This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the WCNC 2010 proceedings.

Distributed Cognitive Sensing for Time Varying Channels: Exploration and Exploitation

Song Gao†, Lijun Qian†, Dhadesugoor R. Vaman†, and Zhu Han*
† ARO/ARL Center for Battlefield Communications Research
Prairie View A&M University, Texas A&M University System, Prairie View, TX 77446
*Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004

***Abstract*— Spectrum under-utilization calls for the open and dynamic spectrum access mechanism, which allows the unlicensed user equipped with cognitive radios to opportunistically sense and access the spectrum that not occupied by primary users. In practice due to the hardware limitations, each cognitive radio user may be only able to sense a portion of the interested wide span spectrum. Hence, a hardware-constrained cognitive MAC to conduct efficient and intelligent spectrum sense decision is desired. In this paper, we formulate the cognitive radio spectrum sensing problem under time-varying channels as an adversarial bandit problem without any assumption of the channel statistics. A fully distributed strategy is proposed to address the fundamental tradeoff between spectrum exploration and spectrum exploitation during the sensing periods. Simulation results demonstrate that significant performance gain can be achieved by the proposed algorithm when the channels are time-varying on small time-scales. A coordination scheme for the multi-user case is also presented and the effectiveness is also demonstrated by the simulation results.**

I. Introduction

Motivated by the bandwidth scarcity and under-utilization of the spectrum, Cognitive Radios (CRs) [1], [2], which are capable of accessing the spectrum opportunistically, are under intensive research recently. In a CR network, CR users periodically sense the spectrum and identify the spectrum opportunities to communicate among themselves without interfering with the *primary* (licensed) users. Challenges arise with such dynamic means of sensing and accessing spectrum, especially for identifying the availability of time-varying channels[1], which in turn imposes design issues to the medium access control (MAC) layer. This work aims at designing efficient distributed cognitive sensing strategies for time-varying channels. Namely, we design a distributed strategy for each cognitive user to decide which channel(s) it should probe such that the most favorable channel(s) can be obtained, and avoid the interference from the primary users.

In this paper, cognitive users are assumed to operate synchronously in a time-slotted system. At the beginning of each time slot, they need to choose a candidate channel set with the best expected throughput to sense and access. In general, wireless channels are time-varying [11] and in many cases the statistics of the channels are not known as a priori knowledge. Furthermore, the behaviors of the primary users may not be known to the cognitive users or may not be predictable. Hence, it is assumed in this paper that the channels are time-varying on small time-scales and the cognitive users have *no* prior knowledge of the statistics of the channels as well as the behaviors of the primary users.

This problem becomes even more challenging in ad hoc networks where there are no centralized infrastructure to coordinate the sensing strategies among adjacent cognitive users. Furthermore, practical CRs have sensing constraints due to hardware limitations. The spectrum opportunities of interest may span a wide range of bandwidth, however, during each sensing action, a cognitive user can only scan a small portion of the spectrum.

Under the above constraints, the goal of the cognitive sensing strategy is to distributely choose which channel(s) each cognitive user should probe at the beginning of each time slot in order to optimize the returned throughput. It should be noted that the sensing decision can be enhanced by taking into account the observation history from the previous channel sensing outcomes, i.e., through seeking the fundamental tradeoff between exploitation and exploration. Exploitation refers to the immediate benefit resulting from accessing the channel with the best estimated reward; whereas exploration is the process by which the cognitive users tend to probe more channels to discover better channel opportunities. Moreover, the competition among different cognitive users over the same channel should be incorporated in the sensing decision when multiple cognitive users present in the adjacent neighborhood.

Several cognitive MAC protocols have been proposed in the literature to address the spectrum sensing issues in cognitive radio networks. Decentralized Cognitive MAC (DC-MAC) in [3] is the first work that assumes the partial sensing ability of cognitive radio in a spectrum management system and studies a joint sensing and transmission decision. However, the influence of sensing overhead for multi-channel sensing is not considered in this work. In [4], hardware-constrained cognitive MAC, HC-MAC is proposed to conduct efficient spectrum sensing and accessing decision by formulating it as an optimal stopping rule problem. The tradeoff between overhead for sensing action and expected throughput is analyzed by assuming the joint channel distributed is known by cognitive users. Collaborative channel spectrum-sensing policy is proposed in [5] for the detection of spectrum opportunities. However, spectrum band is assumed to be ON/OFF binary

[1]In this paper, spectrum and channel are used interchangeably whenever there is no confusion.

