INVITED PAPER

Spectrum Access Games and Strategic Learning in Cognitive Radio Networks for Delay-Critical Applications

This interactive framework aims at efficient use of the communications spectrum by allowing users to compete, within limitations, for network resources.

By Mihaela van der Schaar, *Senior Member IEEE*, and Fangwen Fu, *Student Member IEEE*

ABSTRACT | With the current proliferation of high bandwidth and delay-sensitive multimedia applications and services, each wireless user will try to maximize its utility by acquiring as much spectrum resources as possible unless a preemptive mechanism exists in the network. Thus, emerging solutions for dynamic spectrum access in cognitive radio networks will need to adopt market-based approaches in order to effectively regulate the available resources. In this paper, we show how various centralized and decentralized spectrum access markets can be designed based on a stochastic game framework, where wireless users (also referred to as secondary users) can compete over time for the dynamically available transmission opportunities (spectrum "holes"). When operating in such spectrum access "markets," wireless users become selfish, autonomous agents that strategically interact in order to acquire the necessary spectrum opportunities. We also show how wireless users can successfully compete with each other for the limited and time-varying spectrum opportunities, given the experienced dynamics in the wireless network, by optimizing both their external actions (e.g., the resource bids, power and channel used for transmission, etc.) and their internal actions (e.g., the modulation schemes, etc.). To determine their optimal actions in an informationally decentralized setting, users will need to learn and model directly or indirectly the other users' responses to their external actions. We studied the outcome of various dynamic interactions among self-interested wireless users possessing different knowledge and determine that the proposed framework can lead to multiuser communication systems that achieve new measures of efficiency, rationality and fairness. Lastly, our illustrative results show that the presented game-theoretic solution for wireless resource management enables users deploying enhanced ("smarter") learning and communication algorithms and being able to make efficient use of the spectrum resources can derive higher utilities. This presents the designers of wireless devices and systems with important incentives to endow their next-generation products and services with enhanced capabilities to gather information, learn, and strategically compete for resources in the emerging spectrum resource markets made possible by the cognitive radio network technologies.

KEYWORDS | Channel access games; cognitive radios; knowledge-driven wireless networking; multiuser wireless communication; spectrum access; stochastic games; strategic learning; wireless resource markets

I. INTRODUCTION

A. Motivation

Due to their flexible and low-cost infrastructure, wireless networks are poised to enable a variety of delay-sensitive multimedia transmission applications, such as videoconferencing, emergency services, surveillance, telemedicine, remote teaching and training, augmented reality, and distributed gaming. However, existing wireless networks provide limited, time-varying resources with only limited support for the quality of service (QoS) required by the *delay-sensitive, bandwidth-intense, and loss-tolerant*

Manuscript received November 5, 2008. Current version published April 15, 2009. This work was supported by grants from NSF Career and ONR.
The authors are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: mihaela@ee.ucla.edu).

Digital Object Identifier: 10.1109/JPROC.2009.2013036

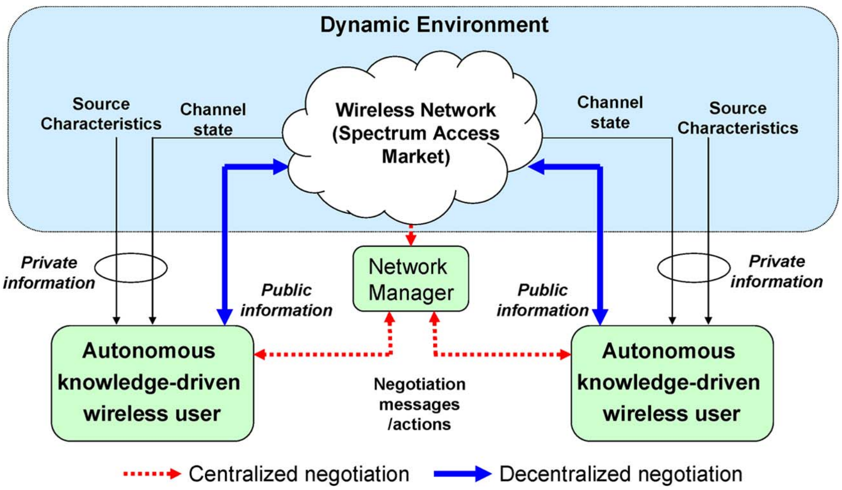

Fig. 1. ***Coupling and information flow between wireless users.***

multimedia applications. The scarcity and variability of resources does not significantly impact delay-insensitive applications (e.g., file transfers) but has considerable consequences for multimedia applications and often leads to an unsatisfactory user experience.

To cope with these challenges, the research focus has been in recent years to adapt the resource allocation methods (e.g., integrated and differentiated services) and real-time transmission (e.g., TCP) strategies and concepts designed for the wired (Internet, ATM) communications to the time-varying and bandwidth-constrained wireless networks. However, such solutions (e.g., QoS-enabled 802.11e solutions) do not provide fair or efficient support for delay-sensitive applications such as multimedia streaming in crowded or dynamic wireless networks [12] because they ignore the coupling between the various users concurrently transmitting traffic over the same infrastructure as well as the wireless system dynamics, including the time-varying source and channel characteristics, the mobility of the wireless sources, the unpredictability of wireless users or interference sources coming or leaving the network, etc. Hence, existing solutions do not provide adequate support for multimedia applications in crowded wireless networks when the interference is high or when the stations are mobile.

To fulfill the necessary QoS requirements under such conditions, wireless stations need to harvest additional resources as well as optimally adapt their transmission strategies to the available resources. A key technology enabling users to harvest additional resources is the emerging cognitive radio networks. One vision for such networks assumes that certain portions of the spectrum will be opened up for secondary users[1] (SUs), such as wireless multimedia applications, to autonomously and opportunistically share the spectrum becoming available once primary users (PUs) are not active [1]–[3]. Importantly, in cognitive radio networks, heterogeneous wireless users (with different utility-rate functions, delay tolerances, traffic characteristics, knowledge, and adaptation abilities) will need to coexist and interact within the same band [5]. However, despite the increased amount of available spectrum resources, to enable the proliferation of multimedia applications over cognitive radio networks, solutions for dynamic spectrum access will need to address the different challenges outlined subsequently.

[1]The secondary users/applications are envisioned in this paper to be a single transmitter–receiver pair.

B. Challenges for Dynamic Spectrum Access in Cognitive Radio Networks

Next-generation networking solutions for spectrum access will need to address, besides other issues related to the coexistence between the PUs and SUs, the following four challenges associated with designing efficient resource management solutions for dynamic and autonomous applications over wireless environments (see Fig. 1).

- A first challenge arises due to the *dynamic, time-varying nature of applications and source and channel characteristics*. As the source characteristics are changing, the delays that are tolerable at the application layer and the derived utility (e.g., quality or fidelity) can vary significantly. This influences the performance of the different transmission strategies at the various layers and, ultimately, the choice of the optimal strategy adopted by the transmitter. Hence, the utility that a user derives from using a certain resource dynamically varies over time, depending on both the "environment" (e.g., application, source and channel characteristics), which is not in the control of the user, and the user's response to this environment, which is the selected cross-layer transmission strategy [at the application, transport, network, media access control (MAC), or physical layers] [7].

- A second challenge associated with multiuser transmission and resource management is that the *wireless users' actions and their performances are coupled* [2], [34], since the transmission strategy of a user impacts and is impacted by the competing users (see Fig. 1). Hence, a user's actions will have a direct impact not only on its own utility but also on the performance of the other wireless users sharing the spectrum. Thus, the interaction among users can be viewed as a game played by the various users for the same spectrum resource. The use of games for dynamic spectrum access in cognitive radio networks was discussed in several early publications, including [2], [25], and [34].
- A third challenge comes from the *informationally decentralized and heterogeneous nature of the multiuser interaction in the wireless resource market* (see Fig. 1). Microeconomics today mainly studies the design of game or system rules to achieve a specific outcome (e.g., mechanism design, etc.), tries to explain the behavior of supply, demand, and prices of existing markets (e.g., general equilibrium theory), or to characterize the equilibrium outcomes of given games (e.g., game theory). It most often neglects how users acquire information; how they interact and successfully compete with each other in repeated or stochastic games based on their local asymmetric information; how they learn over time based on this information; and how they cope with time-varying utilities and environments. Also, it rarely discusses how informationally decentralized games can be constructed and equilibriums be selected such that the users can play a specific, hopefully efficient equilibrium, based solely on their local information. It does not discuss how multiuser interactions can be shaped by network policers making private observations about the users' interactions and deploying only minimal interventions to induce users to adopt desired behaviors [39]. Hence, new theories and solutions need to be developed to address these challenges, which often arise in spectrum access markets.
- Lastly, by imposing rigid rules for wireless spectrum access among users, most existing wireless resource management solutions disregard two important properties of the *autonomous wireless users:* their heterogeneous knowledge (and thus ability to learn and optimize their transmission strategies by anticipating the coupling with the other users and the impact of their actions on both their immediate and their long-term utilities) and their self-interest in maximizing their own utilities. Such rigid regulation may often result in *inefficient resource allocations or even the manipulation of spectrum access rules* [23], [39]. Particularly, in a congested network, if some users inefficiently utilize the spectrum because they deploy old technologies or they inefficiently optimize their cross-layer strategies, the performance of the entire wireless network may significantly degrade [7], [20]. Existing spectrum access rules do not prevent wireless users from inefficiently using resources or even exaggerating their resource requirements at the expense of competing users [20]. This is especially important when multimedia applications are deployed, since these require a high bandwidth. Also, they do not provide incentives to the users to minimize their resource usage in order to limit their impact on the utilities and costs of other users. Hence, the lack of incentives in current wireless networks for users to declare their information truthfully, to use resources efficiently, or to adhere to fairness or courtesy rules will ultimately lead to a tragedy of commons, since there is no incentive, other than the ultimate survival of the system, for users to limit their use [14]. To avoid this, deterrents for spectrum misuse are required, such as charging for spectrum utilization or deploying spectrum policers that can intervene only when users do not adhere to existing spectrum access policies [39].

C. Proposed Knowledge-Driven Multiuser Networking Paradigm

A new networking paradigm is needed to address the above-mentioned challenges for managing, characterizing, and optimizing multiuser communication systems, such that delay-sensitive multimedia applications and services will be able to proliferate over next-generation cognitive radio networks [2].[2] To enable spectrum access to be efficiently and fairly divided among heterogeneous and self-interested users, we propose in this paper a mathematical framework that enables us to design, analyze, and optimize dynamic multiuser environments and applications as "markets."

To create a market-based resource management solution, we introduce in this paper a new way of architecting multiuser communications systems, where the spectrum access is governed by *market-driven* spectrum access rules [2], [9] and where SUs can compete with each other based on their available transmission strategies as well as their

[2] http://www.infoworld.com/article/06/04/06/77219_HNspectrum-frenzy_1.html.

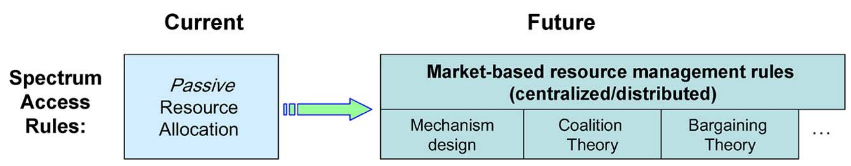

Fig. 2. *Evolution of spectrum access rules to create a dynamic wireless resource market.*

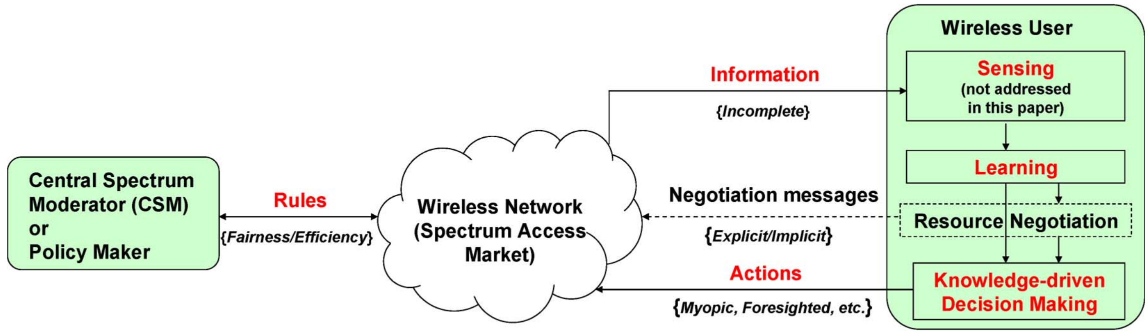

Fig. 3. *Knowledge-driven wireless networking.*

heterogeneous knowledge about the environment and other SUs. Specifically, wireless users can engage in noncollaborative (e.g., mechanism design [13], [20]) or even cooperative[3] (e.g., bargaining or coalition theory [26]) interactions with other users (see Fig. 2). In this case, when cooperation between users emerges, it is based on their self-interest and is self-enforcing rather than in current networks, where cooperation is mandated by a rigid, a priori established protocol.

