Wireless Pers Commun (2017) 96:1887–1905
DOI 10.1007/s11277-017-4274-4

Optimizing Spectrum Sharing in Wireless Mesh Network Using Cognitive Technology

Ayoub Alsarhan[1] · Ahmad N. Quttoum[2] · Yousef Kilani[1]

Published online: 24 April 2017

Abstract In cognitive wireless networks, spectrum owners (primary users, PUs) may lease the unused spectrum to unlicensed users (secondary users, SUs). This spectrum is used to establish a secondary network that serves real time connections. The size of leased spectrum influences both the admitted traffic of SUs and the cost of spectrum. For this spectrum market, we present unsupervised learning paradigm as a means for extracting the optimal control policy for spectrum trading. This policy gives spectrum owner the opportunity to maximize its profit by adapting network resources to the changes in the network status and the market conditions. To meet different requirements, the problem is formulated as reward maximization with penalty for delay. The numerical results show that the proposed machine learning method is able to find an efficient trade-off between profit loss, and average delay for SUs.

Keywords Cognitive radio · Dynamic spectrum access · Spectrum resource management · Spectrum trading · Wireless mesh networks · Markov decision process

1 Introduction

Wireless mesh networks (WMNs) have emerged recently to provide high-bandwidth network, extend internet access, and other networking services. Hence, WMNs are predicted to be a key technology that offer ubiquitous connectivity to the end users [12, 13, 21]. Although WMNs provide better services in term of flexible network

✉ Ayoub Alsarhan
AyoubM@hu.edu.jo

Ahmad N. Quttoum
quttoum@hu.edu.jo

Yousef Kilani
ymkilani@hu.edu.jo

[1] Deptartment of Computer Information System, Hashemite University, Zarqa, Jordan

[2] Deptartment of Computer Engineering, Hashemite University, Zarqa, Jordan

architectures, easy deployment and configuration, and fault tolerance, still the limited available spectrum and the inefficiency in the spectrum usage degrade the network performance significantly.

Nowadays, the scarcity of bandwidth is the main challenge in WMNs technology. Indeed, beside the dramatic increase in the access to the limited bandwidth, fixed spectrum assignment policies prevent users from dynamically utilizing the unused spectrum, which results in very poor utilization of spectrum. To overcome spectrum scarcity problem, Federal Communications Commission (FCC) allows SUs to access unutilized spectrum if they do not interfere with PUs [12, 13, 21].

Dynamic spectrum access techniques enable users to choose operating spectrum on-demand. Spectrum Management is one of the key challenges of dynamic spectrum access techniques [10, 11, 43]. Besides utilizing the unused spectrum, we also need an efficient mechanism that considers maximizing the spectrum owner profit while sharing spectrum.

Presently, new technologies such as cognitive radio (CR) motivate the concept of opportunistic spectrum access using dynamic and adaptable spectrum sharing. Moreover, the social and economic value of spectrum applications is enhanced. CR is a promising technology for next generation wireless networks. In CR, spectrum can be shared among several users to improve spectrum utilization. Cognitive user can select the best channel [13, 21]. Moreover, CR encourages implementing new more flexible spectrum sharing paradigms. These sharing paradigms include: underlay, overlay, and spectrum trading techniques [1, 2]. In overlay technique, the spectrum can be accessed by SU if it is free. Underlay technique allows SUs to transmit concurrently with PU if the signal power of SU is below the interference temperature of PU. Unfortunately, SUs access the licensed spectrum without paying any usage charge to PUs in overlay and underlay approaches. Although these approaches solve spectrum scarcity problem, it is not likely to be accepted in the current spectrum market since the PUs do not have any financial incentive from SUs usage of spectrum.

In order to enhance SUs' satisfaction and generate more revenue, PUs lease free spectrum to SUs. This process is referred to as spectrum trading [1, 2] where spectrum is purchased and sold. In this paradigm, spectrum sharing among PUs is a key challenging problem.

This paper addresses when and how spectrum is shared between PUs and SUs based on economic model and under dynamic traffic load conditions. Economic model includes cost and reward of spectrum trading. Our design objective is to improve spectrum utilization and maximize PUs' profits, while meeting certain predefined constraints. In our work, PUs coordinate among themselves to trade spectrum and maximize their profits. The reinforcement learning (RL) framework [8, 9, 29, 36] is used for extracting the optimal trading policy. RL policy is used to serve spectrum requests in a given single queue using an adaptable amount of spectrum with certain reward and cost parameters. Spectrum requests are served based on the gained profit. In this paper, PU may borrow spectrum if its spectrum is inadequate to serve SUs. However, a request is placed in the queue if there is no spectrum for serving it.

