Efficient Objective Functions for Coordinated Learning in Large-Scale Distributed OSA Systems

MohammadJavad NoroozOliaee, Bechir Hamdaoui, and Kagan Tumer

***Abstract*—In this paper, we derive and evaluate private objective functions for large-scale, distributed opportunistic spectrum access (OSA) systems. By means of any learning algorithms, these derived objective functions enable OSA users to assess, locate, and exploit unused spectrum opportunities effectively by maximizing the users' average received rewards. We consider the elastic traffic model, suitable for elastic applications such as file transfer and web browsing, and in which an SU's received reward increases proportionally to the amount of received service when the amount is higher than a certain threshold. But when this amount is below the threshold, the reward decreases exponentially with the amount of received service. In this model, SUs are assumed to be treated fairly in that the SUs using the same band will roughly receive an equal share of the total amount of service offered by the band. We show that the proposed objective functions are: near-*optimal*, as they achieve high performances in terms of average received rewards; highly *scalable*, as they perform well for small- as well as large-scale systems; highly *learnable*, as they reach up near-optimal values very quickly; and *distributive*, as they require information sharing only among OSA users belonging to the same band.**

***Index Terms*—Objective function design; scalable and distributed opportunistic spectrum access; dynamic bandwidth sharing; cognitive radio networks; coordinated learning.**

I. Introduction

FCC foresees opportunistic spectrum access (OSA) as a potential solution to the spectrum shortage problem [1]. Essentially, OSA improves spectrum efficiency by allowing unlicensed or secondary users (SUs) to exploit unused licensed spectrum, but in a manner that limits interference to licensed or primary users (PUs). During the past few years, due to its apparent promises, OSA has created significant research efforts, resulting in numerous works ranging from protocol design [2–8] to performance optimization [9–13], and from market-oriented access strategies [14–18] to new management and architecture paradigms [19–25]. More recently, research efforts have also been given to the development of optimal channel selection techniques that rely on spectrum availability prediction models to adapt themselves to the environment so as to promote effective OSA [26–35]. Unnikrishnan et al. [26] propose a cooperative, channel selection policy for OSA systems under interference constraints, where the PUs' activities are assumed to be stationary Markovian whose statistics are assumed to be known to all SUs. A centralized approach is considered, where all cooperating secondary users report their observations to a decision center, which makes decision regarding when and which channels to sense and access at each time slot. In [27], the authors develop a prediction model to capture the dynamics of the OSA environment under periodic channel sensing, where a simple two-state Markovian model is assumed for PUs' activities on each channel. Zhao et al. derive an optimal access policy for this model that can be used to maximize channel utilization while limiting interference to PUs. In [29], the authors model PUs' activities as a discrete-time Markov chains, and develop channel decision policies for two SUs in a two-channel OSA system. Liu et al. [30] assume that SUs cannot exchange information among themselves, and consider an OSA system with multiple, non-cooperative SUs. The occupancy of primary channels is modeled as an i.i.d. Bernoulli process, and OSA is formulated as a multi-armed bandit problem where agents are not cooperative with each others. Gai et al. [35] formulate the cognitive radio access problem as a combinatorial multi-armed bandit, in which each arm corresponds to a matching of the users to channels. In this work, the throughput achievable from accessing the available spectrum on each user-channel combination over a decision period is modeled as an arbitrarily-distributed random variable with bounded support but unknown mean. Chen et al. [32, 34] investigate cross-layer approaches for accessing OSA systems that integrate physical-layer's with MAC-layer's sensing and access policy. They establish a separation principle, meaning that physical-layer's sensing and MAC-layer's access policy can be decoupled from MAC-layer's sensing without losing optimality. This framework assumes that spectrum occupancy of PUs also follows a discrete-time ON/OFF Markov process.

In most of these works, the models developed for deriving optimal channel selection policies assume that PUs' activities follow the Markovian process model. Although analytically tractable, Markovian process may not accurately model the PUs' behaviors. The challenge is that the OSA environment has very unique characteristics that make it too difficult to construct models that can capture its dynamics without making assumptions about the environment itself. Such assumptions are often unrealistic, leading to inaccurate prediction of the environment's behaviors.

As a result, there have also been some efforts on developing learning-based techniques that do not require such models, yet can still perform well by learning directly from interaction with the environment [28–31, 36–38]. Instead of using models, these techniques rely on learning algorithms (e.g., reinforcement learners [39, 40] and evolving neuro-controllers [41, 42]) to learn from past and present interaction experience to decide what to do best in the future. For example, Anandkumar et al. [36, 37] propose a distributed learning technique for channel

This work was supported by NSF under NSF CAREER award CNS-0846044.

MohammadJavad NoroozOliaee and Bechir Hamdaoui are with the School of EECS at Oregon State University. Kagan Tumer is with the school of MIME at Oregon State University. Emails: noroozom@eecs.orst.edu; hamdaoui@eecs.orst.edu; kagan.tumer@oregonstate.edu.

access in cognitive radio network with multiple SUs, where channel statistics are not known to SUs but estimated via sensing decisions. They propose policies that minimize regret in distributed learning and access, and show that the total regret is logarithmic in the number of slots when assuming that the number of SUs is known to the policy, and grows slightly faster than logarithmic also in the number of slots when assuming that the number of SUs is fixed but unknown. In essence, learning algorithms allow SUs to learn by interacting with the environment, and use their acquired knowledge to select the proper actions that maximize their own (often selfish) objective functions, thereby "hopefully" maximizing their long-term cumulative received rewards.

The key challenge that we address in this work is that when SUs' objective functions are not carefully coordinated, learning algorithms can lead to poor performances in terms of SUs' long-term received rewards. In other words, when SUs aim at maximizing their intrinsic (not carefully designed) objective functions, their collective behavior often leads to worsening each other's long-term cumulative rewards, a phenomenon known as the "tragedy of the commons" [43]. It is, therefore, imperative that objective functions be designed carefully so that when SUs maximize them, their collective behavior does not result in worsening each other's performance.

With this in mind, in this work, we derive efficient objective functions for SUs that are aligned with system objective in that when SUs aim to maximize them, their collective behaviors also lead to good system-level performance, thereby resulting in increasing each SU's long-term received rewards. We consider the elastic traffic model, in which an SU's received reward increases proportionally to the service it receives from using the spectrum band when the amount of received service is higher than a certain (typically low) threshold, R. But when the amount of received service is below the threshold R, the reward decreases rapidly with the amount of received service; i.e., the QoS satisfaction goes almost immediately to zero when the amount of received service is below R. This service model is suitable for elastic applications, such as file transfer and web browsing, where the higher the amount of received service, the better the quality perceived by these applications. But when the amount of received service is below a certain low threshold (i.e., R), the quality of these applications becomes unacceptable. In this model, we also assume that SUs are treated fairly in that the SUs using the same band will receive roughly an equal share of the total amount of service offered by the band.

To sum up, we propose in this work objective functions for supporting elastic traffic in OSA systems that are: (i) near-*optimal*, in that they allow SUs to achieve rewards close to the maximal achievable rewards, (ii) *scalable*, in that they perform well in systems with a small as well as a large number of users, (iii) *learnable*, in that they allow SUs to reach up high rewards very quickly, and (iv) *distributive*, in that they are implementable in a decentralized manner by relying on local information only. We want to emphasize that the focus of this paper is not on learning, but rather on the design of coordination techniques that can be used by any learning algorithms.

The rest of the paper is organized as follows. Section II presents the system model. Section III states the motivation and the objective of this work. In Section IV, we propose the objective functions. In Section V, we derive upper bounds on the maximal achievable rewards. In Section VI, we evaluate the proposed functions. In Section VII, we discuss the implementation incentives of the proposed techniques. Finally, we conclude the paper in Section IX.

II. System Model

We assume that spectrum is divided into m non-overlapping bands, and that each band is associated with many PUs. We consider a distributed system where PUs can arrive and leave independently at different times. Throughout this paper, we refer to any group of two or more SUs who want to communicate together as an *OSA agent*, or simply *an agent*. In order to communicate with each other, all SUs in the group must be tuned to the same band.

A. Learning Algorithm

We assume that each agent implements a learning algorithm (e.g., a reinforcement learner [39, 40]) to help it find spectrum opportunities. That is, every once in a while (e.g., after every time interval or time episode), each agent relies on its learner to select the "best" available spectrum band, which it will then use until the end of the episode. Each agent does so independently from all other agents, and as long as it needs to access the OSA system. At each episode, each agent receives a service that is passed to it from the environment/system. One possible service metric is the amount of throughput that the visited spectrum band offers the agent. Another possible metric is the reliability of the communication carried on the spectrum band, which can be measured through, for example, SNR (signal to noise ratio), PSR (packet success rate), etc. What service metric to use and how to quantify it are beyond the scope of this work. Here, we assume that once the agent switches to a particular band, the received service level can immediately be quantified by monitoring the metric in question. Hereafter, we then assume that each band j is characterized by a value V_j that represents the maximum/total service level that the band can offer.

B. Network Topology

We model the network composed of all agents as a graph, where each vertex/node corresponds to an agent, and each edge between two agents indicates that these two agents interfere with one another. Two agents interfering with one another must share the spectrum band if both are tuned to it; i.e., an edge between two nodes implies that these two agents contend with each other over each spectrum resource. When there is no edge between two nodes, it means that the corresponding agents do not interfere/contend with one another. Throughout this work, network topologies in which each and every agent contends with all other agents are referred to as *fully connected topologies*. Topologies that are not fully connected (i.e., general topologies) are referred to as *partially connected topologies*.

