Title	Learning Temporal–Spatial Spectrum Reuse
Author(s)	Zhang, Yi; Tay, Wee Peng; Li, Kwok Hung; Esseghir, Moez; Gaiti, Dominique
Citation	Zhang, Y., Tay, W. P., Li, K. H., Esseghir, M., & Gaiti, D. (2016). Learning Temporal–Spatial Spectrum Reuse. IEEE Transactions on Communications, 64(7), 3092-3103.
Date	2016
URL	http://hdl.handle.net/10220/43462
Rights	© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: [http://dx.doi.org/10.1109/TCOMM.2016.2569093].

Learning Temporal-Spatial Spectrum Reuse

Yi Zhang, *Student Member, IEEE*, Wee Peng Tay, *Senior Member, IEEE*, Kwok Hung Li, *Senior Member, IEEE*, Moez Esseghir, *Member, IEEE* and Dominique Gaïti, *Member, IEEE*

***Abstract*—We formulate and study a multi-user multi-armed bandit (MAB) problem that exploits the temporal-spatial opportunistic spectrum access (OSA) of primary user (PU) channels so that secondary users (SUs) who do not interfere with each other can make use of the same PU channel. We first propose a centralized channel allocation policy that has logarithmic regret, but requires a central processor to solve an NP-complete optimization problem at exponentially increasing time intervals. To overcome the high computation complexity at the central processor, we also propose heuristic distributed policies that however have linear regrets. Our first distributed policy utilizes a distributed graph coloring and consensus algorithm to determine SUs' channel access ranks, while our second distributed policy incorporates channel access rank learning in a local procedure at each SU at the cost of a higher regret. We compare the performance of our proposed policies with other distributed policies recently proposed for temporal (but not spatial) OSA. We show that all these policies have linear regrets in our temporal-spatial OSA framework. Simulations suggest that our proposed policies have significantly smaller regrets than the other policies when spectrum temporal-spatial reuse is allowed.**

***Index Terms*—Cognitive radio, spectrum reuse, multi-armed bandit.**

I. Introduction

Static spectrum allocation has been shown to result in spectrum under-utilization [3]. In cognitive radio networks (CRNs), opportunistic spectrum access (OSA) alleviates the problem by allowing unlicensed secondary users (SUs) to identify and exploit the unused spectrum owned by primary users (PUs) opportunistically while limiting the interference to PUs below a predefined threshold. OSA finds application as the underlying communication paradigm in sensor networks and the Internet of Things [4]–[8].

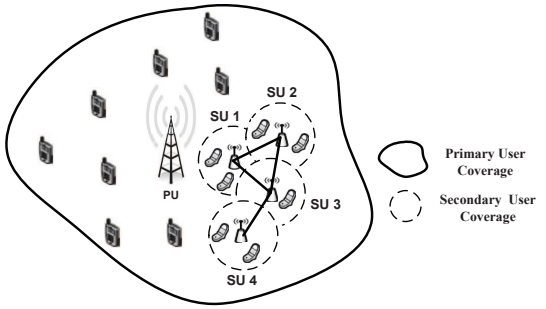
OSA has been extensively studied at the physical (PHY) and medium access control (MAC) layers, and various temporal [9]–[13], spatial [14]–[20], or temporal-spatial [21]–[26] spectrum-sensing algorithms have been proposed to detect and utilize spectrum opportunities temporally and spatially with acceptable interference to PUs. In [20], a two-phase cooperative spectrum sensing algorithm based on low-rank matrix completion is proposed to efficiently detect the spectrum opportunities by utilizing the spatial diversity of multiple SUs. To exploit the temporal and spatial diversities in CRNs, in [23], [24], cooperative spectrum sensing methods are proposed to allow SUs to determine if they fall within the PU radio coverage region, while in [21], [25], [26], collaborative boundary estimation methods are developed to allow SUs to estimate the PU radio coverage region. In [21], a cooperative boundary detection scheme is proposed, which intelligently incorporates the cooperative spectrum sensing concept and the recent advances in support vector machine (SVM) [27]. In [26], a nonlinear SVM algorithm is implemented to perform irregular coverage boundary detection of a licensed digital TV transmitter, which then helps to build a location-specific TV white space database for opportunistic spatial reuse. However, all the before mentioned works do not address the sharing of PU spectrum by SUs inside the PU radio coverage region when the PU is idle. To study the interactions among SUs in a distributed manner, game theory has been used to design efficient distributed OSA mechanisms [28]–[32]. However, most of these works do not exploit *spatial* spectrum reuse and assume that each SU interferes with every other SU in the CRN. To allow for spatial spectrum sharing amongst the SUs, graphical game algorithms have formulated spatial reuse of the spectrum as a local bargaining process [33], a local minority game [34], a pricing game [35], a spatial congestion game [36] or a MAC-layer interference minimization game [37]. However, these works assume that some information about PUs like their locations or PU channel idle probabilities are known by all SUs.

Multi-armed bandit (MAB) techniques have been applied for OSA when PU channel information is unknown. A no-regret learning method was proposed in [29], assuming that the channel selection of each SU is known among all SUs. Accordingly, the learning problem could be simplified as a single-user MAB problem. Several distributed multi-user MAB policies have also been proposed when the reward of each SU is not observable by other SUs. Each SU needs to sense and access channels by learning the channel states independently. In [38], all SUs are assumed to interfere with one another, and a time-division fair share (TDFS) policy was used to orthogonally divide SUs temporally. The TDFS policy was shown to achieve logarithmic regret under spectrum temporal reuse. Instead of partitioning SUs into different time slots, a distributed channel access policy has been proposed in [39] to incorporate adaptive randomization to subdivide SUs into different channels. The total regret is also order-optimal logarithmic. The work in [40] considers the setting where SUs have prioritized rankings and proposed a distributed policy based on the well-known UCB1 policy [41], which yields a uniformly logarithmic regret over time. In [42], the same channel yield different rewards for different SUs. By

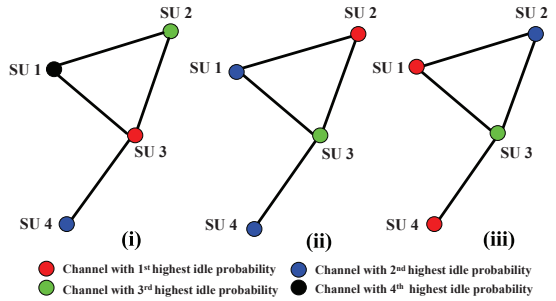
Parts of this paper were presented at the IEEE Global Conference on Signal and Information Processing, Dec. 2014 [1], and the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2016 [2]. This research is supported in part by the Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2014-T2-1-028.

Y. Zhang, W. P. Tay, and K. H. Li are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore (email: yzhang29@e.ntu.edu.sg, {wptay, ekhli}@ntu.edu.sg)

M. Esseghir and D. Gaïti are with the Charles Delaunay Institute, University of Technology of Troyes, France (email: {moez.esseghir, dominique.gaiti}@utt.fr).



(a) An edge between two SUs indicate that they interfere with each other. When a channel is sensed idle, SUs 1, 2 and 3 cannot use it simultaneously. However, SUs 1 and 4 can reuse the same channel.



(b) Three possible channel allocations.

Fig. 1. Spatial spectrum reuse in a CRN with multiple SUs.

embedding a bipartite matching algorithm, an on-line index-based distributed learning policy was proposed to achieve order $\log^2 n$ regret over time horizon n. All these methods assume that all SUs interfere with each other if they use the same channel, and spatial reuse of channels was not addressed.

In this paper, we investigate temporal-spatial spectrum reuse in a CRN, where PU channel idle probabilities are unknown to the SUs. This can be formulated as a multi-user MAB problem. SUs need to perform spectrum sensing for temporally reusing the spectrum without harmful interference to PUs. We also consider the case where SUs are geographically dispersed over a large area, and that not all SUs interfere with each other when using the same channel. If SUs are constrained to using different channels at the same time due to interference between them, as assumed in [29], [38], [39], then the optimal allocation is to assign each SU a different channel with the best availability. However, if SUs are spatially separated so that it is possible for some SUs to share the same channel without significant interference with each other, then it becomes optimal for some SUs to share the same channels with the highest idle probabilities. For example, consider the scenario depicted in Figure 1, where the expected network reward is given by the expected total number of interference-free channel uses by the SUs. Because SU 1 and SU 4 do not interfere with each other, they can be assigned the same PU channel with the highest idle probability. It can be shown that scenario (iii) in Figure 1(b) achieves the highest expected network reward. To the best of our knowledge, this is the first paper to consider temporal-spatial spectrum reuse in a MAB formulation.