We model the dynamic, repeated interactions among the heterogeneous devices as stochastic or repeated games played over time based on the dynamic changes in the network environment and the delay-critical characteristics of the applications. To enable the devices to successfully participate in the resource competition and maximize their performances as well as the overall network performance, we need to carefully "design" the devices' utilities and allow them to strategically compete for resources. This is unlike current communication solutions, which requires them to follow prescribed and rigid behaviors and adaptation rules. In our proposed paradigm, the devices become *cognitive* entities, which can harvest knowledge about each other as well as the environment by proactively learning based on their locally available information in order to strategically maximize their utilities (see Fig. 3). Note that the heterogeneous devices may have different utilities and cost–performance tradeoffs, and that the information acquired by different devices may be asymmetric and may be inferred by the devices, based on their repeated interactions with environment or other devices, or directly exchanged as part of collaborative interactions. In the considered paradigm, the emerging collaborative interactions among devices are self-enforcing rather than being mandated by fixed, predetermined protocol rules, as in current network designs. Also, the cognitive devices will need to deploy online multiagent learning solutions in order to be able to make accurate forecasts about the impact of their actions on the dynamic environment and other devices' behaviors and, based on this knowledge, determine how to interact with the other users. Lastly, in the envisioned network architecture, the cross-layer transmission solutions deployed by the users become the strategies with which network users can interact, influence each other, compete, and/or cooperate (see, e.g., [20] and [23]).

We refer to the paradigm proposed in this paper as *knowledge-driven networking* since the various network entities (spectrum moderators, access points, or wireless users) will need to make their decisions about spectrum division, spectrum negotiation, or cross-layer optimization based on their knowledge about the environment and other network entities (PUs and SUs). Since the decisions that need to be taken are based on the users' *incomplete and asymmetric information* about the environment and other users (see Fig. 3), the knowledge that a network entity possesses will influence its efficiency and performance. By gathering information (private observations or explicit information exchanges with other SUs) and subsequently

[3]Cooperative game theory is a parallel branch to the more widely known topic of noncooperative game theory. The term "cooperative" does *not* mean that users have interests that are aligned; rather, cooperative game theory concepts are relevant in situations where a scarce resource is to be divided *fairly* among competing users. Concepts such as bargaining solutions embody specific notions of fairness and take into account the strategic interests of competing users [26].

learning and reasoning based on this information, network entities can develop beliefs[4] about the current state of the communication system and its evolution over time, based on which they can proactively select the optimal policies for interacting with other entities such that they maximize their utilities. For example, a foresighted user can learn the other entities' responses to its actions, thereby being able to forecast the impact of its actions on the wireless system and, ultimately, to optimize its resulting utility over time rather than just myopically optimize its immediate performance.

Summarizing, in this paper, we show how users can compete for resources in various wireless markets and briefly introduce the necessary principles and methods for:

- designing different dynamic spectrum access rules for a variety of communication scenarios;
- enabling the wireless users to learn the system dynamics based on observations and/or explicit information exchanges and improve their strategies for playing the spectrum access game;
- evaluating the "value" of learning and "value of information" for a user in terms of its utility impact;
- coupling the internal and external actions[5] of the wireless users to allow them to achieve an optimal response to the dynamically changing wireless resource market.

D. Paper Layout

This paper is organized as follows. First, we discuss in Section II several important issues that need to be considered when defining the spectrum access game among the wireless users. In Section III, we define a general framework for constructing the wireless resource game. Section IV presents learning solutions that can be deployed by the users to improve their performance when playing the spectrum access game. This section also presents several illustrative examples for the presented framework. However, note that these results are only illustrative and a significant body of work will need to take place before comprehensive solutions can be implemented based on the presented knowledge-driven networking framework. Section V concludes this paper by highlighting the impact of developing such a knowledge-driven networking framework.

II. ISSUES TO CONSIDER FOR THE DESIGN AND CONSTRUCTION OF SPECTRUM ACCESS GAMES

In this section, we summarize several key issues that need to be considered when designing and constructing any spectrum access "market" for cognitive radio networks.

[4]The beliefs of a user are formally defined in Section IV-B.

[5]The internal and external actions and the coupling between them will be discussed in detail in Section III.

A. Resource Types

In [36], the resources in a certain market were classified according to different criteria—continuous versus discrete, divisible or not, sharable or not, static or not. These classification criteria are also useful and should be considered by both wireless users and spectrum regulators when accessing or determining the spectrum division rules and policies for cognitive radio networks. For instance, the resources can be discrete [e.g., in frequency-division multiple access (FDMA), where a single channel is allocated to each user, or even in the case of time-division multiple access (TDMA), when users are allocated a certain percentage of a service time interval to access the channel] or they can be continuous (e.g., the adjustment of power levels). The resources can be defined as sharable or not, depending on the wireless protocol used. For instance, in FDMA, only one user can share a frequency band, and in TDMA, multiple users can timeshare the same channel access. The resources are static or dynamically varying over time, e.g., depending on the PU access.

B. Stochastic Versus Repeated Games

Stochastic games [16] are dynamic, competitive games with probabilistic transitions played by several SUs. The game is played in a sequence of stages. At the beginning of each stage, the game is in a certain state. The SUs select their actions, and each SU receives a reward that depends on both its current state and its selected external and internal actions. The game then moves to a new state with a certain probability, which depends on the previous state and the actions chosen by the SUs. This procedure is repeated at the new state, and the interaction continues for a finite or infinite number of stages. The stochastic games are generalizations of repeated games, which correspond to the special case where there is only one state.

C. One Shot Versus Multistage Games

The games taking place in the wireless networks can be categorized as one-shot or multistage games, depending on whether the allocation is performed once or repeatedly. For instance, in 802.11e, the resource allocation is usually performed by the wireless access point only once, when a SU joins the network [12]. The advantages of such one-shot allocations are that the complexity associated with implementing any resource allocation is kept limited. However, the disadvantage is that this solution does not consider the time-varying source and channel characteristics of the SUs, and the static allocation may become inefficient over time [12], [20]. In this case, repeated or stochastic games can be defined, where the users repeatedly compete for the available resources at each stage of the wireless resource allocation game.

D. Centralized Versus Decentralized

In the centralized setting, a central spectrum moderator (CSM) such as an access point or a base station is responsible for determining and enforcing the allocation among the competing users. In the decentralized setting, the SUs interact with each other directly, through the actions they perform [27], [28], and there may be no moderator involved in the negotiation. Note that in current ISM bands, the wireless users are using the same spectrum access protocols, and thus distributed solutions can be easily designed and enforced. However, in cognitive radio networks, the SUs will be heterogeneous in terms of protocols, utility-cost functions, etc., and this heterogeneity and information decentralization needs to be explicitly considered when designing distributed solutions/protocols.

E. Budget-Balanced Resource Allocation Solutions

Wireless resource allocation solutions can be budget balanced (i.e., all the taxes that SUs pay to the wireless network are reallocated back to other SUs) or not. For instance, many well-known mechanism implementations, such as the Vickrey–Clarke–Groves (VCG) mechanism [20], charge users for using resources in order to provide them incentives to truthfully declare their resource needs. The transfer (money, tokens, etc.) in the VCG mechanism is only delivered from the SUs to the moderator (i.e., the transfers are not allocated back to other users). Hence, the moderator can either discard these transfers (if they are simply used as a money-proxy) or can use these transfers to maintain or upgrade its service, purchase additional spectrum, etc. However, if there is either no moderator in the system or the participating SUs want to prevent the moderator from behaving as a profit maker [e.g., in some wireless local-area network (LAN) usage scenarios], which may potentially result in the moderator trying to alter the users' allocations in order to maximize its revenue, the SU may deploy a budget-balanced mechanism [13]. Moreover, it should be noted that the wireless resource markets can also be designed and regulated without transfers (see, e.g., [39]).

F. Social Decisions (Fairness Rules)

Various fairness rules can be imposed by the CSM or can be negotiated in a decentralized manner by the wireless users, e.g., using bargaining solutions. Some examples investigated in current wireless networks are weighted-sum maximizations of rates or utilities among the participating users, envy-free fairness solutions, or egalitarian solutions. For a comprehensive discussion of these fairness rules, the interested reader is referred to [36]. Performing the resource allocation in the utility domain rather than the resource domain is vital for multimedia users and can result in significant performance gain over application-agnostic resource allocation solutions [24].

G. Desired Equilibrium Concepts That Correspond to the Informational Constraints of the Users

When playing or designing wireless resource games, SUs or moderators will need to *proactively* negotiate or select their desired equilibrium point. This is in contrast to most game-theoretic literature [21], which is developing descriptive models (in, e.g., social or biological systems) to show that certain equilibrium exist. In wireless communication games, *constructive* models are required, where the equilibrium can be designed or influenced by the participating SUs based on their available information and utilities (e.g., [23] and [39]). It is well known that the Nash equilibrium, which is based only on the local information available to each user, is often inefficient in multiuser communication games. Hence, new equilibrium concepts need to be adopted to characterize the interaction of users having different information availability and knowledge about the environment and other users. These may include correlated equilibriums [33], dominant strategy equilibriums [20], Stackelberg equilibriums [23], [39], conjectural equilibriums [42], etc. For instance, in [23], to characterize the multiuser interaction in the distributed power-control game where a foresighted SU can anticipate the responses of its opponent SUs to its actions, the Stackelberg equilibrium is introduced, which is shown to outperform the well-known Nash equilibrium.

H. Implementation Complexity

An important issue associated with the implementation and adoption of wireless resource markets is the resulting complexity for both the CSM, which needs to implement the different resource allocations, and the SUs, which may adopt strategic learning algorithms to be able to compete against other SUs. Hence, new metrics such as the value of learning or the value of information exchanges (which will be discussed in Section IV) need to be deployed to tradeoff the actual benefit that the network entities can derive by increasing their knowledge against the expense of a higher complexity cost.

III. DYNAMIC MULTIUSER SPECTRUM ACCESS GAMES

While the knowledge-driven framework presented in this paper can be implemented in numerous network settings, we will illustrate in this paper only several specific wireless transmission scenarios. Our main focus in this paper will be on designing solutions for emerging cognitive radio networks in which wireless stations[6] (WSTAs) are able to utilize multiple frequency bands, thereby allowing WSTAs to dynamically harvest additional resources. However, the proposed solutions will also be beneficial when deployed in existing ISM radio bands, dedicated bands, or first-generation versions of cognitive radio networks, which may only rely in their implementations on multiple ISM bands. Thus, the focus of this paper will be on designing new dynamic spectrum access and strategic transmission solutions, and not on detecting primary users and identifying spectrum opportunities for WSTAs. For this

[6]In this paper, the SUs are also termed WSTAs, and these two denominations are used interchangeably.

topic, we refer the interested reader to [4] as well as to several papers in this Special Issue, which are addressing these important issues. In this paper, we assume that the spectrum opportunities can be known by simply accessing a dynamically created spectrum opportunity map [10].

As mentioned in the Introduction, we focus on developing wireless resource markets for secondary networks (SNs). In an SN, the secondary users can opportunistically utilize the network resources that are vacated by the PUs. For illustration purposes, we assume that the SN consists of M SUs, which are indexed by $i \in \{1, \ldots, M\}$. The SUs compete for the dynamically available transmission opportunities based on their own "private" information, knowledge about other WSTAs, and available resources (and/or PUs' behaviors). In each time slot ΔT, the WSTAs compete with each other for spectrum access and, given the allocated transmission opportunities,[7] they deploy optimized cross-layer strategies to transmit their delay-sensitive bitstreams.

During each time slot, a "state" of network resources can be defined to represent the available transmission opportunities in a SN, which is denoted by $w \in \mathcal{W}$, where $\mathcal{W}$ is the set of possible resource states. We can also interpret the state of network resources to reflect the behaviors of PUs in the cognitive radio networks [10]. We can also define "states" for the WSTAs. For instance, the states may represent their private information, which includes the traffic and channel characteristics. The current state of a WSTA i is denoted by $s_i \in \mathcal{S}_i$, where $\mathcal{S}_i$ is the set of possible states of WSTA i.

At each time slot, WSTA i will deploy an action to compete for the network resources. This action is referred to as the external action denoted by $a_i \in \mathcal{A}_i$, where $\mathcal{A}_i$ is the set of possible external actions. An example of external actions in wireless networks is the selected transmit power in interference channels or the declared resource request like the TSPEC in 802.11e WLANs. Besides the external action, WSTA i will also deploy an internal action in order to transmit the delay-sensitive data. The internal action can be an action profile including all or a subset of actions from different layers (e.g., adaptation of the packet scheduling strategy, error-correcting codes or retransmission limits to use, etc.). This action is denoted as the internal action denoted by $b_i \in \mathcal{B}_i$, where $\mathcal{B}_i$ is the set of possible internal actions. Note that the external and internal action selections are coupled together as shown in [20]. Moreover, the actions' adaptation can be driven by cross-layer optimization.