978-1-4244-6398-5/10/$26.00 ©2010 IEEE

channel model without considering the difference of channel condition to cognitive users. In [10], a distributed channel sensing and access policy base on multi-armed bandit problem is presented for cognitive users to opportunistically exploit the available channels under statistical assumption on the channels.

Comparing to the existing approaches, a more realistic scenario and thus a more challenging problem is considered in this paper where it is assumed that the channels are time-varying on small time-scales and the cognitive users have *no* prior knowledge of the statistics of the channels. An efficient cognitive sensing strategy is formulated as a *adversarial bandit* problem. The spectrum sensing scheme for single cognitive user case is presented first and then extended to multi-user scenario with considerations of coordination among peer cognitive radio users. In this work, we assume the cognitive radio users are not aware of any prior knowledge of the statistics of the channels. Simulation results demonstrate that the proposed schemes provide significant gain (up to 40%) over the existing approaches under fast time-varying channels. And comparable performance can be obtained even under slow time-varying or stationary channels.

This paper is organized as follows: In Section II, the system model is given. The proposed sensing algorithm for a single cognitive user is presented in Section III. In Section IV, the coordination algorithm for the two-user case is described. Simulation results and analysis are given in Section V. Section VI contains concluding remarks.

II. System Model

We consider a wireless cognitive radio ad hoc network consisting of J communication pairs. Both transmitter and receiver are indexed by $\mathcal{J} := \{1, 2, ..., J\}$. The entire spectrum is appropriately divided into N independent channels $\mathcal{N} := \{1, 2...N\}$ each with bandwidth B. Throughout this work, we assume that the coherent time is larger than time slot duration T_S and the bandwidth of each channel (B) is greater than the coherence bandwidth, thus the channel is quasi-static and remains invariant during each time slot. The cognitive users operate in a synchronous time-slotted fashion. Each cognitive user explores the spectrum opportunities through channel sensing and reservation on a common control channel. In this work, we use i to index channel, t refers to the time slot and j denotes the index of cognitive user. In each time slot, the emerging cognitive users first choose a candidate chanel set to sense for the spectrum opportunities and then reserve the best one for data transmission. Due to hardware limitation, we assume cognitive user can only sense K channel each time where $K < N$.

For cognitive user j in time slot t, we define $\boldsymbol{X}_j(t) = \{X_{1,j}(t), X_{2,j}(t), ...X_{N,j}(t)\}, j \in \mathcal{J}, t \geqslant 1$ as channel state random variables. $\mathcal{A}_j(t)$ is defined as the action taken by cognitive user j at time slot t, where $\mathcal{A}_j(t) \subset \mathcal{N}, j \in \mathcal{J}$, and $|\mathcal{A}_j(t)| = \Omega$ where Ω is the number of channels that will be sensed at time slot t by cognitive user j. The signal to noise ratio (SNR) observed by cognitive user j at time slot t is defined as $\boldsymbol{x}_j(t) = \{x_{1,j}(t), x_{2,j}(t), ...x_{N,j}(t)\}, j \in \mathcal{J}, t \geqslant 1, x_{i,j}(t) \in \mathbb{R}^+$. The channel state observation indicates the channel quality which determines the corresponding reward $R(x_{i,j}(t))$. If primary user is detected, the observation $x_{i,j}(t)$ will be zero. The obtained reward $R(x_{i,j}(t))$ in terms of channel state $x_{i,j}(t)$ can be expressed as

$$R(x_{i,j}(t)) = B \cdot \log_2 (1 + x_{i,j}(t)) . \tag{1}$$

In addition, if multiple cognitive users choose the same channel i to sense and access at the same time slot t, the collected reward by these cognitive users will be zero. In finite time horizon T, we define $\boldsymbol{\pi}_j$ as the policy adopted by cognitive user j for spectrum opportunity exploration. We have

$$\boldsymbol{\pi}_j = \{\mathcal{A}_j(1), \mathcal{A}_j(2), ..., \mathcal{A}_j(T)\}, j \in \mathcal{J}. \tag{2}$$