We say that a SU has been allocated a channel access rank k if it is assigned to use only the k-th best channel (in terms of idle probability). Our main contributions are the following:

1) We propose a centralized policy that uses a central processor to optimize the channel access ranks of the SUs at exponentially increasing time intervals, based on the idle probability estimates of an arbitrary SU. We call this the Centralized Channel Allocation (CCA) policy.
2) To overcome the requirement for a central processor, we propose a distributed three-stage policy to enable SUs to learn their channel access ranks and the channel idle probabilities. In our proposed policy, we adopt distributed graph coloring, consensus and ϵ-greedy learning approaches. We call this the Collaborative Access Ranking and Learning (CARL) policy.
3) To avoid the need for SU synchronization and the overhead incurred by the CARL policy, we also propose a distributed channel learning and allocation policy that integrates the first two stages of CARL into the ϵ-greedy learning process. We call this the Distributed Access Rank Learning (DARL) policy.
4) We provide theoretical bounds on the regrets achieved by our proposed CCA, CARL and DARL policies, the random access policy [39], the time-division fair sharing (TDFS) policy [38], and the adaptive randomization policy [39]. We show the CCA policy has logarithmic regret, while all other policies have linear regret. We provide simulation results to verify the performance of our proposed CCA, CARL and DARL policies. Our simulation results suggest that CCA, CARL and DARL perform significantly better in terms of average regret than the random access policy, the TDFS policy and the adaptive randomization policy.

This paper focuses on spatial spectrum reuse, in contrast to other MAB formulations for spectrum reuse like those in [38], [39], which only considers temporal spectrum reuse. However, in a practical CRN, both temporal and spatial spectrum reuse need to be implemented in order to maximize the spectrum usage; and in many cases (e.g., when using distributed spectrum sensing methods) spatial spectrum reuse cannot be achieved independently of temporal spectrum reuse due to lack of prior information about interference between different SUs and PU channel idle probabilities. Therefore, our algorithms target *temporal-spatial* spectrum reuse. We note that the algorithms in [38], [39] cannot be easily adapted for spatial spectrum reuse since solving the spatial spectrum reuse problem exactly even when channel idle probabilities are known, is a NP-complete problem, as explained later in Section II.

The rest of this paper is organized as follows. In Section II, we introduce our system model and problem formulation. In Section III, we propose a centralized channel allocation policy and we show that the regret of this policy is order-optimal with logarithmic regret if the channel access rank optimization procedure is performed at exponentially increasing time intervals. In Section IV, we propose two distributed learning policies that enable SUs to find their channel access ranks and

independently learn the channel statistics. We then provide simulation results in Section V and conclude in Section VI.

II. Problem Formulation

Suppose that there are $M \geq 2$ SUs, and let $\mathcal{M}$ denote the set of SUs. We model the SU network as a graph $G = (V, E)$, where V is the set of SUs, and E is a set of edges. Two SUs are connected by an edge if the mutual interference between them is above a predefined threshold. If two SUs are not connected via an edge, then we assume that they can utilize the same PU channel simultaneously.

Let $\mathcal{N}$ be the set of $N \geq 1$ orthogonal PU channels. We divide time into equal intervals. In each time slot n, each channel $j \in \mathcal{N}$ is idle with probability $\mu_j \in (0, 1]$, independent of all other channels. Without loss of generality, we assume that $\mu_1 > \mu_2 > \ldots > \mu_N$ (SUs are not aware of this ordering). For each j, we use $S_j(n)$ to denote the channel state of a channel j in time slot n with $S_j(n) = 1$ if the channel j is idle and 0 otherwise. To simplify our formulation and algorithm descriptions, we include a sufficient number of null channels each with idle probability zero and let $\mathcal{N}^+$ be the set of channels $\mathcal{N}$ augmented with the null channels such that $|\mathcal{N}^+| \geq M$. A SU allocated a null channel simply means that it is not able to perform OSA in practice. This allows us to include cases where there are insufficient number of PU channels to be allocated to all the SUs.

Since the PU channel idle probabilities are unknown to all SUs, each SU needs to learn them through their sensing observations. In each time slot n, each SU can only sense and access one channel. Let $X_{i,j}(n) = 1$ if SU i chooses channel $j \geq 1$ and senses that it is idle, and $X_{i,j}(n) = 0$ otherwise. We assume that channel sensing is perfect for all SUs so that $X_{i,j}(n) = S_j(n)$ if channel j is chosen by SU i at time slot n.

Let $Y_{i,j}(n)$ be the reward of a SU i from accessing a channel $j \in \mathcal{N}^+$ in slot n after sensing it free. Let $\mathcal{M}_i$ be the set of neighboring SUs of the SU i in the graph G, not including SU i itself. If any SU $k \in \mathcal{M}_i$ uses the same channel as that of SU i, packet collisions occur. We adopt the following reward model:

$$Y_{i,j}(n) = \begin{cases} 1 & \text{if channel } j \text{ is idle and no other } k \in \mathcal{M}_i \\ & \text{transmits over it in the same time slot } n, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

This reward model is similar to one used in [38], [39], the only difference being that $\mathcal{M}_i = \mathcal{M}$ for all i in that model, i.e., every SU interferes with each other when using the same channel.

We are interested to design a policy ψ to learn the channel idle probabilities so as to maximize the total expected number of successful transmissions of all SUs by exploiting spatial channel reuse among SUs. The policy ψ is a rule that determines which channel $\psi_i(n) \in \mathcal{N}$ SU i chooses to sense in time slot n. The choice $\psi_i(n)$ can be made based on the sensing results of SU i at previous time slots $1, 2, \ldots, n-1$, and on the previous channel choices $\{\psi_i(l) : \text{ for } l < n\}$. If the channel $\psi_i(n)$ is idle, SU i will transmit over the channel. At the end of each time slot n, each SU is assumed to know whether it has transmitted successfully or not (e.g., through an acknowledgment from the SU receiver). Let $T_{i,j}(n)$ be the total number of time slots that the SU i has sensed the channel j in n time slots, and let $V_{i,j}(n) = \sum_{l \leq n} Y_{i,j}(l)$ be the total number of time slots that the channel j is successfully accessed by SU i up to time slot n. After n time slots, the total network reward is $\sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}^+} \mu_j \mathbb{E}[V_{i,j}(n)]$, which is proportional to the network throughput.

The regret of the policy ψ until time slot n is defined as the difference between the total reward of a genie-aided rule and the expected reward of all SUs given by

$$R(n, \psi) = n \sum_{i \in \mathcal{M}} \mu_{\pi^*(i)} - \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}^+} \mu_j \mathbb{E}[V_{i,j}(n)], \tag{2}$$

where $\pi^* : \mathcal{M} \mapsto \mathcal{N}^+$ is an optimal channel allocation if all channel idle probabilities are known. Let x^*_{ij} be an indicator variable with value 1 iff $\pi^*(i) = j$, i.e., $x^*_{ij} = 1$ iff SU i is allocated channel j. Then $(x^*_{ij})_{i \in \mathcal{M}, j \in \mathcal{N}^+}$ is an optimal solution to the following optimization problem:

$$\text{(P0)} \quad \max_{x_{ij}} \quad \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}^+} x_{ij} \mu_j \tag{3}$$

$$\text{s.t.} \quad x_{ij} + \sum_{k \in \mathcal{M}_i} x_{kj} \leq 1, \ \forall i \in \mathcal{M}, j \in \mathcal{N}^+, \tag{4}$$

$$\sum_{j \in \mathcal{N}^+} x_{ij} \leq 1, \ \forall i \in \mathcal{M}, \tag{5}$$

$$x_{ij} \in \{0, 1\}, \ \forall i \in \mathcal{M}, j \in \mathcal{N}^+. \tag{6}$$

In the above optimization problem, x_{ij} is an indicator variable that takes value one if channel j is allocated to SU i. The constraint (4) ensures that neighboring SUs do not use the same channel, while the constraint (5) ensures that each SU is allocated at most one PU channel.

(P0) is an integer linear program, which corresponds to a NP-complete decision problem (of which finding if there exists an independent set in the graph G for a given size is a special case [43]). In general, even if the channel idle probabilities are known a priori, it is analytically difficult for a genie to find the optimal channel allocations. To ensure that optimization is done within a reasonable amount of time, the genie can adopt an approximate method [44], which however leads to a linear regret as the number of time slots $n \to \infty$. For a distributed policy that does not know the channel idle probabilities a priori, the problem is even harder, and in general we cannot hope to learn the channel probabilities and an optimal channel allocation with sub-linear regret, unlike other MAB problems in which logarithmic regrets are common [38], [39].

