This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

Dynamic Spectrum Sharing Using Learning for Delay-Sensitive Applications

Fangwen Fu, and Mihaela van der Schaar
Electrical Engineering Department, University of California Los Angeles (UCLA)
{fwfu, mihaela}@ee.ucla.edu

Abstract— **In this paper, we model the various wireless users streaming delay-sensitive data in a wireless network as a collection of selfish, autonomous agents that strategically interact in order to acquire the spectrum opportunities. The spectrum allocation is coordinated by a central spectrum moderator which deploys an incentive compatible mechanism. We further model the repeated spectrum competition as a stochastic game through which we are able to characterize the interaction among wireless users. Based on the observed resource allocation and corresponding rewards from previous allocations, we propose a best response learning algorithm that can be deployed by wireless users to improve the accuracy of their private information at each stage. The simulation results show that by deploying the proposed best response learning algorithm, the wireless users can significantly improve their own performance.**

I. Introduction

Wireless networks are envisioned to play a crucial role in the delivery of various delay-sensitive multimedia services to homes, enterprises, and campuses. A fundamental problem in enabling the large scale deployment of such networks is the absence of effective resource allocation schemes, which can arbitrate the division of the scarce wireless resource among competing and strategic delay-sensitive users.

To overcome the resource allocations over informationally decentralized wireless networks, pricing-based distributed resource allocation algorithms have been extensively investigated [1], where the network users are assumed to be "price-takers", i.e. the users accept the price announced by the network and do not consider the effects of their actions on the network price. To prevent the network users from anticipating the effects of their own actions on network performance, mechanism design-based resource allocation schemes have also been extensively studied in static network settings [3][4]. However, these solutions may not be suitable for the networks in which the available resources and network users' traffics are time-varying.

In this paper, we focus on developing solutions that can be employed by the wireless users to improve their performance in the dynamic wireless network. Specifically, we aim at investigating how delay-sensitive applications (e.g. multimedia applications) can efficiently forecast their future utility impact, and then determine their resource requirements and associated transmission strategies over time, based on information about the available spectrum opportunities, their source and channel characteristics, and interactions with the other competing users. Our solutions take into account the "self-interested" behavior of individual users/applications that may try to selfishly influence the resource management.

In our considered wireless network, the users are modeled as rational and strategic ones. We model the spectrum management as a stochastic game [5] in which the users simultaneously and repeatedly compete for the available wireless network resource (e.g. bandwidth). The competition for the resources is assisted by a central spectrum moderator (CSM) (similar to that in existing wireless LAN standards such as 802.11e HCF [6]). We assume that the CSM deploys a Vickery-Clarke-Groves (VCG) mechanism [9] for dynamically allocating resources which is similar to [4]. In order to capture the network dynamics, we allow the CSM to repeatedly allocate the available spectrum. Meanwhile, each user is allowed to strategically reveal to the CSM its private information about its source and channel characteristics, and impacts of the other users' private information.

Using this general stochastic wireless allocation framework, the key focus of this paper is to develop a learning methodology for users to improve their policies for playing the resource allocation game, i.e. the policies for generating and revealing the private information about the dynamics they experience. Specifically, during the repeated multi-user interactions, the users can observe partial historic information of the outcomes of the game, through which the users can estimate the impact on their future rewards and then adopt their best response in order to effectively compete for the spectrum resource. The estimation of the impact on the expected future reward can be performed using reinforcement learning [7] because this allows the users to improve their revealing strategy based only on the knowledge of their own past received payoffs, without knowing the private information and payoffs of the other users. Our proposed best response learning algorithm is inspired from the Q-learning [8] for the single agent interacting with environment. By deploying the best response learning algorithm, the user can strategically predict the impact of current actions on future performance and then optimally reveal its private information.