Our trading approach can be applied in a style of trading where the agent (PU) charges the clients for serving their requests. PUs trade their services on cloud resources for money. Our approach presents a general framework for studying, analyzing, and optimizing other resource trading in the wireless environment.

Hence, the contribution of this paper comes in twofold. First, we describe how the concept of RL is used to obtain a computationally feasible solution to the considered spectrum trading problem. Second, we present an extensive numerical evaluation, based on simulation of the RL-based method.

The rest of the paper is organized as follows. Related works to spectrum trading is reviewed in Sect. 2. Section 3 describes the system model and assumptions. The RL formulation is presented in Sect. 4. Section 5 presents the performance evaluation results. Finally, the paper is concluded in Sect. 6.

2 Related Work

With Dynamic Spectrum Accessing (DSA), spectrum trading is proposed to solve the problem of spectrum scarcity [12, 20]. Nevertheless, the implementation of spectrum trading faces several challenges because of the fluctuating nature of the available spectrum and the changes in the spectrum demand. Some appropriate incentives for spectrum trading must be provided to the PUs for leasing their spectrum and to compensate SUs for waiting time. Spectrum trading is adopted in [20, 21] for spectrum management. The scheme provides PUs incentives (e.g. money) to temporarily lease their spectrum, as well as provides SUs opportunities to access the unused spectrum. So, spectrum trading is used to promote PUs for sharing their unused spectrum. The main approaches for spectrum trading are summarized in [31]. These approaches include game theory, auction theory, and microeconomics.

Multiple PUs sell spectrum to multiple SUs in [32]. Non-cooperative game is used to model the competition among the PUs and evolutionary game is used for modeling SUs' behavior in the spectrum market. New system for spectrum trading is proposed in [27]. In order to react dynamically and locally to the secondary spectrum market, the proposed system [27] combines pricing, spectrum allocation, and billing. A joint power/channel allocation scheme is proposed in [37] for trading free channels. The proposed trading scheme uses a pricing strategy to improve the network's performance. A non-cooperative game based on pricing scheme is proposed in [41] to control the uplink power of the cognitive network. Auction theory is used in [41] for the problem of dynamic spectrum sharing.

An auction mechanism is applied in [39] for spectrum sharing among SUs using spread spectrum signaling. In order to generate additional revenue, multiple auctioneers sell idle spectrum bands for SUs in [23]. A Multiauctioneer Progressive auction mechanism (MAP) is proposed where each auctioneer (PU) raises trading price and each bidder (SU) subsequently chooses one auctioneer for bidding. The problem of maximizing PU's average profit is tackled in [30]. The problem of spectrum pricing is investigated in the presence of PUs and SUs [24]. Stochastic dynamic programming is used to find the optimal spectrum price.

Markov approach is used in [38] for spectrum trading. The interactions between PUs and SUs are modeled using continuous time Markov chains. The spectrum resources can be efficiently and fairly shared among SUs in an opportunistic way without interrupting the spectrum usage of the PUs by studying the optimal spectrum access probabilities of SUs. Authors in [15] present a cross-layer design for reliable data transmission over a cognitive radio network. The new design combines adaptive modulation at the physical layer and hybrid automatic repeat request at the data link layer. The proposed scheme follows the principles of opportunistic spectrum access that utilizes an optimal power adaptation policy for channel allocation. Three-dimensional traffic model is proposed in [42]. This model is used to identify and to utilize those under-utilized channel opportunities efficiently. Furthermore, a new scheme for scheduling free channel is proposed. The scheme provides SUs with Quality of Service (QoS) support.

Learning-based schemes are proposed in [7] to sense multiple access (CSMA) for SUs when the PU operates with the conventional CSMA/CA scheme. The learning algorithms are applied to tune the value of transmission probability to balance the channel idle time

and collision costs due to the fact that both the PU and the SUs are sharing the same wireless channel. The authors in Alsarhan et al. [3, 4] propose a new approach for utilizing the unused spectrum. The new scheme merges three techniques for accessing the spectrum as one combined system. The new combined scheme utilizes the spectrum in an efficient way in the cognitive network. Simulation results show the ability of the new scheme to serve extra traffic. In [19], transmission opportunity-based spectrum access control protocol is proposed with the aim to improve spectrum access fairness and to ensure safe coexistence of multiple heterogeneous unlicensed radio systems. In the proposed scheme, multiple radio systems coexist and they dynamically use the available free spectrum without interfering with PUs.