For any channel allocation $\pi : \mathcal{M} \mapsto \mathcal{N}^+$, we say that $\pi(i)$ is the *channel access rank* of SU i because of the assumption that $\mu_1 > \mu_2 > \ldots > \mu_N$. Our main idea is to learn the optimal channel access rank of each SU and the idle probability of each channel in order to optimize the regret.[1] In the following, we propose a centralized policy that we show to have asymptotic logarithmic regret, but requires solving an

[1]Channel access rank corresponds to spatial spectrum reuse, while the channel idle probability corresponds to temporal spectrum reuse.

analytically difficult optimization problem like (P0) at exponentially increasing time intervals. We also propose heuristic distributed policies, which have linear regret in general. We compare our distributed policies to those in [38], [39] when applied to an incomplete graph G, and show that those policies also have linear regrets since they do not consider spatial reuse of channels. Our simulation results however indicate that our distributed policies have smaller regret than the existing ones in [38], [39].

For the convenience of the reader, we list some commonly used notations in Table I. Some of these notations have been defined in this section, while the remaining ones will be defined formally in the sequel where they first appear. In addition, for non-negative functions f and g, we say that $f(n) \in O(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) < \infty$, $f(n) \in \Omega(g(n))$ if $\liminf_{n\to\infty} f(n)/g(n) > 0$, and $f(n) \in \Theta(g(n))$ if $f(n) \in O(g(n))$ and $f(n) \in \Omega(g(n))$. The notation $\mathcal{P}(A)$ is the probability of the event A.

TABLE I
SUMMARY OF NOTATIONS USED

Symbol	Definition
$\mathcal{M}$	set of M SUs
$\mathcal{N}$	set of N channels
μ_j	idle probability of channel j
$X_{i,j}(n)$	sensing decision of SU i for channel j in time slot n
$Y_{i,j}(n)$	reward of SU i from accessing channel j in time slot n
$R(n,\psi)$	regret of policy ψ until time slot n
$T_{i,j}(n)$	number of time slots SU i has sensed channel j up to time slot n
$V_{i,j}(n)$	number of time slots SU i has successfully accessed channel j up to time slot n
Δ_1	lower bound for $\min_{1\le j<N} \lvert\mu_j - \mu_{j+1}\rvert$
$r_i(n)$	channel access rank of SU i in time slot n
$\rho_i(n)$	channel sensed by SU i in time slot n

III. Centralized Channel Allocation Policy

In this section, we propose a centralized policy ψ^{CCA}, and show that it has asymptotic log regret. We assume that there is a central processor in the CRN capable of solving problem (P0) with the true channel idle probabilities μ_j, $j = 1,\ldots,N$, replaced by empirical estimates from an arbitrary SU. We call this optimization problem $(\widehat{\text{P0}})$. However, since $(\widehat{\text{P0}})$ corresponds to a NP-complete decision problem, we suppose that the central processor only performs this optimization at specific irregular time instances (see Figure 2) instead of at every time slot. For a time horizon n, let t_k, $k = 1,\ldots,\xi(n)$, be the $\xi(n)$ time instances at which the central processor solves $(\widehat{\text{P0}})$ with updated empirical estimates of μ_j, $j = 1,\ldots,N$. We assume $t_1 < \infty$, i.e., there is at least one optimization time instance.

For $k = 0,\ldots,\xi(n)$, let $l_k = t_{k+1} - t_k$, where $t_0 = 1$ and $t_{\xi(n)+1} = n$, be the number of time slots starting from the k-th optimization up to before the next optimization by the central processor. Let $\bar{X}_{i,j}(n) = \sum_{k=1}^{n} X_{i,j}(k)/T_{i,j}(n)$ be the empirical estimate of the idle probability of channel j by SU i.

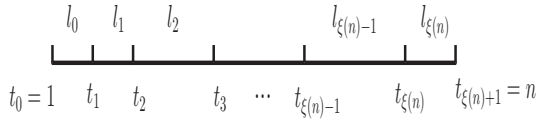


Fig. 2. Optimization time instances for ψ^{CCA}.

At each time instance t_k, $k = 1,\ldots,\xi(n)$, an arbitrarily chosen SU i sends $\{\bar{X}_{i,j}(t_k) : j \in \mathcal{N}\}$ to the central processor, which replaces μ_j with $\bar{X}_{i,j}(t_k)$ in problem (P0), and finds the optimal or near-optimal solution using the branch and bound algorithm [44]. Let $\{r_i(t_k) : i \in \mathcal{M}\}$ be the optimal channel access ranks found by the central processor (i.e., $r_i(t_k) = j$ iff $x_{ij} = 1$ in the solution of $(\widehat{\text{P0}})$). These ranks are then communicated to the SUs, which utilizes their assigned ranks in a local random ϵ-greedy channel learning algorithm: In each time slot $t \in [t_k, t_{k+1})$ of the channel learning period, each SU i chooses to sense a random channel in $\mathcal{N}$ (not $\mathcal{N}^+$ since there is no need to learn the null channels) with probability ϵ_t, and with probability $1-\epsilon_t$ chooses the $r_i(t_k)$-th best channel according to its empirical idle probability estimates $\{\bar{X}_{i,j}(t_k) : j \in \mathcal{N}^+\}$. Let $\rho_i(t)$ be the channel chosen by SU i in time slot t. This learning algorithm is an extension of the work in [41]. The probability ϵ_t is chosen to be decreasing in t with a specific form as shown in Algorithm 1. To do this, we need the following assumption.

Assumption 1: A positive lower bound $\Delta_1 \le \min_{1\le j<N} |\mu_j - \mu_{j+1}|$ is known to the SUs.

Note that for Assumption 1 to hold, SUs do not need to know the channel idle probabilities, but an estimate of how different the PU channel utilization rates are from each other. In practice, one can always choose a sufficiently small Δ_1.

We call the above procedure the Centralized Channel Allocation Policy ψ^{CCA}, and its formal description is given in Algorithms 1 and 2. Note that if for some $i \in \mathcal{M}$, the central processor returns $r_i(t_k) > N$, then SU i is allocated a null channel and it does not engage in OSA for $t \in [t_k, t_{k+1})$.

Algorithm 1 Centralized Channel Allocation policy ψ^{CCA}

1: **Input:** SU interference network.
2: The following two for loops are executed in parallel at the central processor and SUs.
3: **for** $t \in \{t_1, t_2, \ldots, t_{\xi(n)}\}$ **do**
4: Central processor chooses an arbitrary SU i, which sends it $\{\bar{X}_{i,j}(t) : j \in \mathcal{N}\}$.
5: Central processor solves the optimization problem $(\widehat{\text{P0}})$, and for each $i \in \mathcal{M}$, sets $r_i(t) = j$ if $x_{ij} = 1$.
6: Central processor sends $r_i(t)$ to each SU i, $i \in \mathcal{M}$.
7: **end for**
8: **for** each SU $i \in \mathcal{M}$ **do**
9: SU i performs the random ϵ-greedy channel learning algorithm in Algorithm 2 with channel access rank at time slot t set as $r_i(t) = r_i(t_k)$, where $k = \max\{q \ge 1 : t_q \le t\}$.
10: **end for**

In the following proposition, we show that for any policy ψ, the regret is at least $\Omega(\log n)$. The proof is provided in

Algorithm 2 Random ϵ-greedy channel learning at a SU i

Input: Channel access ranks $\{r_i(t) : t \geq 1\}$.
Choose $0 < \gamma < \min\{1, \Delta_1\}$ and $\delta > \max\{2, 5\gamma^2\}$.
for $t \geq 1$ **do**
  Set $\epsilon_t = \min\{1, \frac{\delta N}{\gamma^2 t}\}$.
  With probability $1 - \epsilon_t$, let $\rho_i(t)$ be a channel with the $r_i(t)$-th highest empirical idle probability estimate (with ties broken randomly), otherwise let $\rho_i(t)$ be chosen uniformly at random from the channel set $\mathcal{N}$.
  if channel $\rho_i(t)$ is sensed to be PU-free **then**
    SU i transmits over channel $\rho_i(t)$ and sets $X_{i,\rho_i(t)}(t) = 1$.
  else
    Set $X_{i,\rho_i(t)}(t) = 0$.
  end if
  Set

$$T_{i,\rho_i(t)}(t) = T_{i,\rho_i(t)}(t-1) + 1,$$
$$\bar{X}_{i,\rho_i(t)}(t) = \frac{\bar{X}_{i,\rho_i(t)}(t-1)T_{i,\rho_i(t)}(t-1) + X_{i,\rho_i(t)}(t)}{T_{i,\rho_i(t)}(t)}.$$

  For all $j \neq \rho_i(t)$, set $T_{i,j}(t) = T_{i,j}(t-1)$ and $\bar{X}_{i,j}(t) = \bar{X}_{i,j}(t-1)$.
end for

Appendix A.