The paper is organized as follows. In Section II, we describe a system model for delay-sensitive transmission over wireless networks and a general model for the resource competition among the users. In Section III, we propose a stochastic framework to model the multi-user interactions. In Section IV, we propose a best response learning approach for the users to predict their future

This work was supported in part by NSF CAREER Award CCF-0541867 and in part by grants from UC Micro and ONR.

rewards impact based on the observed historic information. In Section V, we present the simulation results, followed by the conclusions in Section VI.

II. Spectrum Sharing Model for Delay-Sensitive Applications

We consider a situation in which M users, each formed by a single transmitter-receiver pair transmitting delay-sensitive data and denoted by i, coexist to share the same spectrum with bandwidth W Hz. We assume that each user experiences a Gaussian interference channel in discrete time fashion. By assuming that other users' transmitted signals are treated as white Gaussian noise, user i can achieve the transmission rate for a specific power allocations at time t:

$$R_i^t = \int_0^W \log\left(1 + \frac{c_{i,i} p_i^t(f)}{N_0 + \sum_{j \neq i} c_{j,i} p_i^t(f)}\right) df \qquad (1)$$

where $p_i^t(f)$ is the power spectral density of the input signal of user i at time t, N_0 is the power spectral density of the white Gaussian noise, and $c_{i,j}$ is the path gain from user i to user j. The power allocation $p_i^t(f)$ is constrained by:

$$\int_0^W p_i^t(f) df \leq P_i \text{ and } p_i^t(f) \geq 0 \qquad (2)$$

In this paper, we assume that the path gains $c_{i,j}, \forall i, j$ are constant during the whole course of transmission and satisfy a pairwise high interference condition (i.e. $c_{i,j}c_{j,i} > c_{i,i}c_{j,j}$). As shown in [2], when the channel satisfies a pairwise high interference condition, the optimal power allocations are orthogonal (i.e. $p_i^t(f)p_j^t(f) = 0$ for $f \in [0, W]$). Then the power allocations for the M users become the allocations of bandwidth, i.e. $\boldsymbol{w}^t = [w_1^t, ..., w_M^t]$ satisfying $\sum_{i=1}^M w_i^t = W$. In this case, the rate achieved by user i is computed as

$$R_i^t = w_i^t \log\left(1 + \frac{c_{i,i} P_i}{N_0 w_i^t}\right) \qquad (3)$$

To achieve the efficient allocations, the information exchange among the users is required [4]. In this paper, a mechanism for information exchange coordination is proposed which will be detailed in Section II.B. We first present the model for delay-sensitive applications.

A. Modeling for delay-sensitive applications

We consider that each delay-sensitive application associated with user i generates a random number of data in packets, denoted by A_i^t, for transmission at time t. The average packet length is L_i. The data generated at time t will expire at time $t + n_i$ if they are not successfully received. n_i is referred to as the life time of the data from user i and assumed to be constant. The data arrival A_i^t at any time t is assumed to follow the Poisson distribution with average λ_i. The packets arriving at time t will be in the buffer with life time n_i at time $t+1$. We assume that the packet with life time n at time t has utility (i.e. quality contribution) $\psi_{i,n}^t$.

We define a "state" $\boldsymbol{s}_i^t$ for user i at time t as the number of packets remaining for transmission, i.e. $\boldsymbol{s}_i^t = [v_{i,1}^t, ..., v_{i,n_i}^t]$ where $v_{i,n}^t$ represents the number of remaining packets with life time $n (1 \leq n \leq n_i)$. Given the bandwidth allocation w_i, the number of packets which can be transmitted in total is $m_i^t = \lfloor R_i^t / L_i \rfloor$. The state transition of user i is expressed by the following equations:

$$v_{i,n}^{t+1} = \begin{cases} \left[v_{i,n+1}^t - \left[m_i^t - \sum_{j=1}^n v_{i,j}^t \right]^+ \right]^+, & 1 \leq n \leq n_i - 1 \\ A_i^{t+1}, & n = n_i \end{cases}, \qquad (4)$$

where $[x]^+ = \max\{0, x\}$.