The aim is to find an optimal policy $\boldsymbol{\pi}_j^*$ for cognitive user j to maximize the expected reward during the time horizon T

$$V_{\boldsymbol{\pi}_j^*}(\boldsymbol{X}_j, T) = \max \mathbb{E}_{\boldsymbol{\pi}_j} \left[\sum_T \sum_{\mathcal{N}} R(\boldsymbol{x}_{\mathcal{A}_j(t)}) \right] . \tag{3}$$

where $\boldsymbol{x}_{\mathcal{A}_j(t)}$ is denoted as the observation of channel state $\boldsymbol{X}_j(t)$ by taking action $\mathcal{A}_j(t)$ of cognitive radio user j at time slot t, and $R(X_{\mathcal{A}_j(t)})$ is the corresponding collected reward. If channel i is not chosen by action $\mathcal{A}_j(t)$, observation of $X_{i,j}(t)$ is zero, i.e. $x_{i.j}(t) = 0$. From (3), it can be observed that the optimal policy $\boldsymbol{\pi}_j^*$ depends on the channel condition observed from $\boldsymbol{X}_j(t)$. This problem is a paradigmatic example of the trade-off between exploration and exploitation. When the cognitive user exclusively sense the channel(s) that it *thinks* best (exploitation), the cognitive user may fail to discover other channels may actually have better reward. On the other hand, if the cognitive user spends too much time trying out all the channels and gathering statistic information (exploration), the cognitive user may fail to access the best channels often enough to obtain a higher return. If the distribution of $\boldsymbol{X}_j(t)$ is deterministic, the optimal solution is trivial by choosing the channel with maximal expected value and the problem can be solved by the standard dynamic programming procedure. Furthermore, if the distribution of random variable $\boldsymbol{X}_j(t)$ is fixed but unknown, authors in [10] and [8] have proved that the regret over time horizon T, for $T \to \infty$, can approach as small as $O(\ln(T))$. The regret is defined as the performance loss entailed by policy $\boldsymbol{\pi}_j$ compared with the idealistic case where the exact channel state value of $\boldsymbol{X}_j(t)$ is known, in which case the optimal strategy for the cognitive user is always to choose channel(s) with the largest rewards. The regret is defined as:

$$L = G^* - V_{\boldsymbol{\pi}_j^*}(\boldsymbol{X}_j, T) \tag{4}$$

where G^* is the optimal collected reward when the best channel is always selected by the cognitive user at each time slot with gene-aided system. The second term of (4) represents the expected payoff under the current strategy.

It has been proved that this bound is optimal in the sense that there does not exist a strategy with a better asymptotic performance. However, fixed treatment of the channel state $\boldsymbol{X}_j(t)$

may not be adequate to model the primary traffic and channel conditions in cognitive radio ad hoc networks. Therefore, in this work we make no statistical assumptions about $\boldsymbol{X}_j(t)$ associated with each channel. We assume that each channel is initially assigned an arbitrary and unknown sequence of observed state $x_{i,j}(t)$ with corresponding throughput $R(x_{i,j}(t))$. At the beginning of each time slot, the cognitive radio user need to make a decision of choosing the candidate channel set for sensing in order to maximize the collected througput.

III. Single User Channel Probe Algorithm

In this section, we present spectrum sensing scheme for single cognitive radio user case. We assume each cognitive user selects the candidate channel set with the *best* expected throughput and conducts channel sense thereafter individually. Without loss of generality, we assume $K = 1$ in this work. The objective of each cognitive user is to maximize the collected reward (throughput) based on the history of channel selections and observations. The historical information pattern obtained through previous channel sensing is given as

$$\Psi(t) = \{\boldsymbol{x}_{\mathcal{A}_j(1)}, \boldsymbol{x}_{\mathcal{A}_j(2)}, ..., \boldsymbol{x}_{\mathcal{A}_j(t-1)}, \boldsymbol{x}_{\mathcal{A}_j(t)}\}, t \geq 1 \quad (5)$$

For a certain policy $\boldsymbol{\pi}_j$ in finite time horizon T, the expected reward the cognitive user can obtain based on the historical observation $\Psi(t)$ is expressed as

$$\begin{aligned} V_{\pi_j}(\boldsymbol{x}_j) &= \mathbb{E}_{\boldsymbol{\pi}_j}\left[\sum_{T}\sum_{\mathcal{N}} R(x_{\mathcal{A}_j(t)})\right] \\ &= \sum_{T}\sum_{\mathcal{N}} R(x_{\mathcal{A}_j(t)}) p\{\boldsymbol{\pi}_j = i|\Psi(t)\} \end{aligned} \quad (6)$$

where $p\{\boldsymbol{\pi}_j = i|\Psi(t)\}$ is the probability that cognitive user j choose channel i to sense at time slot t based on historical information pattern $\Psi(t)$ under the policy $\boldsymbol{\pi}_j$. And for the entire channel set $\mathcal{N}$, we have

$$\sum_{\mathcal{N}} p\{\boldsymbol{\pi}_j = i|\Psi(t)\} = 1 \quad (7)$$

This single cognitive radio user channel sensing problem under time-varying channels can be formulated as a *adversarial bandit* problem [7], a variant of the bandit problem in which *no* statistical assumptions are made about expected reward. Different from the deterministic bandit problem which can compute the exact expected payoffs that will be received before making the decision, adversarial bandit problem tries to strike a balance between exploration and exploitation in order to improve the returned reward instead of optimizing the expected reward.

