Section III.

III. Problem Formulation

Consider p sniffers monitoring user activities in K channels. A user u operates in one of K channels, $c(u) \in \mathcal{K} = \{1, \ldots, K\}$. Let p_u denote the transmission probability of user u. We represent the relationship between users and sniffers using an undirected bi-partite graph $G = (S, U, E)$, where $S = \{1, \ldots, p\}$ is the set of sniffer nodes and U is the set of users. An edge $e = (s, u)$ exists between sniffer $s \in S$ and user $u \in U$ if u is within the reception range of sniffer s. If transmissions from a user cannot be captured by any sniffer, the user is excluded from G. For every vertex $v \in S \cup U$, we let $N(v)$ denote vertex v's neighbors in G. For users, their neighbors are sniffers, and vice versa. We assume that one sniffer can observe one user at a time. This is consistent with many existing multiple access mechanisms including FDMA, TDMA.

At any point in time, a sniffer can only observe transmissions over a single channel. We will consider *channel assignments* of sniffers to channels, $\mathbf{k} = (k_1, \ldots, k_p)$, where $1 \leq k_i \leq K$. Let $\mathbb{K} = \{\mathbf{k} \mid \mathbf{k} : S \to \{1, .., K\}^p\}$ be the set of all possible assignments. The set of users a sniffer s can observe is given by $N(s) \bigcap \{\, u : c(u) = k_s \,\}$.

A. Optimal channel assignment in the nominal form

We first consider the formulation of the optimal sniffer-channel assignment where the graph G and the user-activity probabilities $(p_u; u \in U)$ are both known. The discussion serves two purposes. First, optimal channel assignment with uncertainty is inherently harder than that without uncertainty. Therefore, determining the complexity of the later provides a baseline understanding of the computational aspect of the former problem. Second, as will become clearer, each instance of the decision problem along the sequential learning process can in fact be cast as the optimal channel assignment with known parameters, where the known parameters in this case are in fact the “best” estimates of these parameters (plus some margins due to insufficient samples).

The objective of optimal channel assignment is to maximize the expected number of active users monitored. Let MAX-EFFORT-COVER (MEC) denote the problem of finding the largest (weight) set of users that can be monitored by a set of sniffers, where each sniffer can monitor one of a set of k channels. Note that in MEC, the weights can in fact be any non-negative values and are not limited to $[0, 1]$. The MEC problem can be cast as the following integer program (IP):

$$\begin{array}{lll} \max & \sum_{u \in U} p_u y_u & \\ \text{s.t.} & \sum_{k=1}^{K} z_{s,k} \leq 1 & \forall s \in S \\ & y_u \leq \sum_{s \in N(u)} z_{s,c(u)} & \forall u \in U \\ & y_u, z_{s,k} \in \{0, 1\} & \forall u, s, k. \end{array} \tag{1}$$

Each sniffer is associated with a set of binary decision variables, $z_{s,k} = 1$ if the sniffer is assigned to channel k; 0, otherwise. Further, y_u is a binary variable (but not a decision variable) indicating whether or not user u is monitored, and p_u is the weight associated with user u. We have proven that MEC is NP-hard in [12]:

Theorem 1 (Theorem 1[12]): The MEC problem is NP-hard with respect to the number of sniffers, even for $K = 2$.
In other words, the computational complexity for a genie to make the optimal choice with the knowledge of all users' activity grows at least exponentially with respect to the number of sniffers, unless $P = NP$. However, when the graphs G have some specific structure, there may exist efficient algorithms. For example, when G is restricted to be a complete bipartite graph, it can be shown that MEC reduces to maximum matching in a transformed bipartite graph, which can be solved in polynomial time.

B. Linear bandit for optimal channel assignment with uncertainty

Now, we turn to the optimal channel assignment when there is uncertainty in both G and p_u's. We first define the structure of instantaneous feedback and payoff of each sniffer.

Let $U_{ik}(t)$ be a nonnegative, integer-valued random variable which denotes the index of the user whose activity sniffer i can observe in channel k at time t, or which takes the value of zero if there is no activity in the chosen channel. For simplicity, we assume that $U(t) = (U_{ik}(t); 1 \leq i \leq p, 1 \leq k \leq K)$ is a sequence of IID random variables. The instantaneous feedback (observations) received under the joint action $\mathbf{k}(t) = (k_1, \ldots, k_p)$ is $Y^{\circ}_{(k_1,\ldots,k_p)}(t) = (U_{1,k_1}(t), U_{2,k_2}(t), \ldots, U_{p,k_p}(t))$. Note that the indicator $\mathbb{I}_{\{U_{i_1,k_{i_1}}(t) = U_{i_2,k_{i_2}}(t) = \ldots = U_{i_s,k_{i_s}}(t) > 0\}}$ is a function of $Y^{\circ}_{(k_1,\ldots,k_p)}(t)$ and hence can be taken as part of the observation $Y_{(k_1,\ldots,k_p)}(t)$, defined as the collection

$$\left[\mathbb{I}_{\{U_{i_1,k_{i_1}}(t) = U_{i_2,k_{i_2}}(t) = \ldots = U_{i_s,k_{i_s}}(t) > 0\}}; 1 \leq s \leq p,\ 1 \leq i_1 < \ldots < i_s \leq p \right]. \tag{2}$$

Note that spatial multiplexing is allowed such that multiple users can be active at the same time in one channel (as long as they are sufficiently far apart geographically). However, we assume one user can be observed by one sniffer at a time. This is consistent with many existing multiple access mechanisms including FDMA, TDMA. As in Section III-A, the payoff upon selecting the joint action is the number of distinct users observed. That is, the joint payoff for selecting channels $\mathbf{k} = (k_1, k_2, \ldots, k_p)$ is

$$\begin{array}{rcl} X_{\mathbf{k}}(t) & = & \left|\{U_{1,k_1}(t), \ldots, U_{p,k_p}(t)\}\right| \\ & & -\mathbb{I}_{\{U_{1,k_1}(t)=0,\ldots,U_{p,k_p}(t)=0\}} \\ & = & \sum_{i=1}^{p} \mathbb{I}_{\{U_{1,k_i}(t)>0\}} \\ & - & \sum_{i,j=1}^{p} \mathbb{I}_{\{U_{i,k_i}(t)=U_{j,k_j}(t)>0\}} \mathbb{I}_{\{k_i=k_j, i\neq j\}} \\ & \ldots & \\ & - & (-1)^p \mathbb{I}_{\{U_{1,k_1}(t)=U_{2,k_2}(t)=\ldots=U_{p,k_p}(t)>0\}} \\ & \times & \mathbb{I}_{\{k_1=k_2=\ldots=k_p\}}. \end{array} \tag{3}$$