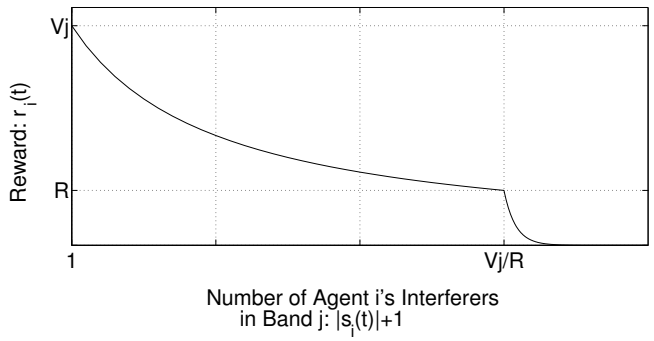


Fig. 1. Elastic reward function: $\beta = 50$ and $V_j/R = 4$.

C. Reward Function

In this paper, we consider the elastic traffic model, which is suitable for elastic applications such as file transfer and web browsing. When considering the elastic traffic, the agent's received reward (i.e., satisfaction) increases proportionally to the amount of service it receives from using the spectrum band. That is, the higher the amount of service, the greater the reward. But when the received QoS level drops below a certain (typically low) threshold R, the agent's reward can become unacceptable quite quickly. In other words, the reward can decrease exponentially with the received QoS level when the received QoS level is below R. The idea here is that even though elastic applications are for e.g. delay tolerant, when the delay becomes significant, the experienced quality can degrade substantially (the reward/satisfaction goes almost immediately to zero). Formally, the reward $r_i(s_i(t))$ (also often referred to simply as $r_i(t)$ for simplicity of notation) contributed by band j to agent i at episode t can be written as:

$$r_i(t) = \begin{cases} V_j/(|s_i(t)|+1) & \text{if} \quad |s_i(t)|+1 \leq V_j/R \\ Re^{-\beta\frac{(|s_i(t)|+1)R-V_j}{V_j}} & \text{otherwise} \end{cases} \quad (1)$$

where $s_i(t)$ denotes the set of agents that interfere with agent i and choose band j at episode t (note: agent i does not belong to the set $s_i(t)$), $|.|$ denotes the cardinality of a set, and β is a reward decaying factor. The parameter β is a design parameter that can be used to tune and adjust the reward model; i.e., β determines how fast the reward decays once the received service drops below the threshold R. The greater the β, the faster the reward decay. In Fig. 1, we show for the sake of illustration the agent reward $r_i(t)$ as a function of the number of agents $|s_i(t)|+1$ (including agent i) when $\beta = 50$ and $V_j/R = 4$.

Note that here each agent assumes that the total amount of service V_j offered by any band j is split equally among all the $|s_i(t)|+1$ interfering agents that use band j at time t. The intuition here is that all agents (i.e. users), by nature, will try to receive as much service as possible once they tune into a spectrum band. This enables each of them to receive roughly the same amount of service, assuming that agents use identical access techniques. For example, when multiple users use a CSMA scheme to share and access a communication medium, they all, on average, roughly receive the same access time share. That is said, we also believe that other scenarios where different agents can receive different amounts of service may exist. However, given that the main contribution of this work lies in the proposed private objective function design (to be presented in later sections), an equally split reward model is used here for simplicity, although variable reward models can also be used by these functions.

From the system's perspective, the system or global reward can be regarded as the sum of all agents' received rewards. Formally, at any episode t, the global reward $G(t)$ is

$$G(t) = \sum_{i=1}^{n} r_i(s_i(t)) \quad (2)$$

where n is the number of all agents in the system. The per-agent average reward $\bar{r}(t)$ at episode t is then

$$\bar{r}(t) = \frac{\sum_{i=1}^{n} r_i(s_i(t))}{n} = \frac{G(t)}{n} \quad (3)$$

III. Motivation and Objective

The goal of this work is to design efficient objective functions for OSA agents, so that when agents aim to maximize them, their collective behaviors lead to good system-level performance, thereby resulting in increasing each agent's long-term received rewards. Hereafter, we let g_i denote agent i's objective function that we will design in this paper.

First, we want to iterate and emphasize that the focus of this work is not on learning, but rather on the design of private objective functions that can be used by any learner. However, in order for us to evaluate the performance of our proposed functions, we had to choose and implement a learning technique. For this, we chose to use the ϵ-greedy Q-learner [39] (with a discount rate of 0 and an ϵ value of 0.05). We chose Q-learning because it is broadly used, can be understood and implemented easily, and fit well with forming the mapping from state-action pairs to the "value" of that state; i.e., it serves well the purpose of evaluating our developed techniques without needing to delve into the complicated mechanics of the learning technique itself.

Therefore, we assume that each agent implements and uses the Q-learner [39]; that is, at each episode t, each agent i aims at maximizing its own private objective function $g_i[t]$ by means of its own ϵ-greedy Q-learner. At the end of every episode, each agent selects and takes the action with the highest entry value with probability $1-\epsilon$, and selects and takes a random action among all possible actions with probability ϵ. After taking an action, the agent then computes the reward that it receives as a result of taking such an action (i.e., as a result of using the selected band), and uses it to update its Q-table. A table entry $Q(a)$ corresponding to action a is updated via $Q(a) \leftarrow (1-\alpha)Q(a)+\alpha u$, where α (here, the value of α is set to 0.5) is the learning rate, and u is the received reward from taking action a. All the results presented in this paper are based on this Q-learner. Readers are referred to [39] for more details on the Q-learning technique.

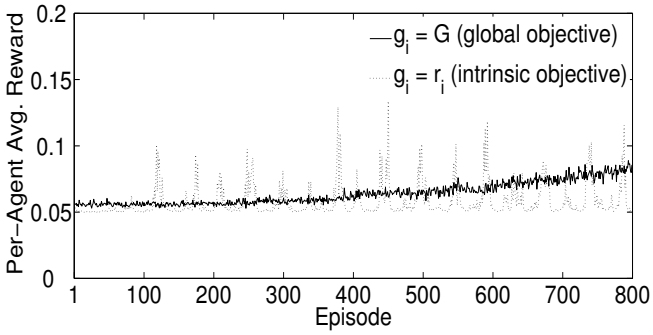


Fig. 2. Per-agent average achieved reward $\bar{r}(t)$ as a function of episode t under the two private objective functions: intrinsic choice ($g_i = r_i$) and global choice ($g_i = G$) for $R = 2$, $\beta = 2$, $V_j = 20$ for $j = 1, 2, \ldots, 10$.

A. Motivation

The key question that arises naturally is which objective function g_i should each OSA agent i aim to maximize so that its received reward is maximized? There are two intuitive choices that one can think of. One possible objective function choice is for each agent i using band j to selfishly go after the intrinsic reward r_i contributed by the band j as defined in Eq. (1); i.e., $g_i = r_i$ for each agent i using band j. A second also intuitive choice is for each agent to maximize the global (i.e., total) rewards received by all agents; i.e., $g_i = G$ for each agent i as defined in Eq. (2). For illustration purposes, we plot in Fig. 2 the per-agent average received reward $\bar{r}(t)$ (measured and calculated via Eq. (3)) under each of these two private objective function choices. In this illustrative experiment, we consider a simple OSA system with a total number of 500 agents and a total number of $m = 10$ bands. There are two important observations that we want to make regarding the performance behaviors of these two objective functions, and that constitute the main motivation of this work. First, note that when agents aim to maximize their own intrinsic rewards (i.e., $g_i = r_i$ for each agent i using band j), the per-agent average received reward presents an oscillating behavior: it ramps up quickly at first but then drops down rapidly too, and then starts to ramp up quickly and drop down rapidly again, and so on, which is explained as follows. With the intrinsic objective function, an agent's reward, by design, is sensitive to its own actions, which enables it to quickly determine the proper actions to select by limiting the impact of other agents' actions, thus learning about good spectrum opportunities fast enough. However, agents' intrinsic objectives are likely not to be aligned with one another, which explains the sudden drop in their received reward right after learning about good opportunities; i.e., right after their received reward becomes high.

The second observation is regarding the second objective function choice, G. Observe that, unlike the intrinsic function, when each agent i sets its objective function g_i to the global reward function G, this results in a steadier performance behavior where the per-agent average received reward increases continuously, but slowly. With this function choice, agents' rewards are aligned with one another by accounting for each other's actions, and thus are less (or not likely to be) sensitive to the actions of any particular agents. The alignedness feature of this function is the reason behind the observed monotonic increase in the average received reward. However, the increase in the received reward is relatively slow due to the function's insensitivity to one's actions, leading then to slow learning rates.

Therefore, it is imperative that private objective functions be designed with two (usually conflicting) requirements in mind: (i) *alignedness*; when agents maximize their own private objectives, they should not end up working against one another; instead, their collective behaviors should result in increasing each agent's long-term received rewards, and (ii) *sensitivity*; objective functions should be sensitive to agents' own actions so that proper action selections allow agents to learn about good opportunities fast enough.

B. Objective

Our aim in this work is to design efficient private objective functions for large-scale, distributed OSA systems. More specifically, we aim to devise private objective functions with the following design requirements and objectives. First, they should be optimal in that they should enable agents to achieve high rewards. Second, they should be scalable in that they should perform well in OSA systems with a small as well as a large number of agents. Third, they should be learnable in that they should enable OSA agents to find and locate spectrum opportunities quickly. Fourth, they should be distributive in that they should be implementable in a decentralized manner. The objective functions that we derive in this work meet all of these design requirements.

IV. Objective Function Design

We first begin by summarizing the concepts of factoredness and learnability, both of which are essential for capturing as well as ensuring the two required design properties: alignedness and sensitivity. Then, we derive and present efficient objective functions that meet the above design requirements by striking a good balance between alignedness and sensitivity.