Based on the *adversarial bandit* problem definition, cognitive spectrum sensing problem can be formulated as follows. To decide the candidate channel for sensing, each cognitive user need to pre-determine $p\{\boldsymbol{\pi}_j\}$ associated with each channel based on the observation history $\Psi(t)$ to maximize the returned reward $V_{\pi_j}(\boldsymbol{x}_j)$. For single user spectrum sensing decision rule, $p\{\boldsymbol{\pi}_j = i|\Psi(t)\}$ is defined as

$$\begin{aligned} p\{\boldsymbol{\pi}_j = i|\Psi(t)\} &= q \cdot \frac{\omega_{i,j}(t)}{\sum_{i=1}^{N}\omega_{i,j}(t)} + (1-q)\cdot\frac{1}{N} \\ \mathbf{P}\{\boldsymbol{\pi}_j|\Psi(t)\} &= \prod_{i\in\mathcal{N}} p\{\boldsymbol{\pi}_j = i|\Psi(t)\} \end{aligned} \quad (8)$$

where $\mathbf{P}\{\boldsymbol{\pi}_j|\Psi(t)\}$ is the channel selection probability distribution. $\omega_{i,j}(t)$ is denoted as the knowledg of channel i cognitive user j obtained from the observation history $\Psi(t)$. And q is the weight factor for balancing spectrum *exploration* and *exploitation*. It can be noticed that distribution $p\{\boldsymbol{\pi}_j\}$ is a mixture of the uniform distribution and a distribution which assigns to each action a probability mass exponential in the estimated cumulative reward for that action. Intuitively, mixing in the uniform distribution is to make sure that the algorithm tries out all N channels and gets good estimates of the reward for each one. Otherwise, the CR user might miss a good channel because the initial reward it observed is low and the better reward that may occur later will not be observed because the channel is less likely to be selected. At the beginning of sensing stage, the action of exploration should be preferred to that of exploitation due to the lack of the knowledge of channels. However, with the increase of the obtained channel knowledge $\Psi(t)$ exploitation should outweight exploration in order to gain better returned rewards. Therefore, the weight factors of *exploration* and *exploitation* should evolve with the increase of time slots. Thus, the proposed algorithm proposed in this work proceeds in *epochs*, where each epoch consists of a sequence of time slots (T_S). In each epoch, the weight factors keep stationary. When the weight factors evolve to new values, the algorithm enters the next epoch. The threshold for the evolution of the weight factor will be defined later. We use $r = 0, 1, 2...$ to index epochs. To maintain the implementation simplicity and tractable complexity of the proposed algorithm, the weight factor evolution rule of the presented algorithm is given as

$$\gamma_r = \alpha^r, 0 < \alpha < 1 \quad (9)$$

We define $\gamma_r = 1 - q$. It can be noted that if $\alpha = 0$, the algorithm is a random selection scheme that cognitive user treats every channel equally for probing decision in each time slot. From the definition of γ_r, it is apparent that γ_r is the tuning parameter to balance the exploitation and exploration for the channel probe decisions. At the initial phase of the algorithm, since cognitive users have no enough knowledge of channels, preference will be given to exploration by increase the weight of γ_r. With the increase of the epochs, γ_r will evolve to smaller values which means cognitive users will tend to make their decision based on historical channel observation information $\Psi(t)$, i.e exploitation.

Each cognitive user make the channel selection decision $\mathcal{A}_j(t)$ based on the probability distribution $\mathbf{P}\{\boldsymbol{\pi}_j|\Psi(t)\} = \{p_{1,j}(t), p_{2,j}(t), ...p_{N,j}(t)\}$ which means the probability that cognitive user j chooses channel i for sensing and acess is determined by $p_{i,j}(t)$. Larger $p_{i,j}(t)$ will result in bigger chance that the channel i will be chosen by cognitive user