At time t, we assume that user i receives utility $(\alpha_i)^{n-1}\psi_{i,n}^t$ if it sends out a packet with life time n. This can be interpreted that the utility $\psi_{i,n}^t$ of a packet with life time n is discounted by $(\alpha_i)^{n-1}$ if it sends out at the current time. The factor $\alpha_i (0 \leq \alpha_i < 1)$ is the discounted factor determined by a specific application. Hence, by obtaining the bandwidth w_i^t, user i can get the gain

$$g_i^t = \sum_{n=1}^{n_i} (\alpha_i)^{n-1} \psi_{i,n}^t \left[\min\left\{ v_{i,n}^t, m_i^t - \sum_{j=1}^{n-1} v_{i,j}^t \right\} \right]^+ \qquad (5)$$

Recall the computation m_i^t and R_i^t. we note that g_i^t is determined by $L_i, c_{i,i}, P_i, N_0, w_i^t, \alpha_i$, $\psi_{i,1}^t, ..., \psi_{i,n_i}^t$ and $v_{i,1}^t, ..., v_{i,n_i}^t$. Since $L_i, c_{i,i}, P_i$ and N_0 are constant in the whole course of transmission, we write g_i^t as a function of w_i^t, $\boldsymbol{s}_i^t$ and $(\alpha_i)^{n-1}\psi_{i,n}^t$, i.e. $g_i^t(w_i^t, \boldsymbol{s}_i^t, \boldsymbol{\phi}_i^t)$ where $\boldsymbol{\phi}_i^t = [\psi_{i,1}^t, ..., (\alpha_i)^{n_i-1}\psi_{i,n_i}^t]$.

B. Bandwidth allocation Mechanism

As mentioned in Section I, in a wireless network, the information is decentralized, and thus, the information exchange between the users needs to be kept limited due to the incurred communication cost. On the other hand, the users competing with each other are selfish and strategic and hence, the information they hold is private and may not be shared with each other. Therefore, one of our key interests in this paper is to determine what information should be exchanged between users and how this information should be exchanged.

In this section, we propose that the users exchange information with central spectrum moderator (CSM) instead of direct communicating with each other. Specifically, we present a mechanism named as VCG[1] [9] for dynamically coordinating the interactions among users.

The VCG mechanism is performed by the CSM during each time slot. At the beginning of the time slot, the users are required to submit their own private information

[1] Other mechanisms like the ones in [3] can also be deployed without modifying our framework.

$\theta_i^t = \{c_{i,i}, P_i, L_i, N_0, \boldsymbol{s}_i^t, \boldsymbol{\phi}_i^t\}$. In general, the submitted version $\tilde{\theta}_i^t$ of the private information can be different from the true value θ_i^t due to the strategic behavior of user i. It has been proved that, in VCG mechanism, the optimal submitted version of the private information is the true value, i.e. $\tilde{\theta}_i^{t,opt} = \theta_i^t$, which is called "truth telling" property of the VCG mechanism [9]. From now on, we assume that user i always reveals the true value θ_i^t. As assumed before, the value of $c_{i,i}, P_i, L_i, N_0$ keeps constant during the transmission, and hence they are only required to submit once to the CSM. Thus, at each time slot (expect the first time slot), user i submits the private information $\theta_i^t = \{\boldsymbol{s}_i^t, \boldsymbol{\phi}_i^t\}$. For the remaining text, we rewrite $g_i^t(w_i^t, \boldsymbol{s}_i^t, \boldsymbol{\phi}_i^t)$ as $g_i^t(w_i^t, \theta_i^t)$.