A. Factoredness and Learnability

First, let g_i denote the private objective function that OSA agent i aims to maximize, which we want to derive in this work. Now, let $z(t)$ characterize the joint move of all OSA agents in the system at time t. Here, the global (total) reward, G, is a function of $z(t)$, which specifies the full system state; hence, $G(t)$ can precisely be written as $G(z(t))$. The system state $z(t)$ basically captures the agent-channel assignment information, which also depends on the actions taken by the agents. Hereafter, we use the notation $-i$ to specify all agents other than agent i, and $z_i(t)$ and $z_{-i}(t)$ to specify the parts of the system state controlled respectively by agent i and agents $-i$ at time t. $z(t)$ can then be written as $z(t) = (z_i(t), z_{-i}(t))$. For simplicity of notation, we often omit throughout the paper the dependency of these states on time t.

For the joint actions of multiple OSA agents to lead to good overall average reward, two requirements must be met. First,

we must ensure that an OSA agent aiming to maximize its own private objective function also leads to maximizing the global (total achievable) rewards, so that its long-term average received rewards are indeed maximized. This means that the agents' private objective functions ($g_i(z)$ for agent i) need to be "aligned" or "factored" with the global reward function ($G(z)$) for a given system state z. Formally, for systems with discrete states, the degree of *factoredness* of a given private objective function g_i is defined as [44]:

$$\mathcal{F}_{g_i} = \frac{\sum_z \sum_{z'} h[(g_i(z) - g_i(z'))\ (G(z) - G(z'))]}{\sum_z \sum_{z'} 1} \tag{4}$$

for all z' such that $z_{-i} = z'_{-i}$, where $h[x]$ is the unit step function, equal to 1 if $x > 0$, and zero otherwise. Intuitively, the higher the degree of factoredness of an agent's private objective function g_i, the more likely it is that a change of state will have the same impact on both the agent's (local) and the system (global) received rewards. A system is fully factored when $\mathcal{F}_{g_i} = 1$.

Second, we must ensure that each OSA agent can discern the impact of its own actions on its private objective function, so that a proper action selection allows the agent to quickly learn about good spectrum opportunities. This means that the agent's private objective function should be more sensitive to its own actions than the actions of other agents. Formally, the level of sensitivity or *learnability* of a private function g_i, for agent i at z, can be quantified as [44]:

$$\mathcal{L}_{i,g_i}(z) = \frac{E_{z'_i}[|g_i(z) - g_i(z_{-i} + z'_i)|]}{E_{z'_{-i}}[|g_i(z) - g_i(z'_{-i} + z_i)|]} \tag{5}$$

where $E[\cdot]$ is the expectation operator, and in this definition, $|.|$ denotes the absolute value operator, z'_i's are parts of the system states, controlled only by agent i, that are resulting from agent i's alternative actions at z, and z'_{-i}'s are parts of the system states, controlled by agent $-i$, that are resulting from agent $-i$'s alternative joint actions. So, at a given state z, the higher the learnability, the more $g_i(z)$ depends on the move of agent i. Intuitively, higher learnability means that it is easier for an agent to achieve higher rewards.

Unfortunately, these two requirements are often in conflict with one another [44]. Therefore, the challenge in designing objective functions for large-scale OSA systems is to find the best tradeoff between factoredness and learnability. Doing so will ensure that agents can learn to maximize their own objectives while doing so will also lead to good overall system performance; i.e., their collective behaviors will not result in worsening each other's received rewards.

B. Efficient Objective Functions

The selection of an agent's private objective function that provides the best performance hinges on balancing the degree of factoredness and the level of learnability. In general and as discussed in the previous section, a highly factored private objective function will experience low learnability, and a highly learnable function will have low factoredness [44].

To provide some intuition on our proposed function design, we will again revisit the behaviors of the global reward function, illustrated earlier in Section III-A. Recall that when agents set the global reward G as their objective functions (i.e., $g_i = G$ for each agent i), their collective behaviors did indeed result in increasing the total system achievable rewards (i.e., did result in a fully factored system), because agents' private objectives are aligned with system objective. The issue, however, is that because G depends on all the components of the system (i.e., all agents), it is too difficult for agents (using G as their objective functions) to discern the effects of their own actions on their objectives, resulting then in low learnability rates.

The key observation leading to the design of our functions is that by removing the effects of all agents other than agent i from the function G, the resulting agent i's private objective function will have higher learnability than G, yet without compromising its alignedness quality. Formally, these functions can be written as

$$D_i(z) \quad \equiv \quad G(z) - G(z_{-i}) \tag{6}$$

where z_{-i} again represents the parts of the state on which agent i has no effect. These difference functions have been shown to lead to good system performance in other domains, such as multi-robot control [45] and air traffic flow regulation [46]. First, note that these proposed functions (D_i for agent i) are fully factored, because the second term of Eq. (6) does not depend on agent i's actions. On the other hand, they also have higher learnability than G, because subtracting this second term from G removes most of other agents' effects from agent i's objective function. Intuitively, since the second term evaluates the value of the system without agent i, subtracting it from G provides an objective function (i.e., D_i) that essentially measures agent i's contribution to the total system received rewards, making it more learnable without compromising its factoredness quality.

By substituting Eq. (2) into Eq. (6), where the implicit dependence on the full state $z(t)$ is replaced for clarity and simplicity with the time t, the objective function D_i for agent i selecting band j at time t can then be written as:

$$\begin{aligned} D_i(t) &= \sum_{k=1}^{n} r_k(s_k(t)) - \left(\sum_{k=1, k\neq i}^{n} r_k(s_k(t) - \{i\}) \right) \\ &= \sum_{k:i\in s_k(t)} r_k(s_k(t)) - \left(\sum_{k:i\in s_k(t)} r_k(s_k(t) - \{i\}) \right) + r_i(s_i(t)) \end{aligned} \tag{7}$$

C. Function Computation Method: A Discussion

Before delving into the study of the effectiveness of the proposed objective functions in terms of their ability to achieve high long-term rewards, their ability to scale well with the number of agents, and their ability to quickly learn about good spectrum opportunities, we want to discuss and shed some light on their practical/implementation aspects (even though the implementation methods are beyond the scope of this work, and are in themselves a different interesting problem).

Note that, by taking away agent i from the second term of the function D_i, the terms corresponding to all spectrum bands k, except the band that agent i is using, cancel out.

This is important because it makes the proposed function D_i easier and simpler to compute than the global function. More specifically, this makes the proposed function implementable in a decentralized manner; agents can implement the function by relying on local information that can be observed by the agent itself (in the case of fully-connected topologies), or gathered and shared among the agents that belong to the same band only (in the case of partially-connected topologies). For e.g. when considering fully-connected networks, Eq. (7) can be written as

$$D_i(t) = (|s_i(t)| + 1)r_i(s_i(t)) - |s_i(t)|r_i(s_i(t) - \{i\}) \quad (8)$$

and as a consequence, $D_i(t)$ becomes dependent only on $s_i(t)$, the set of agents that happen to also be contending for the spectrum band j with agent i at time t. In fact and more precisely, the function $D_i(t)$ depends on $|s_i(t)|$, the number of agents that happen to be contending with agent i (because the function $r_i(t)$ given in Eq. (1) depends on $|s_i(t)|$), and not on the set $s_i(t)$. For the sake of illustration, one can think of $s_i(t)$ as $z(t)$, and of $(s_i(t) - \{i\})$ as $z_{-i}(t)$. Now in order to compute (or precisely estimate) D_i, one needs to estimate $|s_i(t)|$ given the information that agent i observes locally.

Now recall that an agent i using band j can quantify (by relying on local information only) the amount of service/reward it receives once it uses the OSA system, which can for e.g. be measured in terms of the amount of throughput the agent receives. For illustration purposes, let this received throughput be $a_i(t)$. Now assuming that the total amount of throughput/service V_j each bands offers is known, and all agents sharing the same band will roughly receive the same amount of throughput, the number of agents, $|s_i(t)| + 1$, using band j at time t can be estimated to $V_j/a_i(t)$, which is all what is needed for agent i to be able to estimate/compute its function D_i via Eq. (8).

Partially-connected topologies are more challenging, and we are currently in the process of developing distributed methods to compute the proposed functions in such topologies.

V. Optimal Achievable Rewards

In this section, we want to derive the optimal achievable rewards. We first show, through a reduction from a known graph coloring problem, that finding the optimal achievable rewards for general network topologies is an NP-hard problem. We then derive the optimal achievable rewards for fully connected network topologies, which will serve as the basis for our performance comparison, to be shown later in Section VI.

A. The Problem of Finding Optimal/Maximum Achievable Rewards is NP-Hard

We now claim and prove that the problem of finding the optimal/maximum achievable rewards in our studied system is NP-hard for general network topologies.

Proposition 5.1: The problem of finding the optimal achievable system rewards ($maxsys$) is NP-hard.

Proof: First, we do a reduction from graph coloring as a decision problem (GC); i.e., whether a graph is colorable with k colors. GC is a known NP-complete problem. The input to GC(G,k) is a graph G and an integer k, and the output is *'YES'* if the input graph G is colorable with k colors, and *'NO'*[6] otherwise.