TABLE I
PROPOSED SINGLE USER ALGORITHM

Algorithm:
Initialization: $\hat{G}_i(1) = 0$ for $i = 1, ..., N$ at time slot $t = 1$
Initialization: $\omega_i(1) = 1$ for $i = 1, ...N$ at time slot $t = 1$
Repeat for $r = 0, 1, 2...$
1. Set g_r according to (12)
2. Set $\gamma_r = \alpha^r$
3. While $\max_i \hat{G}_i(t) \leq g_r$do:
   (a). Set $p_i(t) = (1-\gamma_r)\frac{\omega_i(t)}{\sum_{i=1}^{N}\omega_i(t)} + \frac{\gamma_r}{N}$ for $i = 1, ...N$
   (b). Take action $\mathcal{A}(t)$ according to $p_1(t), p_2(t)..., p_N(t)$
   (c). Collect reward $R(x_{\mathcal{A}(t)})$ for $i \in \mathcal{N}$
   (d). For $i = 1, ...N$ set
   $\hat{R}_i(t) = \frac{R(x_i(t))}{p_i(t)}$ if $i \in \mathcal{A}(t)$ and 0 for $i \notin \mathcal{A}(t)$
   $\omega_i(t+1) = \omega_i(t) \cdot \exp(\gamma_r \cdot \frac{\hat{R}_i(t)}{N})$
   (e). $\hat{G}_i(t+1) = \hat{G}_i(t) + \sum_{i=1}^{N} \hat{R}_i(t)$ for $i = 1, ...N$
   (f). $t := t + 1$
4. Otherwise, enter the next epoch $r := r + 1$

j. Given the channel selection action $\mathcal{A}_j(t)$ and observation $x_{\mathcal{A}_j(t)}$, the update algorithm of $\omega_{i,j}(t)$ is given as

$$\begin{aligned} \omega_{i,j}(t+1) &= \omega_i(t) \cdot \exp(\gamma_r \cdot \frac{\hat{R}_i(t)}{N}) \quad \text{for} \quad t > 1 \\ \hat{R}_{i,j}(t) &= \frac{R(x_{i,j}(t))}{p_{i,j}(t)} \end{aligned} \tag{10}$$

At the first time slot, $\omega_i(1) = 1$ for $i = 1, 2, ...N$ since no knowledge of the channels is available. For action $\mathcal{A}_i(t)$ in time slot t, the presented algorithm sets the estimated reward $\hat{R}_i(t)$ to $R(x_i(t))/p_i(t)$. Dividing the observed gain by the probability that the channel was chosen compensates the reward of the channel that are unlikely to be chosen. This choice of estimated rewards guarantees that their expectations are equal to the actual rewards for each channel, that is,

$$\mathbb{E}[\hat{R}_i(t)|\mathcal{A}_j(1), \mathcal{A}_j(2), ...\mathcal{A}_j(t-1)] = R(x_i(t)). \tag{11}$$

where the expectation is taken with respect to the action of $\mathcal{A}_j(t)$ at trial t given the actions $\mathcal{A}_j(1), \mathcal{A}_j(1), ..., \mathcal{A}_j(1)$ in the previous $t-1$ trials.

For each epoch r, the algorithm needs to set a reward bound g_r to tune the parameter γ_r at the beginning of the epoch. In this work, we adopt the definition of g_r in [7] which is given as

$$g_r = \frac{N \cdot \ln N}{(e-1)} 4^r. \tag{12}$$

g_r is the expected reward which acts as the performance threshold for epoch r. The proposed algorithm for the single user channel probe problem is illustrated in Table I (For simplicity, we drop user index j in the algorithm).

In Table I, the estimation of the reward $\hat{G}_i(t)$ for channel i is unbiased in the sense that

$$\mathbb{E}[\hat{G}_i(t+1)] = \mathbb{E}\left[\sum_{s=1}^{t} \hat{R}(x_i(s))\right] = \sum_{s=1}^{t} R(x_i(t)). \tag{13}$$

Using this estimation, the proposed algorithm can detect when the actual gain goes beyond g_r. When it happens, the proposed algorithm enters the next epoch and resets the tuning parameter γ_r and g_r with a larger bound on the maximal gain.

Lemma 1: The proposed algorithm yields the expected regret of $\mathcal{O}(\sqrt{G^*})$ uniformly over T, i.e. $\mathbb{E}[L(\boldsymbol{X}_j; \boldsymbol{\pi}_j)] \sim \mathcal{O}(\sqrt{G^*})$

Proof: Lemma 1 can be directly derived from theorem (3.1) and (4.1) in [7]. ■

Since $\sqrt{G^*}$ is usually difficult to estimate, a loose bound for the expected regret is given in Lemma 2. To estimate the upper bound of the returned reward over time horizon T, we define the maximal reward can be obtained by cognitive user at each time slot as $\tilde{R}$ which corresponds to the best channel quality cognitive user can expect

$$\tilde{R} = R(\max\{x_i(t)\}), t \in T, i \in \mathcal{N} \tag{14}$$

Thus, $\sum_{t=1}^{T} \mathbf{R}(\boldsymbol{X}_{\boldsymbol{A}(t)}) \leq G^* \leq T \cdot \Omega \cdot \tilde{R}$.