The expected payoff for channels $\mathbf{k} = (k_1, k_2, \ldots, k_p)$ is given by,

$$\begin{array}{rl} & \mathbb{E}\left[X_{\mathbf{k}}(t)\right] \\ = & \sum_{i=1}^{p} \mathbb{P}\left(U_{1,k_i}(t) > 1\right) \\ - & \sum_{i,j=1}^{p} \mathbb{P}\left(U_{i,k_i}(t) = U_{j,k_j}(t) > 0\right) \mathbb{I}_{\{k_i=k_j, i\neq j\}} \\ & \ldots \\ - & (-1)^p \mathbb{P}\left(U_{1,k_1}(t) = \ldots = U_{p,k_p}(t) > 0\right) \\ & \times \mathbb{I}_{\{k_1=k_2=\ldots=k_p\}} \end{array} \tag{4}$$

Define an unknown vector θ with the following elements:

$$\begin{array}{rl} \mathbb{P}\left(U_{i,k} > 0\right), & 1 \leq i \leq p, 1 \leq k \leq K, \\ \mathbb{P}\left(U_{i_1,k} = U_{i_2,k} > 0\right), & 1 \leq i_1 < i_2 \leq p, 1 \leq k \leq K, \\ \vdots & \\ \mathbb{P}\left(U_{1,k} = U_{2,k} = \ldots = U_{p,k} > 0\right), & 1 \leq k \leq K. \end{array} \tag{5}$$

We introduce the "arm features", $\phi_{\mathbf{k}} \in \mathbb{R}^M$ as (6), where $M = K(2^p - 1)$. Note that the jth arm feature $\phi_{\mathbf{k},j}$ is uniquely determined by the arm $\mathbf{k} = (k_1, k_2, \ldots, k_p)$. Let $\mathcal{M}_{\mathbf{k}} = \{\, i : 1 \leq i \leq M, \phi_{\mathbf{k},i} \neq 0 \,\}$ be the set of nonzero components of feature vector $\phi_{\mathbf{k}}$ and let $M_{\mathbf{k}} = |\mathcal{M}_{\mathbf{k}}|$.

To this end, we can rewrite the expected payoff as a linear function of the arm feature $\phi_{\mathbf{k}}$,

$$\mathbb{E}\left[X_{\mathbf{k}}(t)\right] = \theta^T \phi_{\mathbf{k}}, \tag{7}$$

where $(\cdot)^T$ denotes transposition.

Knowing θ suffices to play optimally: An arm with maximal payoff is given by $\mathbf{k}^* = \operatorname{argmax}_{\mathbf{k}\in\mathbb{K}} \theta^\top \phi_{\mathbf{k}}$ (here, and in what follows, for the sake of simplicity, we assume that there is a unique optimal arm). A reasonable way to estimate the parameter vector θ is to keep a running average for the components of θ. If at time t the agent chose $\mathbf{k}(t) \in \mathbb{K}$ then the current estimate, $\hat{\theta}(t-1)$, can be updated by

$$\begin{array}{l} \hat{\theta}_i(t) = \hat{\theta}_i(t-1) + \dfrac{1}{N_i(t)} \left(Y_i(t) - \hat{\theta}_i(t-1)\right) \mathbb{I}_{\{i \in \mathcal{M}_{\mathbf{k}(t)}\}}, \\ N_i(t) = N_i(t-1) + \mathbb{I}_{\{i \in \mathcal{M}_{\mathbf{k}(t)}\}}. \end{array} \tag{8}$$

Here $N_i(0) = 0$, $\hat{\theta}_i(0) = 0$. Thus, $N_i(t)$ counts the number of times data for component i was observed up to time t.

Example 1 (Co-located sniffers): When the sniffers are "co-located" or are deployed at close proximity, their observations are identical. Therefore, $U(t)$ will be such that if $k_i = k_j$ then $U_{i,k_i}(t) = U_{j,k_j}(t)$.[2] Then, the expected payoff is maximized by putting different sniffers to different channels, i.e., $k_i \neq k_j$, $1 \leq i < j \leq p$. It can be proved that it is strictly better to put different sniffers to different channels. In this case it suffices to estimate $P(U_{ik} > 0)$, i.e., a total of $K \cdot p$ parameters. The problem then becomes essentially the multi-armed bandit problem with multiple plays considered in a number of previous works [23], [19], [20].

Example 2 (Independent sniffers): The opposite case is when $U_{i,k_i}(t) \neq U_{j,k_j}(t)$ whenever $i \neq j$ and when one of $U_{i,k_i}(t)$ and $U_{j,k_j}(t)$ is nonzero. In words, all sniffers are guaranteed to observe distinct users (e.g., they are far away from one another). Then, $\mathbb{I}_{\{U_{i_1,k_{i_1}}=U_{i_2,k_{i_2}}=\ldots=U_{i_s,k_{i_s}}>0\}} = 0$, $2 \leq s \leq p$, $1 \leq i_1 <, \ldots, < i_s \leq p$. Therefore, the number of parameters are reduced to $K \cdot p$ and each sniffer can decide independently which channel to monitor. Thus the, problem reduces to p independent K-arm bandit problems.

In practice, sniffers are deployed distributedly. Their observations are typically correlated but non-identical. This motives us to consider the optimal channel assignment in general configurations. An optimal monitoring policy π determines a sequence of actions in $\mathbb{K}$ over time such that the expected *regret* is minimized:

$$R_n^\pi = \mathbb{E}\left[\sum_{t=1}^{n} \left\{\max_{\mathbf{k}\in\mathcal{A}} \phi_{\mathbf{k}}^T \theta - \phi_{\mathbf{k}_t}^T \theta\right\}\right].$$

Here, $\mathbf{k}_t$ denotes the joint action selected at time t.

C. Relationships between p_u and θ

Theorem 2 states the one-to-one mapping between p_u and θ with G properly defined. As such, we can apply optimization solutions to (1) to determine the "best" arm to play at each instance.

Theorem 2: Let $\bar{p}$ be the vector denoting the user-activity probabilities. Under some mild non-limiting conditions, there exists a full rank matrix A such that $\log(1-\theta) = \log(1-\bar{p}) \cdot A$.

Proof: See Appendix A. ■

D. Spanners

Since some arms reveal information about other arms, it might be possible to identify a restricted set $\mathcal{E} \subset \mathbb{K}$, which might be much smaller than $\mathbb{K}$, so that playing only arms in $\mathcal{E}$ gives sufficient information to identify the optimal arm. A sufficient condition for this is that $\cup_{\mathbf{k}\in\mathcal{E}} \mathcal{M}_{\mathbf{k}} = \{1, \ldots, M\}$. This condition ensures that by choosing an appropriate arm in $\mathcal{E}$ any component of $X(t)$ can be observed, which is clearly sufficient to identify θ. Since exploration is generally costly, the set $\mathcal{E}$ is ideally chosen to be small. In the monitoring problem $\mathcal{E}$ can be chosen to be $\mathcal{E} = \{\, (k, \ldots, k) : 1 \leq k \leq K \,\}$, i.e. all the sniffers assigned to the same channel to cover $(2^p - 1)$ parameters, whose cardinality is $K \ll K^p = |\mathbb{K}|$. The set $\mathcal{E}$ is called a *spanning set* or a *spanner* and its elements are called *spanner arms*.