Proposition 1: For any policy ψ, $R(n, \psi) \in \Omega(\log n)$.

The next result, whose proof is in Appendix B, shows that the regret using ψ^{CCA} is order-optimal for appropriately chosen optimization time instances.

Theorem 1: If $l_k > l_{k-1}$ for $1 < k < \xi(n)$ and $l_k \leq c l_{k-1}$ for all $k \geq 2$ and some $c > 0$, then $R(n, \psi^{\text{CCA}}) \in \Theta(\log n)$.

Theorem 1 shows that the central processor needs to perform a re-optimization of $(\widehat{\text{P0}})$ only at exponentially increasing time intervals to achieve order optimality. However, since $(\widehat{\text{P0}})$ is an NP-complete problem, the central processor incurs high computation cost at each optimization time instance if the number of SUs is large. Furthermore, the CRN is also vulnerable to the failure of the central processor. In the next section, we propose heuristic distributed policies that do not have such drawbacks, but which are no longer order optimal.

IV. Distributed Channel Access Ranking and Learning

In this section, we propose two heuristic distributed policies to perform channel access ranking and learning. Note that to optimize (P0), neighboring SUs in the interference graph have to be allocated different channels. Therefore, the genie-aided channel allocation in (P0) is a graph coloring problem in which we wish to partition the graph into disjoint maximal independent sets $I_1, \ldots, I_{\chi(G)}$, where $\chi(G)$ is known as the *chromatic number* of the graph G [45]. The SUs assigned to the same independent set are allocated the same channel, with a larger independent set being assigned a channel with a higher idle probability. Partitioning the graph G into $\chi(G)$ independent sets is not unique, and some partitions are non-optimal for (P0). Finding an optimal partition is again difficult, and we have to resort to heuristics.

Our first policy, which we call CARL, is a three-stage distributed channel access ranking and learning policy, which however incurs overheads in SU synchronization. The second policy, which we call DARL, does not require SU synchronization, but is expected to have higher regret than CARL. We also derive performance bounds for both policies.

A. CARL policy

The CARL policy is a three-stage distributed channel access ranking and learning policy denoted by ψ^{CARL}, and which performs the following procedures:

1) distributed graph coloring algorithm (Algorithm 3);
2) distributed channel access rank determination method (Algorithm 4); and
3) random ϵ-greedy channel learning (Algorithm 2) at each SU.

The first two stages enable each SU to find its optimal channel access rank, which will be utilized in the last stage to find the optimal channel allocation for maximizing the total network reward.

The graph coloring algorithm is described in Algorithm 3, which aims to cluster SUs into a minimal number of maximal independent sets so that channels with higher idle probabilities are spatially reused by more SUs. We adopt a distributed greedy graph coloring algorithm [46], [47] to color the graph G using the smallest number of colors. Suppose that SUs are colored using a set of colors $\{1, \ldots, |\mathcal{N}^+|\}$, with SUs labeled the same color belonging to the same maximal independent set. In the graph G, let $d(i)$ denote the degree of SU i and $u(i)$ be the palette of forbidden colors, which are used by its colored neighbors. Let $\alpha(i) = |u(i)|$ be the number of different colors used by the SUs in the neighborhood of SU i. A color is said to be legal for SU i if it is not contained in $u(i)$. In each round, each SU generates a random value λ_i and the SU with the highest $\alpha(i)$ will be colored. If two SUs have the same $\alpha(i)$, the SU with a higher λ_i will be colored first.

The graph coloring algorithm in Algorithm 3 is heuristic and is not guaranteed to find an optimal coloring for a general graph. It can use up to $\triangle(G) + 1$ colors to color a general graph, where $\triangle(G)$ is the maximum vertex degree. However, in practice, CRNs are usually sparse satisfying $\triangle(G)+1 \ll M$ and Algorithm 3 has been shown to find the chromatic number of such graphs effectively [47]. Furthermore, it produces optimal colorings for complete graphs and bipartite graphs like trees [48]. Since in each time slot, at least one uncolored node in the graph is colored, Algorithm 3 requires at most $O(M)$ rounds to color all nodes. In each round, at most $O(M^2)$ messages are exchanged among nodes so that the message complexity of Algorithm 3 is $O(M^3)$.

After performing the distributed graph coloring step, SUs are clustered into different color groups identified by $c(i) \in \{1, \ldots, |\mathcal{N}^+|\}$ for each $i \in \mathcal{M}$. To maximize the total network reward, the group having more SUs should be assigned a smaller channel access rank so that more SUs can access a

Algorithm 3 Distributed graph coloring algorithm

Initialization: for each $i = 1 : M$, define
- $u(i) = \emptyset$,
- $d(i) =$ degree of SU i,
- $\alpha(i) = d(i)$.

for $n = 1 : T_1$ **do**
  for each SU i **do**
    if SU i is uncolored **then**
      Set $c(i) = \min(\{1, \ldots, N\} \backslash u(i))$.
      Generate a random number λ_i uniformly distributed in $[0, 1]$.
      Broadcast to $\mathcal{M}_i$ the parameters $\alpha(i)$, λ_i and $c(i)$.
      if $\Big(\alpha(i) > \alpha(j)$ OR $(\alpha(i) = \alpha(j)$ AND $\lambda_i > \lambda_j)$ for all uncolored $j \in \mathcal{M}_i\Big)$ **then**
        SU i is colored with $c(i)$.
      else
        SU i updates $u(i) = u(i) \cup \{c(i)\}$ and $\alpha(i) = |u(i)|$.
      end if
    end if
  end for
end for

PU channel with higher idle probability. Therefore, we need to relabel the group identities according to their group sizes. In each time slot n, each SU i maintains two variable vectors $\mathbf{z}_{i,n} = [z_{i,n}(1), \ldots, z_{i,n}(|\mathcal{N}^+|)]$ and $\mathbf{w}_{i,n}$, where we initialize $\mathbf{w}_{i,1} = \mathbf{z}_{i,1}$, and set $z_{i,1}(c(i)) = 1$ and all other entries in $\mathbf{z}_{i,1}$ to 0. We use an average consensus algorithm [49] to compute $\overline{\mathbf{z}} = \sum_{i=1}^{M} \mathbf{z}_{i,1}/M$ in a distributed manner (see Algorithm 4). From the time slot $n > 1$, each SU i updates $\mathbf{z}_{i,n}$ and $\mathbf{w}_{i,n}$ according to:

$$\mathbf{w}_{i,n+1} = \mathbf{z}_{i,n} + \frac{1}{2} \sum_{j \in \mathcal{M}_i} \frac{\mathbf{z}_{j,n} - \mathbf{z}_{i,n}}{\max(d(i), d(j))},$$

$$\mathbf{z}_{i,n+1} = \mathbf{w}_{i,n+1} + \left(1 - \frac{2}{9M+1}\right)(\mathbf{w}_{i,n+1} - \mathbf{w}_{i,n}).$$

By ordering the entries of $\mathbf{w}_{i,n}$ at the end of Algorithm 4, each SU i can then determine its channel access rank r_i. Note that if $r_i > N$, then SU i does not engage in OSA. In general, because of spatial spectrum reuse, the number of PU channels required for all M SUs to perform OSA is $\chi(G) \leq M$. The value $\chi(G)$ depends on the sparsity of the interference graph.

From Theorem 1.1 in [49], we have

$$\|\mathbf{w}_{i,n} - \overline{\mathbf{z}}\|_2^2 \leq 2\left(1 - \frac{1}{9M}\right)^{n-1} \|\mathbf{w}_{i,1} - \overline{\mathbf{z}}\|_2^2 \leq 2M\left(1 - \frac{1}{9M}\right)^{n-1}. \quad (7)$$

Therefore, to ensure that each SU's local estimate of $\mathbf{w}_{i,n}$ converges to $\overline{\mathbf{z}}$, we can set the upper bound in (7) to $1/(2M)^2$ to obtain an upper bound on the number of iterations T_2 required for Algorithm 4 to be $3\log(2M)/\log(9M/(9M-1))$.