In this paper, we formulate the multiuser wireless resource competition as a stochastic game. Formally, the stochastic game is defined as a tuple $(\mathcal{I}, \mathcal{S}, \mathcal{W}, \mathcal{A}, \mathcal{B}, \boldsymbol{P_s}, P_w, \mathcal{R})$, where $\mathcal{I}$ is the set of agents (SUs), i.e., $\mathcal{I} = \{1, \ldots, M\}$, $\mathcal{S}$ is the set of state profiles of all SUs, i.e., $\mathcal{W}$ with $\mathcal{S}_i$ being the state set of SU i, and $\mathcal{W}$ is the set of network resource states. $\mathcal{A}$ is the joint external action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_M$, with $\mathcal{A}_i$[8] being the external action set available for SU i to play the resource sharing game, and $\mathcal{B}$ is the joint internal action space $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_M$, with $\mathcal{B}_i$ being the internal action set available for SU i to transmit delay-sensitive data. $\boldsymbol{P_s}$ is a transition probability function defined as a mapping from the current state profile $\boldsymbol{s} \in \mathcal{S}$, corresponding joint external actions, $\boldsymbol{a} \in \mathcal{A}$ and internal actions $\boldsymbol{b} \in \mathcal{B}$ and the next state profile $\boldsymbol{s}' \in \mathcal{S}$ to a real number between zero and one, i.e., $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S} \mapsto [0, 1]$. P_w is a transition probability function defined as a mapping from the current resource state $w \in \mathcal{W}$ and the next state $w' \in \mathcal{W}$ to a real number between zero and one, i.e., $P : \mathcal{W} \times \mathcal{W} \mapsto [0, 1]$. This will be discussed subsequently in more detail. $\mathcal{R}$ is a reward vector function defined as a mapping from the current state profile $\boldsymbol{s} \in \mathcal{S}$ and corresponding joint external and internal actions $\boldsymbol{a} \in \mathcal{A}$ and $\boldsymbol{b} \in \mathcal{B}$ to an M-dimensional real vector with each element being the reward to a particular agent, i.e., $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}^M$.

The state transition for the network resource state is determined by the PUs and not by the SUs. In other words, the SUs' actions will not affect the network resource state transition. This structure actually creates an opportunity to allow the PUs to be agents with higher priorities in this stochastic game. Moreover, multiple parallel games can be easily defined in this way for different priority users of the same wireless infrastructure (see [38] for more details). According to how the WSTAs compete for spectrum access and exchange information about (and access) the available spectrum opportunities, we consider two types of stochastic games for wireless resource "markets": centralized stochastic games and distributed stochastic games.

A. Centralized Stochastic Game

In the centralized stochastic game, the competition between WSTAs is coordinated by a CSM, which can be an access point, base station, or selected leader. Specifically, at each stage, the WSTAs perform the external actions a_i (e.g., resource requirement, competition bids) and send the CSM a message m_i representing the selected actions. An example of a wireless infrastructure where such a centralized stochastic game can be implemented is wireless LANs (802.11a PCF or 802.11e HCF), where the CSM role is played by the access point.

After receiving the messages $\boldsymbol{m} = [m_1, \ldots m_M]$ from all the WSTAs, the CSM performs the resource allocation according to a certain rule, i.e.,

$$[\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M] = f(\boldsymbol{m}, w) \tag{1}$$

[7] Note that the resource competition and data transmission may take place concurrently.

[8] Note that the action set may depend on the state of the SU. For simplicity, we assume that the actions sets are the same for all the states of the SU.

where $\boldsymbol{r}_i$ is the resource allocation to WSTA i and $f(\cdot,\cdot)$ represents the resource allocation rule based on the announced message $\boldsymbol{m}$ and network resource state w.

After receiving the resource allocation $\boldsymbol{r}_i$, WSTA i performs its own internal action b_i to transmit the delay-sensitive data based on its current state s_i. Note that in the centralized game, the resource allocation $\boldsymbol{r}_i$ for each SU i is computed by the CSM based on the external actions of all the WSTAs. The state transition can be represented by

$$s_i^+ = g_i(s_i, \boldsymbol{r}_i, b_i) \tag{2}$$

and the reward function is computed as

$$R_i = h_i(s_i, \boldsymbol{r}_i, b_i). \tag{3}$$

The states and reward functions for an SU as well as the coupling with the other SUs will be discussed in subsequent sections.

The centralized stochastic game for the cognitive radio network is illustrated in Fig. 4. This can be employed across multiple channels (frequency bands) simultaneously. Each WSTA plays the centralized stochastic game against the other WSTAs by not only selecting its external actions but also by selecting and implementing its internal actions for data transmission. Moreover, the state transition of each WSTA i is directly impacted by its own internal actions and indirectly impacted by the external actions of all WSTAs through the resource competition. The same holds true for the reward. The cross-layer transmission strategies constitute the internal actions deployed by a WSTA. When determining its external actions, a WSTA will need to predict not only what will be the evolution of the source and channel characteristics over time but also the cross-layer strategy that the user will select given the current environment condition. As shown in [20], the cross-layer transmission strategy will impact not only the immediate reward derived by the WSTA based on transmitting the current packets but also the future states and rewards. This is because the current cross-layer strategy will determine which packets get transmitted and, thus, what are the remaining packets to be transmitted etc., which affects the future states. Hence, as shown in [20], the ability of a WSTA to adopt more efficient transmission algorithms at the various layers as well as optimize its cross-layer transmission strategies significantly impacts the performance of both the WSTA and that of its competing WSTAs.

B. Distributed Stochastic Game

In the distributed stochastic game, the SUs simultaneously compete for the spectrum opportunities in the absence of a CSM that coordinates their interactions. In the distributed game, no moderator exists. However, a network policer may be able to intervene if the users are misbehaving [39]. Examples of such distributed games are the power control games in interference channels. (For instance, the distributed power control games in, e.g., [35] and [37] or the contention games in [39] can be represented using the stochastic game formulation presented here.) In the distributed stochastic game, the WSTAs simultaneously implement the internal and external actions. However, the interactions between WSTAs are realized through the external actions. From the perspective of each WSTA, the impact from other WSTAs is aggregated into the experienced channel interference $e(\boldsymbol{s}_{-i}, \boldsymbol{a}_{-i})$. In power control games, the

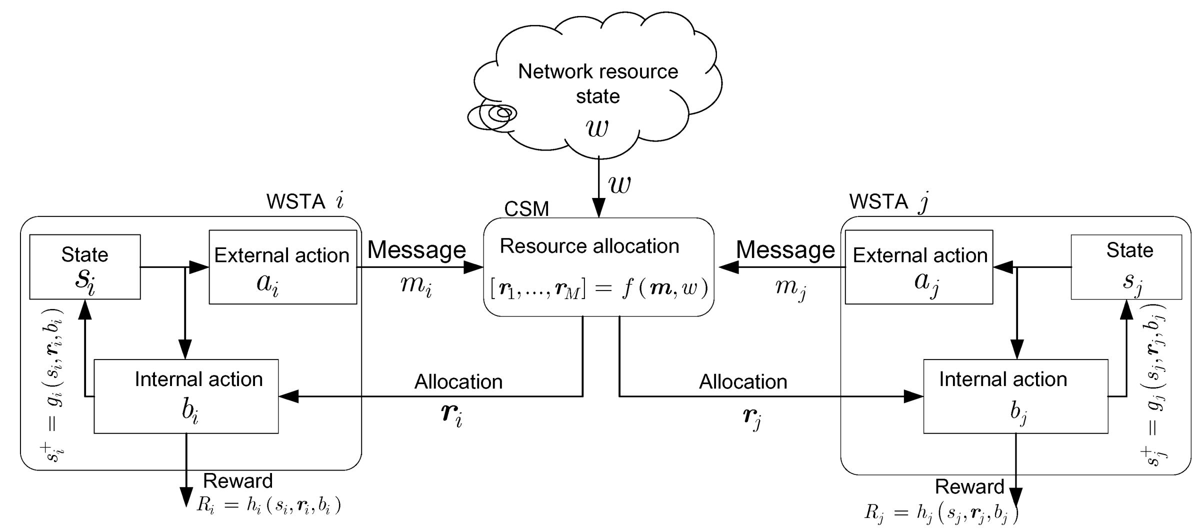

Fig. 4. *Message exchange between WSTAs and CSM, and the actions performed by WSTAs.*

external action can be the power allocation, while the internal action can be the modulation and channel coding scheme. Hence, in distributed stochastic games, the reward of each WSTA i is determined by

$$R_i = h_i(s_i, a_i, b_i, e(\boldsymbol{s}_{-i}, \boldsymbol{a}_{-i}), w) \tag{4}$$

where $-i$ is the set of WSTAs except WSTA i. The state transition is determined by

$$s_i^+ = g_i(s_i, a_i, b_i, e(\boldsymbol{s}_{-i}, \boldsymbol{a}_{-i}), w). \tag{5}$$

The states and reward functions for an SU will be discussed in subsequent sections. The distributed stochastic game for the cognitive radio network is illustrated in Fig. 5.

C. Specification of the Centralized Stochastic Game

As illustration, we consider that the SN can be formed across N channels, each indexed by $j \in \{1, \ldots, N\}$. At each time slot, each channel is assumed to be in one of the following two states: ON (this channel is currently used by the PUs) or OFF (this channel is not used by the PUs and hence can be used by the SUs). Within each time slot, the channel is only OFF or ON [10]. At time slot $t \in \mathbb{N}$, the availability of each channel j is denoted by $w_j^t \in \{0, 1\}$, with w_j^t being zero if the channel is in the ON state and being one if it is in the OFF state. The channel availability profile for the N channels is represented by $\boldsymbol{w}^t = [w_1^t, \ldots, w_N^t]$, which is the state of the network resource at time slot t. As mentioned before, this can be characterized using a spectrum opportunity map [10], provided by the CSM. If the CSM performs imperfect spectrum sensing (as in [4]), this imperfect detection only affects the common system state $\boldsymbol{w}^t = [w_1^t, \ldots, w_N^t]$, which is announced to the secondary users. In this case, instead of announcing the exact (perfect) common resource state, the CSM can announce the probability of the channel being available to the secondary users or not. In this case, the secondary users will compete for the resource based on the probability of the channel availability. However, this relaxation does not affect how the secondary users compete for the available resources with each other or the presented stochastic game formulation. Hence, for simplicity, we focus in this paper on the case in which the CSM performs the perfect spectrum sensing and efficiently allocates the detected spectrum among the competing secondary users.

As in [12] and [20], we assume that a polling-based medium access protocol is deployed in the secondary network, which is arbitrated by a CSM. The polling policy is changed only at the beginning of every time slot. For simplicity, we assume that each SU can access a single channel and that each channel can be accessed by a single SU within the time slot. The SUs can switch the channels only when crossing time slots. Note that this simple medium access model used for illustration in this paper can be easily extended to more sophisticated cognitive radio models [11], where each SU can simultaneously access multiple channels or the channels are being shared by multiple SUs, etc.

1) Wireless Stations States: We assume that WSTAs need to transmit delay-sensitive applications. The bitstream at the application layer is packetized with an average packet length ℓ. In this paper, we consider multimedia applications, where the application packets have a hard delay deadline, i.e., the packets will expire J stages after they are ready for transmission. Then, we can define the state of the buffer as $\boldsymbol{v}_i^t = [v_{i1}^t, \ldots, v_{iJ}^t]^T$, where $v_j^t (1 \leq j \leq J)$ is the number of packets waiting for transmission that have a remaining life time of j time slots. The condition of channel j experienced by WSTA i is represented by the signal-to-noise ratio (SNR) and is denoted as c_{ij}^t in decibels. The channel condition profile is given by $\boldsymbol{c}_i^t = [c_{i1}^t, \ldots, c_{iN}^t]$. To model the dynamics experienced by WSTA i at time t in the cognitive radio network, we

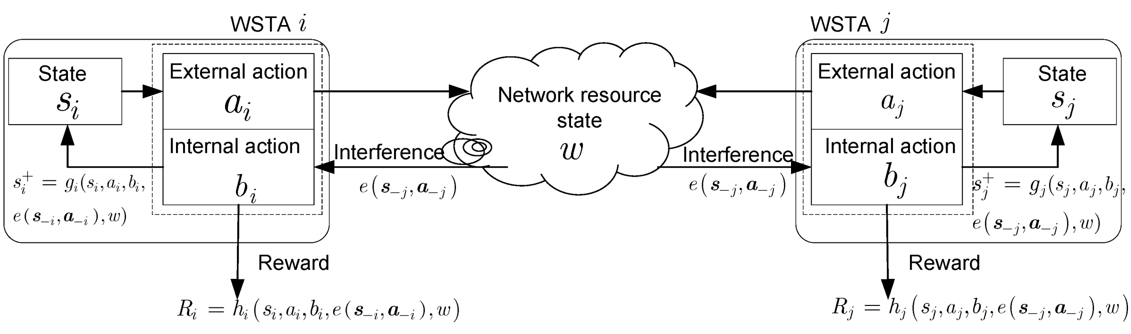

Fig. 5. *Actions performed by WSTAs in the distributed stochastic game.*

define a "state" $s_i^t = (\boldsymbol{v}_i^t, \boldsymbol{c}_i^t) \in \mathcal{S}_i$, which encapsulates the current buffer state as well as the state of each channel.

The environment experienced by each WSTA is characterized by the packet arrivals from the (multimedia) source (i.e., source/traffic characterization) connected with the transmitter, the spectrum opportunities released by PUs, and the channel conditions. Different models can be used by a WSTA to characterize the environment. However, the accuracy of the deployed models will only affect the performance of the solution and not the general framework for multiuser interaction presented here.