Lemma 2: The proposed algorithm yields the expected regret of $\mathcal{O}(\sqrt{T \cdot \Omega \cdot \tilde{R}})$ uniformly over T, i.e. $\mathbb{E}[L(\boldsymbol{X}_j; \boldsymbol{\pi}_j)] \sim \mathcal{O}(\sqrt{T \cdot \Omega \cdot \tilde{R}})$ which is always loose than $\mathcal{O}(\sqrt{G^*})$

Proof: From (14) and Lemma 1, $\tilde{R}$ is the maximal collected reward for cognitive user to choose any channel at any time slot t, which is always larger or equal to the actual channel reward $R(x_i(t))$. Thus during the finite horizon time T, the normalized collected reward is $\sum_T \mathbf{R}(\boldsymbol{X}_{\boldsymbol{A}(t)}) \leq G^* \leq T \cdot \Omega \cdot \tilde{R}$, which leads to the conclusion. ■

IV. MULTI-USER COOPERATIVE CHANNEL PROBE SCHEME

In this section, we consider the scenario that multiple cognitive users present in adjacent area. If multiple cognitive users make channel probe decision independently at the beginning of each time slot, they may choose same channel(s) to sense and access at the same time slot. When such collision happens, cognitive users who made same decision will not be able to obtain any throughput (we assume the channel can not be shared among cognitive users in this work). Therefore, a cooperative channel probe scheme is desired in multi-user environment, especially when the multiple cognitive users possess similar spectrum perception.

In this work, we assume a common control channel is available where the cognitive users collaborate with each other as described later. In order to avoid possible *collision* among peer cognitive users, cognitive users start a reservation phase before the beginning of sensing period in each time slot. In the reservation period, cognitive users will start to broadcast their action and the corresponding channel knowledge $\{\mathcal{A}_j(t), \omega_{i,j}(t)\}$ in a reservation package on the common control channel. The peer cognitive users who have made the same channel selection decision will compare the received

TABLE II
MULTI-USER CHANNEL PROBE SCHEME

Algorithm:
...
...
While $\max_i \hat{G}_{i,j}(t) \leq g_r$ do:
(a). Set $p_{i,j}(t) = (1-\gamma_r)\frac{\omega_{i,j}(t)}{\sum_{i=1}^{N}\omega_{i,j}(t)} + \frac{\gamma_r}{N}$ for $i = 1, ...N$
(b). Take action $\mathcal{A}_j(t)$ according to $p_{1,j}(t), ...p_{N,j}(t)$, for $i \in \mathcal{N}$
(c). Broadcast $\{\mathcal{A}_j(t), \omega_{i,j}(t)\}$ on common control channel
(d). Listen on the common control channel for $\{\mathcal{A}_j(t), \omega_{k,j}(t)\} k \neq i$
(e). For the same selection from peer cognitive users
If received $\omega_{k,j}(t) > \omega_{i,j}(t)$, then $\mathcal{N} = \mathcal{N} - \mathcal{A}_j(t)$, go to step (b)
Otherwise, go to step (d)
(f). Enter the spectrum sensing and access phase
(g). Collect reward $R(x_{\mathcal{A}(t)})$ for $i \in \mathcal{N}$
...
...
Otherwise, enter the next epoch $r := r + 1$

$\omega_{k,j}(t), k \neq i$ with its own $\omega_{i,j}(t)$. If the peer cognitive user has better estimated channel reward, it will broadcast its own action and the corresponding channel knowledge $\{\mathcal{A}_j(t), \omega_{i,j}(t)\}$ on the common control channel, otherwise it will restrain from current decision and re-select the candidate channel based on $\mathbf{P}\{\boldsymbol{\pi}_j|\Psi(t)\}$. Therefore, before cognitive users switch to the selected channel for sensing and access, all of them have broadcasted their decision and are aware of the decisions from their neighbor cognitive users. Note that there may be multiple transmissions from different cognitive users in the reservation phase, and the contention can be resolved by using the backoff mechanism as in the conventional IEEE 802.11 DCF mode.