Finally, each SU i sets for all time slots t, $r_i(t) = r_i$ from Algorithm 4 as the random access rank in the ϵ-greedy channel learning process in Algorithm 2.

Algorithm 4 Distributed channel access rank determination

Initialization: for each $i = 1 : M$, let $\mathbf{w}_{i,1} = \mathbf{z}_{i,1}$, where $\mathbf{z}_{i,1}$ is a vector of length $|\mathcal{N}^+|$ consisting of all zeros except for a 1 in the $c(i)$-th entry.
for $n = 1 : T_2$ **do**
  for $i = 1 : M$ **do**
    $\mathbf{w}_{i,n} = \mathbf{z}_{i,n-1} + \frac{1}{2}\sum_{j \in \mathcal{M}_i} \frac{\mathbf{z}_{j,n-1} - \mathbf{z}_{i,n-1}}{\max(d(i), d(j))}$
    $\mathbf{z}_{i,n} = \mathbf{w}_{i,n} + (1 - \frac{2}{9M+1})(\mathbf{w}_{i,n} - \mathbf{w}_{i,n-1})$
    SU i broadcasts $\mathbf{z}_{i,n}$ and $d(i)$ to all its neighbors $j \in \mathcal{M}_i$.
  end for
end for
for $i = 1 : M$ **do**
  Set $r_i = r$ if group $c(i)$ has the r-th largest value in $\mathbf{w}_{i,T_2}$.
end for

B. DARL policy

The policy ψ^{CARL} uses its first two stages to determine the appropriate channel access ranks to assign to each SU. This requires that SUs are synchronized between each of its three stages, and may incur significant communication overhead if the number of SUs is large. To mitigate this problem, we propose another distributed policy DARL, denoted as ψ^{DARL}, which embeds the channel access rank determination procedure in the channel statistics learning process (see Algorithm 5). Since there is a higher likelihood for DARL to assign incorrect channel access ranks to the SUs, we expect DARL to have higher regret than CARL, as verified by the simulation results in Section V.

At the start of DARL, the channel access ranks of SUs $r_i(1)$, $i \in \mathcal{M}$ are all set to be 1. In subsequent time slots $n > 1$, if there is no collision for SU i in the previous time slot, it continues to use the same channel access rank as $r_i(n-1)$. Otherwise, it generates a random number λ_i uniformly distributed in $[0, 1]$ and keeps on using the same channel access rank if λ_i has the largest value among all its neighbors who also have collisions in the previous time slot. If its random number λ_i is not the largest value, SU i is allocated the smallest channel access rank not used by its neighbors. DARL is somewhat similar to the adaptive randomization policy proposed in [39] for temporal spectrum reuse (see Section IV-C for a brief description), and can be viewed as a greedy version of the latter since each SU tries to acquire the smallest channel access rank available instead of choosing a random channel. SU i then performs the ϵ-greedy channel learning process as given in Algorithm 2.

Through this process, with an increasing likelihood over time, each SU is allocated the minimal available channel access rank that is different from its neighbors and the channel statistics are learned at the same time. The DARL policy does not require two separate initial stages to learn the access ranks. However, since the access ranks are assigned somewhat randomly and do not take into consideration the number of

Algorithm 5 Distributed access rank learning (DARL) ψ^{DARL}

1: **Initialization:**
- Choose $0 < \gamma < \min\{1, \Delta_1\}$ and $\delta > \max\{2, 5\gamma^2\}$.
- Set channel access rank $r_i(1) = 1$, for all $i \in \mathcal{M}$.

2: **for** $t \geq 1$ **do**
3: **if** there was a collision in previous time slot $t-1$ **then**
4: Broadcast $r_i(t-1)$ to all $j \in \mathcal{M}_i$.
5: Generate a random number λ_i uniformly distributed in $[0, 1]$ and broadcasts λ_i to all $j \in \mathcal{M}_i$.
6: Let $\bar{\mathcal{M}}_i$ be the set of SUs $j \in \mathcal{M}_i$ that also have collisions in time slot $t-1$.
7: **if** $\lambda_i \geq \max_{j \in \bar{\mathcal{M}}_i} \lambda_j$ **then**
8: Set $r_i(t) = r_i(t-1)$.
9: **else**
10: Set $r_i(t) = \min\left\{\mathcal{N}^+ \backslash \{r_j(t-1) : j \in \bar{\mathcal{M}}_i\}\right\}$.
11: **end if**
12: **else**
13: Set $r_i(t) = r_i(t-1)$.
14: **end if**
15: Execute lines 4 – 12 of Algorithm 2.
16: **end for**

SUs with the same access rank, it is likely to have a higher regret compared to the CARL policy.

C. Regret bounds

In this section, we derive theoretical bounds on the regret (2) achieved by our proposed CARL and DARL policies. As benchmark comparisons, we also derive bounds under our spatial spectrum reuse framework for the random access policy [39], the TDFS policy [38] and the adaptive randomization policy [39], which we denote by ψ^{rand}, ψ^{TDFS} and ψ^{adapt}, respectively. We first describe these policies briefly below.

1) Random access policy ψ^{rand}: In each time slot, each SU randomly chooses a channel $j \in \mathcal{N}$ to sense. The SU transmits if the channel is found to be idle.
2) TDFS policy ψ^{TDFS}: As describing the policy accurately requires some technical details, we refer the reader to [38]. We summarize the policy here briefly. Without going into the technical details, the TDFS policy is intuitively similar to a policy in which each SU accesses the M best channels in a round-robin fashion. A different offset, based on each SU's identity, in the channel access sequence is used to ensure that every SU is assigned a different channel access rank in each time slot. Each SU determines the M best channels by running parallel Lai-Robbins single-player policies [50]. To determine the k-th best channel, a SU removes the $k-1$ channels it considers to be the best (these are channels it has attempted to access in the previous $k-1$ time slots). It then considers the subsequence of time slots that also have the same set of $k-1$ best channels removed, and performs a Lai-Robbins single-player policy on this subsequence of time slots, using only the remaining $N-k+1$ channels to find the best channel amongst these. For the purpose of our discussion, the main point to note is that each SU is assigned the channel access ranks $1, \ldots, M$ in a round-robin fashion over time slots. This ensures "fairness" in accessing the best channel, which we do not address in this paper.
3) Adaptive randomization policy ψ^{adapt}: Every SU is initially assigned the channel access rank of 1. Then, in each time slot, each SU i randomizes its channel access rank only if there is a collision in the previous slot, otherwise it retains its channel access rank from the previous time slot. Suppose the channel access rank is r. The r-th highest channel based on SU i's sample-mean statistics [41] given by

$$\bar{X}_{i,j}(n) + \sqrt{\frac{2\log n}{T_{i,j}(n)}},$$

for each channel $j \in \mathcal{N}$, is chosen to be sensed. The SU transmits if the channel is found to be idle.

It is clear that since there is a positive probability that the optimal channel access ranks are not found in ψ^{CARL} and ψ^{DARL}, these policies have $\Theta(n)$ regret in general. The same can be said of our benchmark policies as shown below.

Proposition 2: Under spatial spectrum reuse, if the graph G is incomplete and has a connected component of size at least two, then ψ^{rand}, ψ^{TDFS}, and ψ^{adapt} each has $\Theta(n)$ regret.

Proof: See Appendix C. ■

Although ψ^{CARL} has $\Theta(n)$ regret, it is able to achieve $\Theta(\log n)$ regret if its distributed graph coloring step in Algorithm 3 produces an optimal graph coloring. The proof of Proposition 3 is similar to that for Theorem 1 and is omitted here. As explained in Section IV, Algorithm 3 is likely to produce an optimal graph coloring for a sparse graph G. This explains our observation in our simulation results that ψ^{CARL} has better regret performance than the benchmark policies although all of them have worst-case linear regret.

Proposition 3: Suppose that the output from Algorithm 4 satisfy $r_i = \pi^*(i)$ for all $i \in \mathcal{M}$, then the policy ψ^{CARL} has $\Theta(\log n)$ regret.

V. Simulation Results

In this section, we verify the performance of our proposed policies by simulations. We first consider small size simple graphs, and then provide simulation results on large size random graphs.

A. Small size graphs

Suppose that there are $M = 9$ SUs and $N = 9$ orthogonal PU channels. The idle probabilities of the PU channels are evenly spaced values from 0.1 to 0.9. We apply our proposed policy in three connected interference graphs, which are the nine-node ring graph, a grid graph and a randomly generated graph (See Figure 3). Similar types of graphs have been adopted to evaluate spatial channel reuse performance in [36].