The concept of CR is proposed in [34] for large-scale wireless systems, which opportunistically utilize network resources including both spectrum bandwidth and radio availability. Free resources cannot be predetermined in large-scale wireless systems, due to various reasons such as interference and dynamic traffic load. The proposed CR not only establishes dynamic wireless networks, but also provides reliable network QoS. A MAC-layer QoS provisioning protocol is proposed for CR in [22]. The proposed protocol combines adaptive modulation and coding with dynamic spectrum access.

A novel spectrum trading system is proposed in [33] and a theoretical study on the optimal session based spectrum trading problem is presented under multiple cross-layer constraints in multi-hop CRs. A general secondary spectrum trading framework is presented in [25]. Using this framework a PU can sell access to its unused or under-utilized spectrum resources in the form of certain fine-grained spectrum-space–time unit. PUs lease free spectrum for SUs with QoS guarantees in [3, 4]. Free spectrum is used to establish the links of secondary network. The Markov decision process is used to derive the spectrum adaption scheme. Free spectrum is used to establish the links of secondary network for SUs. Generally, the leased spectrum for SUs influences the QoS for the PU and the gained rewards. The main concern of the proposed spectrum sharing scheme in [1] is maximizing a PU's reward and maintaining QoS for the PUs and for the different classes of SUs. Authors propose cooperative scheme for spectrum sharing among PUs in [2]. In this scheme, PUs exchange channels dynamically based on the availability of neighbor's idle channels. Simulation results show the ability of this cooperative scheme to maximize the profit of PUs and utilize the spectrum efficiently. Authors in [5] design new dynamic auction where spectrum is periodically auctioned off to meet SUs demands over time. The proposed auction scheme determines the size of spectrum to be auctioned for each session.

Each PU attracts SUs to lease spectrum by setting a lower price than the other PUs in [26]. Game theory is used to analyze the price competition scenario and seek a Nash Equilibrium (NE). Novel matching-based multi-radio multi-channel spectrum trading (M^3-STEP) scheme is proposed in [40]. Authors employ conflict graphs to characterize the interference relationship among SUs with PUs. M^3 -STEP algorithm is suggested to maximize the revenue of PUs. Authors [18] study spectrum trading problem in a self-organized and two-tier heterogeneous cellular network. The problem is formulated as a Stackelberg game. The main objectives of the proposed scheme are: maximizing the revenue of macro eNodeB (MeNB), affording minimum required bandwidth for each home eNodeB (HeNB), enhancing per femto-user throughput, and providing better quality of service for macro-users nearby each femto-cell. The designed discounting strategy is applied for the extra bandwidth request of HeNBs to encourage them in supporting nearby macro-users.

Authors propose new scheme for SUs in [28] to access the unused spectrum of PUs. SUs act autonomously and fast in order to detect vacant communication channels. Reinforcement learning scheme is proposed to determine the sensing order of the available channels employing two alternative update rules. Authors propose new scheme for spectrum trading

in [6]. Trading scheme allows PU's to be efficiently shared with the SUs in exchange for a monetary cost. The scheme is based on demand and supply economics wherein the highest bidder for spectrum is awarded with getting access to the offered spectrum. Authors consider a three-layered spectrum trading market consisting of the PU, service providers, and SUs in [14]. They jointly study the strategies of the three parties. PU determines the auction scheme and spectrum supplies to maximize its revenue. However, most of these studies focused only on optimizing the pricing policy without considering admission of worthy spectrum requests which could further increase the PU's profit. Moreover, they neglect the waiting time for SUs' requests.

3 Network Overview

In this section, we present our assumptions. As seen in Fig. 1, a wireless mesh network has several mesh routers (MRs) and each MR serves several mesh clients (MCs) that jointly form a *cluster*. The WMN structure consists of several clusters, and all uplink/downlink flows are directed from the MCs towards MRs. Each cluster can be imagined as a single WLAN system. In this secondary network, a MR plays the role of access point to serve the MCs. While MRs have fixed locations, MCs move and change their places arbitrarily.

Each router and client is equipped with a single IEEE 802.11b based transceiver. The spectrum is divided into non-overlapping channels which is the basic unit of allocation. We define a PU as a spectrum owner that leases free spectrum to MRs. PU offers K channels for the secondary network. MRs use this spectrum to serve MCs. The jth class of SUs is characterized by:

- Required number of channels.
- Request arrival rate λ_j
- Exponentially distributed service time with mean $1/\mu_j$.
- Price parameter p_m that SU of jth class pays for PU to lease a channel m.