2) Internal and External Actions: At the beginning of each time slot, each WSTA deploys an external action a_i to compete for the spectrum opportunities with other WSTAs. The selection of external actions will be discussed in Section III-C5. After receiving the resource allocation $\boldsymbol{r}_i$ from the CSM, the WSTA will deploy the internal action b_i^t. The internal action in this example includes the modulation scheme $\gamma_i^t \in \Gamma_i$ in the physical layer and retransmission limit $\zeta_i^t \in \mathbb{N}$ in the MAC layer, i.e., $b_i^t = (\gamma_i^t, \zeta_i^t)$. Here, Γ_i is the set of possible modulation schemes. For more sophisticated examples of actions, see, e.g., [7] for application layer actions.

3) State Transition and Stage Reward: Since the network resource state is not affected by the actions performed by the WSTAs, the transition of $\boldsymbol{w}^t$ can be modeled as a finite state Markov chain (FSMC) [22]. The transition probability is denoted by $q(\boldsymbol{w}^{t+1}|\boldsymbol{w}^t)$. In this section, we assume that the transition probability $q(\boldsymbol{w}^{t+1}|\boldsymbol{w}^t)$ is known by all the WSTAs and CSM. However, more complicated models for the network resource state transition can also be involved in our stochastic game framework.

When WSTA i receives the resource allocation $\boldsymbol{z}_i^t$, it deploys the internal action b_i^t and can transmit n_i^t packets during time slot t, which is computed as

$$n_i^t = \left\lfloor \frac{\Psi\left(\boldsymbol{c}_i^t, \boldsymbol{z}_i^t, \gamma_i^t, \zeta_i^t\right) \triangle T}{\ell} \right\rfloor \tag{6}$$

where $\Psi(\cdot)$ is the effective rate function, the form of which depends on the protocols implemented at the WSTA. Then, the buffer state can be updated as

$$\begin{bmatrix} v_{i1}^{t+1} \\ \vdots \\ v_{ij}^{t+1} \\ \vdots \\ v_{iJ}^{t+1} \end{bmatrix} = \begin{bmatrix} v_{i2}^t - \max\left(n_i^t - v_{i1}^t, 0\right) \\ \vdots \\ v_{i(j+1)}^t - \max\left(n_i^t - \sum_{m=1}^{j} v_{im}^t, 0\right) \\ \vdots \\ Y_i^t \end{bmatrix} \tag{7}$$

where Y_i^t is a random variable representing the number of packets arriving at time slot t having lifetime J. The distribution of Y_i^t is denoted by $p_{Y_i^t}(l)$. Hence, the transition probability is given by

$$p\left(v_i^{t+1}|v_i^t, z_i^t, b_i^t\right) = \begin{cases} p_{Y_i^t}(l), & \text{if } v_i^{t+1} \text{ satisfies (7) and } Y_i^t = l \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

The channel condition $\boldsymbol{c}_i^t$ depends on the channel gain and the power level for transmission. The channel gain is generally modeled as an FSMC. In this example, we also consider that the power allocation is constant during the data transmission and, hence, the channel condition $\boldsymbol{c}_i^t$ can be formulated as a FSMC with transition probability $p(\boldsymbol{c}_i^{t+1}|\boldsymbol{c}_i^t)$. Details about such transition probability computations can be found in [7].

The state transition probability for WSTA i is given by

$$p\left(s_i^{t+1}|s_i^t, \boldsymbol{z}_i^t, b_i^t\right) = p\left(\boldsymbol{v}_i^{t+1}|\boldsymbol{v}_i^t, z_i^t, b_i^t\right) p\left(\boldsymbol{c}_i^{t+1}|\boldsymbol{c}_i^t\right). \tag{9}$$

Here, we assume that the transition of the channel condition is independent of the transition of the buffer state. The utility for the delay-sensitive application at time slot t is defined here as

$$u\left(s_i^t, \boldsymbol{z}_i^t, b_i^t\right) = \min\left(n_i^t, \sum_{j=1}^{J} v_{ij}^t\right) - \lambda_g \max\left\{v_{i,1}^t - n_i^t, 0\right\} \tag{10}$$

where λ_g is the parameter to trade off the received and lost packets (see [20] for details). More sophisticated utility formulations for multimedia transmission, which consider the explicit impact on the multimedia quality [e.g., peak SNR (PSNR)], can be found in [20].

4) Resource Allocation Rule: We model the multiuser wireless resource allocation as an auction [6], [8], [19] for spectrum opportunities held by the CSM during each time slot. The WSTAs calculate the external action a_i^t based on the information about the network resources, their own private information about the environment they experience, and their anticipated internal actions [20]. In this auction game, the external action is the competition bid, i.e., $m_i^t = a_i^t$. Next, we use the terms external action and bid interchangeably. Subsequently, each WSTA submits the bid a_i^t to the CSM. After receiving the bid vectors from the WSTAs, the CSM computes the channel allocation $\boldsymbol{z}_i^t$ for each WSTA i based on the submitted bids. To compel the WSTAs to declare their bids truthfully [20], the CSM also computes the payment $\tau_i^t \in \mathbb{R}_-$ that the WSTAs have to pay for the use of resources during the current stage of the

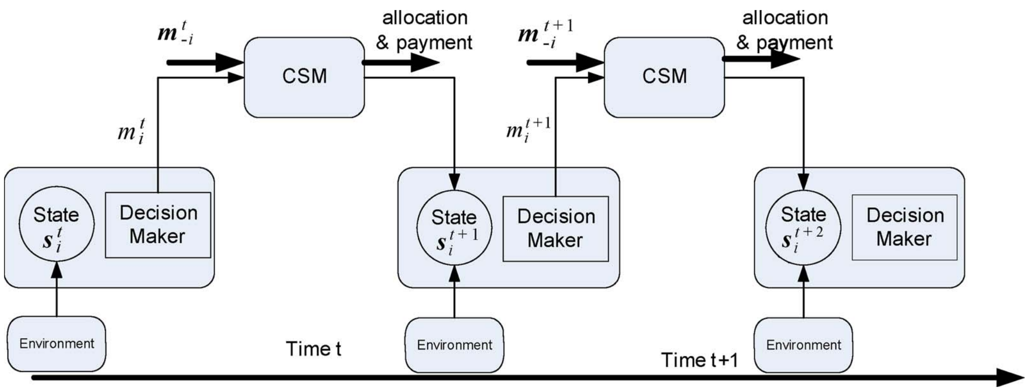

Fig. 6. ***Evolution of multiuser interaction.***

game. The negative value of the payment represents the absolute value that WSTA i has to pay the CSM for the used resources. The auction result is then transmitted back to the WSTAs, which can deploy their transmission strategies in different layers and send data over the assigned channel. After the data transmission, another auction starts at the next time slot $t+1$. A schematic of the evolution of the multiuser interaction is portrayed for illustration in Fig. 6.

The computation of the allocation $\boldsymbol{z}_i^t$ and payment τ_i^t is described as follows. After each WSTA submits the bid vector, the CSM performs two computations: i) channel allocation and ii) payment computation. During the first phase, the CSM allocates the resources to WSTAs based on its adopted fairness rule, e.g., maximizing the total "social welfare"[9]

$$\boldsymbol{z}^{t,opt} = \arg\max_{\boldsymbol{z}^t} \sum_{i=1}^{M} \tilde{h}_i\left(a_i^t, \boldsymbol{z}_i^t, w\right) \tag{11}$$

where $\tilde{h}_i(\cdot)$ is the utility function of WSTA i seen by the CSM. Note that this utility can be represented by either the effective rate or time on the network allocated to each user or can be determined in the utility domain by considering the utility-rate functions of the deployed multimedia coders [24].

We will consider in this paper, for illustration, a second price auction mechanism [18] for determining the tax that needs to be paid by WSTA i based on the above optimal channel assignment $\boldsymbol{z}^{t,opt}$, i.e.,

$$\tau_i^t = \sum_{\substack{j=1,\\ j\neq i}}^{M} \tilde{h}_j\left(a_j^t, \boldsymbol{z}_j^{t,opt}, w\right) - \max_{\boldsymbol{z}_{-i}^t} \sum_{\substack{j=1,\\ j\neq i}}^{M} \tilde{h}_j\left(a_i^t, \boldsymbol{z}_i^t, w\right). \tag{12}$$

[9]Note that other social welfare solutions [38], [41] could be adopted, and this will not influence our proposed solution.

For simplicity, we can denote the output of the resource allocation game as $\boldsymbol{r}^t = (\boldsymbol{z}^t, \boldsymbol{\tau}^t) = \Omega(\boldsymbol{a}^t, w^t)$.

Note that, as mentioned in Section II, the CSM can design different resource allocation games using different mechanisms, leading to different social decisions, allocations, equilibriums, etc. Moreover, the taxation does not need to be implemented and can be omitted. However, we would like to point out that, unless the taxation is implemented, the WSTA will have no incentive to efficiently optimize their cross-layer strategies, upgrade their systems, or truthfully and optimally declare their requirements.

5) *Selecting the Policy for Playing the Resource Management Game:* In the cognitive radio network, we assume that the stochastic game is played by all WSTAs for an infinite number of stages. This assumption is reasonable for multimedia applications, which usually have a long duration. In our network setting, we define a history of the stochastic game up to time t as $\boldsymbol{h}^t = \{\boldsymbol{s}^0, w^0, \boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{z}^0, \boldsymbol{\tau}^0, \ldots, \boldsymbol{s}^{t-1}, w^{t-1}, \boldsymbol{a}^{t-1}, \boldsymbol{b}^{t-1}, \boldsymbol{z}^{t-1}, \boldsymbol{\tau}^{t-1}, \boldsymbol{s}^t\} \in \mathcal{H}^t$, which summarizes all previous states and the actions taken by the WSTAs as well as the outcomes at each stage of the auction game and $\mathcal{H}^t$ is the set of the entire history up to time t. However, during the stochastic game, each WSTA i cannot observe the entire history but rather part of the history $\boldsymbol{h}^t$. The observation of WSTA i is denoted as $\boldsymbol{o}_i^t \in \mathcal{O}_i^t$ and $\boldsymbol{o}_i^t \subset \boldsymbol{h}^t$. Note that the current state $\boldsymbol{s}_i^t$ can be always observed, i.e., $\boldsymbol{s}_i^t \in \boldsymbol{o}_i^t$. Then, a bidding policy $\pi_i^t : \mathcal{O}_i^t \mapsto \mathcal{A}_i \times \mathcal{B}_i$ for WSTA i at the time t is defined as a mapping from the observations up to the time t into the specific action, i.e., $[a_i^t, b_i^t] = \pi_i^t(\boldsymbol{o}_i^t)$. Furthermore, a policy profile $\boldsymbol{\pi}_i$ for WSTA i aggregates the bidding policies about how to play the game over the entire course of the stochastic game, i.e., $\boldsymbol{\pi}_i = (\pi_i^0, \ldots, \pi_i^t, \ldots)$. The policy profile for all the WSTAs at time slot t is denoted as $\boldsymbol{\pi}^t = (\pi_1^t, \ldots, \pi_M^t) = (\pi_i^t, \pi_{-i}^t)$.

The reward for WSTA i at the time slot t is $R_i^t(s_i^t, \boldsymbol{r}_i^t, b_i^t) = u(s_i^t, \boldsymbol{z}_i^t, b_i^t) + \tau_i^t$. Since the resource allocation also depends on other SUs' states and external actions as well as the channel state, the reward can be estimated based on the observation $\boldsymbol{o}_i^t$; thus, the reward used by a WSTA will be $R_i^t(s_i^t, \boldsymbol{o}_i^t, b_i^t)$. The reward $R_i^k(s_i^k, \boldsymbol{o}_i^k, b_i^k)$ at stage k is discounted by factor $(\alpha_i)^{k-t}$ at time t. The factor $\alpha_i (0 \leq \alpha_i < 1)$ is the discounted factor determined by a specific application (for instance, for video streaming applications, this factor can be set based on the tolerable delay). The total discounted sum of rewards $Q_i^t(\pi_i^t, \boldsymbol{\pi}_{-i}^t | \boldsymbol{s}^t, w^t)$ for SU i can be calculated at time t from the state profile $\boldsymbol{s}^t$ as

$$Q_i^t\left(\left(\pi_i^t, \boldsymbol{\pi}_{-i}^t\right) | \boldsymbol{s}^t, w^t\right) = \sum_{k=t}^{\infty} (\alpha_i)^{k-t} R_i^k\left(s_i^k, \boldsymbol{o}_i^k, b_i^k\right) \tag{13}$$

where $\boldsymbol{\pi}_{-i}^t(\boldsymbol{s}_{-i}^t) = ([a_j^t, b_j^t]_{j \neq i})$. We assume that the SUs implement the policy $\boldsymbol{\pi}^t$ in the subsequent time slots. The total discounted sum of rewards in (13) consists of two parts: i) the current stage reward and ii) the expected future reward discounted by α_i. Note that SU i cannot independently determine the above value without explicitly knowing the policies and states of other SUs. The SU maximizes the total discounted sum of future rewards in order to select the bidding policy, which explicitly considers the impact of the current bid vector on the expected future rewards.