In Algorithms 2, 3 and 4, we set $T_1 = 9$, $T_2 = 300$, $\delta = 5.1$ and $\gamma = 0.1$. For ψ^{CCA}, we let $l_0 = 2$ and $l_k = 2l_{k-1}$ for $k \geq 1$. The following results are averaged over 500 simulation runs.

In Figures 4, 5 and 6, we observe that ψ^{CCA}, ψ^{CARL} and ψ^{DARL} outperform the policies ψ^{rand} [39], ψ^{TDFS} [38] and

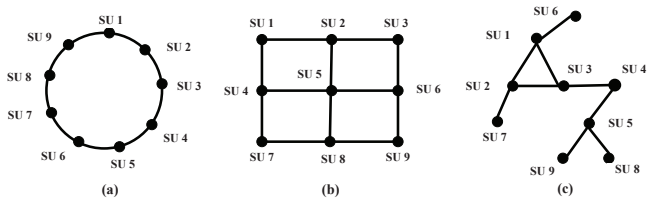


Fig. 3. Interference graphs of three CRNs (a) a ring graph (b) a grid graph (c) a randomly generated graph.

ψ^{adapt} [39] in all three interference graphs, where the regrets of ψ^{CCA} and ψ^{CARL} are approximately a constant multiple of $\log n$. Moreover, the regret of ψ^{CARL} is close to that of ψ^{CCA} because Algorithm 4 manages to find optimal or near-optimal channel access ranks for all SUs in these graphs. We also note that ψ^{DARL} has worse regret than ψ^{CARL}, but still performs better than the benchmark algorithms.

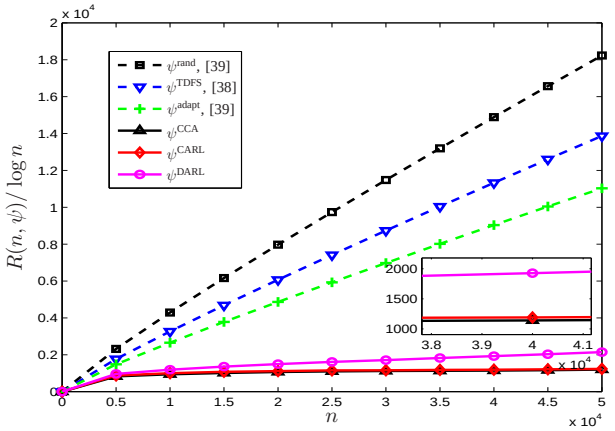


Fig. 4. Normalized regret $\frac{R(n,\psi)}{\log n}$ vs. time slot n on ring graph.

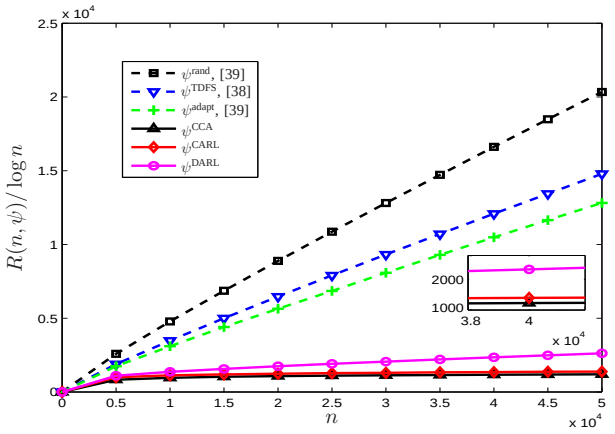


Fig. 5. Normalized regret $\frac{R(n,\psi)}{\log n}$ vs. time slot n on grid graph.

B. Large size random graphs

In this subsection, we consider large size random graphs that have $M = 100$ SUs and $N = 100$ orthogonal

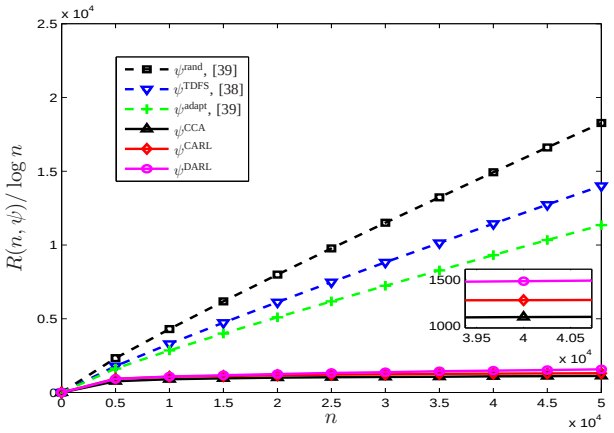


Fig. 6. Normalized regret $\frac{R(n,\psi)}{\log n}$ vs. time slot n on randomly generated graph.

PU channels. The idle probabilities of the PU channels are $[0.9, 0.8, 0.7, 0.6, 0.5, 0.495, 0.490, \ldots, 0.025]$. In Algorithms 2, 3 and 4, we set $T_1 = 100$, $T_2 = 3500$, $\delta = 5.1$ and $\gamma = 0.1$. We evaluate the performance of our proposed policies and that of the benchmark policies on the following:

(i) Erdös-Rényi (ER) graphs: 500 instances of Erdos-Renyi random graphs with M nodes and different probabilities of attachment [51]

(ii) Random connection (RC) graphs: 500 random graphs with M nodes and different number of edges. Edges are generated sequentially, and each edge is formed by choosing two distinct nodes uniformly at random and connecting them if they are not already connected.

Table II shows the percentage of simulation runs in which ψ^{CARL} found the correct graph chromatic number $\chi(G)$. We can see that ψ^{CARL} estimates the chromatic number $\chi(G)$ better when the network is sparse.

TABLE II
PERFORMANCE OF ψ^{CARL} IN FINDING THE CORRECT CHROMATIC NUMBER.

ER attachment probability	0.05	0.1	0.2
Percentage instances where correct $\chi(G)$ is found	**71.1**%	53.7%	34%
RC number of edges	200	500	1000
Percentage instances where correct $\chi(G)$ is found	**60.4**%	51%	36%

For the instances in which ψ^{CARL} estimated the correct chromatic number, let

$$D = \frac{1}{M}\sum_{i=1}^{M} |r_i - \pi^*(i)|,$$

be the average error in finding the correct channel access ranks. Table III shows that policy ψ^{CARL} could allocate channel access ranks for sparse random graphs with a small error.

Finally, we show the regrets in Figure 7 and Figure 8 when the graph is a randomly generated ER graph with attachment probability 0.05 and a RC graph with 200 edges. We compare the average regrets using 500 trials for all the policies. We

TABLE III
PERFORMANCE OF ψ^{CARL} IN FINDING THE CORRECT CHANNEL ACCESS RANKS.

ER attachment probability	0.05	0.1	0.2
D	**0.268**	0.379	0.664
RC number of edges	200	500	1000
D	**0.219**	0.398	0.658

observe that the regret of ψ^{CCA} is again approximately a constant multiple of $\log n$ and regrets using other policies on both types of random graphs increase linearly over time. However, our policies ψ^{CARL} and ψ^{DARL} give a much smaller regret than the benchmark policies.

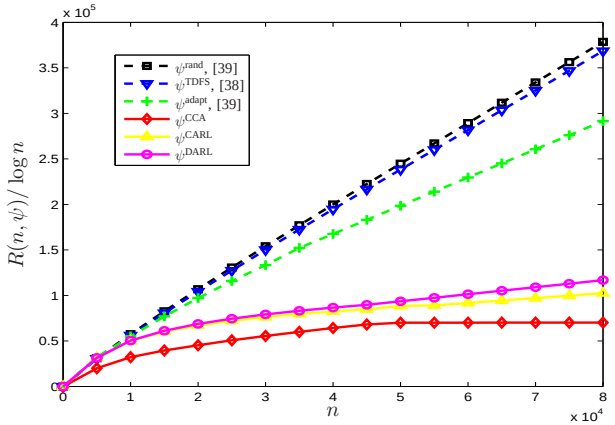


Fig. 7. Normalized regret $\frac{R(n,\psi)}{\log n}$ vs. time slot n on ER graphs.