The price parameter p_m is a control parameter specified by PU and it can be used to achieve different conflicting objectives for the PU. It can be used to maximize the PU's reward and to enforce fairness among different classes of users. In this work, the price parameter is used for maximizing the PUs' reward.

Each PU has one finite FIFO queue for SUs requests. PU receives spectrum requests from SUs (MRs) and serves them either using its own spectrum or using the borrowed one. The PU borrows spectrum for SUs if their requests are worthy. The signaling protocol which was suggested for spectrum borrowing in [1] is used in our scheme. Spectrum request is added to the queue if the available spectrum is insufficient to accommodate it and the PU fails to borrow spectrum from other PUs.

The request is served when the PU has sufficient bandwidth and SU accepts to pay for the spectrum. The request is rejected if the SU refuses the offered price of spectrum. Moreover, if the queue is full the request is rejected. The network is assumed to consist of N PUs. In our model, we define the following components:

- Spectrum status pool S at PU i:
  $S = \{s_{i,m} | s_{i,m} \in \{0, 1\}\}$ is a binary matrix of spectrum status. If $s_{i,m} = 1$ then channel m is occupied by PU i.
- Interference constraint among PUs:
  let $I = \{l_{i,j} | l_{i,j} \in \{0, 1\}\}$ is $N \times N$ binary matrix that represents the interference among

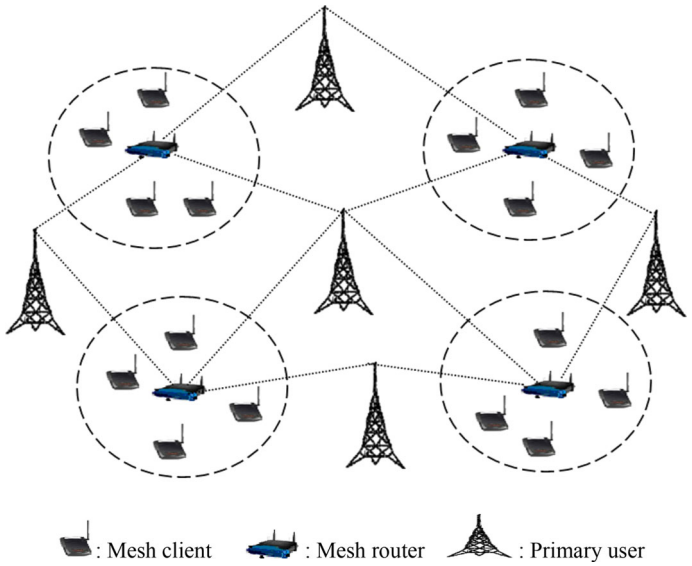

Fig. 1 Cognitive network architecture

PUs; if $l_{i,j} = 1$ then PUs i and j cannot use the same channel at the same time because they would interfere with each other.

- Channel reward R:
  let $R = \{r_{j,m}\}$ describes the reward that a PU gets successfully by leasing channel m to the SU of class j. Assume the price of leasing channel m is p_m per time slot, t is the holding time of channel m; the reward that a PU gets from leasing channel m to the SU of jth class is computed as follows:

$$r_{j,m} = tp_m \tag{1}$$

  Each PU specifies p_m to maximize its reward. Let T_i represent the current reward of PU i. T_i is computed as follows:

$$T_i = \sum_{\forall j \in J, \forall m \in K} r_{j,m} \cdot s_{i,m} \tag{2}$$

- Borrowable channel set B:
  our scheme allows two neighbors to exchange channels to maximize their reward while complying with the conflict constraint from the other neighbors. The set of channels that PU i can borrow from PU j are expressed as:

$$B(i,j) = L(j) \backslash \mathrm{U}_{w \in NG(i)} L(w) \tag{3}$$

  where $L(j)$ represent the list of channel of PU j, $NG(i)$ is a list of neighbors of a PU i. The symbols used in this paper are listed in Table 1.