After receiving the announced private information from the users, CSM computes the optimal bandwidth allocation $w_i^{t,opt}$ for each user i based on the submitted information $\{\theta_1, ..., \theta_M\}$ by maximizing the sum of gain at time t:

$$\boldsymbol{w}^{t,opt} = \arg\max_{\boldsymbol{w}^t} \sum_{i=1}^{M} g_i^t(w_i^t, \theta_i^t) \quad (6)$$

To compel the users to declare their private information truthfully, the CSM also computes the payment $\tau_i^t \in \mathbb{R}_-$ that the users have to pay for the use of resources during the current stage of the game as:

$$\tau_i^t = \sum_{j \neq i} g_j^t(w_j^{t,opt}, \theta_j^t) - \max_{\boldsymbol{w}_{-i}^t} \sum_{j \neq i} g_j^t(w_j^t, \theta_j^t). \quad (7)$$

where $-i = \{1, ..., i-1, i+1, ..., M\}$. It is easy to see that $\tau_i^t \leq 0$. The absolute value of the payment is the amount of "money" or tokens that user i has to pay the CSM for the used resources. The allocation result is then transmitted back to the users which can deploy their transmission strategies in different layers and send data over the assigned spectrum. After the data transmission, another competition starts at the next time slot $t+1$.

C. Private information for users in repeated games

The variations of the source characteristics of delay-sensitive applications are characterized by the current "states" as shown in Section II.A. At the various state s_i^t, user i will announce different private information θ_i^t. As discussed in Section II.B, for one stage of allocation game induced by the VCG mechanism, user i takes the optimal announcement $\tilde{\theta}_i^{t,opt} = \theta_i^t$. When the user is aware of the sequential games it has to play, it also has to take into account the future impact of current announcement as its private information in the current stage. In other words, the key questions to determine private information in the repeated allocation games among the users are: (i) what state each user experiences in each time slot in the dynamic network; (ii) during the repeated competition, how the interactions among the users are modeled; and (iii) how the users forecast the impact of the current announcement on the future performance.

To overcome the above addressed problems, we present in the next section a stochastic framework for modeling the dynamic interaction among users.

III. Stochastic Model for Users Interaction

The users announce their own private information in the repeated games given their dynamically changing states which they experience. The evolution of users' indirect interactions across the various time slots can be modeled as a stochastic game [5]. In the stochastic game, every user has its own state and its own action space for that state. The time slot corresponds to the "stage" commonly used in the stochastic game. In the remainder of the paper, we use the time slot and stage interchangeably. The users choose their own actions independently and simultaneously at each time slot. Next, they receive their rewards and transit to the next states. It is worth noting that the reward received by each user, and state transition also depend on other users' states and actions.

Formally, a stochastic game is a tuple $(\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{S}}, \boldsymbol{\Theta}, \boldsymbol{P}, \boldsymbol{\mathcal{R}})$, where $\boldsymbol{\mathcal{I}}$ is the set of users, i.e. $\boldsymbol{\mathcal{I}}=\{1, ..., M\}$, $\boldsymbol{\mathcal{S}}$ is the set of state profiles of all users, i.e. $\boldsymbol{\mathcal{S}}=\mathcal{S}_1 \times \cdots \times \mathcal{S}_M$ with $\mathcal{S}_i$ being the state set of user i, and $\boldsymbol{\Theta}$ is the joint action space $\boldsymbol{\Theta}=\Theta_1 \times \cdots \times \Theta_M$, with Θ_i being the action (revealing) set available for user i to play the game. $\boldsymbol{P}$ is a transition probability function defined as a mapping from the current state profile $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$, corresponding joint actions $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the next state profile $\boldsymbol{s}' \in \boldsymbol{\mathcal{S}}$ to a real number between 0 and 1, i.e. $\boldsymbol{P}: \boldsymbol{\mathcal{S}} \times \boldsymbol{\Theta} \times \boldsymbol{\mathcal{S}} \mapsto [0,1]$. $\boldsymbol{\mathcal{R}}$ is a reward vector function defined as a mapping from the current state profile $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$ and corresponding joint actions $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ to an M-dimensional real vector with each element being the reward to a particular user, i.e. $\boldsymbol{\mathcal{R}}: \boldsymbol{\mathcal{S}} \times \boldsymbol{\Theta} \mapsto \mathbb{R}^M$.