Let us refer to the problem of finding the maximum/optimal achievable system rewards as $maxsys$. We assume that we have a $maxsys$ solver (a black box), and try to solve GC using this black box. We need to formulate the GC problem to fit into $maxsys$, but before doing so, we need to define the input and output of the black box accurately. The input of $maxsys$(G, m, c) consists of a graph G (i.e., the network topology), an integer m (i.e., the number of bands/channels), and an integer c (i.e., the channel capacity). The output of $maxsys$ is the value of the maximum achievable system rewards. Now that these parameters are defined, we can explain the reduction. We show that GC(G, k) returns *'YES'* iff $maxsys$(G, k, 1) returns nR which is the product of the number of nodes in the graph G and the threshold R. Note that the channel capacity is set to one in $maxsys$, since no adjacent nodes can have the same color in GC. We show the proof for both necessary and sufficient conditions as follows:

- GC(G, k) returns *'YES'* if $maxsys$(G, k, 1) returns nR: Suppose that $maxsys$(G, k, 1) returns nR, meaning that the system reward is nR. From Eq. (2), we know that the system reward is the sum of n terms, where each term corresponds to the reward value of an agent. Since the channel capacity is set to 1, in order to get a system reward of value nR, each term must be R; i.e., each agent must receive R. This means that in our $maxsys$ problem, each agent is assigned to a channel that has at most $c - 1 = 0$ adjacent agents that share the same channel with it; i.e., there are no adjacent agents (nodes) in the same channel. Bringing this to a graph coloring problem, it follows that no two adjacent nodes have the same color. Hence, GC's output is *'YES'*.
- $maxsys$(G, k, 1) returns nR if GC(G, k) returns *'YES'*: This direction is easier than the first one. Assume that GC returns *'YES'*, meaning that the graph G is k-colorable. Hence, no two adjacent nodes have the same color. This means that no two adjacent agents are assigned to the same channel, and regardless of channel capacity, all agents receive the highest reward since there is no sharing. It follows that the value of service level for each agent is R, and hence, the system reward is nR. Thus, the output of $maxsys$ is nR.

Up to now, we have proven that $maxsys$ is at least as hard as GC. So, $maxsys$ is as hard as all NP-Complete problems. Since $maxsys$ is not a decision problem, it cannot be a NP-complete problem. As a result, $maxsys$ is NP-hard. ■

B. Special Case: Optimal Achievable Rewards in Fully Connected Network Topologies

In this section, we derive a theoretical upper bound on the maximum/optimal achievable rewards for a special case where the network topology is a fully connected graph. This upper bound will serve as a means of assessing how well the developed objection functions perform when compared not only with intrinsic ($g_i = r_i$ for each agent i) and global (g_i=G for

each agent i) functions, but also with the optimal achievable performances.

Without loss of generality and for simplicity, let us assume that $V_j = V$ for $j = 1, 2, \cdots, m$. Let n denote the total number of agents in the system at any time. First, note that when $n \leq m\frac{V}{R}$, the maximum global achievable reward is simply equal to mV (assume $n \geq m$), which corresponds to having each band contain no more than $\frac{V}{R}$ agents. Therefore, in what follows, we assume that $n > m\frac{V}{R}$, and let $c = \frac{V}{R}$, which denotes the capacity (in terms of number of supported OSA agents) of each spectrum band. Now, we start by proving the following lemma, which will later be used for proving our main result.

Lemma 5.2: When the number of agents in each channel is higher than the channel capacity c, the global received reward of an OSA system reduces less when a new OSA agent joins a more crowded spectrum band than when it joins a less crowded band. Otherwise, the global received reward does not decrease.

Proof: First, let us define $G_j(n')$ to be band j's reward where n' is the number of agents using band j. That is,

$$G_j(n') = \begin{cases} V_j & \text{if} \quad n' \leq V_j/R \\ n'Re^{-\beta\frac{n'R-V_j}{V_j}} & \text{otherwise} \end{cases}$$

Note that since the network is fully connected, n' corresponds to the number of all agents using band j. Now when band j has n' agents, if a new agent joins it, the new reward becomes $G_j(n'+1) = (n'+1)Re^{-\beta(\frac{n'+1}{c}-1)}$. It can easily be shown that when $n' > c \geq 1$, $G_j(n') > G_j(n'+1)$; i.e., the reward when joining band j decreases by $\Delta_j(n') \equiv G_j(n') - G_j(n'+1)$. Now we can easily see that $\Delta_j(n')$ decreases when n' increases. Hence, the greater the number n' (i.e., the more crowded the band), the smaller the decrease in reward. For the case $n' < c$, band j reward is $\Delta_j(n') \equiv G_j(n') - G_j(n'+1) \equiv V_j - V_j \equiv 0$, and hence, in this case, if an agent joins a channel, the global received reward does not decrease. ∎

Theorem 5.3: When there are n agents in the system, the global reward reaches its maximal only when $m-1$ bands (out of the total m bands) each has exactly c agents, and the m-th band has the remaining $n - c(m-1)$ agents.

Proof: Let $k = n - mc$, and let us refer to the agent distribution stated in the theorem as C. Note that C corresponds to when $m-1$ bands each has exactly c agents and the other m-th band has the remaining $c+k$ agents (since $n - c(m-1) = c + k$). We proceed with the proof by comparing C with any possible distribution C' among all possible distributions. Let $c + k_1$ be the number of agents in the most crowded band in C', $c+k_2$ be the number of agents in the second most crowded band in C', and so forth. We just need to deal with the bands that each contains more than c agents. If there are p bands each containing more than c agents, then we know that $\sum_{i=1}^{p} k_i \geq k$.

For each band having $c + k'$ agents, let ϵ_i be the amount by which the global reward is reduced when agent i joins the band for $i = 1, 2, \cdots, k'$. From Lemma 5.2, it follows that $\epsilon_i > \epsilon_{i+1} > 0$, for all $i = 1, 2, \cdots, k' - 1$. Note that for the distribution C, the global reward is reduced by $t = \sum_{i=1}^{k} \epsilon_i$, and for C', it is reduced by $t' = \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i$. It remains to show that $t' - t > 0$ for any $C' \neq C$. We consider three different scenarios:

- $k_1 > k$: Here, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \sum_{i=k}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i \end{aligned}$$

  which is greater than zero.
- $k_1 = k$: In this scenario, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i \end{aligned}$$

  which is also greater than zero.
- $k_1 < k$: In this scenario, we have

$$\begin{aligned} t' - t &= \sum_{i=1}^{k_1} \epsilon_i + \sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i - \sum_{i=1}^{k} \epsilon_i \\ &= \underbrace{\sum_{i=1}^{k_2} \epsilon_i + \cdots + \sum_{i=1}^{k_p} \epsilon_i}_{part\ a} - \underbrace{\sum_{i=k_1}^{k} \epsilon_i}_{part\ b} \end{aligned}$$

  Since $k_1 + k_2 + \cdots + k_p \geq k$, the number of ϵ_i terms in $part\ a$ is greater than the number of terms in $part\ b$. From Lemma 5.2, we know that the largest term in $part\ b$ is ϵ_{k_1}, which is smaller than the smallest term ϵ_{k_2} in $part\ a$. Hence, $part\ a$ is greater than $part\ b$, and thus $t' - t$ is greater than zero.

In all scenarios, we showed that $t' - t > 0$. Therefore, the global reward for any distribution C' is smaller than that for the distribution C; i.e., C is the distribution that corresponds to the maximal global achievable reward. ∎

Corollary 5.4: The per-agent average achievable reward is at most $(m-1)V/n + (R - (m-1)V/n)e^{-\beta(\frac{nR}{V}-m)}$.

Proof: The proof follows straightforwardly from Theorem 5.3 by calculating the global achievable reward for the derived optimal agent distribution. ∎

Note that this upper bound (that we derived and stated in Corollary 5.4) is the maximum/optimal average reward that an agent can achieve (it is a theoretical upper bound). In the next section, we will evaluate the performances of the proposed objective functions in terms of their achievable rewards, and compare them against these optimal achievable performances.

VI. Performance Evaluation

In this section, we evaluate and compare the performances of the proposed objective functions in terms of the per-agent average achievable rewards with the optimal achievable rewards calculated through Corollary 5.4 as well as with those achievable under each of the two functions: r_i and G.

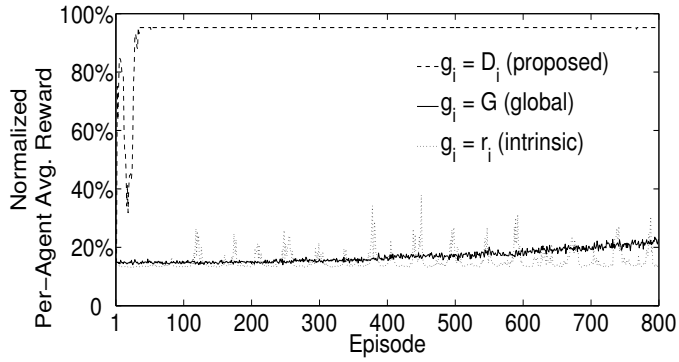


Fig. 3. Per-agent average achieved reward normalized w.r.t. maximum achievable reward under intrinsic function ($g_i = r_i$), global function ($g_i = G$), and proposed function ($g_i = D_i$): $R = 2$, $\beta = 2$, $V = 20$, fully connected topology.

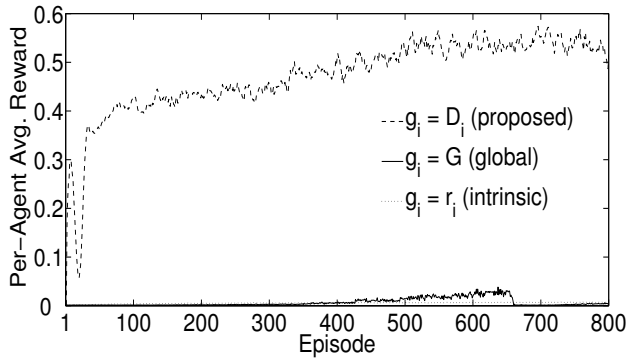


Fig. 4. Per-agent average achieved reward for partially connected network under intrinsic function ($g_i = r_i$), global function ($g_i = G$), and proposed function ($g_i = D_i$): $R = 2$, $\beta = 2$, $V = 20$, partially connected topology.