The channel probe algorithm in multi-user scenario is illustrated in Table II. In Table II, only the difference between single user probe algorithm and multi-user probe algorithm is highlighted due to the page limitations.

V. SIMULATION RESULTS

In this section, we evaluate performance of the proposed channel probe algorithm under different channel conditions. We compared the performance of the proposed algorithm with three other schemes, including UCB algorithm presented in [10], random selection scheme, and optimal upper bound obtained through gene-aided system which can be taken as the optimal upper bound. Moreover, the performance of the cooperative sensing scheme in multi-user scenario is also evaluated, the performance gain compared with non-cooperative approach is demonstrated.

In Figure 1 and Figure 2, we show the accumulated returns of the four algorithms over 5000 time slots for doppler shift equal to 10Hz and 110Hz, respectively. The corresponding system parameters adopted for these two figures are summarized in Table III. It can be noted from Figure 2 which is the fast fading scenario ($f_m = 110$Hz), the proposed algorithm has almost 40% performance gain over the algorithm proposed in [10] due to the fast adaptation of the proposed algorithm to

TABLE III
UNITS OF SYSTEM PARAMETERS

Symbols	*Description*	*Value*
f_m	Doppler shift	110Hz / 10 Hz
σ_τ	rms delay spread	10^{-7}s
N	Number of channels	128
B	Bandwidth of each channel	10KHz
f_S	Symbol rate	10^6 symbol/s

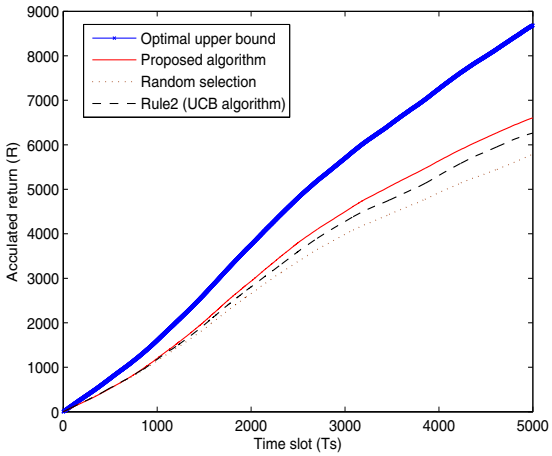

Fig. 1. Performance Comparison, Doppler Shift=10Hz

the varying channels which results from the evolving factor γ_r. And when the doppler shift is large, the algorithm in [10] performs no much better than the random selection due to the fact that the sensing decision of the algorithm in [10] is largely determined by the observation history (exploitation) instead of seeking the balance of exploration and exploitation. As shown in Figure 1, the proposed algorithm still has comparable performance with the UCB in [10] in the slow fading scenario ($f_m = 10$Hz) even though the performance gain is not as significant as in the fast fading scenario. Moreover, random selection always performs worst which is taken as lower bound in the simulation, since the cognitive user treats each channel equally without any consideration of the observation history. Compared with the optimal upper bound which is obtained through gene-aided system, the proposed scheme still has some performance loss, which is the inevitable cost for distributed learning.

In Figure 3, we show a window for the collected reward over time slots. We can see that the optimal selection scheme and UCB algorithm in [10] have a big performance gap and the proposed algorithm always outperforms UCB algorithms. The propose algorithm tries to track the channel changes by striking a balance between exploration and exploitation, while occasionally collapses due to the cost of the exploration learning.

In Figure 4, we show the performance evaluation of the multi-user cooperative algorithm in two user case. Compared with the non-cooperative case in which two users will not exchange the selection information on the common channel,