The state profile $\boldsymbol{s} \in \boldsymbol{\mathcal{S}}$ can be sometimes rewritten as $\boldsymbol{s} = (\boldsymbol{s}_i, \boldsymbol{s}_{-i})$ to distinguish the state of user i and the states of other users. Similarly, the joint action $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can also be represented as $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$. In the subsequent sections, we specify the elements of the stochastic game model for the interactions among the users in the considered wireless network.

A. State transition

We will now discuss the state transition process. Remember that the state of user i includes the buffer state $v_{i,n}^t$ with the life time n. Given the bandwidth allocation w_i^t, user i can transmit m_i^t packets at most. Remember that the number of packets arriving at time t is A_i^t according to Poisson distribution. Then the next state $\boldsymbol{s}_i^{t+1}$ at time $t+1$ is computed as Eq. (4). The state transition probability is expressed by

$$q_i(\boldsymbol{s}_i^{t+1} \mid \boldsymbol{s}_i^t, w_i^t) = \begin{cases} \dfrac{\lambda^{v_{i,n_i}^{t+1}} e^{-\lambda}}{(v_{i,n_i}^{t+1})!} & if\ v_{i,n}^{t+1}\ satisfies\ eq(4) \\ 0 & otherwise \end{cases} \quad (8)$$

B. Stage reward

By playing the allocation game in the current stage, user i receives the bandwidth w_i^t and needs to pay the corresponding payment $|\tau_i^t|$. Based on the allocated bandwidth, the user transmits the available packets in the buffer. In the next time slot, new packets arrive into the buffer. The reward at time t for user i is expressed using the quasi-linear form $r_i^t = g_i^t + \tau_i^t$. Note that the gain g_i^t and payment τ_i^t depend on the states and revealed information of all the competing users in the network.

C. Determining the private information

In the wireless network, we assume that the stochastic game is played by all users for an infinite number of stages. This assumption is reasonable for applications having a long duration, e.g. video streaming. We assume that the revealing policy is to submit the private information θ_i^t to the CSM always. Unlike the private information for one-shot game discussed in Section II.B, the private information θ_i^t in the repeated game should include the impact of the current announcement on the expected future rewards.

The reward for user i at the time slot t is $r_i((\boldsymbol{s}_i^t, \boldsymbol{s}_{-i}^t),(\theta_i^t, \boldsymbol{\theta}_{-i}^t))$. This reward $r_i((\boldsymbol{s}_i^k, \boldsymbol{s}_{-i}^k),(\theta_i^k, \boldsymbol{\theta}_{-i}^k))$ of the stage k is discounted by a factor $(\alpha_i)^{k-t}$ at time t $(t \le k)$. The discounted factor is the same as before. The total discounted sum of rewards $Q_i^t(\boldsymbol{s}^t)$ for user i can be calculated at time t from any state profile $\boldsymbol{s}^t$ as:

$$Q_i^t(\boldsymbol{s}^t) = \sum_{k=t}^{\infty} (\alpha_i)^{k-t} r_i((\boldsymbol{s}_i^k, \boldsymbol{s}_{-i}^k),(\theta_i^t, \boldsymbol{\theta}_{-i}^t)) =$$

$$\underbrace{g_i(s_i^t, w_i^t) + \tau_i^t}_{\text{stage reward at time t}} + \alpha_i \underbrace{\sum_{\boldsymbol{s}^{t+1} \in \boldsymbol{\mathcal{S}}} \prod_{k=1}^{M} q_k(\boldsymbol{s}_k^{t+1} \mid \boldsymbol{s}_k^t, w_i^t) Q_i^{t+1}(\boldsymbol{s}^{t+1})}_{\text{expected future reward}}. \quad (9)$$