[2] Clock synchronization among sniffers can be achieved online or offline using methods such as in [22].

$$\phi_{\mathbf{k},i} = \begin{cases} \mathbb{I}_{\{k_1=i\}}\,, & \text{if } 1 \le i \le K;\\ \dots & \\ \mathbb{I}_{\{k_2=i-l\cdot K\}}\,, & \text{if } l\cdot K+1 \le i \le (l+1)\cdot K;\\ \dots & \\ -\mathbb{I}_{\{k_1=k_2=i-p\cdot K\}}\,, & \text{if } p\cdot K+1 \le i \le (p+1)\cdot K;\\ \dots & \\ -(-1)^p\mathbb{I}_{\{k_1=k_2=\ldots=k_p=i-K(2^p-2)\}}\,, & \text{if } K(2^p-2)+1 \le i \le K(2^p-1) \end{cases} \tag{6}$$

IV. An Upper Confidence Bound (UCB)-based Policy

The first policy that we consider is similar to UCB1 [3] with the difference that in the initialization stage, we only play each of the spanners once. Formally, the algorithm first plays each arm in $\mathcal{E}$ once and then at time $t \ge |\mathcal{E}|+1$ chooses

$$\mathbf{k}(t) = \operatorname*{argmax}_{\mathbf{k}\in\mathcal{E}} V_{\mathbf{k}}(t-1),$$

where

$$\begin{aligned} V_{\mathbf{k}}(t-1) &= \hat{\mu}_{\mathbf{k}}(t-1) + \sum_{i\in\mathcal{M}_{\mathbf{k}}} \sqrt{\frac{\rho\log t}{N_i(t-1)}},\\ \hat{\mu}_{\mathbf{k}}(t-1) &= \hat{\theta}(t-1)^\top \phi_{\mathbf{k}}. \end{aligned}$$

After playing $\mathbf{k}(t)$ and observing $(Y_i(t); i \in \mathcal{M}_{\mathbf{k}(t)})$ the parameter estimate is updated using (8). Then, the process is repeated.

Theorem 3: Choose any ρ that satisfies $\rho > 1/1.99$. Then, there exists a constant $C > 0$ (which may depend on ρ) such that for all $n \ge 1$, the expected regret of UCB1 satisfies

$$R_n^{\mathrm{UCB1}} \le 4M\Delta_{\max}\left(\max_{\mathbf{k}:\Delta_{\mathbf{k}}>0}\frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho\log n + C,$$

where $\Delta_{\max} = \max_{\mathbf{k}} \Delta_{\mathbf{k}}$.
The exact dependence of C on the problem parameters can be extracted from the proof. In particular, C scales linearly with $|\mathbb{K}|$.

Proof: The proof is similar to the original proof that given by Auer *et al*[3], with some elements borrowed from the analysis technique of Audibert *et al*[24]. (see also, [25]) and Gai *et al* [21].

We start by introducing the necessary notation. We denote by $T_{\mathbf{k}}(n)$ the number of times arm $\mathbf{k}$ is chosen up to time n (including time n): $T_{\mathbf{k}}(n) = \sum_{t=1}^n \mathbb{I}_{\{\mathbf{k}(t)=\mathbf{k}\}}$. We let $\mu^* = \max_{\mathbf{k}} \mu_{\mathbf{k}}$, $\Delta_k = \mu^* - \mu_{\mathbf{k}}$. Then, it is easy see that $\mathbb{E}\left[R_n^{\mathrm{UCB1}}\right] = \sum_{\mathbf{k}} \Delta_{\mathbf{k}}\mathbb{E}\left[T_{\mathbf{k}}(n)\right] \le (\max_{\mathbf{k}}\Delta_{\mathbf{k}})\,\mathbb{E}\left[\sum_{\mathbf{k}:\Delta_{\mathbf{k}}>0} T_{\mathbf{k}}(n)\right]$. Our goal is to develop a bound on $\mathbb{E}\left[\sum_{\mathbf{k}:\Delta_{\mathbf{k}}>0} T_{\mathbf{k}}(n)\right]$ which scales linearly with M rather that with $|\mathbb{K}|$.

Let $I(t) = \operatorname{argmin}_{j\in\mathcal{M}_{\mathbf{k}(t)}} N_j(t-1)$ (ties can be broken, say, in favor of the smallest index), $Z_i(t) = \mathbb{I}_{\{\mathbf{k}(t)\neq\mathbf{k}^*, I(t)=i\}}$, $\tilde{T}_i(t) = \tilde{T}_i(t-1) + Z_i(t)$.[3] Note that $\sum_{\mathbf{k}\neq\mathbf{k}^*} T_{\mathbf{k}}(n) = \sum_i \tilde{T}_i(n)$, since exactly one of the counters is incremented on both sides when a suboptimal arm is chosen. Thus, it suffices to bound $\tilde{T}_i(n)$.

Therefore pick any index $1 \le i \le M$ and let u be an integer to be chosen later. We have $Z_i(t) = Z_i(t)\mathbb{I}_{\{\tilde{T}_i(t-1)>u\}} + Z_i(t)\mathbb{I}_{\{\tilde{T}_i(t-1)\le u\}}$. Since $\sum_{t=1}^n Z_i(t)\mathbb{I}_{\{\tilde{T}_i(t-1)\le u\}} \le u+1$, it suffices to deal with the first term, which we bound as follows:

$$\begin{aligned} &Z_i(t)\mathbb{I}_{\{\tilde{T}_i(t-1)>u\}}\\ &\le \mathbb{I}_{\{V_{\mathbf{k}(t)}(t-1)>\mu^*, \tilde{T}_i(t-1)>u, I(t)=i\}} + \mathbb{I}_{\{V_{\mathbf{k}^*}(t-1)\le\mu^*\}}. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}\left[\tilde{T}_i(n)\right] &\le u+1\\ &+\sum_{t=1}^n \mathbb{P}\left(V_{\mathbf{k}(t)}(t-1)>\mu^*, \tilde{T}_i(t-1)>u, I(t)=i\right)\\ &+\sum_{t=1}^n \mathbb{P}\left(V_{\mathbf{k}^*}(t-1)\le\mu^*\right). \end{aligned}$$

We will now show that both sums are finite, provided that u is sufficiently large.