Table 1 List of symbols used in this paper

Parameter	Symbol
Number of pus	N
Service time for SUs of the jth class	μ_j
Reward parameter for jth class of SUs	$r_{j,m}$
Number of channels per a PU	K
Request arrival rate for jth class of SUs	λ_j
Spectrum status pool	S
Interference matrix among pus	I
Reward vector	R
Price of leasing channel m	p_m
The total reward of PU i	T_i
Borrowable channel set	B
The list of channels for jth PU	$L\ (j)$
A list of neighbors for ith PU	$NG\ (i)$
State space	Z
The number of jth class requests	Z_j
The length of queue	l
The maximum queue size	Q
Action space	A
Spectrum size at state Z	$f(Z)$
Leasing policy	π
The reward of leasing channel m for the SU of class j	$r_{j,m}$
The mean value of reward	$\overline{R}$
The action that is taken at time t	a_t
The size of the leased spectrum	$\Delta f(Z)$
The cost of leasing spectrum	C
The cost of one spectrum unit	β
The average delay of SUs' requests	$\overline{D}$
The cost of delay for each second	α
The mean value of reward with penalty of delay	$\overline{R}_t$
The average length of the transition time	τ
The expected reward at state Z_j under policy π	$R\left(Z_j(t), \pi, T\right)$
The approximation of the average reward	$\overline{R}_m$
The value function at stat Z_j under policy π	$V\left(Z_j(t), \pi\right)$
The rate of reward at state Z	$q(z)$
The required accuracy	ϵ
The discount reward	γ
Likelihood of reward loss	L_r

4 Description of Reinforcement Learning Based Model

RL is a sub-area of machine learning that is concerned with the way a system administrator takes actions at different circumstances in work environment to maximize some notions of long-term reward [8, 9, 29, 36]. In order to maximize the reward, RL is used to extract an optimal control policy that maps states of the system to the actions that should be taken by

the manager. The work environment is typically represented as a finite-state Markov decision process (MDP). In this section, we introduce the elements of the RL model.

4.1 Basic Formulation of Markov Decision Process

For the basic formulation, we describe the elements of RL that include the objective function, events, and states of the system. In our model, we have an adaptable free spectrum size $f(Z)$ that is dynamically increased according to the queue length and the gained reward.

The state of the considered system can be described by a matrix $Z = \{Z_j\}$ where Z_j denote the number of jth class requests. For state Z, $f(Z)$ is the spectrum size that is used to serve the requests with a service rate $f(Z)\mu$. State transition takes place when a new request is arrived or a request is served. All possible states are limited by the following constraints:

- $\sum_{j \in J} Z_j < NK$.
- $l < Q$.

where l is length of queue that represents spectrum demand and Q is the maximum queue size. At each decision epoch, the system administrator has to take an action among all the possible actions. When a new request is arrived, the PU should choose one of the following actions:

- Add the request to the queue and adapt spectrum size by borrowing spectrum from neighbors.
- Serve the request using the borrowed spectrum.
- Add the request to the queue without borrowing spectrum.
- Serve the request using PU's spectrum without borrowing.

The set of the possible actions available to PU in state Z is denoted by A.

4.2 RL for Extracting the Optimal Policy

Optimal policy is a policy that gives the maximum reward when the PU adopts it. It specifies for each state the optimal spectrum size for each class. Basically, in our model the optimal policy is specified according to the average reward value obtained for each transition with the offered spectrum size.

For each state, the gained reward depends on the following three parameters: action reward, cost of spectrum, and cost of waiting time (delay). The PU borrows spectrum for a new request if it is worthy otherwise the request is queued.

The PU may decrease the offered spectrum (based on reward) when SU departs the system. Although reducing the size of the spectrum decreases customers' satisfaction (since their waiting time increases accordingly), with certain constraints, the PU takes the action that maximizes its reward. The main focus of this work is to study the long term average behavior of PU that evolves in time. In our work, spectrum trading problem is solved within the framework of the theory of continuous-time Markov Decision Processes (MDP). From MDP theory [36], our system is ergodic, the optimal policy is deterministic and can be found by applying policy iteration algorithm.

In our work, RL is used to extract the policy, $\pi : Z \rightarrow A$, for choosing next action a_t based on the current stat Z. For each new event in the system, PU senses current state Z and selects next action. The policy π specifies the set of actions that the PU can take. In our leasing system, the value $f(Z)\mu$ indicates the optimal service rate provided by the allocated

spectrum $f(Z)$ at state Z that maximizes the mean value of reward. The mean value of reward for policy π^* is defined as:

$$\overline{R} = \sum_{\forall j \in J, \forall m \in K} r_{j,m} * f(Z) * \mu_j \tag{4}$$

The PU borrows spectrum from its neighbors if its spectrum is inadequate to accommodate SUs requests and it is profitable to serve new SUs in terms of profit. Let $\Delta f(Z)$ denote the size of the borrowed spectrum. The mean value of reward for policy π^* after borrowing $\Delta f(Z)$ is computed as follows:

$$\overline{R} = \sum_{\forall j \in J, \forall m \in K} r_{j,m} * (f(Z) + \Delta f(Z)) * \mu_j \tag{5}$$