We define the *best response* β_i for SU i to other WSTAs' policies $\boldsymbol{\pi}_{-i}^t$ as

$$\beta_i\left(\boldsymbol{\pi}_{-i}^t\right) = \arg\max_{\pi_i} Q_i^t\left(\left(\pi_i^t, \boldsymbol{\pi}_{-i}^t\right) | \boldsymbol{s}^t, w^t\right). \tag{14}$$

The central issue in such stochastic game in cognitive radio networks is how the best response policies can be determined by the SUs. This will be the topic of Section IV.

D. Specification of Distributed Stochastic Game

An example of a distributed game is the power-control game played by SUs in the interference channels in cognitive radio network. There are M SUs, each of which comprises one transmitter and one receiver. There are N channels potentially vacated by the PUs for SU transmission. At time slot t, the network resource state is $\boldsymbol{w}^t = [w_1^t, \ldots, w_N^t] \in \{0,1\}^N$. The channel gain of SU i at channel $j \in \{1, \ldots, N\}$ is H_{ii}^j and the cross-channel gain from transmitter i (belonging to SU i) to receiver i' (belonging to SU i') at channel j is $H_{ii'}^j$. We assume that the (cross) channel gains for all the SUs are constant.

In this game, the state of SU i is defined as a vector $s_i^t \in \{0,1\}^N$, with each element indicating whether SU i selects that channel (corresponding to 1) or not (corresponding to 0). The external action a_i^t of SU i includes two components: channel selection κ_i^t and power allocation φ_i^t, i.e., $a_i^t = (\kappa_i^t, \varphi_i^t)$, where $\kappa_i^t \subseteq \{1, \ldots, N\}$. For each external action a_i^t of SU i, there is power constraint imposed on the power allocation, i.e.,

$$\sum_{j \in \kappa_i^t} \varphi_{ij}^t \leq P_i. \tag{15}$$

In this power-control game, at the beginning of each time slot, the SUs simultaneously choose the channels over which they will transmit delay-sensitive data and allocate the power on the selected channels under the power constraints. In order not to interfere with the PUs, the SUs are not allowed to transmit any data over those channels with $w_j^t = 0$ (i.e., channel j is occupied by the PUs). For simplicity, we consider the case that the SUs are free to choose any channels. Hence, the state of SU i equals the channel selection action, i.e., $s_i^t = \kappa_i^t$. The internal actions for the SUs are empty. The effective transmission rate can be computed as

$$\begin{aligned} T_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) &= h_i\left(s_i^t, a_i^t, b_i^t, e\left(\boldsymbol{s}_{-i}^t, \boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) \\ &= h_i\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) \\ &= \sum_{\substack{j \in \kappa_i^t \\ w_j^t = 1}} \frac{1}{2} \log_2\left(1 + \frac{H_{ii}^j \varphi_{ij}^t}{N_{0j} + \sum_{i' \in C_j^t} H_{i'i}^j \varphi_{i'j}^t}\right) \end{aligned} \tag{16}$$

where C_j^t is the set of SUs who select channel j in time slot t, $e(\boldsymbol{a}_{-i}^t) = \sum_{i' \in C_j^t} H_{i'i}^j \varphi_{i'j}^t$, and N_{0j} represents the noise level in the selected channel j. In this power-control game, the stage reward function for SU i can be defined as effective transmission rate per joule, similarly to [37], i.e.,

$$R_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) = \frac{T_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right)}{\sum_{j \in \kappa_i^t} \varphi_{ij}^t}. \tag{17}$$

However, such a reward function cannot satisfy the QoS requirements of multimedia applications. Hence, the following stage reward function can be adopted for such applications:

$$R_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) = \frac{\Lambda_i \times \left\{1 - P_i\left(T_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right), d_i\right)\right\}}{\sum_{j \in \kappa_i^t} \varphi_{ij}^t} \tag{18}$$

where Λ_i represents the arrival source rate of the applications of SU i and $P_i(T_i^t(a_i^t, e(\boldsymbol{a}_{-i}^t), \boldsymbol{w}^t), d_i)$ represents the packet error rate, which is a function of the effective transmission rate T_i^t [in (16)] and the delay deadline d_i of the applications of SU i. More complicated utility-resource functions as in [20] can also be employed.

Since the state of each SU i is the same as the channel selection and no internal actions are considered, the channel-selection and power-control game is reduced to a repeated game [32]. The essential goal of SU i is to find the best response to the aggregated interference $e(\boldsymbol{a}_{-i}^t)$ under various network resource states, i.e.,

$$a_i^{t,*}\left(e\left(\boldsymbol{a}_{-i}^t, \boldsymbol{w}^t\right)\right) = \arg\max_{a_i^t} R_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right). \tag{19}$$

This will be discussed in the next section.

IV. STRATEGIC LEARNING SOLUTIONS IN MULTIUSER WIRELESS SYSTEMS

A. Why Learn?

In the previous section, it was shown that in order for an SU to derive its own transmission policy, it needs to know how its decision process and resulting performance are coupled to that of other SUs. In a stochastic game framework, the goal for each SU is to find a policy π_i such that its own utility is maximized. However, as discussed in Section III, SU i's policy π_i depends on other SUs' policies, which is formulated as

$$\pi_i^* = \arg\max_{\pi_i} Q_i(\pi_i, \boldsymbol{\pi}_{-i} | s_i, \boldsymbol{s}_{-i}, \boldsymbol{w}). \tag{20}$$

To solve this optimization, the following information is required by SU i:

- the state transition model of SU i $p(s_i^{t+1} | s_i^t, a_i^t, \boldsymbol{a}_{-i}^t, b_i)$;
- the state transition model of other SUs $p(s_j^{t+1} | s_j^t, a_j^t, \boldsymbol{a}_{-j}^t, b_j), \forall j \neq i$;
- the state of other SUs s_{-i};
- the policy of other SUs $\boldsymbol{\pi}_{-i}$;
- the network resource state w.

This coupling among SUs is due to the shared nature of the wireless resources [2]. However, an SU may not exactly know the other SUs' actions and models, and it cannot know their private information. Thus, an SU can only predict these dynamics (uncertainties) caused by the competing SUs based on its observations from past interaction. In the cognitive radio network, there are different levels of information availability.

- *Private information:* this includes the characteristics of the application traffic, channel gain, or channel conditions [signal-to-(interference plus noise) ratio, etc.].
- *Network information:* this refers to the network resource states or the behaviors of PUs.
- *Opponents information:* this includes the states and possible actions of the opponents. This information can be, for instance, known when all the SUs adopt the same protocol, having the same set of states and actions.

To reduce the uncertainty and increase the knowledge about the environment when selecting an action, an SU can deploy learning in games algorithms [21]. Depending on the information availability, different learning solutions can be deployed by a WSTA. The existing learning in games literature provides a broad spectrum of analytical and practical results on learning algorithms and underlying game structures for a variety of competitive interaction scenarios. In general, the main issue considered has been to characterize long-term behavior in terms of a generalized equilibrium concept or characterize the *lack* of convergence for general classes of learning dynamics. However, when selecting learning solutions for wireless networks games, the specific constraints and features of wireless systems will need to be considered. For instance, the learning algorithm that should be deployed by a user in a wireless environment strongly depends on what information an SU can observe about the other SUs, given the adopted protocols or spectrum regulation rules. Moreover, unlike a majority of work in learning in games solutions [21], where the main focus is on proving the existence of equilibriums or where the only goal of the agents is to achieve different equilibrium conditions, learning solutions in wireless networks are deployed by self-interested and heterogeneous users, which have as only goal to improve their own performance. Thus, a learning algorithm $\mathcal{L}_i$ adopted by SU i to efficiently play the spectrum allocation game will be evaluated based on the information that can be acquired (i.e., the observed information o_i^t) and exchanged I_{-i}^t, the complexity requirements, and the resulting (long-term or short-term) utility U_i.

B. Definitions of Learning Algorithms and Beliefs

The goal of learning for an SU in the multiuser games is to update its own policy and belief about the other SUs' states and policies. Specifically, by learning from the observed and exchanged information, a user can build its *belief* on the other users' strategies and determine its own best response policy. In our stochastic game framework, the SU also needs to update its knowledge about the network resource state using learning. We note that a learning algorithm is built based on the observation o_i^t and exchanged information I_{-i}^t; hence, it is denoted as $\mathcal{L}_i(\boldsymbol{o}_i, \boldsymbol{I}_{-i})$, where $\boldsymbol{o}_i, \boldsymbol{I}_{-i}$ are all the observation and exchanged information obtained by SU i.

A learning algorithm $\mathcal{L}_i$ can be defined using the following equations:

$$[a_i^t, b_i^t] = \pi_i^t\left(s_i^t, B_{s_{-i}}^t, B_{\pi_{-i}}^t, B_w^t\right) \tag{21}$$
$$\Omega^t = Game(s^t, \boldsymbol{a}^t, w^t) \tag{22}$$
$$o_i^t = O(s_i^t, \Omega_i^t, b_i^t) \tag{23}$$
$$\pi_i^{t+1} = \mathcal{F}_i(\pi_i^t, o_i^t, I_{-i}^t) \tag{24}$$
$$B_{\pi_{-i}}^{t+1} = \mathcal{F}_{\pi_{-i}}\left(B_{\pi_{-i}}^t, o_i^t, I_{-i}^t\right) \tag{25}$$
$$B_w^{t+1} = \mathcal{F}_w(B_w^t, o_i^t, I_{-i}^t) \tag{26}$$
$$B_{s_{-i}}^{t+1} = \mathcal{F}_{s_{-i}}\left(B_{s_{-i}}^t, o_i^t, I_{-i}^t\right) \tag{27}$$

where $B_{\boldsymbol{s}_{-i}}^t$, $B_{\boldsymbol{\pi}_{-i}}^t$, and B_w^t are the belief about the other SUs' states $\boldsymbol{s}_{-i}$, policies $\boldsymbol{\pi}_{-i}$, and the network resource state w, respectively; Ω^t is the output of the multiuser interaction game ($\Omega^t = \boldsymbol{r}^t$ in the centralized stochastic game and $\Omega^t = \{a_i^t, e(\boldsymbol{a}_{-i}^t), w^t\}$ for the distributed stochastic game or repeated game); o_i^t is the observation of SU i and O is the observation function, which depends on the current state, the current game output, and the current internal action taken; $\mathcal{F}$ is the update function about the belief and policies; and I_{-i}^t is the exchanged information with the other SUs. The learning process is pictorially depicted in Fig. 7.

Equation (21) shows that SU i generates the external actions based on its own states, the belief about the other SUs' states, policies, and network resource state. After each SU executes its external actions, a multiuser spectrum access game is played and the results of the game are produced as shown in (22). The results of the multiuser game may or may not be fully observed by the SUs based on the game form or the implemented network protocol. Equation (23) represents the observation function which depends on the network protocols and the SUs' measurement methods. Different (accurate or inaccurate) observations may lead to different learning algorithms, which will be discussed in subsequent sections. Hence, an SU may have incentives to exchange information with other SUs. The exchanged information I_{-i}^t may be used to update the belief about the other SUs' states, policies, and network resource state. Equations (24)–(27) represent the updates of the beliefs.

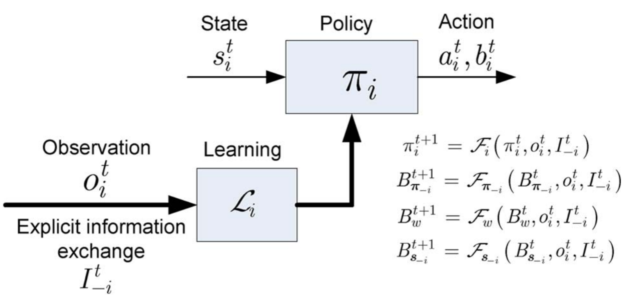

Fig. 7. ***Strategic learning in the multiuser wireless communication game.***

In a wireless communication game, we differentiate two types of users based on their response strategies.

- *Myopic users:* users that always act to maximize their immediate achievable reward. They are myopic in the sense that, at each decision stage, they treat other users' actions as fixed, ignore the impact of their competitors' reactions over their own performance, and determine their responses to gain the maximal immediate rewards.
- *Foresighted users:* users that behave by taking into account the long-term impacts of their actions on their rewards. They avoid shortsighted (myopic) actions, anticipate how the other users will react, and maximize their performance by considering the responses of the other users [7], [23]. Note that such foresighted users require additional knowledge about the other users to assist their decision making. We will discuss this in more detail later in this section.

Before we proceed in detail with discussing how a learning algorithm is built, we discuss first how we can evaluate a learning algorithm for the cognitive radio network.