A. Static OSA Systems

We first consider evaluating the proposed objective functions under the same experiment, conducted in Section III-A, where again the total number of agents considered in this experiment is equal to 500, and that of bands is equal to $m = 10$. Here, the simulated network is static, in that all agents are assumed to enter and leave the OSA system at the same time, and fully connected, in that all agents contend with one another. In this section, we ignore the PUs' activities. Dynamic OSA systems as well as PUs' activities will be considered later in Section VI-B.

Fig. 3 shows the per-agent average achievable reward normalized w.r.t. the optimal achievable reward under each of the three functions: intrinsic ($g_i = r_i$), global ($g_i = G$), and proposed ($g_i = D_i$). The figure clearly shows that the proposed function D_i achieves substantially much better performances than the other two. In fact, when using D_i, an agent can achieve up to about 90% of the total possible, achievable reward, whereas it only can achieve up to about 20% when using any of the other two functions. Another distinguishing feature of the proposed D_i function lies in its learnability; that is, not only does D_i achieve good rewards, but also it does so quite fast, as the received rewards ramp up rapidly, quickly reaching near-optimal performance.

In Fig. 4, we also show the same behaviors when considering partially connected topologies with a total number of agents equal to 1000. In this experiment, the per-agent average number of non-interfering agents is set to 500. Recall that as stated in Proposition 5.1, finding the optimal achievable rewards is an NP-hard problem (i.e., computationally unfeasible), and hence, we cannot compare the performance of our proposed objective functions with the optimal achievable rewards. Instead, we compare them with those achievable under the two existing functions. The figure shows that the proposed objective functions D_i also outperform the other two functions when considering a partially connected topology. For completeness, we plot in Fig. 5 the same performance results but when considering partially connected topologies with varying agent connectivity ratios, where the agent connectivity ratio is defined as the

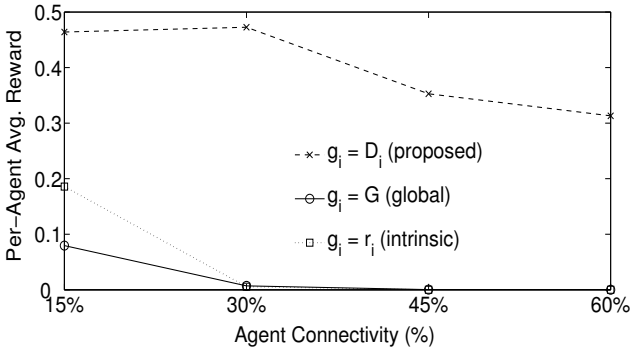


Fig. 5. Per-agent average achieved reward under intrinsic ($g_i = r_i$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of interfering/neighboring agents: $R = 2$, $\beta = 2$, $V = 20$, partially connected topology.

ratio of the per-agent average number of interfering/neighboring agents to the total number of agents. In this experiment, the total number of agents is kept equal to 1600, and the connectivity ratio is varied from 15% to 65%. As it can be seen from the figure, the proposed function outperforms the other two functions regardless of the connectivity ratio of agents.

We also study the proposed function with regard to another performance metric: scalability. For this, we plot in Fig. 6 the per-agent average achievable reward under each of the three studied objective functions when varying the total number of OSA agents in a fully connected topology from 100 to 800. The number of bands m is set to 10. Observe that D_i outperforms the other two functions substantially when it also comes to scalability. Note that D_i achieves high rewards, even for large numbers of agents, whereas the achievable reward under either of the other two functions drops dramatically with the number of agents. This trend is also observed in the case of partially connected topologies as shown in Fig. 5, where the per-average achievable reward under the proposed function is kept relatively high regardless of the per-agent average number of interfering/neighboring agents. We therefore conclude that the proposed function D_i is very scalable, and works well in systems with small as well as large numbers of agents.

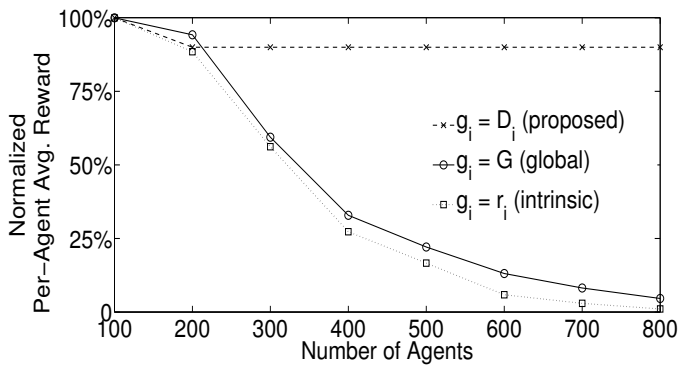


Fig. 6. Per-agent average achieved reward normalized w.r.t. maximum achievable reward under intrinsic ($g_i = r_i$), global ($g_i = G$), and proposed ($g_i = D_i$) functions for various numbers of agents: $R = 2$, $\beta = 2$, $V = 20$, fully connected topology.

To summarize, the obtained results show that the proposed objective function (i) allows agents to achieve high rewards, (ii) is very scalable as it performs well for small- as well as large-scale systems, (iii) is highly learnable as rewards reach up high values very quickly, and (iv) is distributive, as it requires information sharing only among agents belonging to the same band.

B. Dynamic OSA Systems

In this section, we further assess how well these objective functions perform (i) by considering dynamic OSA networks (both with and without PUs' activities), and (ii) by investigating another important performance metric, fairness, in addition to the optimality and scalability metrics. Hereafter, we consider fully connected topologies. Recall that in the previous section, we considered an OSA system in which all agents enter the system, use the available spectrum bands, and then leave the system all at the same time. That is, the number of agents does not change over time, and remains the same during the course of the entire system's lifetime. In this section, we want to investigate how well these obtained results hold when considering dynamic systems, in which agents enter and leave at different independent times. OSA agents can then be viewed as data sessions/flows that start and finish at different times, and independently from one another. We will study these dynamic systems (i) without as well as (ii) with the presence of PUs' activities, and see how well these functions behave under such systems.

1) Without PUs' Activities: To mimic the dynamic behaviors of OSA agents, we assume that agents (e.g., data sessions) arrive according to a Poisson process with arrival rate λ, and stay in the system for an exponentially distributed duration of mean τ. Let $\kappa = \lambda\tau$ represent the average number of agents that are using the system at any time. In this section, we first begin by studying dynamic OSA systems without considering the presence of PUs. The impact of PUs will be investigated in the next section.

Fig. 7 shows the normalized per-agent average received reward for $\frac{\lambda}{\tau} = 1$ when κ equals 500 (Fig. 7(a)), 750 (Fig. 7(b)), and 1000 (Fig. 7(c)). The figure shows that the proposed function D_i still possesses its distinguishing performance features/trends even under dynamic behaviors. These trends are as follows. First, note that the function D_i receives near-optimal rewards, which are more than 90% of the maximal achievable rewards. Second, it is highly learnable as it reaches up near-optimal behaviors quite fast. Third, it is highly scalable as the achievable rewards do not drop below $\approx 90\%$ of the maximal achievable reward even when the average number of agents in the system is increased from 500 to 1000. Fourth, it outperforms the other two functions significantly, and this is regardless of the number of agents.

For completeness, we also study these same performances under different values of the ratio $\frac{\lambda}{\tau}$. Recall that for a given number of agents, the higher the ratio $\frac{\lambda}{\tau}$, the shorter the sessions' durations. For example, when $\kappa = 500$, $\frac{\lambda}{\tau} = 1$ implies that the sessions' average duration τ and arrival rate λ are both equal to ≈ 22.3, whereas $\frac{\lambda}{\tau} = 5$ implies that $\tau = 10$ and $\lambda = 50$. Figs. 8 and 9 show the normalized per-agent average received reward for two more ratios: $\frac{\lambda}{\tau} = 5$ and $\frac{\lambda}{\tau} = 20$. Observe that the performances of D_i are still close to optimal, and are much higher than those obtained under r_i and G regardless of the ratio $\frac{\lambda}{\tau}$; i.e., whether $\frac{\lambda}{\tau}$ equals 1, 5, or 20.

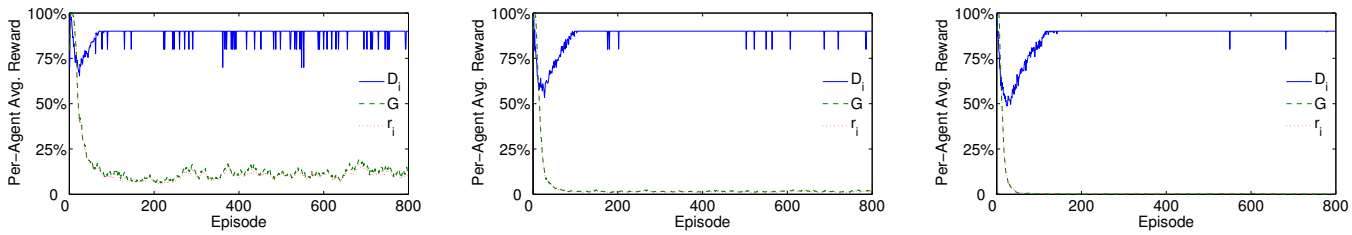
Fairness is also another important performance metric to evaluate. We want to assess how fair D_i is when compared with the other two functions. Fig. 10 depicts the coefficient of variations (CoV)[1] of the per-agent average received rewards under the three studied functions for various combinations of arrival rates λ and durations τ of OSA agents. Observe that D_i achieves CoV values similar to those achievable under any of the other two functions, and this is regardless of sessions' durations and arrival rates. In other words, these results show that the proposed function, when used in practical, dynamic network settings, not only does it achieve good performances in terms of optimality, scalability, and learnability, but it does so while reaching a fairness quality as good as those reached through the other two functions. Next, we show that this is also true in the presence of primary users.