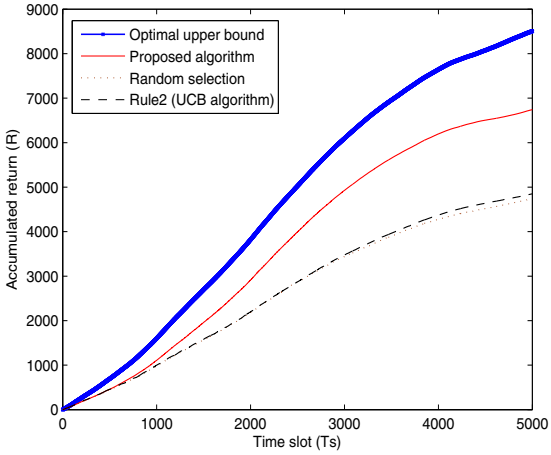

Fig. 2. Performance Comparison, Doppler Shift=110Hz

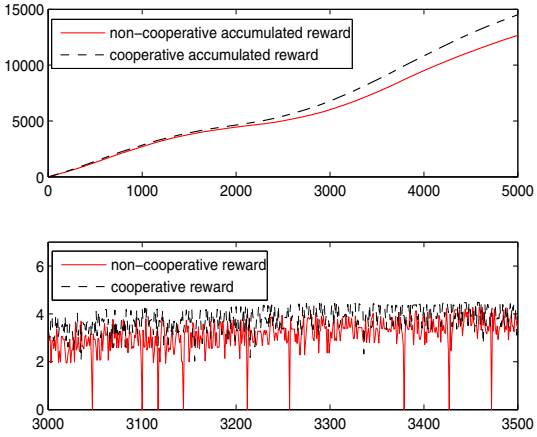

Fig. 4. Performance Evaluation between Single User Algorithm and Two User Cooperative Algorithm

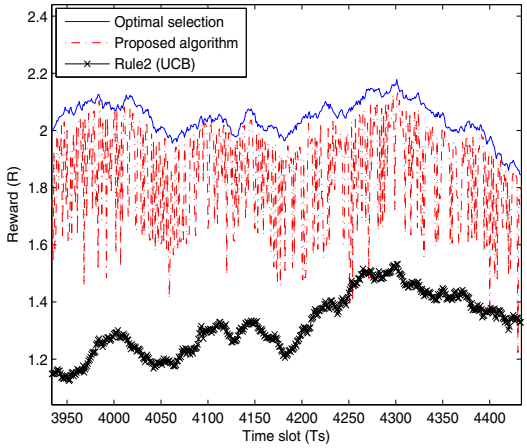

Fig. 3. Collected Reward Return over Time Slots

the proposed cooperative multi-user algorithm has better performance. This is because with the limited information shared between two users, possible collision will be avoided and the cognitive user with better estimated channel reward will obtain the candidate channel for sensing and access. We also show a window for the rewards over time.

VI. Conclusion

In this paper, we propose a fully distributive cognitive radio spectrum sensing algorithm under time varying channels. The single-user algorithm is formulated as an adversarial bandit problem, where the sensing and access decision is made through seeking a balance between exploration and exploitation. The multi-user cooperative scheme is also proposed to address the possible collision that may occurred among peer cognitive users through the exchange the channel selection information on common control channel. From the simulation results, the proposed single-user scheme has significant performance gain under fast fading channels compared with the existing algorithms. And the effectiveness of multi-user cooperative scheme is also demonstrated in the simulation. A large scale simulation experiment for the multi-user case will be carried out soon.

VII. Acknowledgment

This research work is supported in part by the U.S. Army Research Office under Cooperative Agreement No. W911NF-04-2-0054. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U. S. Government

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal," *IEEE Pers. Commun.*, vol. 6, pp. 13–18, Aug. 1999.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, pp. 201–220, Feb. 2005.

[3] Q. Zhao, L. Tong, and A. Swami, "Decentralized Cognitive MAC for Dynamic Spectrum Access", *IEEE Dyspan*, November 2005.

[4] J. Jia, Q. Zhang, and X. Shen, "HC-MAC: A Hardware-Constrained Cognitive MAC for Efficient Spectrum Management", *IEEE JSAC*, vol. 26, NO. 1, pp. 106-117, January 2008.

[5] H. Su, and X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Prvovisionings Over Cognitive Radio Wireless Networks", *IEEE JSAC*, vol. 26, NO. 1, pp. 118-129, January 2008.

[6] E. Hossain, D. Niyato, and Z. Han, *Dynamic Spectrum Access in Cognitive Radio Networks*, Cambridge University Press, UK, 2009.

[7] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The non-stochastic multi-armed bandit problem", *SIAM Journal on Computing*, to appear. Available: http://www.cs.ucsd.edu/ yfreund/papers/bandits.pdf

[8] H. Robbins, "Some aspects of the sequential design of experiments", *American Mathematical Society*, vol. 55, pp. 527-535, 1952.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite time analysis of the multiarmed bandit problem", *Machine Learning*, vol. 47, pp. 235-256, 2002. Kluwer Academic Publishers.

[10] L. Lai, H. E. Gamal, H. Jiang, and H. V. Poor, "Cognitive Medium Access: Exploration, Exploitation and Competition", *IEEE/ACM Transactions on Networking*, revised July 2008

[11] T.S. Rappaport, *Wireless Communications*, Prentice Hall, 2002.