C. Value of Learning, Value of Information, and Regret Computation

As mentioned previously, the performance of a learning algorithm will depend on the resulting SU reward. We denote a policy generated by the learning algorithm $\mathcal{L}_i$ as $\pi_i^{\mathcal{L}_i}$. An SU will learn in order to improve its policy and its rewards from participating in the spectrum access game. The performance of SU i when adopting the learning algorithm $\mathcal{L}_i$ is defined as the time average reward obtained in a time window with length T in which this learning algorithm was used

$$\mathcal{V}^{\pi_i^{\mathcal{L}_i(\boldsymbol{o}_i, \boldsymbol{I}_{-i})}}(T) = \frac{1}{T}\sum_{t=1}^{T} R_i^t\left(\pi_i^{\mathcal{L}_i(\boldsymbol{o}_i, \boldsymbol{I}_{-i})}\right) \tag{28}$$

where the reward R_i^t depends on both the learning approach $\mathcal{L}_i$ and the observation o_i^t and information exchanged I_{-i}^t. Thus, using this definition, the "value of a learning scheme" can be determined. For instance, given the same observation o_i^t and exchanged information I_{-i}^t, if the time average rewards of two algorithms $\mathcal{L}_i'$ and $\mathcal{L}_i''$ satisfy $\mathcal{V}^{\pi_i^{\mathcal{L}_i'(\boldsymbol{o}_i, \boldsymbol{I}_{-i})}}(T) > \mathcal{V}^{\pi_i^{\mathcal{L}_i''(\boldsymbol{o}_i, \boldsymbol{I}_{-i})}}(T)$, then we say that learning algorithm $\mathcal{L}_i'$ is better than $\mathcal{L}_i''$. The "value of information exchange" with respect to a learning algorithm $\mathcal{L}_i$ also can be similarly computed. This value of information will play a significant role in what information should be exchanged among WSTAs and how WSTAs should negotiate in a cooperative[10] setting (e.g., in a bargaining or coalition setting). The value of making various observations and learning based on them can be similarly computed. Moreover, we define a generalized "regret" for the stochastic game at each time slot t as

$$\Delta_{\mathcal{L}_i} \triangleq \max_{\pi_i} Q_i^t\left(\pi_i, \boldsymbol{\pi}_{-i} | s_i^t, \boldsymbol{s}_{-i}^t, w^t\right) - Q_i^t\left(\pi_i^{\mathcal{L}_i}, B_{\boldsymbol{\pi}_{-i}}^t | s_i^t, B_{\boldsymbol{s}_{-i}}^t, B_w^t\right). \quad (29)$$

When the stochastic game is reduced to a repeated game, the regret can be computed as

$$\Delta_{\mathcal{L}_i} \triangleq \max_{a_i^t} R_i^t\left(a_i^t, e\left(\boldsymbol{a}_{-i}^t\right), \boldsymbol{w}^t\right) - R_i^t\left(a_i^{\mathcal{L}_i}, B_{\boldsymbol{a}_{-i}}^t, B_w^t\right). \quad (30)$$

The regret is computed as the reward loss due to the lack of knowledge about the network resource and components' states and actions. The regret can be computed and used by the SUs in order to adjust their learning strategies and improve their strategies for playing the game.

D. Learning Framework for Wireless Stochastic Games

One possible simplification for the stochastic learning is to assume that other SUs perform a fixed policy. This is a good assumption especially for the case when WSTAs adopt the same protocols, which implement the same policies. Hence, SU i does not need to update its belief about other SUs' policies (i.e., $B_{\boldsymbol{\pi}_{-i}}^t$). Instead, SU i needs to update its belief about other SUs' states and state transition probability. However, to observe other SUs' states in a cognitive radio network is also difficult, and even impossible in some cases. To solve this problem, an SU can classify the states of other SUs based on the output of the game. For simplicity, we assume that the network resource state is common information, which is known by all the participating SUs. However, the learning algorithm discussed in this section can also be extended to the case in which the network resource state and the corresponding state transition are unknown to the SUs. In this case, the WSTA needs to learn the state transition probability for each channel's state based on its observations [40].

1) What Information to Learn From?: First, let us consider what information the SU can observe while playing the stochastic game in our cognitive radio network. As shown in Fig. 3, at the beginning of time slot t, the SUs submit the bids $a_i^t, \forall i$. Then, the CSM returns the channel allocation $z_i^t, \forall i$ and $\tau_i^t, \forall i$. In a cognitive radio network, if SU i is not allowed to observe the bids, the channel allocations, and the payments for other SUs, then the observation of SU i becomes $\boldsymbol{o}_i^t = \{\boldsymbol{s}_i^0, w^0, \boldsymbol{a}_i^0, \boldsymbol{b}_i^0, \boldsymbol{z}_i^0, \boldsymbol{\tau}_i^0, \ldots, \boldsymbol{s}_i^{t-1}, w^{t-1}, \boldsymbol{a}_i^{t-1}, \boldsymbol{b}_i^{t-1}, \boldsymbol{z}_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$. If the information is fully exchanged among SUs or broadcast and overheard by all SUs, the observed information by SU i becomes $\boldsymbol{o}_i^t = \boldsymbol{h}^t$. Now, the problem that needs to be solved by SU i is how it can improve its own policy for playing the game by learning from the observation $\boldsymbol{o}_i^t$. In this paper, we assume that SU i observes the information $\boldsymbol{o}_i^t = \{\boldsymbol{s}_i^0, w^0, \boldsymbol{a}_i^0, \boldsymbol{b}_i^0, \boldsymbol{z}_i^0, \boldsymbol{\tau}_i^0, \ldots, \boldsymbol{s}_i^{t-1}, w^{t-1}, \boldsymbol{a}_i^{t-1}, \boldsymbol{b}_i^{t-1}, \boldsymbol{z}_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$.

2) What Needs to be Learned?: A key question is what needs to be learned within a wireless stochastic game in order to improve the policy of an SU. We focus here on the learning procedure for the external policy (generating external actions, i.e., bidding actions).

In Section IV-A, we discussed the information that SU i needs to learn in order to be able to solve the optimization in (20). However, SU i can only observe the information $\boldsymbol{o}_i^t = \{\boldsymbol{s}_i^0, w^0, \boldsymbol{a}_i^0, \boldsymbol{b}_i^0, \boldsymbol{z}_i^0, \boldsymbol{\tau}_i^0, \ldots, \boldsymbol{s}_i^{t-1}, w^{t-1}, \boldsymbol{a}_i^{t-1}, \boldsymbol{b}_i^{t-1}, \boldsymbol{z}_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$ from which SU i cannot accurately infer the other SUs' state space (i.e., $\mathcal{S}_{-i}$), the current state of other SUs (i.e., $\boldsymbol{s}_{-i}^t$), and the transition probability of other SUs [i.e., $\prod_{k \neq i} q_k(\boldsymbol{s}_k^{t+1} | \boldsymbol{s}_k^t, \boldsymbol{z}_k^t)$]. Moreover, capturing the exact information about other SUs requires heavy computational and storage complexity. Instead, we allow SU i to classify the space $\mathcal{S}_{-i}$ into H_i classes, each of which is represented by a representative state $\tilde{s}_{-i,h}, h \in \{1, \ldots, H_i\}$. By dividing the state space $\mathcal{S}_{-i}$, the transition probability $\prod_{k \neq i} q_k(\boldsymbol{s}_k^{t+1} | \boldsymbol{s}_k^t, \boldsymbol{z}_k^t)$ is approximated by $q_{-i}(\tilde{s}_{-i}^{t+1} | \tilde{s}_{-i}^t, \boldsymbol{z}_i^t)$, where $\tilde{s}_{-i}^t$ and $\tilde{s}_{-i}^{t+1}$ are the representative states of the classes to which $\boldsymbol{s}_{-i}^t$ and $\boldsymbol{s}_{-i}^{t+1}$ belong. This approximation is performed by aggregating all other SUs' states into one representative state and assuming that the transition depends on the resource allocation $\boldsymbol{z}_i^t$. Note that the classification on the state space $\mathcal{S}_{-i}$ and approximation of the transition probability and discounted sum of rewards affects the learning performance. Hence, a user should trade off an increased learning complexity for an increased value of learning.

In this setting, to find the approximated optimal bidding policy, we need to learn the following from the past observations: i) how the space $\mathcal{S}_{-i}$ is classified; ii) the transition probability $q_{-i}(\tilde{s}_{-i}^{t+1} | \tilde{s}_{-i}^t, z_i^t)$; and iii) the average future rewards $V_i^{t+1}((\boldsymbol{s}_i^{t+1}, \tilde{s}_{-i}^{t+1}))$.

3) How to Learn?: In this section, we develop a learning algorithm to estimate the terms listed in the above section.

Step 1) Decomposition of the space $\mathcal{S}_{-i}$: As discussed in Section IV-D1, only $\boldsymbol{o}_i^t = \{\boldsymbol{s}_i^0, w^0, \boldsymbol{a}_i^0, \boldsymbol{b}_i^0, \boldsymbol{z}_i^0, \boldsymbol{\tau}_i^0, \ldots, \boldsymbol{s}_i^{t-1}, w^{t-1}, \boldsymbol{a}_i^{t-1}, \boldsymbol{b}_i^{t-1}, z_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$ are observed. From the

[10]Note that the communications society refers to collaborative solutions in cases where users have a common objective. In the game-theoretic society, the term *collaborative* refers instead to users that are collaboratively exchanging information and make agreements, but they are still behaving strategically and aiming to optimize their own utilities. In our proposed research, we use the term collaborative in its game-theoretic meaning, since we always consider self-interested users.

auction mechanism presented in Section III-D, we know that the value of the tax τ_i^t is computed based on the inconvenience that SU i causes to the other SUs. In other words, a higher value of $|\tau_i^t|$ indicates that the network is more congested.[11] Based on the bid vector b_i^t, the channel allocation $\boldsymbol{z}_i^t$, and the tax τ_i^t, SU i can infer the network congestion and thus, indirectly, the resource requirements of the competing SUs. Instead of knowing the exact state space of other SUs, SU i can classify the space $\mathcal{S}_{-i}$ as follows. We assume the maximum absolute tax is Γ. We split the range $[0, \Gamma]$ into $[\Gamma_0, \Gamma_1), [\Gamma_1, \Gamma_2), \ldots, [\Gamma_{H_i-1}, \Gamma_{H_i}]$ with $0 = \Gamma_0 \leq \Gamma_1 \leq \cdots \leq \Gamma_{H_i} = \Gamma$. Here, we assume that the values of $\{\Gamma_1, \ldots, \Gamma_{H_i-1}\}$ are equally located in the range of $[0, \Gamma]$. (Note that more sophisticated selection for these values can be deployed, and this forms an interesting area of future research.)

We need to consider three cases to determine the representative state $\tilde{s}_{-i}^t$ at time t.

1) If the resource allocation $\boldsymbol{z}_i^t \neq \boldsymbol{0}$, then the representative state of other SUs is chosen as

$$\tilde{s}_{-i}^t = h, \text{ if } \left|\tau_i^t\right| \in [\Gamma_{h-1}, \Gamma_h). \tag{31}$$

2) If the resource allocation $\boldsymbol{z}_i^t = \boldsymbol{0}$ but $\boldsymbol{w}^t \neq \boldsymbol{0}$, the tax is zero. In this case, we cannot use the tax to predict the network congestion. However, we can infer that the congestion is more severe than the minimum bid for those available channels, i.e., $\min_{j\in\{l:y_l^t\neq 0\}} \{a_{ij}^t\}$. This is because, in this current stage of the auction game, only SU i' with $a_{i'j}^t \geq a_{ij}^t$ can obtain channel j, which indicates that $|\tau_i^t| \geq \min_{j\in\{l:y_l^t\neq 0\}} \{a_{ij}^t\}$, if SU i is allocated any channel. Then the representative state of other SUs is chosen as

$$\tilde{s}_{-i}^t = h, \text{ if } \min_{j\in\left\{l:y_l^t\neq 0\right\}} \left\{a_{ij}^t\right\} \in [\Gamma_{h-1}, \Gamma_h). \tag{32}$$

3) If the resource allocation $\boldsymbol{z}_i^t = \boldsymbol{0}$ and $\boldsymbol{w}^t = \boldsymbol{0}$, there is no interaction among the SUs in this time slot. Hence, $\tilde{s}_{-i}^t = \tilde{s}_{-i}^{t-1}$.

Step 2) Estimating the transition probability: To estimate the transition probability, SU i maintains a table F with size $H_i \times H_i \times (N+1)$. Each entry $f_{h',h'',j}$ in the table F represents the number of transitions from state $\tilde{s}_{-i}^t = h''$ to $\tilde{s}_{-i}^{t+1} = h'$ when the resource allocation $\boldsymbol{z}_i^t = \boldsymbol{e}_j$ (or $\boldsymbol{0}$ if $j = 0$). Here, $\boldsymbol{e}_j$ is a N-dimensional vector with the jth element being one and otherwise being zero. It is clear that H_i will influence significantly the complexity and memory requirements etc. of SU i. The update of F is simply based on the observation $\boldsymbol{o}_i^t$ and the state classification in the above section. Then, we use the frequency to approximate the transition probability [15], i.e.,

$$q_{-i}\left(\tilde{s}_{-i}^{t+1} = h' | \tilde{s}_{-i}^t = h'', \boldsymbol{e}_j\right) = \frac{f_{h',h'',j}}{\sum_{h'} f_{h',h'',j}}. \tag{33}$$

Step 3) Learning the future reward: By classifying the state space $\mathcal{S}_{-i}$ and estimating the transition probability, SU i can now forecast the value of the average future reward $V_i^{t+1}((\boldsymbol{s}_i^{t+1}, \tilde{s}_{-i}^{t+1}))$ using learning. Equation (13) can be approximated by

$$\begin{aligned} Q_i^t\left(\boldsymbol{s}_i^t, \tilde{s}_{-i}^t\right) = \Big\{ & g_i\left(\boldsymbol{s}_i^t, \boldsymbol{z}_i^t\right) + \tau_i^t \\ & + \alpha_i \sum_{\left(\boldsymbol{s}_i^{t+1}, \tilde{s}_{-i}^{t+1}\right)\in\mathcal{S}} q_i\left(\boldsymbol{s}_i^{t+1} | \boldsymbol{s}_i^t, \boldsymbol{z}_i^t\right) q_{-i} \\ & \times \left(\tilde{\boldsymbol{s}}_{-i}^{t+1} | \tilde{s}_{-i}^t, \boldsymbol{z}_i^t\right) V_i^{t+1}\left(\left(\boldsymbol{s}_i^{t+1}, \tilde{s}_{-i}^{t+1}\right)\right)\Big\}. \end{aligned} \tag{34}$$

The received rewards are used to update the estimation of future rewards, similarly to Q-learning [17]. However, the main difference between this algorithm [7] and Q-learning is that the former explicitly considers the impact of other SUs' bidding actions through the state classifications and transition probability approximation.