2) With PUs' Activities: We now study the impact of the presence of PUs on the performance of the studied objective functions. In this study, we consider the same dynamic model used in Section VI-B1 to mimic the behavior of SUs, but while assuming and accounting for PUs' activities. Here, PUs' activities are: $a)$ mimicked via computer simulation by modeling the PUs' behavior via renewal ON/OFF process, and $b)$ obtained via real data traces collected by monitoring and measuring primary users' activities in real networks [47][2].

a) Evaluation based on mimicking PUs' activities via simulation: We assume that each band is associated with a set of PUs that enter and leave the band at random times. We model PUs' activities on each band as a renewal process alternating between ON and OFF periods, which represent the time during which PUs are respectively present (ON) and absent

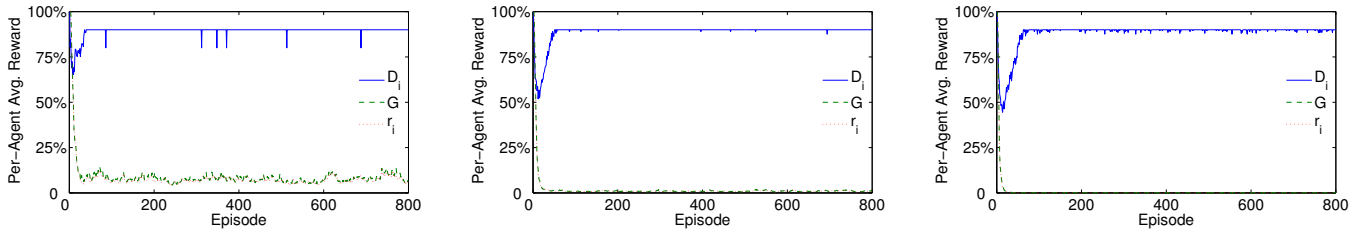
[1]CoV is the ratio of the standard deviation to the mean of the agents' received rewards; we use this metric as a means of assessing the fairness, which reflects how close agents' received rewards are to one another.

[2]The data traces were collected and sent to us by the authors of [47]. We thank them for providing us with this valuable data



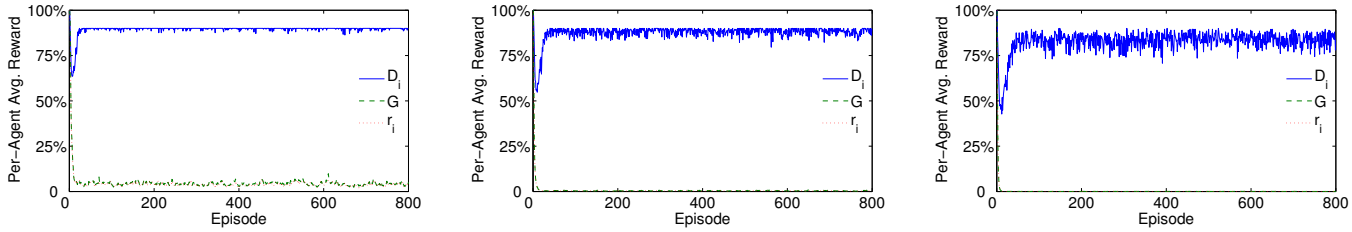
(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 7. Normalized per-agent average received reward under Poisson arrival traffic model: $\lambda/\tau = 1$



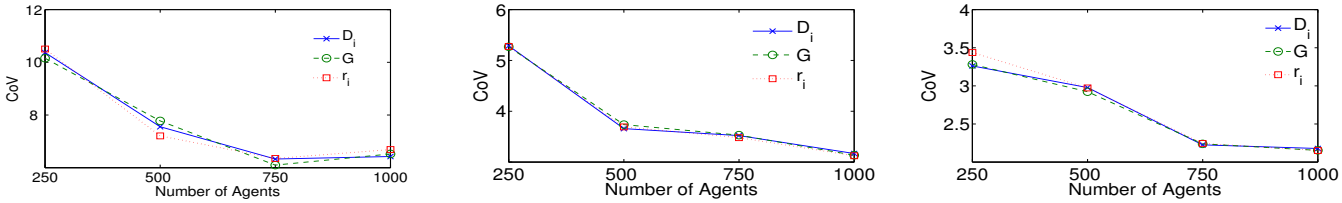
(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 8. Normalized per-agent average received reward under Poisson arrival traffic model: $\lambda/\tau = 5$



(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 9. Normalized per-agent average received reward under Poisson arrival traffic model: $\lambda/\tau = 20$



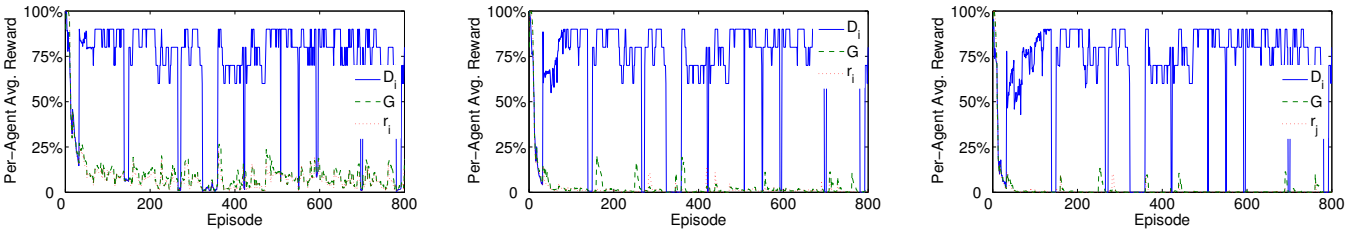
(a) $\lambda/\tau = 1$ (b) $\lambda/\tau = 5$ (c) $\lambda/\tau = 20$

Fig. 10. Coefficient of variation (CoV) of normalized per-agent average reward under r_i, G, and D_i: dynamic OSA system without PUs' presence

(OFF). For each spectrum band j, we assume that ON and OFF durations are exponentially distributed with means ν_j^{ON} and ν_j^{OFF}, respectively[3]. We use $\eta_j \equiv \nu_j^{ON}/(\nu_j^{OFF} + \nu_j^{ON})$ to denote the *PU traffic load* on band j.

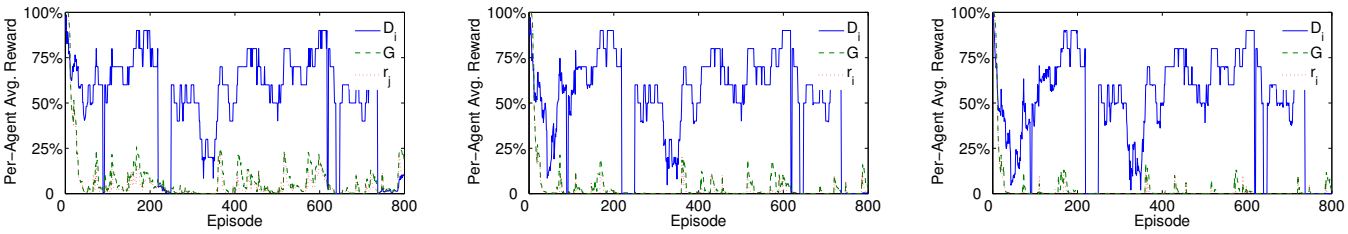
Figs. 11, 12, and 13 show the normalized per-agent average reward for PU traffic loads η of 10%, 30%, and 50%, respectively ($\eta = \eta_j \ \forall j$). The subfigures in each figure each corresponds to a different average number of agents (i.e., for $\kappa = 500$, $\kappa = 750$, and $\kappa = 1000$). There are three observations that we can make out of these results. First, observe that the proposed function D_i achieves rewards higher than those achieved under the other two functions, and for all combinations of number of agents κ and primary user loads η. Second, unlike in the case of dynamic OSA without primary users, as expected, the achievable rewards reach zero when primary users are present, but quickly reach up high values as soon as PUs leave their bands. Third, also as expected, when the PU traffic load increases (i.e., η is increased), the total achievable average reward decreases, since rewards will also be taken away by PUs themselves. For example, Fig. 13 shows that when the PU load $\eta = 50\%$, D_i reaches up to only about 50% of the maximal achievable reward. This is because 50% of the total reward has already been received by the primary users (i.e., $\eta = 50\%$).

[3]Recall that learners do not actually need prior knowledge of primary users' traffic behavior. Here, the exponential distributions will be used to generate samples so as to be able to mimic the OSA environment.