The summand of the first sum is bounded as follows:

$$\begin{aligned} p_{1t} &\overset{\text{def}}{=} \mathbb{P}\left(V_{\mathbf{k}(t)}(t-1)>\mu^*, \tilde{T}_i(t-1)>u, I(t)=i\right)\\ &\le \mathbb{P}\Big\{\hat{\mu}_{\mathbf{k}(t)}(t-1) > \mu_{\mathbf{k}(t)} + \Delta_{\mathbf{k}(t)} - c_{\mathbf{k}(t),t-1},\\ &\qquad\qquad \tilde{T}_i(t-1)>u, I(t)=i\Big\} \end{aligned}$$

where $c_{\mathbf{k},t-1} = \sqrt{\rho\log t}\sum_{i\in\mathcal{M}_{\mathbf{k}}}\sqrt{\frac{1}{N_i(t-1)}} \overset{\text{def}}{=} \sqrt{\rho\log t}\;W_{\mathbf{k}}(t-1)$. Now,

$$\begin{aligned} &\Delta_{\mathbf{k}} - c_{\mathbf{k},t-1}\\ &= \sum_{i\in\mathcal{M}_{\mathbf{k}}}\left(\frac{\Delta_{\mathbf{k}}}{W_{\mathbf{k}}(t-1)} - \sqrt{\rho\log t}\right)\sqrt{\frac{1}{N_i(t-1)}}. \end{aligned}$$

We claim that under the condition that $\tilde{T}_{I(t)}(t-1) > u$ the largest value $W_{\mathbf{k}(t)}(t-1)$ can take is bounded from above by $M_{\mathbf{k}(t)}/\sqrt{u}$. To see this note that $\tilde{T}_i(t-1) \le N_i(t-1)$ holds for any i and t, because $N_i(\cdot)$ is always incremented when $\tilde{T}_i(\cdot)$ is incremented. Further, since $I(t) = \operatorname{argmin}_{j\in\mathcal{M}_{\mathbf{k}(t)}} N_j(t-1)$, $N_{I(t)}(t-1) \le N_j(t-1)$ holds for any $j \in \mathcal{M}_{\mathbf{k}(t)}$. Thus, for arbitrary $j \in \mathcal{M}_{\mathbf{k}(t)}$, $u < \tilde{T}_{I(t)}(t-1) \le N_{I(t)}(t-1) \le N_j(t-1)$. The claim then follows from the definition of $W_{\mathbf{k}(t)}(t-1)$.

[3] We are using the assumption that there is a unique optimal arm $\mathbf{k}^*$. Note that this is assumed just for the sake of simplicity and the proof, at the price of a more complicated presentation, works without it.

Hence,

$$\begin{aligned}\Delta_{\mathbf{k}(t)} &- c_{\mathbf{k}(t),t-1} \\ &\geq \sum_{i\in\mathcal{M}_{\mathbf{k}(t)}} \left(\frac{\Delta_{\mathbf{k}(t)}\sqrt{u}}{M_{\mathbf{k}(t)}} - \sqrt{\rho\log t}\right)\sqrt{\frac{1}{N_i(t-1)}}.\end{aligned}$$

Further, $\frac{\Delta_{\mathbf{k}(t)}\sqrt{u}}{M_{\mathbf{k}(t)}} - \sqrt{\rho\log t} \geq \sqrt{\rho\log n}$ holds for $1 \leq t \leq n$ if

$$u \geq \left(2\max_{\mathbf{k}:\Delta_{\mathbf{k}}>0}\frac{M_{\mathbf{k}}}{\Delta_{\mathbf{k}}}\right)^2 \rho\log n.$$

Then, $\Delta_{\mathbf{k}(t)} - c_{\mathbf{k}(t),t-1} \geq \sqrt{\rho\log n}\ W_{\mathbf{k}(t)}(t-1)$ and thus

$$\begin{aligned}p_{1t} &\leq \mathbb{P}\left(\hat{\mu}_{\mathbf{k}(t)}(t-1) > \mu_{\mathbf{k}(t)} + \sqrt{\rho\log n}\ W_{\mathbf{k}(t)}(t-1)\right) \\ &\leq \sum_{\mathbf{k}} M_{\mathbf{k}}\left\lceil 4\log n\right\rceil \exp\left(-1.99\rho\log n\right),\end{aligned}$$

where the last inequality follows from the union bound and Lemma 6, which is presented in Appendix B. Thus, $\sum_{t=1}^{n} p_{1t} \leq \sum_{\mathbf{k}} M_{\mathbf{k}}\lceil 4\log n\rceil\, n^{1-1.99\rho}$. Hence, if ρ is such that $1.99\rho > 1$, we get that $\sum_{t=1}^{n} p_{1t} = o(1)$ (note that p_{1t} does depend on n through u, although this dependence was chosen not to be shown in the notation).

Using Lemma 6 again, we get that $p_{2t} = \mathbb{P}\left(V_{\mathbf{k}^*}(t-1) \leq \mu^*\right) \leq M_{\mathbf{k}^*}\lceil 4\log n\rceil t^{-1.99\rho}$. Hence, $\sum_{t=1}^{n} p_{2t} \leq M_{\mathbf{k}^*}\lceil 4\log n\rceil(1.99\rho-1)^{-1}$, again, under the assumption that $1.99\rho > 1$.

Putting together the inequalities obtained we get the desired result. ■

Note that in linear bandit problems there exist similar regret bounds, see [26], [27]. The (problem dependent) bound developed in [27] takes the form $(M^2/\min_{\mathbf{k}}\Delta_{\mathbf{k}})\log^3(n)$, i.e., it is in general incomparable to our bound: Our bound scales better as a function of n and M when $\max_{\mathbf{k}} M_{\mathbf{k}}^2$ is "small". However, the scaling of our bound as a function of $\Delta_{\min} = \min_{\mathbf{k}:\Delta_{\mathbf{k}}>0}\Delta_{\mathbf{k}}$ is worse. In general, one expects the algorithm presented here to perform better than the ones developed for the linear bandit problem since those algorithms do not take advantage of the potentially richer feedback. However, this remains to be proven.

Our result is more directly comparable to that of Gai *et al* [21]. In fact, their problem is a special case of the problem studied here (when we allow arbitrary $\phi_{\mathbf{k}} \in \{-1,0,1\}^M$). The scaling behavior of our bound (for their problem) is essentially the same as a function of n and $\Delta_{\min}$ (after bounding $\max_{\mathbf{k}:\Delta_{\mathbf{k}}>0} M_{\mathbf{k}}/\Delta_{\mathbf{k}}$ by $(\max_{\mathbf{k}} M_{\mathbf{k}})/\Delta_{\min}$) but our bound scales better as a function of $\max_{\mathbf{k}} M_{\mathbf{k}}$ (their bound scales with $\max_{\mathbf{k}} M_{\mathbf{k}}^3$, whereas ours scales with $\max_{\mathbf{k}} M_{\mathbf{k}}^2$).

Note that in the proof no attempt was made to optimize the constants. The major issue with this algorithm is that apart from the initialization phase in its "exploration" it does not take full advantage of the correlations between the payoffs of the arms, at least when it is exploring. One idea to overcome the algorithm's potential insensitivity to the correlation structure is to modify the algorithm so that the arms in $\mathcal{E}$ are explored (with uniform probabilities) in the explicit "exploration steps", i.e., when $\mathbf{k}(t) \neq \operatorname{argmax}_{\mathbf{k}}\hat{\mu}_{\mathbf{k}}(t-1)$. We conjecture that this algorithm indeed overcomes the above mentioned handicap, i.e., its regret would scale with $|\mathcal{E}|$ and not with the number of the parameters.

In the next section we explore a similar idea in the context of a simpler algorithm, ε-greedy.

V. An ϵ-Greedy Algorithm

The policy considered here is a variant of ε-greedy. The standard ε-greedy algorithm for bandit problems chooses with probability ε uniformly at random some arm (i.e., it "explores" with probability ε) and it chooses the arm with the highest estimated payoff otherwise. When ε is appropriately scheduled (basically, one needs $\varepsilon = \varepsilon_n = c/n$ with an appropriately selected constant $c > 0$) this policy can also achieve a logarithmically bounded expected regret just like UCB1 [3].