The total discounted sum of rewards in Eq. (9) consists of two parts: (i) the current stage reward and (ii) the expected future reward discounted by α_i. Note that user i cannot independently determine the above value without explicitly knowing the private information of other users. We define the impact of w_i^t as

$$F_i^t(w_i^t, \boldsymbol{s}^t) = \sum_{\boldsymbol{s}^{t+1} \in \boldsymbol{\mathcal{S}}} \prod_{k=1}^{M} q_k(\boldsymbol{s}_k^{t+1} \mid \boldsymbol{s}_k^t, w_i^t) Q_i^{t+1}(\boldsymbol{s}^{t+1}) - \sum_{\boldsymbol{s}^{t+1} \in \boldsymbol{\mathcal{S}}} \prod_{k=1}^{M} q_k(\boldsymbol{s}_k^{t+1} \mid \boldsymbol{s}_k^t, 0) Q_i^{t+1}(\boldsymbol{s}^{t+1}), \quad (10)$$

which is interpreted as the extra expected future reward when allocating w_i^t comparing to that when no bandwidth is allocated. It is clear that $F_i^t(w_i^t, \boldsymbol{s}^t)$ varies as w_i^t increases. By receiving the allocation w_i^t, user i obtains utility $g_i^t(w_i^t, \boldsymbol{s}_i^t) + \alpha_i F_i^t(w_i^t, \boldsymbol{s}^t)$ now.

From Eq. (9) and (10), we observe that the private information θ_i^t should be modified to include the information $\boldsymbol{s}_i^t$, $\boldsymbol{\phi}_i^t$, and the parameters for computing $\alpha_i F_i^t(w_i^t, \boldsymbol{s}^t)$. However, to reveal the entire function to the CSM, it may require a large number of parameters which characterize the function and thereby, resulting in huge communication cost. To avoid this difficulty, we approximate the function $F_i^t(w_i^t, \boldsymbol{s}^t)$ using the same form as g_i^t. That is,

$$F_i^t(w_i^t, \boldsymbol{s}^t) \approx \sum_{n=1}^{n_i} \varphi_{i,n}^t \left[\min \left\{ v_{i,n}^t, m_i^t - \sum_{j=1}^{n-1} v_{i,j}^t \right\} \right]^+ \quad (11)$$

where $\varphi_{i,n}^t$ is determined using least square approximation [10]. We should note that the approximation trades off the revealing complexity and performance of playing the allocation game. In this way, the private information required to reveal is $\theta_i^t = \{\boldsymbol{s}_i^t, \alpha_i \boldsymbol{\varphi}_i^t + \boldsymbol{\phi}_i^t\}$ where $\boldsymbol{\varphi}_i^t = [\varphi_{i,1}^t, ..., \varphi_{i,n}^t]$.

However, from Eq. (10), we know the computation of $F_i^t(w_i^t, \boldsymbol{s}^t)$ depends on other users' states and revealing strategies which, in general, is unknown to user i. In next section, we develop a simple learning algorithm for user i to estimate the value of $F_i^t(w_i^t, \boldsymbol{s}^t)$.

IV. Learning for Future Reward Impact

In this paper, we assume that user i observes the information $\{\boldsymbol{s}_i^0, \theta_i^0, w_i^0, \boldsymbol{\tau}_i^0, ..., \boldsymbol{s}_i^{t-1}, \theta_i^{t-1}, w_i^{t-1}, \boldsymbol{\tau}_i^{t-1}, \boldsymbol{s}_i^t\}$. We introduce learning as a tool to predict the impacts of current announcement and hence, current private information. However, a key question is what needs to be learned. Here we simply assume that user i models the dynamics of other users as a stationary process. Hence the state of other users is degenerated into a stationary state (i.e. one state). By this approximation, the state transition probability for other users becomes 1, and the computation of $Q_i^t(\boldsymbol{s}^t)$ and $F_i^t(w_i^t, \boldsymbol{s}^t)$ are simplified, respectively, as

$$Q_i^t(s_i^t) = g_i(s_i^t, w_i^t) + \tau_i^t + \alpha_i \sum_{\boldsymbol{s}_i^{t+1} \in \boldsymbol{\mathcal{S}}_i} q_i(\boldsymbol{s}_i^{t+1} \mid \boldsymbol{s}_i^t, w_i^t)) Q_i^{t+1}(s_i^{t+1}). \quad (12)$$