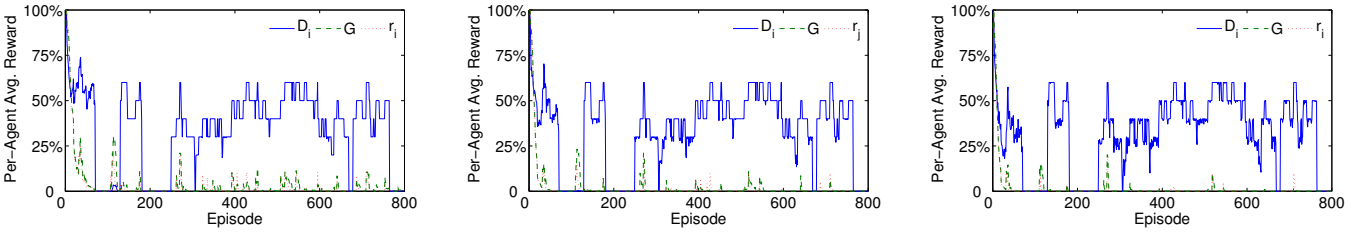
(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 11. Normalized per-agent average reward under OSA agent traffic with Poisson arrival of $\frac{\lambda}{\tau} = 1$ and with simulated PUs traffic load of $\eta = 10\%$



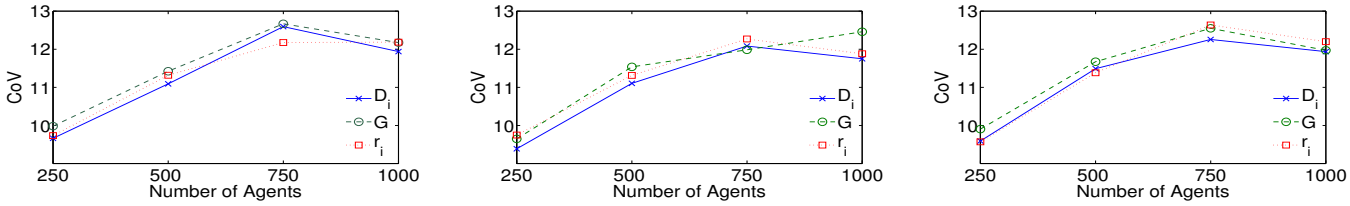
(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 12. Normalized per-agent average reward under OSA agent traffic with Poisson arrival of $\frac{\lambda}{\tau} = 1$ and with simulated PUs traffic load of $\eta = 30\%$



(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 13. Normalized per-agent average reward under OSA agent traffic with Poisson arrival of $\frac{\lambda}{\tau} = 1$ and with simulated PUs traffic load of $\eta = 50\%$



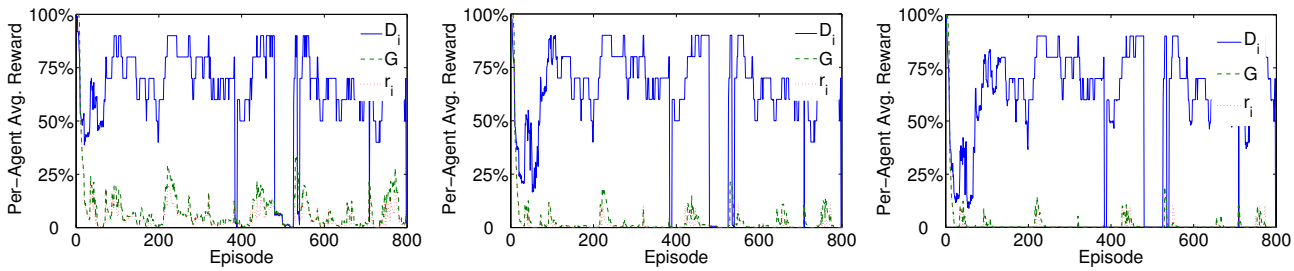
(a) $\eta = 10\%$ (b) $\eta = 30\%$ (c) $\eta = 50\%$

Fig. 14. Coefficient of variation (CoV) of normalized per-agent average achieved reward under r_i, G, and D_i, under various simulated PUs traffic loads

However, this achieved reward under D_i is still considered as about 90% of the total available/achievable rewards. Thus, we conclude that the proposed function yields performances that are close to optimal even in the presence of PUs, and this is true regardless of the number of OSA agents and/or the PU traffic load.

We now show in Fig. 14 the coefficient of variations (CoV) of the per-agent average received rewards under the three studied functions for various PU traffic loads. Like when PUs are absent, when PUs are present, we also observe that the proposed objective function achieves CoVs similar to those achievable under any of the other two functions, independently of PU traffic loads. We also observe that CoV increases with the number of agents.

b) Evaluation based on real data trace of PUs' activities: In this section, we evaluate the performance of the proposed functions by relying on real data traces collected by monitoring and measuring the activities of real PUs. The authors in [47] collected and sent us this data traces, which were measured over a period of 100 minutes through spectral measurements of primary users' activities in the 850-870MHz band at every 0.01 second with a frequency resolution of 8.333kHz [47]. The data measurements are collected over 60 channels each having a bandwidth of 25kHz, and a channel is considered idle when



(a) $\kappa = 500$ (b) $\kappa = 750$ (c) $\kappa = 1000$

Fig. 15. Normalized per-agent average reward obtained through real data traces of PUs' activities: Poisson arrival of OSA agent traffic with $\frac{\lambda}{\tau} = 1$.

it exhibits a low power; that is, when the measured energy level is below a threshold. More details on these traces can be found in [47].

Fig. 15 shows the normalized per-agent average reward under each of the studied objective functions where here PUs' activities are based on the above mentioned real data trace. The subfigures in each figure each corresponds to a different average number of agents (i.e., for $\kappa = 500$, $\kappa = 750$, and $\kappa = 1000$). Note that the performances of the proposed functions under real PUs traffic are as good as those obtained when PUs' activities are mimicked via simulation (as shown in the previous section). Precisely, the proposed functions are shown to achieve high rewards regardless of OSA agents' traffic load, and this is whether mimicking the PUs' activities via renewal ON/OFF process or considering real data traces of PUs' activities.

To summarize, the proposed objective functions are shown to achieve good performances in terms of optimality, scalability, and learnability while also reaching a fairness quality as good as those reached through the other two functions.

VII. Discussion: Implementation Incentives and Selfish Behaviors

One key question that arises naturally is: how would one impose an objective function on an independent entity in the system or why would an entity follow a "new" extrinsic objective function? There are three types of answers to this philosophical question: one from a theoretical, one from a practical, and one from an algorithmic perspectives. First, from a theoretical perspective, one can view the extrinsic objectives as a combination of the intrinsic objectives and an incentive. The issue then becomes how to devise an incentive that makes the intrinsic objective as close to the desired utility as possible. This line of research moves into mechanism design [48–50], and in fact there has been work on reconciling the incentive design with difference objectives [51]. Second, from a practical perspective, maximizing the extrinsic objective may in some cases lead to higher values of the intrinsic objective. That is, by aiming to maximize the provided extrinsic objective, the entity may not only help maximize a global objective function, but also do better in terms of its intrinsic objective [52]. In such cases, there is a direct incentive for an entity to accept the extrinsic objective function. Finally, from an algorithmic perspective, one can simply provide a "swapping" layer for the objectives of each entity where the objective functions are randomly changed from one entity to another. In such a system, each entity may receive the gains of another entity at different time steps, creating a strong incentive to take actions to benefit the full system.

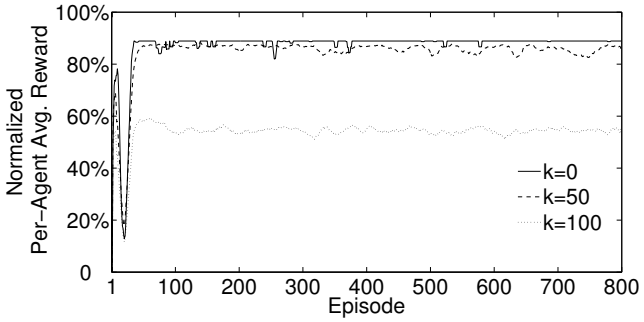


Fig. 16. Per-agent average achieved reward normalized w.r.t. maximum achievable reward under the proposed function (D_i) with $k = 0, 50, 100$ selfish agents among a total of 500 agents: $R = 2$, $\beta = 2$, $V = 20$.

For the sake of illustration, we run an experiment where a number k of agents (among 500 agents) choose their intrinsic objectives while all other agents choose the proposed extrinsic objectives, and show in Fig. 16 the per-agent average received rewards for various values of k. Note that having selfish users/agents in the system reduces the per-agent average rewards, and this reduction can be very substantial, especially when the number of selfish agents, k, is high. Clearly and as demonstrated in this work, when all agents go after their intrinsic objectives (i.e., $k = 500$), their collective behavior will degrade the overall system performance substantially, leading to the degradation of each agent's average received rewards as well. It is therefore the fraction of selfish agents in the system what determines the achievable performances, and as such, it is what gives the incentives for whether an agent should behave selfishly. That is, when this fraction is small, selfish agents may receive higher rewards than the overall per-agent average received rewards, but this is only when a few agents behave selfishly while the rest does not.

VIII. Acknowledgment

The authors would like to thank Omar Alsaleh for his help and the authors of [47] for the data traces.

IX. Conclusion

In this paper, we propose and evaluate efficient private objective functions that OSA users can use to locate the best spectrum opportunities. OSA users can rely on any learning algorithms to maximize these proposed objective functions, thereby ensuring near-optimal performances in terms of the long-term average received rewards. We showed that these proposed functions (i) receives near-*optimal* rewards, (ii) are highly *scalable* as they perform well for small- as well as large-scale systems, (iii) are highly *learnable* as rewards reach up near-optimal values very quickly, and (iv) are *distributive* as they require information sharing only among users belonging to the same spectrum band.

References

[1] M. McHenry and D. McCloskey, "New York city spectrum occupancy measurements," *Shared Spectrum Conf.*, Sept. 2004.

[2] H. Su and X. Zhang, "Cross-layer based oppotrunitic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Communications*, January 2008.

[3] J. Jia, Q. Zhang, and X. Shen, "HC-MAC: a hardware-constrainted cognitive MAC for efficient spectrum management," *IEEE Journal on Selected Areas in Communications*, January 2008.

[4] B. Hamdaoui and K. G. Shin, "OS-MAC: An efficient MAC protocol for spectrum-agile wireless networks," *IEEE Transactions on Mobile Computing*, August 2008.

[5] C. Zou and C. Chigan, "On game theoretic DSA-driven MAC for cognitive radio networks," *Computer Communications*, vol. 32, no. 18, 2009.

[6] M. Ma and D. H. K. Tsang, "Joint design of spectrum sharing and routing with channel heterogeneity in cognitive radio networks," *Physical Communication*, vol. 2, no. 1-2, 2009.