To satisfy its clients, the PU tries to minimize the average delay of requests $\overline{D}$. Our system can be modeled as an M\M\c\k queuing system. Hence, $\overline{D}$ is computed as in [17, 35]. The cost of leasing spectrum is denoted by C and it is computed as follows:

$$C = f(Z) \cdot \beta \tag{6}$$

where β is the cost of one spectrum unit. In general the delay and the reward are conflicting objectives, that is when the delay increases the reward also increases since PU can lease more spectrum for the clients in the queue. On the other hand, the likelihood of losing the reward increases for large values of $\overline{D}$ because more SUs may wish to leave the queue. For the objective function, we select a linear combination of these objectives, which can be also interpreted as reward maximization with penalty of waiting time and spectrum cost. The objective function is expressed as follows:

$$\overline{R}_t = \overline{R} - \left(C + \alpha \overline{D}\right) \tag{7}$$

where α is the delay cost for each time slot which determines the trade-off value between the reward and average delay. The rate of reward at state Z is given by:

$$q(z) = \sum_{\forall j \in J, \forall m \in K} r_{j,m} * Z_j * \mu_j - \left(C + \alpha \overline{D}\right) \tag{8}$$

4.3 Using the Policy Iteration Algorithm for Extracting Optimal Policy

In our work, we apply policy iteration algorithm to extract the optimal policy for spectrum leasing. Since this algorithm is applied only for discrete Markov processes, we use a uniformization technique with certain average length of the transition time τ [36] to convert continuous-time Markov into discrete Markov chain where all states have identical sojourn time without losing the information of state sojourn time. Discrete Markov is easy to analyze and setup.

Policy iteration algorithm is a recursive method and it is used to calculate the expected reward at state until the calculated reward in two successive steps are close enough [8]. Let us define the expected reward, $R\left(Z_j(t_0), \pi, T\right)$, obtained in interval $(t_0, t_0 + T)$ of length T:

$$R\left(Z_j(0), \pi, T\right) = E\left[\int_{t_0}^{t_0+T} q\left(Z_j(0)\right) dt\right] \tag{9}$$

The mean value of the reward can be calculated mathematically as follows:

$$\overline{R}(\pi) = \frac{\lim_{T\to\infty} \sum_{t=1,\forall j\in J}^{T} R\big(Z_j(t),\pi,T\big)}{T} \tag{10}$$

The relative value is expressed as follows:

$$V\big(Z_j(0),\pi\big) = \lim_{T\to\infty} \big(R\big(Z_j(t),\pi,T\big) - R\big(Z_j(0),\pi,T\big)\big) \tag{11}$$

The relative value in state $Z_j(0)$ is defined as the difference in the future gained reward when starting at state $Z_j(0)$ compared to reference state $Z_j(t)$. To apply the iteration policy, we apply the following algorithm:

Policy Iteration Algorithm

INPUTS:

$\mathbf{V}(\boldsymbol{Z_j(t)},\boldsymbol{\pi})$ is the expected reward at state $\boldsymbol{Z_j}$ under policy $\boldsymbol{\pi}$.
$\mathbf{Z}$ is state space.
$\boldsymbol{\epsilon}$ is the required accuracy.
$\boldsymbol{\gamma}$ is the discount reward.

RETURNS:

$\boldsymbol{\pi}$: approximately optimal policy
V: value function

BEGIN
for all $\boldsymbol{Z_j} \in \boldsymbol{Z}$
$\boldsymbol{V\big(Z_j(0),\pi\big) = 0}$
$\boldsymbol{t = 1}$
Repeat

$$\boldsymbol{V(Z_j(t),\pi) = max_{a\in A} E\left\{a\, R\left(Z_j(t),\pi\right) + \gamma V(Z_j(t+1),\pi)\right)\right\}} \tag{12}$$

$$\boldsymbol{\overline{R}(\pi) = q(z) + \sum_{j\in J} \lambda_j\big[\, V(Z_j + \Delta_j(Z_j,\pi),\pi) \; - \; V(Z_j,\pi)\big] + \sum_{j\in J} Z_j * \mu_j\big[V(Z_j - \delta_j,\pi) - V(Z_j,\pi)\big]} \tag{13}$$

$$\boldsymbol{\acute{\pi} = arg\, max_{a\in A} E\left\{a\, R\left(Z_j(t),\pi\right) + \gamma V(Z_j(t+1),\pi)\right)\right\}} \tag{14}$$

$$\boldsymbol{U(t) = max_{Z_j\in Z}\,(\,V(Z_j(t),\pi) - V(Z_j(t-1),\pi))} \tag{15}$$

$$\boldsymbol{L(t) = min_{Z_j\in Z}(V(Z_j(t),\pi) - V(Z_j(t-1),\pi))} \tag{16}$$

$\boldsymbol{t = t + 1}$
until $\boldsymbol{0 \le U(t) - L(t) \le \epsilon L(t)}$
$\boldsymbol{\pi = \acute{\pi}}$
$\boldsymbol{V\big(Z_j,\pi\big) = V\big(Z_j(t),\pi\big)}$
end for
return $\boldsymbol{\pi, V}$
end