$$F_i^t(w_i^t, \boldsymbol{s}_i^t) = \sum_{\boldsymbol{s}_i^{t+1} \in \boldsymbol{\mathcal{S}}_i} q_i(\boldsymbol{s}_i^{t+1} \mid \boldsymbol{s}_i^t, w_i^t) Q_i^{t+1}(\boldsymbol{s}_i^{t+1}) - \sum_{\boldsymbol{s}_i^{t+1} \in \boldsymbol{\mathcal{S}}_i} q_i(\boldsymbol{s}_i^{t+1} \mid \boldsymbol{s}_i^t, 0) Q_i^{t+1}(\boldsymbol{s}_i^{t+1}). \quad (13)$$

Since the state transition $q_i(\boldsymbol{s}_i^{t+1} \mid \boldsymbol{s}_i^t, w_i^t)$ is known, user i only needs to estimate $Q_i^{t+1}(\boldsymbol{s}_i^{t+1})$ to predict $F_i^t(w_i^t, \boldsymbol{s}_i^t)$. Inspired by the Q-learning [8], we can estimate $Q_i^{t+1}(\boldsymbol{s}_i^{t+1})$ in the similar way which is described as follows.

We use a table to store the value $V_i(s_i)$ representing $Q_i^{t+1}(\boldsymbol{s}_i^{t+1})$ with $s_i \in \boldsymbol{\mathcal{S}}_i$. User i updates the value of $V_i(s_i)$ at time t according to the following rules:

$$V_i^t(s_i) = \begin{cases} (1-\gamma_i^t) V_i^{t-1}(\boldsymbol{s}_i) + \gamma_i^t Q_i^t(\boldsymbol{s}_i^t) & \text{if } \boldsymbol{s}_i^t = \boldsymbol{s}_i \\ V_i^t(\boldsymbol{s}_i) & \text{otherwise} \end{cases} \quad (14)$$

where $\gamma_i^t \in [0,1)$ is a learning rate factor satisfying

$\sum_{t=1}^{\infty}\gamma_i^t = \infty$ and $\sum_{t=1}^{\infty}(\gamma_i^t)^2 < \infty$ [8] and $Q_i^t(\boldsymbol{s}_i^t)$ is computed as in Eq. (12) by replacing $Q_i^{t+1}(\boldsymbol{s}_i)$ with $V_i^{t-1}(\boldsymbol{s}_i)$. In summary, the learning procedure that is developed for a user is shown in Table 1.

Table 1. Learning Procedure

Initializing: $V_i^0((s_i)) \Leftarrow 0$ for all possible states $s_i \in \boldsymbol{\mathcal{S}}_i$.

Learning: At time t, user i:

1) Observes the current state $\boldsymbol{s}_i^t$;
2) Compute $F_i^t(w_i^t)$ in Eq. (13) by replacing $Q_i^{t+1}(\boldsymbol{s}_i^{t+1})$ with $V_i(\boldsymbol{s}_i^t)$, approximate $F_i^t(w_i^t, \boldsymbol{s}_i^t)$ using Eq. (11) to produce $\boldsymbol{\varphi}_i^t$ and announce θ_i^t;
3) Receives the allocation w_i^t and payment τ_i^t;
4) Computes the expected total discounted sum of the rewards $Q_i^t(\boldsymbol{s}_i^t)$ as in Eq. (12);
5) Updates the future reward table $V_i(\boldsymbol{s}_i)$ at the state $\boldsymbol{s}_i^t$ using the learning rate factor γ_i^t, according to Eq. (14).

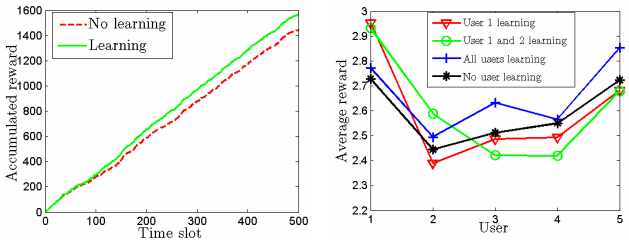
V. Simulation Results

In this simulation, we aim at verifying that the proposed learning algorithm predicts the impact of current revealing action (determining the private information) on future rewards and hence, improve the users' performance in terms of gained utility.