Since in our case the arms are correlated and when an arm is chosen one receives some additional information in addition to the payoffs, one may restrict the set of arms explored to a spanner $\mathcal{E}$. We expect that performance will improve if $|\mathcal{E}| \lll |\mathbb{K}|$ since then one "pays less" for the exploration steps.

Formally, the algorithm works as follows: Choose a spanner $\mathcal{E} \subset \mathbb{K}$ and a sequence $(\varepsilon_t; t \geq 1)$, $\varepsilon_t \in [0,1]$. In the initialization phase explore each arm in $\mathcal{E}$ once and initialize the parameter estimates $\hat{\theta}(\cdot)$ based on the information received. After the exploration phase, at time $t \geq |\mathcal{E}|$, the arm to be played is decided by first drawing a random number U_t from the uniform distribution over $[0,1]$. If $U_t \leq \varepsilon_t$ then $\mathbf{k}(t)$ is chosen uniformly at random from $\mathcal{E}$. Otherwise, $\mathbf{k}(t) = \operatorname{argmax}_{\mathbf{k}}\hat{\mu}_{\mathbf{k}}(t-1)$, where $\hat{\mu}_{\mathbf{k}}(t-1) = \hat{\theta}(t-1)^\top\phi_{\mathbf{k}}$. After playing $\mathbf{k}(t)$ and observing the feedback, the parameters are updated using (8).

The next theorem gives a bound on the regret of this policy:

Theorem 4: Let

$$\epsilon_n = \min\left\{1, \frac{c}{n}\right\},\ n > |\mathcal{E}|, \tag{9}$$

where $c > 0$ is a tuning parameter. Then, assuming that $c > \min(10|\mathcal{E}|, \frac{4|\mathcal{E}|}{d^2})$, where $d = \min_{\mathbf{k}:\Delta_{\mathbf{k}}>0}\Delta_{\mathbf{k}}$, the expected regret of ε-greedy satisfies

$$\mathbb{E}\left[R_n^{\varepsilon-\text{greedy}}\right] \leq c\log(n+1) + O(1). \tag{10}$$

From the point of view of minimizing the leading term, the best choice of c is $\min(10|\mathcal{E}|, \frac{4|\mathcal{E}|}{d^2})$. With such a choice, we see that the leading term of regret scales linearly with $|\mathcal{E}|$, and not with $|\mathbb{K}|$. This is the main difference between the bound in this theorem and in the previous result. This can be a major advantage when $|\mathcal{E}| \lll |\mathbb{K}|$ (e.g., in the monitoring problem). The disadvantage of this algorithm is that in practice tuning c might be difficult, since, typically, d is unknown. One remedy then is to replace c with a slowly growing sequence c_n (e.g., $c_n = \log\log n$, i.e., use $\varepsilon_n = \min(1, \frac{\log\log n}{n})$. This would result in a regret that grows in the order of $c_n\log n$, but the proof of this result is omitted for brevity.

Proof: The proof follows the steps of the proof in [3] with some modifications (and slight improvements). We will use the notation introduced in the proof of Theorem 3.

Without the loss of generality, we may assume that $\varepsilon_n = 0$ if $n \leq |\mathcal{E}|$ (note that the algorithm does not depend on the values of $\varepsilon_1, \ldots, \varepsilon_{|\mathcal{E}|}$ and this assumption allows us to shorten the proof). Clearly, it suffices to bound $\mathbb{E}\left[T_{\mathbf{k}}(n)\right]$. For this purpose we will bound $\mathbb{P}\left(\mathbf{k}(n) = \mathbf{k}\right)$, where $\mathbf{k}$ is any suboptimal action.

For $n > |\mathcal{E}|$, the probability of choosing $\mathbf{k}$ is bounded by

$$\begin{aligned}\mathbb{P}\left(\mathbf{k}(n) = \mathbf{k}\right) \leq \frac{\epsilon_n \, \mathbb{I}_{\{\mathbf{k} \in \mathcal{E}\}}}{|\mathcal{E}|} \\ + \left(1 - \epsilon_n\right) \mathbb{P}\left(\hat{\mu}_{\mathbf{k}}(n-1) \geq \hat{\mu}_{\mathbf{k}^*}(n-1)\right).\end{aligned}$$

We have

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_{\mathbf{k}}(n-1) \geq \hat{\mu}_{\mathbf{k}^*}(n-1)\right) \leq \mathbb{P}\left(\hat{\mu}_{\mathbf{k}}(n-1) \geq \mu_{\mathbf{k}} + \frac{\Delta_{\mathbf{k}}}{2}\right) \\ + \mathbb{P}\left(\hat{\mu}_{\mathbf{k}^*}(n-1) \leq \mu^* - \frac{\Delta_{\mathbf{k}}}{2}\right).\end{aligned}$$

We bound the first term as follows: Define $\delta_{\mathbf{k}} = \frac{\Delta_{\mathbf{k}}}{M_{\mathbf{k}}}$. Then,

$$\begin{aligned}&\mathbb{P}\left(\hat{\mu}_{\mathbf{k}}(n-1) \geq \mu_{\mathbf{k}} + \frac{\Delta_{\mathbf{k}}}{2}\right) \\ &\quad \leq \sum_{i \in \mathcal{M}_{\mathbf{k}}} \mathbb{P}\left(\hat{\theta}_i(n-1)\phi_{\mathbf{k},i} \geq \theta_i \phi_{\mathbf{k},i} + \frac{\delta_{\mathbf{k}}}{2}\right)\end{aligned}$$

Pick $i \in \mathcal{M}_{\mathbf{k}}$. Define $x_0 = \frac{1}{2|\mathcal{E}|}\sum_{t=1}^{n-1} \epsilon_t$. By Lemma 7,

$$\begin{aligned}&\mathbb{P}\left(\hat{\theta}_i(n-1)\phi_{\mathbf{k},i} \geq \theta_i \phi_{\mathbf{k},i} + \frac{\delta_{\mathbf{k}}}{2}\right) \\ &\quad \leq \mathbb{P}\left(N_i(n-1) \leq x_0\right) + \frac{2}{\delta_{\mathbf{k}}^2} \exp\left(-\frac{\lceil x_0 \rceil \delta_{\mathbf{k}}^2}{2}\right).\end{aligned}$$