[7] H. A. Bany Salameh, M. Krunz, and O. Younis, "Cooperative adaptive spectrum sharing in cognitive radio networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 4, 2010.

[8] M. Timmers, S. Pollin, A. Dejonghe, L. Van der Perre, and F. Catthoor, "A distributed multichannel MAC protocol for multihop cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 1, 2010.

[9] S. Srinivasa and S. A. Jafar, "Cognitive radio networks: how much spectrum sharing is optimal?," in *Proceedings of IEEE GLOBECOM*, 2007.

[10] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, Februray 2008.

[11] S. Stotas and A. Nallanathan, "Overcoming the sensing-throughput tradeoff in cognitive radio networks," in *Proceedings of IEEE ICC*, 2010.

[12] Y. Ma, D. I. Kim, and Z. Wu, "Optimization of OFDMA-based cellular cognitive radio networks," *IEEE Transactions on Communications*, vol. 58, no. 8, 2010.

[13] X. Liu, B. Krishnamachari, and H. Liu, "Channel selection in multi-channel opportunistic spectrum access networks with perfect sensing," in *Proceedings of IEEE DySPAN*, 2010.

[14] Z. Ji and K. J. R. Liu, "Multi-stage pricing game for collusion-resistant dynamic spectrum allocation," *IEEE Journal on Selected Areas in Communications*, January 2008.

[15] L. Chen, S. Iellamo, M. Coupechoux, and P. Godlewski, "An auction framework for spectrum allocation with interference constraint in cognitive radio networks," in *Proceedings of IEEE INFOCOM*, 2010.

[16] L. Duan, J. Huang, and B. Shou, "Competition with dynamic spectrum leasing," in *Proceedings of IEEE DySPAN*, 2010.

[17] G. S. Kasbekar and S. Sarkar, "Spectrum auction framework for access allocation in cognitive radio networks," in *Proceedings of ACM MobiHoc*, 2009.

[18] J. Jia, Q. Zhang, and M. Liu, "Revenue generation for truthful spectrum auction in dynamic spectrum access," in *Proceedings of ACM MobiHoc*, 2009.

[19] S. Delaere and P. Ballon, "Flexible spectrum management and the need for controlling entities for reconfigurable wireless systems," in *Proceedings of IEEE DySPAN*, 2007.

[20] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE Journal on Selected Areas in Communications*, January 2008.

[21] R. W. Thomas, L. A. DaSilva, M. V. Marathe, and K. N. Wood, "Critical design decisions for cognitive networks," in *Proceedings of IEEE ICC*, 2007.

[22] X. Jing and D. Raychaudhuri, "Global control plane architecture for cognitive radio networks," in *Proceedings of IEEE ICC*, 2007.

[23] S. Yarkan and H. Arslan, "Exploiting location awareness toward improved wireless system design in cognitive radio," *IEEE Communications Magazine*, January 2008.

[24] G. Gur, S. Bayhan, and F. Alagoz, "Cognitive femtocell networks: an overlay architecture for localized dynamic spectrum access," *IEEE Wireless Communications*, vol. 17, no. 4, 2010.

[25] T. R. Newman, S. M. S. Hasan, D. Depoy, T. Bose, and J. H. Reed, "Designing and deploying a building-wide cognitive radio network testbed," *IEEE Communications Magazine*, vol. 48, no. 9, 2010.

[26] J. Unnikrishnan and V. V. Veeravalli, "Dynamic spectrum access with learning for cognitive radio," in *Proc. of Asilomar Conference on Signals Systems and Computers*, Oct. 2008.

[27] Q. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Transactions on Signal Processing*, February 2008.

[28] Z. Han, C. Pandana, and K. J. R. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proceedings of IEEE WCNC*, 2007.

[29] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *Proceedings of IEEE ICC*, 2008.

[30] K. Liu and Q. Zhao, "Distributed learning in cognitive radio networks: multi-armed brandit with distributed multiple players," in *Proceddings of IEEE Int. Conf. on Acousitcs Speech and Signal Processing*, March 2010.

[31] U. Berthold, M. Van Der Schaar, and F. K. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *Proceedings of IEEE DySPAN*, 2008, pp. 1–5.

[32] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access," in *Proceedings of the SPIE Conf. on Advanced Signal Processing Algorithms, Architectures, and Implementations*, August 2007.

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

[34] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. on Inf. Theory*, May 2008.

[35] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *Proceddings of IEEE DySPAN*, 2010.

[36] A. Anandkumar, N. Michael, and A. K. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proceddings of IEEE INFOCOM*, 2010.

[37] A. Anandkumar, N. Michael, A.K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE JSAC*, vol. 29, no. 4, April 2011.

[38] J. Unnikrishnan and V. V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, August 2010.

[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.

[40] G. Tesauro, "Practical issues in temporal difference learning," *MLJ*, vol. 8, pp. 257–277, 1992.

[41] A. Agogino and K. Tumer, "Unifying temporal and structural credit assignment problems," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, New York, NY, July 2004.

[42] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[43] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, pp. 1243–1248, 1968.

[44] A. K. Agogino and K. Tumer, "Analyzing and visualizing multiagent rewards in dynamic and stochastic environments," *Journal of Autonomous Agents and Multi Agent Systems*, vol. 17, no. 2, pp. 320–338, 2008.

[45] A. K. Agogino and K. Tumer, "Efficient evaluation functions for evolving coordination," *Evolutionary Computation*, vol. 16, no. 2, pp. 257–288, 2008.

[46] K. Tumer and A. Agogino, "Distributed agent-based air traffic flow management," in *Proceedings of the Sixth International Joint Conference*

on Autonomous Agents and Multi-Agent Systems, Honolulu, HI, May 2007, pp. 330–337.
[47] D. Xu, E. Jung, and X. Liu, "Optimal bandwidth selection in multi-channel cognitive radio networks: how much is too much?," in *Proceedings of IEEE DySPAN*, 2008.
[48] T. Groves, "Incentives in teams," *Econometrica*, vol. 41, pp. 617–631, 1973.
[49] D. Parkes and S. Singh, "An MDP-based approach to online mechanism design," in *NIPS 16*, 2004, pp. 791–798.
[50] D. Parkes and J. Shneidman, "Distributed implementations of Vickrey-Clarke-Groves mechanisms," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems*, New York, NY, July 2004, pp. 261–268.
[51] K. L. Milkman, J. Burns, D. C. Parkes, G. Barron, and K. Tumer, "Testing a purportedly more learnable auction mechanism," *Applied Economics Research Bulletin, Special Issue (Auctions)*, vol. 1, pp. 107–140, 2008.
[52] K. Tumer, Z. T. Welch, and A. Agogino, "Aligning social welfare and agent preferences to alleviate traffic congestion," in *Proc. of the Seventh Int'l Joint Conf. on Autonomous Agents and Multi-Agent Systems*, Estoril,Portugal, May 2008.

PLACE PHOTO HERE

MohammadJavad NoroozOliaee received the BS degree in Computer Engineering from Sherif University of Technology, Iran, in 2009. He is currently working toward the PhD degree in Computer Science at Oregon State University. His research focus is on the design and development of distributed coordination techniques for wireless networking systems with dynamic spectrum access capabilities.

PLACE PHOTO HERE

Bechir Hamdaoui received the Diploma of Graduate Engineer (1997) from the National School of Engineers at Tunis, Tunisia. He also received M.S. degrees in both Electrical and Computer Engineering (2002) and Computer Sciences (2004), and the Ph.D. degree in Computer Engineering (2005) all from the University of Wisconsin at Madison. In September of 2005, he joined the RTCL Lab at the University of Michigan at Ann Arbor as a postdoctoral researcher. Since September of 2007, he has been with the School of EECS at Oregon State University as an assistant professor. His research focus is on cross-layer protocol design, system performance modeling and analysis, adaptive and learning technique development, and resource and service management for next-generation wireless networks and communications systems. He has won the NSF CAREER Award (2009), and is presently an Associate Editor for IEEE Transactions on Vehicular Technology (2009-present) and Wireless Communications and Computing Journal (2009-present). He also served as an Associate Editor for Journal of Computer Systems, Networks, and Communications (2007-2009). He served as the chair for the 2011 ACM MobiCom's Student Research Competition, and as the program chair/co-chair of the Pervasive Wireless Networking Workshop (PERCOM 2009), the WiMAX/WiBro Services and QoS Management Symposium (IWCMC 2009), the Broadband Wireless Access Symposium (IWCMC 2010), the Cooperative and Cognitive Networks Workshop (IWCMC 2011 and 2012), and the Internet of Things, Machine to Machine, and Smart Services Applications Workshop (CTS 2012). He also served on program committees of several IEEE conferences. He is a member of IEEE, IEEE Computer Society, IEEE Vehicular Society, and IEEE Communications Society.

PLACE PHOTO HERE

Kagan Tumer is a professor at Oregon State University. He received his PhD (1996) and MS (1992) in Electrical and Computer Engineering at The University of Texas, Austin. Prior to joining OSU, he was a senior research scientist at NASA Ames Research Center (1997-2006). Dr. Tumer's research interests are learning and control in large autonomous systems with a particular emphasis on multiagent coordination. Applications of his work include coordinating multiple robots, controlling unmanned aerial vehicles, reducing traffic congestion and managing air traffic.

His work has led to over one hundred publications, including three edited books, one patent, and a best paper award at the 2007 Conference on Autonomous Agents and Multiagent Systems. He is an associate editor of the Journal on Autonomous Agents and Multiagent Systems. He was the program co-chair of the 2011 Autonomous Agents and Multiagent Systems Conference (AAMAS 2011). Dr. Tumer is a member of AAAI and a senior member of IEEE.