We consider multiple (five) users streaming delay-sensitive data over the wireless network. The signal to noise ratio ($P_i / N0$) is 30dB for all users. The bandwidth W is normalized to 1. The path gain $c_{i,i} = 1, \forall i$. The packet arrive rate λ_i is 10 packets/slot and the slot length is 0.01s. The discounted factor $\alpha_i = 0.2$. Similar parameters are used for the five users in order to clearly illustrate the performance differences obtained based on the different strategies. The similar observations are obtained in other settings.

We first compare two scenarios: (1) no user is deployed with the proposed learning algorithm; (2) user 1 uses the proposed algorithm to determining its private information. Figure 1.(a) shows the accumulated gained rewards under the two scenarios. The average rewards gained per time slot are 2.71 and 2.98, respectively. The learning algorithm improves the average reward by around 10%. This improvement is due to the successful prediction on the future reward impact and more accurate private information.

We further consider the case where multiple users learn simultaneously. The average rewards under different scenarios with various learning users are illustrated in Figure 1.(b). Interestingly, when part of users ((e.g. user 1 and 2)) are deployed with the learning algorithm, they can improve their own average rewards by around 10% but penalize other users by around 2%~8%. While all users start learning, all users obtain benefits from 0.5%~6%, comparing to the case of no users learning. This can be briefly explained as follows: when part of users learn, only these users can accurately evaluate their private information and announce it properly. While those users without learning misrepresent their private information and thereby, losing performance. When all the users start learn, they are able to announce their accurate private information and hence, obtaining various benefits. More investigation on multi-user learning will be conducted in the future.



(a) (b)

Figure 1. (a) Accumulated rewards of user 1 when no user learns or only user 1 learns; (b) average rewards of users under the scenarios with various users learning.

VI. Conclusions and Future Work

In this paper we model the wireless resource allocation problem as a "stochastic game" played among strategic users. The users are allowed to simultaneously and independently determine and reveal their own private information about the dynamics of networks. To improve the revealing strategy at each stage, we propose a best response learning algorithm to predict the possible future reward at each state. The simulation results show that our proposed learning algorithm can significantly improve the users' performance. The proposed stochastic game framework can further allow wireless users to compete for the time-varying available network resources, e.g. in cognitive radio networks by deploying more complicated strategies, e.g. cross-layer optimization. How the network dynamics is further exploited during multi-user learning falls into our future research.

References

[1] F. Kelly *et al,* "Rate control for communication networks: shadow prices, proportional fairness and stability," *Operational Research Society,* vol. 49, pp.237-252, 1998.

[2] R. Etkin *et al*, "Spectrum sharing for unlicensed bands," *IEEE J. Sel. Areas Commun.,* vol. 25, no. 3, pp. 517-528, Apr. 2007.

[3] J. Huang *et al*, "Auction-based Spectrum Sharing", ACM Mobile Networks and Applications Journal (MONET), vol. 11, no. 3, pp. 405-418, June 2006.

[4] F. Fu *et al*, "Non-collaborative resource management for wireless multimedia applications using mechanism design," *IEEE Transaction on Multimedia*, vol. 9, no. 4, pp. 851-868, Jun. 2007.

[5] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, 1095-1100, 1953.

[6] "IEEE 802.11e/D5.0, wireless medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) enhancements for Quality of Service (QoS), draft supplement," June 2003.

[7] M. Bowling *et al*, "Rational and convergent learning in stochastic games," *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)* , 1021--1026, August 2001.

[8] C. Watkins *et al*, "Q-learning." *Technical Note, Machine Learning,* vol. 8, 279-292, 1992.

[9] M. Jackson, "Mechanism theory," *In the Encyclopedia of Life Support Systems*, 2003.

[10] S. P. Boyd *et al*, "Convex optimization," Cambridge University Press, 2004.