A two-dimensional table can be used to store the value $V_i(\boldsymbol{s}_i, \tilde{s}_{-i})$ with $\boldsymbol{s}_i \in \mathcal{S}_i$, $\tilde{s}_{-i} \in \tilde{S}_{-i}$, where $\tilde{S}_{-i}$ is the set of representative states for the other SUs. The total number of entries in V_i is $|\mathcal{S}_i| \times |\tilde{\mathcal{S}}_{-i}|$. SU i updates the value of $V_i((\boldsymbol{s}_i, \tilde{s}_{-i}))$ at time t according to the following rules:

$$V_i^t(\boldsymbol{s}_i, \tilde{s}_{-i}) = \begin{cases} \left(1 - \beta_i^t\right) V_i^{t-1}(\boldsymbol{s}_i, \tilde{s}_{-i}) & \\ \quad + \beta_i^t Q_i^t\left(\boldsymbol{s}_i^t, \tilde{s}_{-i}^t\right), & \text{if } \left(\boldsymbol{s}_i^t, \tilde{s}_{-i}^t\right) = (\boldsymbol{s}_i, \tilde{\boldsymbol{s}}_{-i}) \\ V_i^t(\boldsymbol{s}_i, \tilde{s}_{-i}), & \text{otherwise} \end{cases} \tag{35}$$

where $\beta_i^t \in [0, 1)$ is the learning rate factor. An interesting area of research is determining how the learning rate factor should be determined (and possibly adapted) in various cognitive radio settings, where different dynamics are experienced.

4) Complexity of the Learning Algorithm: In this section, we quantify the complexity of learning in terms of computational and storage requirements. We use the "flop" (floating-point operation) as a measure of complexity, which will provide us an estimation of the computational

[11] When the CSM deploys a mechanism without tax for the resource management, the space classification for other SUs can also be done based on the announced information and corresponding resource allocation.

complexity required for performing the learning algorithm. Also, based on this, we can determine how the complexity grows with an increasing number of SUs. At each stage, the SU performs the classification of other SUs' states, which, in the worst case, requires a number of "flops" of approximately N. The number of "flops" for estimating the transition probability of other SUs' states is approximately $(H_i + 1)$. The number of "flops" for learning the future reward is approximately $(2|S_i|H_i + 6)$. Therefore, the total number of "flops" incurred by the SU is $N + H_i + 2|S_i|H_i + 7$, from which we can note that the complexity of learning for each SU is proportional to the number of possible states of that SU and the number of classes in which the other SUs' state space is decomposed. To perform the learning algorithm, the SU needs to store two tables (i.e., transition probability table and state-value table), which have in total $(H_i^2(N+1) + 2^N|S_i|H_i)$ entries. We note that the storage complexity is also proportional to the number of possible states of that SU and the number of classes in which the other SUs' state space is decomposed.

E. Illustration of Various Bidding and Learning Strategies

In this section, we highlight the performance of the learning framework presented in the previous section in a centralized stochastic game (introduced in Section III). We assume that the SUs compete for the available spectrum opportunities in order to transmit delay-sensitive multimedia data. The SUs can deploy different bidding strategies to generate their bid vector.

- *Fixed bidding strategy* π_i^{fixed}: this strategy generates a constant bid vector during each stage of the auction game, irrespective of the state that SU i is currently in and of the states other SUs are in. In other words, π_i^{fixed} does not consider any source and channel dynamics.
- *Source-aware bidding strategy* π_i^{source}: this strategy generates various bid vectors by considering the dynamics in source characteristics (based on the current buffer state), but not the channel dynamics.
- *Myopic bidding strategy* π_i^{myopic}: this strategy takes into account both the environmental disturbances and the impact caused by other SUs. However, it does not consider the impact on its future rewards.
- *Bidding strategy based on best response learning* $\pi_i^{\mathcal{L}_i}$: This strategy is produced using the learning algorithm presented in the previous section, which considers both the environmental dynamics and the interaction impact on the future reward.

In this simulation, we consider the cognitive radio network as an extension of wireless LANs with cognitive radio capability [10]. (More simulation details can be found in [7].) To highlight the impact on the multimedia quality, in this illustrative simulation, we assume that both users are streaming to their receivers the *Coastguard* video sequence and both tolerate an application layer delay of 500 ms. For illustration, the following four scenarios are considered. In scenarios 1–4, SU 1 deploys a fixed bidding strategy π_1^{fixed}, source-aware bidding strategy π_1^{source}, myopic bidding strategy π_1^{myopic}, and best response learning based bidding strategy $\pi_1^{\mathcal{L}_1}$, respectively, and SU 2 always deploys the myopic bidding strategy π_2^{myopic}. The average video quality (PSNR), average tax, and average reward per time slot (see Section III-C5) are presented in Table 1.

From this simulation, we observe that when SU 2 deploys the myopic strategy, SU 1 increases its own reward by adopting advanced learning algorithms (from fixed bidding strategy π_i^{fixed} to best response learning based bidding strategy $\pi_1^{\mathcal{L}_1}$). On the other hand, SU 2 starts to have an increased cost as SU 1 starts to deploy increasingly advanced learning algorithms.

It is also worth noting that the improvement in video quality for SU 1 in scenarios 1–4 comes from two parts: one is the advanced bidding strategies, which allows the SU to take into consideration more information about its own states and the other SUs' states and, based on this, better forecast the impact of various actions; the other is the increase in the amount of resources consumed by SU 1, which corresponds to higher tax charged by the CSM, as shown in Table 1.

F. Learning in Repeated Games

A simplification of the stochastic game is the case where each SU has only one state. In this case, the stochastic game is reduced to a repeated game. In this case, the policy for each SU becomes the same as the action that each SU selected. Thus, an SU only needs to update its belief about the other SUs' actions.

Table 1 Performance of SU 1 and 2 With Various Bidding Strategies by the Two Competing SUs

	Bidding Strategies	SU 1			SU 2		
		Video Quality (PSNR)	Average tax	Average reward	Video Quality (PSNR)	Average tax	Average reward
Scenario 1	$p_1^{fixed}, p_2^{myopic}$	25 dB	0.1222	2.6337	36 dB	0.5495	1.5105
Scenario 2	$p_1^{source}, p_2^{myopic}$	26 dB	0.3147	2.4915	33 dB	0.6048	1.6116
Scenario 3	$p_1^{myopic}, p_2^{myopic}$	29 dB	0.4669	1.9767	30 dB	0.3763	1.7837
Scenario 4	$p_1^{L_1}, p_2^{myopic}$	35 dB	0.6923	1.7428	27 dB	0.4197	2.2967

1) Myopic Adaptation: In wireless communication, a simple learning (or adaptation) method is myopic adaptation, where the SU does not update its belief about the other SUs' actions. Instead, it maximizes its utility based on the aggregated observation of other SUs' actions during the previous round of game, i.e.,

$$a_i^{t,\text{myopic}} = \arg\max_{a_i^t} R_i^t\left(a_i^t, \mathrm{O}\left(\boldsymbol{a}_{-i}^{t-1}\right), w^t\right) \tag{36}$$

where $\mathrm{O}(\boldsymbol{a}_{-i}^{t-1})$ represents the aggregated observation of other SUs' actions in time slot $t-1$.

In power control games among WSTAs in interference channels, the myopic adaptation has been proven to converge to the Nash equilibrium point [37], which generally leads to a lower system performance for the user than the collaborative case, where a moderator will compel the WSTAs to operate on the Pareto surface.

2) Simple Reinforcement Learning in Repeated Games: In the reinforcement learning solution, an SU does not need to know the actions of the other SUs. Hence, this method is very suitable in a variety of repeated wireless games, including the above-mentioned power control games [31]. In this learning, the SU establishes a preference for each action. The preference is updated based on the utility that it obtains during the different stages of the game, without trying to explicitly model the other SUs' actions. Then, based on its preference, the SU determines a mixed action to perform during each time slot. Formally, when adopting the reinforcement learning algorithm, SU i computes its best response mixed action A_i^t as

$$A_i^t(a_i) = \frac{\phi\left(\rho_i^t(a_i)\right)}{\sum\limits_{a_i \in \mathcal{A}_i} \phi\left(\rho_i^t(a_i)\right)} \tag{37}$$

where $\rho_i^t(a_i)$ represents the *preference* of SU i choosing an action a_i at time slot t; $\phi(\cdot)$ is a nondecreasing positive function (e.g., $\phi(x) = e^x$); and $A_i^t(a_i)$ is the mixed action. When an action a_i is adopted by SU i at time slot t, the reward $R_i^t(a_i, a_{-i})$ is obtained. This reward is used to update the preference as follows:

$$\rho_i^t(a_i) = \rho_i^{t-1}(a_i) + \alpha\left[R_i^t(a_i, \boldsymbol{a}_{-i}) - \rho_i^{t-1}(a_i)\right] \tag{38}$$

where α is a update step size. An adaptive reinforcement (AR) technique can also be implemented, in which an SU can adapt its preference with various frequencies corresponding to different learning speeds, based on a cost–benefit tradeoff. A faster learning speed provides more accurate belief updates [in (25)–(27)]; however, it also requires a slightly higher computational cost and higher private information feedback overheads associated with the increased observations [in (23)].

3) Action-Based Learning in Repeated Games: In this setting, an SU explicitly models the exact actions of other SUs by directly exchanging information with other SUs (i.e., I_{-i}^t) about their taken actions. In this case, fictitious play and regret matching solutions can be used [36]. For instance, SU i can adopt an adaptive fictitious play algorithm, where it maintains a set of strategy vectors $a_{-i}^t[a_{-i}|a_i] = \{a_j^t[a_j \in \mathcal{A}_j|a_i],$ for all SUs $j \neq\}$ for all possible actions $a_i \in \mathcal{A}_i$, with $a_j^t[a_j \in \mathcal{A}_j|a_i]$ representing the estimated strategy of the other users $j \neq$ given that SU i took action a_i at time slot t. The adaptive fictitious play algorithm models the actions of other SUs $j \neq$ as

$$A_j^t(a_j|a_i) = \frac{\phi\left(\rho_j^t(a_j|a_i)\right)}{\sum\limits_{a_j \in \mathcal{A}_j} \phi\left(\rho_j^t(a_j|a_i)\right)} \tag{39}$$

where $\rho_j^t(a_j|a_i)$ represents the *anticipating preference* of SU j's choosing an action a_j at time slot t, given that the anticipator SU i takes an action a_i. The preference can be updated similarly as in the reinforcement learning case. Moreover, adaptive versions of this action learning, which we refer to as adaptive action (AA) learning, can also be adopted, where an SU is modeling other SUs with different accuracies in order to reduce the informational overhead and the computational overhead. This is especially important in the dynamic power/spectrum management games, where the neighboring SUs can be classified by an SU based on their impact on its utility. For instance, a neighboring SU with a larger channel gain will have higher impact on its utility.

G. Illustrative Results for Different Learning Approaches in Repeated Games

Next, we show several illustrative results using the learning schemes discussed in the previous sections in the distributed power control repeated games. We assume that five SUs (distinct transmitter-receiver pairs) are in the network and share three frequency channels. Each user can choose its power level from a set $\mathcal{P} = \{20, 40, 60, 80, 100\}$ (mW). Hence, there are a total of 15 actions for users to select. For the application layer parameters, we set the average packet length $L_v = 1000$ bytes, input rate $R_v = 500$ Kbps ($\Lambda_v = R_v/L_v$), and delay deadline $d_v = 200$ ms for all the users.

Besides the AR scheme mentioned in Section IV-F2 and the AA scheme mentioned in Section IV-F3, we also consider the myopic best response without learning discussed in Section IV-F1, which leads to a Nash equilibrium (NE). We

Table 2 Simulation Results for Various Repeated Games, Using Different Learning Techniques

Adopted schemes	SU	Reward (Kbit/joule)	Average reward
Myopic scheme	1	519.0	890.15
	2	195.2	
	3	530.6	
	4	2073.0	
	5	1132.9	
AR learning scheme	1	555.2	1005.6
	2	113.5	
	3	345.6	
	4	2830.2	
	5	1183.7	
AA learning scheme	1	529.3	1069.3
	2	475.6	
	3	476.8	
	4	2831.2	
	5	1033.3	

select SU 1 to be the user who learns from the observed information.