Here, $Z_j + \Delta_j\big(Z_j,\pi\big)$ denotes the state after accepting the jth class requests, recommended by policy π, in state Z_j. In cases where the queue is full, the decision is to reject new requests and this is defined by $\Delta_j\big(Z_j,\pi\big) = 0$. So, if new requests arrive, the state transition is described as $Z_j + \Delta_j\big(Z_j,\pi\big)$. In the case of requests departure the state transition is described as $Z_j \to Z_j - \delta_j$; where $Z_j - \delta_j$ denotes new state after the departure of jth class requests in state Z_j. The rates of the transitions are λ_j and $Z_j * \mu_j$, respectively.

Initially, the system starts at state $Z_j(t_0)$ where queue is empty and all spectrum is available. Upon request arrival, the extracted policy π selects an action a_0 to maximize rate reward $q(Z_j(t_0))$ in Eq. (8). As a result of taking action a_0, the state of the system transits to new state $Z_j(t_1)$. Then, policy π picks another action a_1. As a result of this action, the system transits to new state $Z_j(t_2)$, Policy π pick a_2, and so on and so forth. This process can be represented as follows:

$$Z_j(t_0) \xrightarrow{a_0} Z_j(t_1) \xrightarrow{a_1} Z_j(t_2) \xrightarrow{a_3} Z_j(t_3) \ldots$$

Upon taking these actions under policy π, we can define the reward as follows

$$R\big(Z_j(t_0), \pi\big) + R\big(Z_j(t_1), \pi\big) + R\big(Z_j(t_2), \pi\big) \ldots$$

The complexity of our algorithm is measured using the size of the search space $Z \times A$. It is clear that $|Z \times A| \leftarrow$ increases exponentially as $N \times K\psi$ or J increases, because $|Z| \leftarrow = 2^{N \times K}\psi$ and $|A| \leftarrow = 2^J\psi$ in the worst case. Note that the optimal policy can be extracted off-line before a PU starts spectrum leasing, so the extracted policy is stored in a table. Using the policies table, the PU can trade its spectrum in real time by looking up the table.

In order to determine how PU should determine the price parameter p_m for leasing channel m, we must specify how SUs in a spectrum market react to various values of price. The most common method of characterizing SUs behavior is by specifying a demand function that determines the quantity of demanded spectrum for each price offered by the PU. We assume that the demand function for spectrum is downward sloping where the quantity of the spectrum demanded decreases as the price for leasing spectrum increases. The demand function of spectrum in our model is defined as follows:

$$Q(p_m) = a p_m^{-\omega} \tag{17}$$

where $\psi a \psi$ is a market scaling parameter and ω is the price elasticity of spectrum demand. We assume both parameters are positive.

Definition 1 Price elasticity of spectrum demand (ω) is the percentage change in the quantity of demanded spectrum divided by the percentage change in the price for leasing spectrum.

The price elasticity of spectrum demand is negative since the demand function is downward-sloping. The quantity of spectrum demanded and price are inversely proportional. The price elasticity of spectrum demand (ω) is computed as follows [16]:

$$\omega = -\frac{\partial Q(p_m)}{\partial p_m} \frac{p_m}{Q(p_m)} \tag{18}$$

In order to maximize its reward, the PU has to select the optimal price for leasing channel m. The optimal price for leasing channel m is computed as follows:

$$p_m^* = \arg\max \big[p_m Q(p_m) - \big(C + \alpha \overline{D}\big)\big] \tag{19}$$

Theorem 1 *The price elasticity of spectrum demand (ω) for the optimal price p_m^* is computed as follows:*

$$\omega = \frac{p_m^*}{p_m^* Q(p_m^*) - (\acute{C} + \alpha \overline{D})} \tag{20}$$

5 Performance Evaluation

In this section, we present the simulation results that demonstrate the ability of our spectrum scheme to adapt to different network conditions. The system of PUs and SUs is implemented as a discrete event simulation. The simulation is written using MATLAB. We uniformly distribute 4 PUs and each PU is randomly assigned 20 channels. For the mesh network, 100 MCs are distributed uniformly in the transmission region of the MRs. The results are presented for several system settings scenarios in order to show the effect of changing some of the control parameters. Both analytical and experimental results are presented in this section to illustrate the performance of the proposed leasing scheme. The network parameters chosen for evaluating the algorithm and the methodology of the simulation are shown in Table 2.