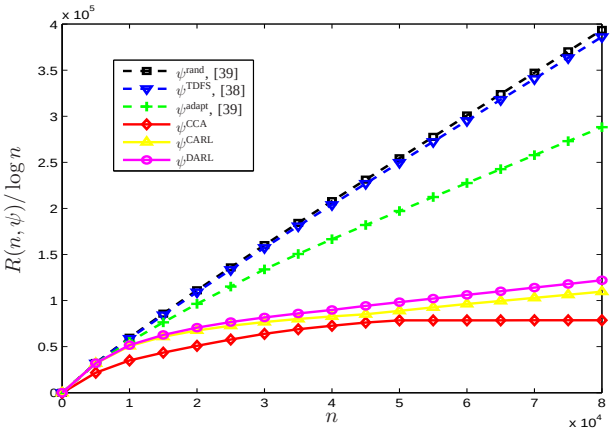


Fig. 8. Normalized regret $\frac{R(n,\psi)}{\log n}$ vs. time slot n on RC graphs.

VI. CONCLUSION

In this paper, we have investigated temporal-spatial channel reuse in cognitive radio networks using a multi-user MAB approach. We have proposed a centralized channel allocation policy for finding an optimal channel allocation and learning the statistics of the channels. We showed that this policy is order-optimal with logarithmic regret, but requires solving a NP-complete optimization problem at exponentially increasing time intervals. To avoid that and to overcome the requirement of centralized processing, we also proposed heuristic distributed policies, which however have linear regrets. The first distributed policy utilizes a distributed greedy graph coloring method and a distributed average consensus method in the learning process to find the optimal channel access rank for each SU. This requires synchronization amongst the SUs. In the second distributed policy, we let each SU determine their channel access ranks locally, which removes the requirement for synchronization. Simulation results suggest that our proposed policies outperform current policies in the literature, which do not consider spatial channel reuse amongst SUs. Future work includes designing policies for mobile SUs and where channel availabilities differ across the SU network. Another interesting future research direction is to allow SUs to not only cooperatively learn channel access ranks, but also channel statistics.

APPENDIX A
PROOF OF PROPOSITION 1

If two SUs use the same channel in the same time slot, a collision occurs and both SUs have rewards of 0. Therefore, $R(n,\psi) \geq R'(n)$, where $R'(n)$ denotes the regret if SUs are still rewarded with a reward of 1 even if collisions occur. This regret $R'(n)$ is then equivalent to M times the regret of a single-user MAB problem, which has been shown to be $\Omega(\log n)$ in [39].

APPENDIX B
PROOF OF THEOREM 1

The following lemma can be shown using the exact same argument in Theorem 3 of [41], and its proof is thus omitted here.

Lemma 1: Suppose in Algorithm 2, $\delta > 5$ and $0 < \gamma < \min\{1, \Delta\}$ for some $\Delta > 0$. Then, for all $i \in \mathcal{M}$, $n \geq \frac{\delta N}{\gamma}$ and $j \in \mathcal{N}$, we have

$$\mathcal{P}(|\bar{X}_{i,j}(n) - \mu_j| \geq \frac{\Delta}{2}) \leq \frac{a}{n^{1+\varepsilon}},$$

where a and ε are positive constants.

Let $R(k)$ be the total regret under ψ^{CCA} in the time slots $[t_k, t_{k+1})$. Let $A(t_k) = \{r_i(t_k) = \pi^*(i), \forall i \in \mathcal{M}\}$ be the event that the central processor returns the optimal channel allocation at time t_k. Then, for each time $m \in [t_k, t_{k+1})$, conditioned on $A(t_k)$, a packet collision occurs only if for some $i \in \mathcal{M}$, either $|\bar{X}_{i,\pi^*(i)}(m) - \mu_{\pi^*(i)}| \geq \Delta_1/2$ or SU i chooses a random channel $\rho_i(m) \neq \pi^*(i)$. On the other hand, the event $A^c(t_k)$ occurs only if $|\bar{X}_{i,\pi^*(i)}(t_k) - \mu_{\pi^*(i)}| \geq \Delta_2/2$ for some $i \in \mathcal{M}$, where Δ_2 is the minimum sensitivity range of the μ_j coefficients, $j = 1, \ldots, N$, in (P0) with respect to the optimal solution. Note that $\Delta_2 > 0$ since (3) is a continuous

function. Therefore, we have

$$R(k) \le \mu_1 \sum_{m=t_k}^{t_{k+1}-1} \Big(\mathcal{P}\Big(A(t_k) \text{ and } \Big\{ |\bar{X}_{i,\pi^*(i)}(m) - \mu_{\pi^*(i)}| \ge \frac{\Delta_1}{2} \text{ or } \rho_i(m) \ne \pi^*(i) \Big\} \Big) + \mathcal{P}(A^c(t_k)) \Big)$$
$$\le \mu_1 \sum_{m=t_k}^{t_{k+1}-1} \sum_{i \in m} \Big(\mathcal{P}(|\bar{X}_{i,\pi^*(i)}(m) - \mu_{\pi^*(i)}| \ge \frac{\Delta_1}{2}) + 2\epsilon_m \Big)$$
$$+ \mu_1 \sum_{m=t_k}^{t_{k+1}-1} \sum_{i \in m} \mathcal{P}(|\bar{X}_{i,\pi^*(i)}(m) - \mu_{\pi^*(i)}| \ge \frac{\Delta_2}{2}), \quad (8)$$

where the last inequality follows from the union bound. From Lemma 1, the right hand side of (8) is upper-bounded by

$$c_1 \left(\sum_{m=t_k}^{t_{k+1}-1} (\frac{1}{m^{1+\epsilon}} + \frac{1}{m}) + \sum_{m=t_k}^{t_{k+1}-1} \frac{1}{m^{1+\epsilon}} \right) \le c_2 \frac{l_k}{t_k},$$

for some constants c_1 and c_2 sufficiently large. Therefore, we have for some constant $c_3 > 0$,

$$R(n, \psi^{\text{CCA}}) \le c_3 t_1 + c_2 \sum_{k=1}^{\xi(n)} \frac{l_k}{t_k} \le c_3 t_1 + c_2 \sum_{k=1}^{\xi(n)} \frac{l_k}{l_{k-1}} \le c_3 t_1 + c_2 c \xi(n), \quad (9)$$

where the second inequality follows because $t_k = \sum_{m=0}^{k-1} l_m \ge l_{k-1}$. Since $l_k > l_{k-1}$ for $1 < k < \xi(n)$ and $l_k \le c l_{k-1}$ for all $k \ge 2$, we have $\xi(n) \le c_4 \log(n)$ for some constant $c_4 > 0$. Then from (9), we have $R(n, \psi^{\text{CCA}}) \in O(\log n)$, and together with Proposition 1, the theorem follows.

Appendix C
Proof of Proposition 2

For any policy ψ, we have

$$R(n, \psi) = n \sum_{i \in \mathcal{M}} \mu_{\pi^*(i)} - \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}^+} \mu_j \mathbb{E}[V_{i,j}(n)] \ge \sum_{i \in \mathcal{M}} \sum_{j > \pi^*(i)} (\mu_{\pi^*(i)} - \mu_j) \mathbb{E}[T_{i,j}(n)]. \quad (10)$$

In the policy ψ^{rand}, each SU randomly chooses a channel to sense in each time slot. Consider a SU i' with $\pi^*(i') = 1$. From (10), we have

$$R(n, \psi^{\text{rand}}) \ge (\mu_1 - \mu_2) \sum_{j=2}^{N} \mathbb{E}[T_{i',j}(n)] = (\mu_1 - \mu_2) \frac{N-1}{N} n.$$

Therefore, ψ^{rand} has $\Theta(n)$ regret.

We next consider the policy ψ^{TDFS}. Every SU uses the Lai-Robbins single-player policy to determine the best channel. For all time slots n sufficiently large, the probability that all SUs identify channel 1 as the best channel is bounded away from zero. From the proposition assumptions, there is a SU i in G with degree less than $M-1$. Since the TDFS policy assigns to each SU channel access ranks in a round-robin fashion over time slots, there exists at least one time slot out of every M slots such that all SUs in $i \cup \mathcal{M}_i$ do not have channel access rank 1. Consider a policy ψ' that is the same as ψ^{TDFS} but which assigns channel access rank 1 to SU i in each of these time slots starting from some time n sufficiently large. Then, $R(n, \psi^{\text{TDFS}})/n \ge R(n, \psi')/n + c(\mu_1 - \mu_2)/M$ for some positive constant $c > 0$. This implies that ψ^{TDFS} has $\Theta(n)$ regret.