Let us now bound the first term of the right-hand side. Let $\mathbf{k}_e \in \mathcal{E}$ be such that $i \in \mathcal{M}_{\mathbf{k}_e}$. Let $N_i^R(n)$ be the number of times $\mathbf{k}_e$ was selected up to time n in an exploration step: $N_i^R(n) = \sum_{t=1}^{n} \mathbb{I}_{\{\mathbf{k}(t) = \mathbf{k}_e, U_t \leq \varepsilon_t\}}$. Clearly, $N_i^R(n-1) \leq N_i(n-1)$. Hence, $\mathbb{P}\left(N_i(n-1) \leq x_0\right) \leq \mathbb{P}\left(N_i^R(n-1) \leq x_0\right)$. Furthermore, $\mathbb{E}\left[N_i^R(n-1)\right] = \frac{1}{|\mathcal{E}|}\sum_{t=1}^{n-1} \epsilon_t = 2x_0$, and $\mathrm{Var}[N_i^R(n-1)] \leq \frac{1}{|\mathcal{E}|}\sum_{t=1}^{n-1} \epsilon_t = 2x_0$. Therefore, by Bernstein's inequality (for details see [3]), we have

$$\mathbb{P}\left(N_i^R(n-1) \leq x_0\right) \leq e^{-x_0/5}. \tag{11}$$

Since $x_0 = \frac{1}{2|\mathcal{E}|}\sum_{t=1}^{n-1} \epsilon_t \geq \frac{c}{2|\mathcal{E}|}\log n$, we have

$$\mathbb{P}\left(N_i^R(n-1) \leq x_0\right) \leq e^{-x_0/5} \leq n^{-\frac{c}{10|\mathcal{E}|}}.$$

Thus,

$$\mathbb{P}\left(\hat{\mu}_{\mathbf{k}}(n-1) \geq \mu_{\mathbf{k}} + \frac{\Delta_{\mathbf{k}}}{2}\right) \leq M_{\mathbf{k}} n^{-\frac{c}{10|\mathcal{E}|}} + \sum_{i \in \mathcal{M}_{\mathbf{k}}} \frac{2}{\delta_{\mathbf{k}}^2} n^{\frac{c\delta_{\mathbf{k}}^2}{4|\mathcal{E}|}}. \tag{12}$$

The same bound holds for $\mathbb{P}\left(\hat{\mu}_{\mathbf{k}^*}(n-1) \leq \mu^* - \frac{\Delta_{\mathbf{k}}}{2}\right)$. Therefore, combining the inequalities obtained so far, we get

$$\mathbb{P}\left(\mathbf{k}(n) = \mathbf{k}\right) \leq \frac{c\,\mathbb{I}_{\{\mathbf{k} \in \mathcal{E}\}}}{n|\mathcal{E}|} + 2M_{\mathbf{k}} n^{-\frac{c}{10|\mathcal{E}|}} + \sum_{i \in \mathcal{M}_{\mathbf{k}}} \frac{4}{\delta_{\mathbf{k}}^2} n^{\frac{c\delta_{\mathbf{k}}^2}{4|\mathcal{E}|}}.$$

Now, $\mathbb{E}\left[R_n^{\varepsilon-\mathrm{greedy}}\right] \leq |\mathcal{E}| + \sum_{t=|\mathcal{E}|+1}^{n} \sum_{\mathbf{k}} \Delta_{\mathbf{k}} \mathbb{P}\left(\mathbf{k}(n) = \mathbf{k}\right) \leq |\mathcal{E}| + c\log n + \left(\sum_{\mathbf{k}} 2\Delta_{\mathbf{k}} M_{\mathbf{k}}\right) \sum_{t=1}^{n} t^{-\frac{c}{10|\mathcal{E}|}} + \sum_{\mathbf{k}:\Delta_{\mathbf{k}}>0} \frac{4M_{\mathbf{k}}^2}{\Delta_{\mathbf{k}}} \sum_{t=1}^{n} t^{\frac{c\delta_{\mathbf{k}}^2}{4|\mathcal{E}|}}$. If $c > \min(10|\mathcal{E}|, \frac{4|\mathcal{E}|}{\delta_{\mathbf{k}}^2})$ holds for any suboptimal $\mathbf{k}$ then the sum of the last two terms over $t = 1, \ldots, n$ becomes finite. This finishes the proof of the result. ■

VI. Numerical Results

We have implemented the proposed UCB and ϵ-Greedy algorithms, and a naive extension of the UCB scheme proposed by Gai *et al* [21] in Matlab. In extending [21] to deal with correlated arms, since the dimension of an arm is the number of non-zero elements in the arm-features, we used

$$\begin{aligned}V_{\mathbf{k}}(t-1) &= \hat{\mu}_{\mathbf{k}}(t-1) + |\mathcal{M}_{\mathbf{k}}|\sqrt{\frac{(|\mathcal{M}_{\mathbf{k}}|+1)\log t}{\min_{\mathcal{M}_{\mathbf{k}}} N_i(t-1)}}, \\ \hat{\mu}_{\mathbf{k}}(t-1) &= \hat{\theta}(t-1)^{\top}\phi_{\mathbf{k}}.\end{aligned}$$

as the index for the UCB scheme in [21]. In the simulations, we vary the number of sniffers $p = \{1, 2, \ldots, 5\}$, and the number of channels $K \in \{1, 2, \ldots, 8\}$. Each channel has one user associated to it. The users are active with probability $[0.1 - 0.8]$ respectively in channel $[1 - 8]$. The adjacency matrix G of all 5 sniffers is given by

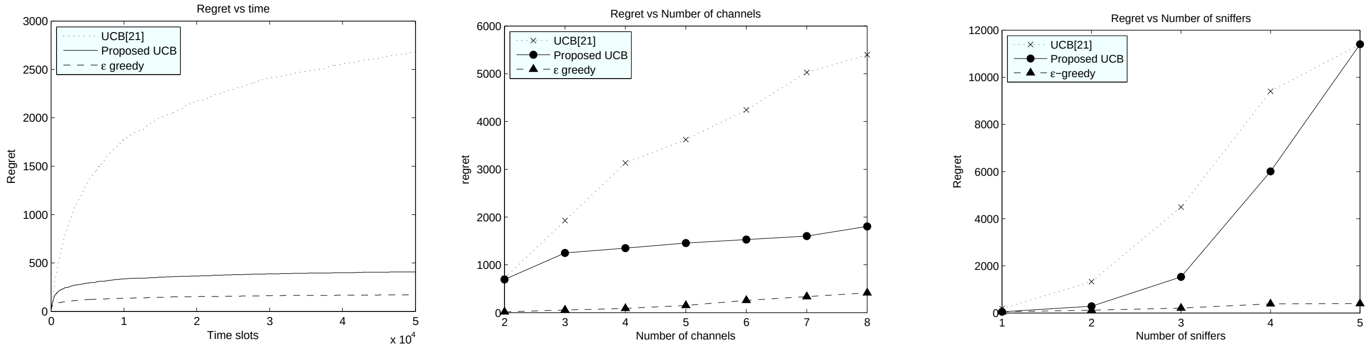
$$\begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

where $g_{i,j} = 0$ indicates that the user j is out of the reception range of sniffer i. θ can be obtained from the adjacency matrix and the user active probability due to Theorem 2.

From Figure 1, we see that for all schemes the regret tends to flatten out over time. We used the exact value of the parameter d for ϵ-greedy algorithm. Among the three schemes, ϵ-greedy has the fastest convergence followed by the proposed UCB as second, and then UCB scheme of Gai *et al* [21]. This is because ϵ-greedy utilizes the spanners during the exploration phases and can gain "most" information regarding the unknown parameters and it also avoids using confidence bounds in making decisions. In contrast, both UCB policies update their confidence bounds quite conservatively, and thus exhibit slow convergence. Similar observations can be made from Figure 1(b)(c) showing the regret after 5000 time slots.