The results are presented in Table 2, where the reward is defined as in (18). From the results, it is interesting to see how the resulting reward of SU 1 improves when this user starts learning (when using AA and AR scheme) as opposed to the case that it is merely adopting a myopic best response (when using NE scheme). Using the AA scheme, users are able to exploit the spectrum more efficiently, due to the ability that the users can better model the strategies of other interference sources in the network. However, this requires significant information overhead, which results in a worse performance than using the AR scheme. Note that although only SU 1 is learning, the average reward of using interactive learning schemes outperforms the myopic NE scheme. Thus, as discovered in [23], this foresighted user benefits both itself as well as the overall system performance.

H. Future Research Directions for Learning in Communication Networks

Learning in games offers significant potential as a paradigm for shaping the dynamic interactions of wireless users and the resulting system efficiency [23], [38], [39]. As stated earlier, the majority of the research literature in this topic was aimed at proving that different types of equilibriums exist [21]. However, in wireless networks, the focus is on constructing adaptive algorithms and protocols that allow SUs to interact with each other based on their knowledge level in order to improve their performance. Accordingly, there are important research directions remaining to be addressed to enable the SUs and the wireless system to achieve the optimal performance. In particular, typical assumptions on knowledge of utility functions in multiagent learning are of the "all or nothing" type. That is, an agent either knows the utility function fully or can only measure payoffs online. A middle ground is the case where there is partial knowledge of the functional form but subject to uncertain parameters that may be estimated online (e.g., [39]). For instance, in the discussed communication setting, users sharing the same protocol have the same states and actions. The only difference is that they experience different private information. Thus, model-based learning approaches can be deployed that take advantage of the fact that users in the same protocol class adopt the same utility functions. These methods allow a user to learn more effectively, since they only need to learn the model parameters. Moreover, this can be also extended to the case where both the parameters and the models are unknown.

V. CONCLUSION

In this paper, we provided a unifying framework for dynamic multiuser spectrum access and strategic learning, which can be used to *architect* next-generation algorithms and implementations for competitive, heterogeneous, and dynamic cognitive radio networks. The presented framework can serve as a guideline for designing spectrum access solutions that are concerned with the tensions and relationships among autonomous adaptation by secondary (unlicensed) users, the explicit and implicit competition among these users, and the interaction of these users with spectrum moderators having their own goals (e.g., making money, imposing fairness rules, ensuring compliance to FCC [1], etc.).

The proposed knowledge-driven framework can be used to design efficient solutions for the usage of the spectrum under a broad set of operating scenarios. These scenarios include "fresh" spectrum, where all radios are cognitive, interactions of cognitive radios with licensed (nonadaptive, high-priority) users, and interactions of cognitive radios with legacy radios in the ISM bands. This framework provides incentives for the secondary users to deploy advanced transmission strategies to effectively gather information about the environment, learn from on it, and, lastly, maximize their own performance. Interestingly, our preliminary research [20], [23], [38], [41] has shown that even though users act competitively, this knowledge-driven approach to multiuser access may actually improve the performance of all or at least a majority of the users operating in the network.

We would like to note though that a large body of research and development work will still need to take place before such a knowledge-driven framework can be implemented in practical systems. For instance, enhanced online learning solutions that make optimal tradeoffs between the resulting utility and implementation costs need to be developed. Moreover, the various solutions for both dynamic spectrum access and learning will need to be tested in heterogeneous and highly dynamic cognitive radio systems, where a variety of SUs are competing for resources. Also, spectrum owners and wireless users will need to decide whether to adopt centralized or distributed solutions for managing the resources, whether they would like to make money, what type of fairness rules they would like to enforce, etc. Moreover, such solutions will also need to be developed for multihop cognitive radio networks (see [41] for some preliminary results on this topic).

Lastly, we believe that such cognitive radio networking solutions, which are based on stochastic interactions among users rather than the fixed, predetermined solutions and regulations used in the current networks, will ultimately lead to a *new generation of cyberinfrastructures and also next-generation applications, services, and intelligent devices*. Such solutions are especially necessary to ensure the proliferation of delay-sensitive high-bandwidth multimedia applications and services because these are most impacted by the inefficient spectrum use. ■

Acknowledgment

The authors would like to acknowledge the support of this work by grants from NSF Career and ONR.

REFERENCES

[1] FCC, "Spectrum policy task force," ET Docket 02-135, Nov. 2002.

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

[3] F. A. Ian, W. Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless network: A survey," *Comput. Netw.*, vol. 50, no. 13, Sep. 2006.

[4] A. Ghasemi and E. S. Sousa, "Collaborative spectrum sensing for opportunities access in fading environments," in *Proc. Dyspan 2005*, Nov. 2005, pp. 131–136.

[5] C. Doerr, M. Neureld, J. Fifield *et al.*, "MultiMAC—An adaptive MAC framework for dynamic radio networking," in *Proc. Dyspan 2005*, Nov. 2005, pp. 548–555.

[6] C. Kloeck, H. Jaekel, and F. Jondral, "Auction sequence as a new resource allocation mechanism," in *Proc. VTC'05*, Dallas, TX, Sep. 2005.

[7] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Trans. Veh. Technol.*, to be published.

[8] J. Huang, R. Berry, and M. L. Honig, "Auction-based spectrum sharing," *ACM Mobile Netw. Applicat. J. (MONET)*, vol. 11, no. 3, pp. 405–418, Jun. 2006.

[9] L. Berlemann, S. Mangold, G. R. Hiertz, and B. H. Walke, "Policy defined spectrum sharing and medium access for cognitive radios," *J. Commun. Acad.*, vol. 1, no. 1, Apr. 2006.

[10] C. T. Chou, N. S. Shankar, H. Kim, and K. Shin, "What and how much to gain by spectrum agility?" *IEEE J. Sel. Areas Commun.*, vol. 25, pp. 576–588, Apr. 2007.

[11] S. Shankar, C. T. Chou, K. Challapali, and S. Mangold, "Spectrum agile radio: Capacity and QoS implications of dynamic spectrum assignment," in *Proc. Global Telecommun. Conf.*, Nov. 2005.

[12] M. van der Schaar, Y. Andreopoulos, and Z. Hu, "Optimized scalable video streaming over IEEE 802.11 a/e HCCA wireless networks under delay constraints," *IEEE Trans. Mobile Comput.*, vol. 5, pp. 755–768, Jun. 2006.

[13] F. Fu, T. M. Stoenescu, and M. van der Schaar, "A pricing mechanism for resource allocation in wireless multimedia applications," *IEEE J. Sel. Topics Signal Process.*, vol. 1, pp. 264–279, Aug. 2007.

[14] R. W. Lucky, "Tragedy of the commons" *IEEE Spectrum*, vol. 43, no. 1, p. 88, Jan. 2006.

[15] R. G. Gallager, *Discrete Stochastic Processes*. Norwell, MA: Kluwer Academic, 1996.

[16] L. S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. USA*, vol. 39, pp. 1095–1100, 1953.

[17] C. Watkins and P. Dayan, "Q-learning," *Tech. Note Machine Learn.*, vol. 8, pp. 279–292, 1992.

[18] P. Klemperer, "Auction theory: A guide to the literature," *J. Econ. Surv.*, vol. 13, no. 3, pp. 227–286, Jul. 1999.

[19] J. Sun, E. Modiano, and L. Zheng, "Wireless channel allocation using an auction algorithm," *IEEE J. Sel. Areas Commun.*, vol. 24, May 2006.

[20] F. Fu and M. van der Schaar, "Non-collaborative resource management for wireless multimedia applications using mechanism design," *IEEE Trans. Multimedia*, vol. 9, pp. 851–868, Jun. 2007.

[21] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1999.

[22] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, Nov. 1999.

[23] Y. Su and M. van der Schaar, "A new perspective on multi-user power control games in interference channels, UCLA Tech. Rep. [Online]. Available: http://arxiv.org/abs/0709.3880

[24] Y. Su and M. van der Schaar, "Multi-user multimedia resource allocation over multicarrier wireless networks," *IEEE Trans. Signal Process.*, vol. 56, pp. 2102–2116, May 2008.

[25] A. Larcher, H. Sun, M. van der Schaar, and Z. Ding, "Decentralized transmission strategy for delay-sensitive applications over spectrum agile networks," in *Proc. 13th Int. Packet Video Workshop (PV 2004)*, 2004.

[26] A. B. MacKenzie and L. A. DaSilva, "Game theory for wireless engineers," in *Synthesis Lectures on Communications*. San Rafael, CA: Morgan and Claypool, 2006.

[27] H. Ji and C.-Y. Huang, "Non-cooperative uplink power control in cellular radio systems," *Wireless Netw.*, vol. 4, no. 3, pp. 233–240, 1998.

[28] D. Goodman and N. Mandayam, "Power control for wireless data," *IEEE Pers. Commun. Mag.*, vol. 7, pp. 48–54, Apr. 2000.

[29] M. Xiao, N. Schroff, and E. Chong, "Utility based power control in cellular radio systems," in *Proc. IEEE INFOCOM*, Anchorage, AK, 2001.

[30] A. B. MacKenzie and S. B. Wicker, "Game theory in communications: Motivation, explanation, and application to power control," in *Proc. IEEE GLOBECOM*, 2001, pp. 821–826.

[31] D. Vengerov, N. Bambos, and H. Berenji, "A fuzzy reinforcement learning approach to power control in wireless transmitters," *IEEE Trans. Systems, Man, Cybern. B*, vol. 35, no. 4, pp. 768–778, 2005.

[32] C. Long, Q. Zhang, B. Li, H. Yang, and X. Guan, "Non-cooperative power control for wireless ad hoc networks with repeated games," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1101–1112, 2007.

[33] Z. Han, C. Pandana, and K. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proc. Wireless Commun. Netw. Conf. (WCNC 2007)*, 2007, pp. 11–15.

[34] J. Neel, R. M. Buehrer, J. H. Reed, and R. P. Gilles, "Game theoretic analysis of a network of cognitive radios," in *Proc. 45th Midwest Symp. Circuits Syst.*, Aug. 2002, vol. 3, pp. 409–412.

[35] W. Yu, G. Ginis, and J. Cioffi, "Distributed multiuser power control for digital subscriber lines," *IEEE J. Sel. Areas Commun.*, vol. 20, pp. 1105–1115, Jun. 2002.

[36] Y. Chevaleyre, P. E. Dunne, U. Endriss, J. Lang, M. Lemaître, N. Maudet, J. Padget, S. Phelps, J. A. Rodríguez-Aguilar, and P. Sousa, "Issues in multiagent resource allocation," *Informatica*, vol. 30, pp. 3–31, 2006.

[37] F. Meshkati, H. V. Poor, and S. C. Schwartz, "Energy-efficient resource allocation in wireless networks," *IEEE Signal Process. Mag.*, vol. 24, pp. 58–68, May 2007.

[38] H. P. Shiang and M. van der Schaar, "Queuing-based dynamic channel selection for heterogeneous multimedia applications over cognitive radio networks," *IEEE Trans. Multimedia*, vol. 10, no. 5, pp. 896–909, Aug. 2008.

[39] J. Park and M. van der Schaar, "Stackelberg Contention Games in Multi-User Networks," *EURASIP Journal on Advances in Signal Process., Special issue on Game Theory in Signal Processing and Communications*, to appear.

[40] Y. B. Reddy, "Detecting primary signals for efficient utilization of spectrum using Q-learning," in *Proc. 5th Int. Conf. Inf. Technol.: New Gen. (ITNG 2008)*, 2008, pp. 360–365.

[41] H. P. Shiang and M. van der Schaar, "Distributed resource management in multi-hop cognitive radio networks for delay sensitive transmission," *IEEE Trans. Veh. Technol.*, 2008.

[42] Y. Su and M. van der Schaar, "Conjectural Equilibrium in Water-filling Games." [Online]. Available: http://arxiv.org/ftp/arxiv/papers/0811/0811.0048.pdf

ABOUT THE AUTHORS

Mihaela van der Schaar (Senior Member, IEEE) received the Ph.D. degree from Eindhoven University of Technology, The Netherlands, in 2001.

She is currently an Associate Professor with the Department of Electrical Engineering, University of California, Los Angeles. Since 1999, she has been an active participant in the ISO MPEG standard, to which she made more than 50 contributions. She is an Editor (with P. Chou) of *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems* (New York: Academic, 2007). She has received 30 U.S. patents.

Prof. van der Schaar received the National Science Foundation CAREER Award in 2004, the IBM Faculty Award in 2005, 2007, and 2008, the Okawa Foundation Award in 2006, the Best IEEE Transactions on Circuits and Systems for Video Technology Paper Award in 2005, the Most Cited Paper Award from the *EURASIP Journal Signal Processing: Image Communications* from 2004 to 2006, and three ISO Recognition Awards. She was on the editorial board of several IEEE Journals and Magazines.

Fangwen Fu (Student Member, IEEE) received the bachelor's and master's degrees from Tsinghua University, Beijing, China, in 2002 and 2005, respectively. He is currently pursuing the Ph.D. degree in the Department of Electrical Engineering, University of California, Los Angeles.

During the summer of 2006, he was an Intern with the IBM T. J. Watson Research Center, Yorktown Heights, NY. In 2008, he was selected by IBM Research as one of the 12 top Ph.D. students to participate in the 2008 Watson Emerging Leaders in Multimedia Workshop. His research interests include wireless multimedia streaming, resource management for networks and systems, applied game theory, video processing, and analysis.