In the policy ψ^{adapt}, each SU randomizes the channel selection from $\mathcal{N}^+$ only if there is a collision in the previous time slot. Since the graph G has a connected component of size at least two, not all feasible solutions in (P0) are optimal. There is a positive probability that at the second iteration of ψ^{adapt}, the SUs are assigned channel access ranks corresponding to a feasible but suboptimal solution of (P0). Subsequently, the channel access ranks do not change. Therefore, ψ^{adapt} has $\Theta(n)$ regret. The proof of the proposition is now complete.

References

[1] Y. Zhang, W. P. Tay, K. H. Li, M. Esseghir, and D. Gaïti, "Distributed opportunistic spectrum access with spatial reuse in cognitive radio networks," in *Proc. IEEE Global Conf. on Signal and Information Processing*, 2014.

[2] ——, "Opportunistic spectrum access with temporal-spatial reuse in cognitive radio networks," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing*, 2016.

[3] Spectrum Efficiency Working Group, "FCC spectrum policy task force," Federal Communications Commission (FCC), Tech. Rep., 2002. [Online]. Available: http://transition.fcc.gov/sptf/reports.html

[4] W. P. Tay, J. N. Tsitsiklis, and M. Z. Win, "On the impact of node failures and unreliable communications in dense sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2535 – 2546, Jun. 2008.

[5] G. Kortuem, F. Kawsar, D. Fitton, and V. Sundramoorthy, "Smart objects as building blocks for the Internet of Things," *IEEE Internet Computing*, vol. 14, no. 1, pp. 44–51, 2010.

[6] L. Xie, D.-H. Choi, S. Kar, and H. V. Poor, "Fully distributed state estimation for wide-area monitoring systems," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1154–1169, 2012.

[7] Y. Ding, Y. Jin, L. Ren, and K. Hao, "An intelligent self-organization scheme for the Internet of Things," *IEEE Computational Intelligence Magazine*, vol. 8, no. 3, pp. 41–53, 2013.

[8] W. P. Tay, "Whose opinion to follow in multihypothesis social learning? A large deviations perspective," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 2, pp. 344 – 359, Mar. 2015.

[9] J. Unnikrishnan and V. V. Veeravalli, "Cooperative sensing for primary detection in cognitive radio," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 18–27, 2008.

[10] D. Duan, L. Yang, and J. C. Principe, "Cooperative diversity of spectrum sensing for cognitive radio systems," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3218–3227, 2010.

[11] M. Z. Shakir, A. Rao, and M.-S. Alouini, "Generalized mean detector for collaborative spectrum sensing," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1242–1253, 2013.

[12] C. Jiang, Y. Chen, K. Liu, and Y. Ren, "Renewal-theoretical dynamic spectrum access in cognitive radio network with unknown primary behavior," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 406–416, 2013.

[13] G. Ozcan, M. cenk Gursoy, and S. Gezici, "Error rate analysis of cognitive radio transmissions with imperfect channel sensing," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1642–1655, 2014.

[14] B. L. Mark and A. O. Nasif, "Estimation of maximum interference-free power level for opportunistic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 8, no. 5, pp. 2505–2513, 2009.

[15] S. M. Mishra, "Maximizing available spectrum for cognitive radios," Ph.D. dissertation, UC Berkeley, 2010.

[16] S. M. Yu and S.-L. Kim, "Optimal detection of spatial opportunity in wireless networks," *IEEE Commun. Lett.*, vol. 15, no. 4, pp. 395–397, 2011.

[17] W. P. Tay, "The value of feedback in decentralized detection," *IEEE Trans. Inf. Theory*, vol. 58, no. 12, pp. 7226 – 7239, Dec. 2012.

[18] D. Tsolkas, N. Passas, and L. Merakos, "Spatial spectrum reuse for opportunistic spectrum access in infrastructure-based systems," *Wireless Pers Commun*, vol. 69, no. 4, pp. 1749–1772, 2013.

[19] M. Leng, W. P. Tay, T. Q. S. Quek, and H. Shin, "Distributed local linear parameter estimation using Gaussian SPAWN," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 244 – 257, Jan. 2015.

[20] Z. Qin, Y. Gao, M. Plumbley, and C. Parini, "Wideband spectrum sensing on real-time signals at sub-nyquist sampling rates in single and cooperative multiple nodes," *IEEE Trans. Signal Process.*, accepted.

[21] Y. Yang, Y. Liu, Q. Zhang, and L. Ni, "Cooperative boundary detection for spectrum sensing using dedicated wireless sensor networks," in *Proc. IEEE Int. Conf. on Computer Communications*, 2010.

[22] A. W. Min, X. Zhang, and K. G. Shin, "Detection of small-scale primary users in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, pp. 349–361, 2011.

[23] Q. Wu, G. Ding, J. Wang, and Y. Yao, "Spatial-temporal opportunity detection for spectrum heterogeneous cognitive radio networks: Two-dimensional sensing," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 516–526, 2013.

[24] G. Ding, J. Wang, Q. Wu, F. Song, and Y. Chen, "Spectrum sensing in opportunity-heterogeneous cognitive sensor networks: How to cooperate?" *IEEE Sensors J.*, vol. 13, no. 11, pp. 4247–4255, 2013.

[25] Y. Zhang, W. P. Tay, K. H. Li, and D. Gaïti, "Distributed boundary estimation for spectrum sensing in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 11, pp. 1–13, 2014.

[26] G. Ding, J. Wang, Q. Wu, Y. D. Yao, F. Song, and T. A. Tsiftsis, "Cellular-base-station-assisted device-to-device communications in TV white space," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 107–121, 2016.

[27] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[28] M. Félegyházi, M. Čagalj, and J.-P. Hubaux, "Efficient MAC in cognitive radio systems: A game-theoretic approach," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1984–1995, 2009.

[29] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Decentralized dynamic spectrum access for cognitive radios: cooperative design of a non-cooperative game," *IEEE Trans. Commun.*, vol. 57, no. 2, pp. 459–469, 2009.

[30] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y. D. Yao, "Opportunistic spectrum access in unknown dynamic environment: a game-theoretic stochastic learning solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380–1391, 2012.

[31] C. Jiang, Y. Chen, Y. Gao, and K. Liu, "Joint spectrum sensing and access evolutionary game in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2470–2483, 2013.

[32] C. Jiang, Y. Chen, and K. Liu, "Multi-channel sensing and access game: Bayesian social learning with negative network externality," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2176–2188, 2014.

[33] L. Cao and H. Zheng, "Distributed spectrum allocation via local bargaining," in *Proc. IEEE Int. Conf. on Sensing, Communication, and Networking*, 2005.

[34] M. Azarafrooz and R. Chandramouli, "Distributed learning in secondary spectrum sharing graphical game," in *Proc. IEEE Global Telecomm. Conf.*, 2011.

[35] G. S. Kasbekar and S. Sarkar, "Spectrum pricing games with spatial reuse in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 153–164, 2012.

[36] X. Chen and J. Huang, "Distributed spectrum access with spatial reuse," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 593–603, 2013.

[37] Y. Xu, Q. Wu, L. Shen, J. Wang, and A. Anpalagan, "Opportunistic spectrum access with spatial reuse: graphical game and uncoupled learning solutions," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4814–4826, 2013.

[38] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, 2010.

[39] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, 2011.

[40] Y. Gai and B. Krishnamachari, "Distributed stochastic online learning policies for opportunistic spectrum access," *IEEE Trans. Signal Process.*, 2014.

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

[42] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multi-player multiarmed bandits," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[43] R. M. Karp, "Reducibility among combinatorial problems," *Complexity of Computer Computations*, pp. 85–103, 1972.

[44] L. A. Wolsey, *Integer Programming*. Wiley-Interscience, 1998.

[45] N. Deo, *Graph Theory with Applications to Engineering and Computer Science*. PHI Learning Pvt. Ltd., 2004.

[46] D. Brélaz, "New methods to color the vertices of a graph," *Commun. ACM*, vol. 22, no. 4, pp. 251–256, 1979.

[47] E. Ruzgar and O. Dagdeviren, "Performance evaluation of distributed synchronous greedy graph coloring algorithms on wireless ad hoc and sensor networks," *Int. J. of Comput. Netw. and Commun.*, vol. 5, no. 2, pp. 169–179, 2013.

[48] A. Kosowski and K. Manuszewski, "Classical coloring of graphs," *American Mathematical Society*, pp. 1–19, 2004.

[49] A. Olshevsky, "Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control," 2015, preprint. [Online]. Available: http://arxiv.org/abs/1411.4186

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

[51] P. Erdös and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.