VII. Conclusion and future work

In this paper, we considered the problem of optimally assigning p sniffers to K channels to monitor the transmission activities in a multi-channel wireless network. Two policies were proposed that learn sequentially the user activities while making channel assignment decisions. Both policies were shown to achieve logarithmic regret in the number of time slots with a term sub-linear in cardinality of the action space.



(a) Regret vs. time $p = 2, K = 5$

(b) Regret vs. # of chan. $p = 3, t = 5000$

(c) Regret vs. sniffers $K = 5, t = 5000$

Fig. 1: Comparison of regrets of three schemes

The generalization of our theorems to the following cases is trivial: *(i)* $X_i(t)$ is sub-Gaussian with known tail behavior (e.g., $X_i(t)$ are bounded with known bounds), *(ii)* $\phi_{\mathbf{k}} \in \mathbb{R}^M$. Other possible future work includes extension to non-stationarity environments, which could be done, e.g., along the line of work of [28], the consideration of an adversarial setting [16], [29], and/or switching costs [30].

Appendix A
Equivalence between the two models

In this section, we establish the equivalence between the two models to describe user activities. In the first model, the restriction of which sniffer can observe which users is modeled as a bi-partite graph $G = (S, U, E)$, and the user activity is encoded by a vector $p = (p_u)_{u \in U}$. In the second model, a $K \cdot 2^p$-dimension vector θ is defined with the following elements:

$$\begin{array}{rl} \mathbb{P}\left(U_{i,k} > 0\right), & 1 \leq i \leq p, 1 \leq k \leq K, \\ \mathbb{P}\left(U_{i_1,k} = U_{i_2,k} > 0\right), & 1 \leq i_1 < i_2 \leq p, 1 \leq k \leq K, \\ \vdots & \\ \mathbb{P}\left(U_{1,k} = U_{2,k} = \ldots = U_{p,k} > 0\right), & 1 \leq k \leq K. \end{array}$$

When the set of sniffers that can observe u and v are identical in the same channel, namely, $N(u) = N(v)$, we treat u and v as a single user. In another word, in the first model, we only consider *distinct* users, each connecting to a different set of sniffers over a specific channel. Note that the total number of distinct users is at most $K \cdot 2^p$. Let us consider users on channel k without loss of generality. Each user u is represented by y_u, a vector of length p, where $y_u(i) = 1$ if $u \in N(i)$, $i = 1, \ldots, p$. Denote a partial order between two binary vectors $y_u \preceq y_v$ if $y_u \neq y_v$, and $\forall l$, s.t., $y_u(l) = 1$, $y_v(l) = 1$. Let $p^k = [p_1^k, p_2^k, \ldots, p_{2^p-1}^k]$, i.e., the vector of probabilities of the individual user being active in channel k.

Construct a matrix $A(k)$ as follows. $A_{u,v}(k) = 1$ if $y_v \preceq y_u$. Clearly, with proper permutation, $A(k)$ is an upper diagonal matrix with all 1 entries at the diagonal. Furthermore, $\log(1 - \theta) = \log(1 - p^k) \cdot A(k)$. Since $A(k)$ is of full rank, we have $\log(1 - p^k) = \log(1 - \theta) \cdot A(k)^{-1}$. Now let $A(k)$ be the kth diagonal block of A and $\bar{p} = [p^1, p^2, \ldots, p^k]$. We have $\log(1 - \theta) = \log(1 - \bar{p}) \cdot A$ and $\log(1 - \bar{p}) = \log(1 - \theta) \cdot A^{-1}$.

Appendix B
Tail probability bounds

The following lemma generalizes Hoeffding's inequality to sums with a random number of terms. The lemma in the form presented here can be found as Theorem 18 of [28] (a similar statement, generalizing Bernstein's inequality can be extracted from [24]).

Lemma 5: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in some interval of length B. Let $\varepsilon_t \in \{0, 1\}$ be a binary sequence. Assume that X_t is $\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable ($t \geq 1$). Let $N_n = \sum_{t=1}^n \varepsilon_t$, $\overline{X}_n = \sum_{t=1}^n \varepsilon_t X_t / N_n$. Then, for any $n \geq 1$, $\eta > 0$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + z\sqrt{\frac{1}{N_n}}, N_n \geq 1\right) \leq \frac{\log n}{\log(1+\eta)} \exp\left(-\frac{2z^2}{B^2}\left(1 - \frac{\eta^2}{16}\right)\right).$$

In particular, when $\eta = 0.3$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}\left[X_1\right] + \frac{z}{\sqrt{N_n}}, N_n \geq 1\right) \leq \lceil 4 \ln n \rceil \exp\left(-\frac{1.99z^2}{B^2}\right).$$

Now, we consider a multi-dimensional generalization of this result:

Lemma 6: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in $\mathbb{R}^M$ such that X_{ti}, the i^{th} component of X_t, takes values in some interval of length B. Define $\mu = \sum_{i=1}^M \mathbb{E}\left[X_{1i}\right]$. Let $\varepsilon_t \in \{0, 1\}^M$ be an M-dimensional binary sequence. Assume that X_t is $\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable ($t \geq 1$). Let $N_{ni} = \sum_{t=1}^n \varepsilon_{ti}$, $\overline{X}_{ni} = N_{ni}^{-1} \sum_{t=1}^n \varepsilon_{ti} X_{ti}$ and $\overline{X}_n = \sum_{i=1}^M \overline{X}_{ni}$. Then, for any $n \geq 1$,

$$\mathbb{P}\left(\overline{X}_n > \mu + z\sum_{i=1}^M \sqrt{\frac{1}{N_{ni}}},\ N_{n1}, \ldots, N_{nM} \geq 1\right) \leq M \lceil 4 \ln n \rceil \exp\left(-\frac{1.99z^2}{B^2}\right).$$

Proof: Let p denote the probability to be bounded and let $\mu_i = \mathbb{E}\left[X_{1i}\right]$. Then, $p \leq \sum_{i=1}^M \mathbb{P}\left(\overline{X}_{ni} > \mu_i + z\sqrt{\frac{1}{N_{ni}}},\ N_{ni} \geq 1\right)$. The result then follows by applying Lemma 5 to each of the M terms on the right-hand side. ■

The next result can be extracted from [3] (with a slight improvement). The setting is similar to that of Lemma 5 with the deviation from the mean as a deterministic number.

Lemma 7: Let $(\mathcal{F}_t; t \geq 0)$ be a filtration. Let $(X_t; t \geq 1)$ be an i.i.d. sequence taking values in some interval of length 1. Let $\varepsilon_t \in \{0, 1\}$ be a binary sequence. Assume that X_t is $\mathcal{F}_t$-measurable and ε_t is $\mathcal{F}_{t-1}$-measurable $(t \geq 1)$. Let $N_n = \sum_{t=1}^{n} \varepsilon_t$, $\overline{X}_n = \sum_{t=1}^{n} \varepsilon_t X_t / N_n$. Then, for any $n \geq 1$, $x > 0$, $z > 0$,

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right) \leq \mathbb{P}(N_n < x) + \frac{2}{z^2} \exp\left(-\frac{\lceil x \rceil z^2}{2}\right).$$

Proof: We have

$$\mathbb{P}\left(\overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right) \leq \mathbb{P}(N_n < x) + \mathbb{P}\left(N_n \geq x, \overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right).$$

Now,

$$\mathbb{P}\left(N_n \geq x, \overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right) = \sum_{s=\lceil x \rceil}^{n} \mathbb{P}\left(N_n = s, \overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right).$$

Let $S_n = \sum_{t=1}^{n} \varepsilon_t X_t$. Define $\tau(s)$ as the first time when s values of X are observed: $\tau(s) = \min\{\, t \geq 1 : N_t = s \,\}$. Further, let $S^{(1)} = S_{\tau(1)}$, $S^{(2)} = S_{\tau(2)}$, ... Note that $S^{(k)}$ has exactly k terms and $S^{(k)}$ is an $\mathcal{F}^{(k)}$-adapted martingale, where $\mathcal{F}^{(k)} = \mathcal{F}_{\tau(k)-1}$ (the so-called the "optional skipping process"). Now, $\overline{X}_n = S_n / N_n = S^{(N_n)} / N_n$. Hence,

$$\begin{aligned} &\mathbb{P}\left(N_n = s, \overline{X}_n > \mathbb{E}[X_1] + \frac{z}{2}\right) \\ &= \mathbb{P}\left(N_n = s, S^{(N_n)}/N_n > \mathbb{E}[X_1] + \frac{z}{2}\right) \\ &= \mathbb{P}\left(N_n = s, S^{(s)}/s > \mathbb{E}[X_1] + \frac{z}{2}\right) \\ &\leq \mathbb{P}\left(S^{(s)}/s > \mathbb{E}[X_1] + \frac{z}{2}\right). \end{aligned}$$

By the Hoeffding-Azuma inequality, $\mathbb{P}\left(S^{(s)}/s > \mathbb{E}[X_1] + \frac{z}{2}\right) \leq \exp(-s\, z^2/2)$. Using $\sum_{s=u}^{\infty} e^{-\kappa u} \leq \kappa^{-1} e^{-\kappa u}$, which holds for any integer u and $\kappa > 0$, we obtain the desired result. ■

References

[1] Robbins H., "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, pp. 527–535, 1952.

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

[3] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[4] Anand Balachandran, Geoffrey M. Voelker, Paramvir Bahl, and P. Venkat Rangan, "Characterizing user behavior and network performance in a public wireless LAN," *SIGMETRICS Perform. Eval. Rev.*, vol. 30, no. 1, pp. 195–205, 2002.

[5] Tristan Henderson, David Kotz, and Ilya Abyzov, "The changing usage of a mature campus-wide wireless network," in *Mobicom*, 2004, pp. 187–201.

[6] Jihwang Yeo, Moustafa Youssef, and Ashok Agrawala, "A framework for wireless LAN monitoring and its applications," in *WiSe '04: Proceedings of the 3rd ACM workshop on Wireless security*, 2004, pp. 70–79.

[7] Jihwang Yeo, Moustafa Youssef, Tristan Henderson, and Ashok Agrawala, "An accurate technique for measuring the wireless side of wireless networks," in *the 2005 workshop on Wireless traffic measurements and modeling*, 2005, pp. 13–18.

[8] Maya Rodrig, Charles Reis, Ratul Mahajan, David Wetherall, and John Zahorjan, "Measurement-based characterization of 802.11 in a hotspot setting," in *Proceedings of the 2005 ACM SIGCOMM workshop on Experimental approaches to wireless network design and analysis*, 2005, pp. 5–10.

[9] Yu-Chung Cheng, John Bellardo, Péter Benkö, Alex C. Snoeren, Geoffrey M. Voelker, and Stefan Savage, "Jigsaw: solving the puzzle of enterprise 802.11 analysis," in *SIGCOMM*, 2006.

[10] Dong-Hoon Shin and Saurabh Bagchi, "Optimal monitoring in multi-channel multi-radio wireless mesh networks," in *MobiHoc*, 2009, pp. 229–238.

[11] Chandra Chekuri and Amit Kumar, "Maximum coverage problem with group budget constraints and applications," in *APPROX*, 2004, pp. 72–83.

[12] Arun Chhetri, Huy Nguyen, Gabriel Scalosub, and Rong Zheng, "On quality of monitoring for multi-channel wireless infrastructure networks," in *The ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2010.

[13] Rajeev Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, 1995.

[14] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," *Foundations of Computer Science, Annual IEEE Symposium on*, vol. 0, pp. 322, 1995.

[15] Paat Rusmevichientong and John N. Tsitsiklis, "Linearly parameterized bandits," *Math. Oper. Res.*, vol. 35, no. 2, pp. 395–411, 2010.

[16] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz, "Regret minimization under partial monitoring," *Mathematics of Operations Research*, vol. 31, pp. 562–580, 2006.

[17] Lifeng Lai, Hesham El Gamal, Hai Jiang, and H. Vincent Poor, "Cognitive medium access: Exploration, exploitation and competition," *CoRR*, vol. abs/0710.1385, 2007.

[18] Lifeng Lai, Hai Jiang, and H.V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, 26-29 2008, pp. 98 –102.

[19] Keqin Liu and Qing Zhao, "Decentralized multi-armed bandit with multiple distributed players," *CoRR*, vol. abs/0910.2065, 2009, informal publication.

[20] A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition," in *Proc. of IEEE INFOCOM*, San Deigo, USA, Mar. 2010.

[21] Yi Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, 6-9 2010, pp. 1 – 9.

[22] Jeremy Elson, Lewis Girod, and Deborah Estrin, "Fine-grained network time synchronization using reference broadcasts," *SIGOPS Oper. Syst. Rev.*, vol. 36, no. SI, pp. 147–163, 2002.

[23] 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," *Automatic Control, IEEE Transactions on*, vol. 32, no. 11, pp. 968 – 976, nov 1987.

[24] J-Y Audibert, R. Munos, and Cs. Szepesvári, "Tuning bandit algorithms in stochastic environments," in *Algorithmic Learning Theory - 18th International Conference, ALT 2007*. 2007, pp. 150–165, Springer.

[25] J.-Y. Audibert, R. Munos, and Cs. Szepesvári, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[26] P. Auer, "Using upper confidence bounds for online learning," in *Proceedings of the 41th Annual Symposium on Foundations of Computer Science*. 2000, pp. 270–293, IEEE Computer Society.

[27] V Dani, TP Hayes, and SM Kakade, "Stochastic linear optimization under bandit feedback," in *COLT-2008*, 2008, pp. 355–366.

[28] A Garivier and E Moulines, "On upper-confidence bound policies for non-stationary bandit problems," Tech. Rep., LTCI, Dec 2008.

[29] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," in *COLT 2009*, S. Dasgupta and A. Klivans, Eds., 2009.

[30] R. Agrawal, M.V. Hedge, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," *IEEE Transactions on Automatic Control*, vol. 33, no. 10, pp. 899–906, 1988.