Note that some of these parameters are varied according to the evaluation scenarios. The key performance measures of interest in the simulations are::

- Mean value of reward for the PU which is computed using Eq. (7). It is worth mentioning that simulation results for the mean value of reward are found to closely match the analytical results for the approximation of the average reward that is computed using Eq. (10).
- Likelihood of reward loss L_r for a PU i which computed as follows:

$$L_r = 1 - \frac{\overline{R}_t}{T_i} \tag{21}$$

- Average request delay which is the time a request waits in the queue until it can be served.

Table 2 Simulation parameters

Parameter	Value
Number of mesh routers	10
Number of clients	100
Number of primary users	4
Number of channels per a PU	20
Total number of channels	80
Number of messages per client	Random
Type of interface per node	802.11 b
MAC layer	IEEE 802.11 b
Transmission power	0.1 W
Packet size	512
λ_1 (arrival rate of SUs class 1)	1
λ_2 (arrival rate of SUs class 2)	1
λ_3 (arrival rate of SUs class 3)	1
λ_4 (arrival rate of SUs class 4)	1
Blocking probably constraint for a PU	0.015
α	0.4
β	4

5.1 The Impact of Delay Penalty on the Likehood of Reward Loss

Simulations are done to explore the effect of delay penalty on the likelihood of reward loss. Figure 2 shows the proposed leasing scheme performance (reward loss) as a function of the delay penalty weight α. We assume the maximum size of queue is $Q = 5$ and all SUs classes have the same arrival rates (spectrum demand). It is clear from the figure that the reward loss increases with the delay penalty weight α. The figure shows that the likelihood of losing the reward increases as the delay penalty increase, resulting in more reward loss. PUs compete with each other on the basis of the QoS their clients experience. PUs realize that time is money for the SUs.

The likelihood of reward loss increases as the traffic load increases in the secondary network. As time elapses, more requests arrive to the system. To reduce the risk of reward loss, the leasing scheme should allocate as much possible spectrum as it can for SUs to decrease their waiting time. Figure 3 shows the reported delay for different spectrum sizes. It is clear that the delay decreases as the number of offered spectrum increases. PUs may increase the size of the offered spectrum by borrowing spectrum from other PUs. High demand means consumers would simply line up to get service regardless of waiting time. The chance of losing the reward decreases as the demand increases.

5.2 The Likelihood of Reward Loss as a Function of the Spectrum Cost and Service Demand

The aim of this experiment is to simulate the behavior of our RL leasing scheme under different rates of service cost. Figure 4 shows the likelihood of reward loss for each service cost. It is clear that there is a direct correlation between the service cost and the likelihood of reward loss. PUs increase the price of service as the service cost increases. Clearly, SUs are not interested in leasing the spectrum for higher prices of service. For higher spectrum demand, SUs become more and more interested in the service and the likelihood of reward

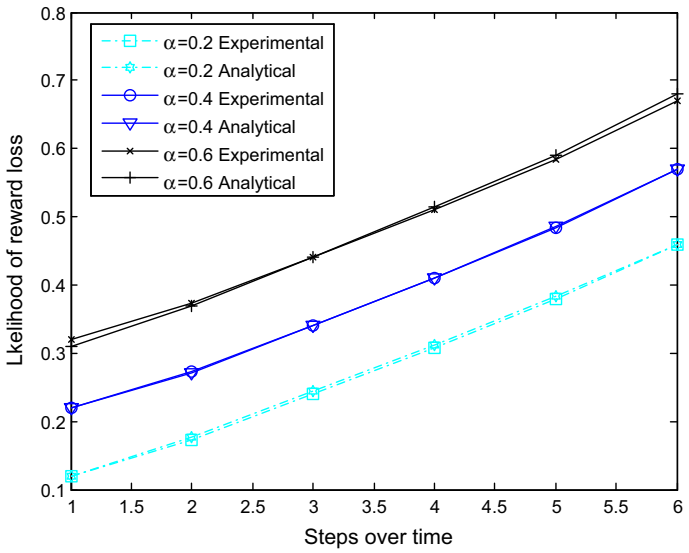

Fig. 2 Likelihood of reward loss for different delay penalties

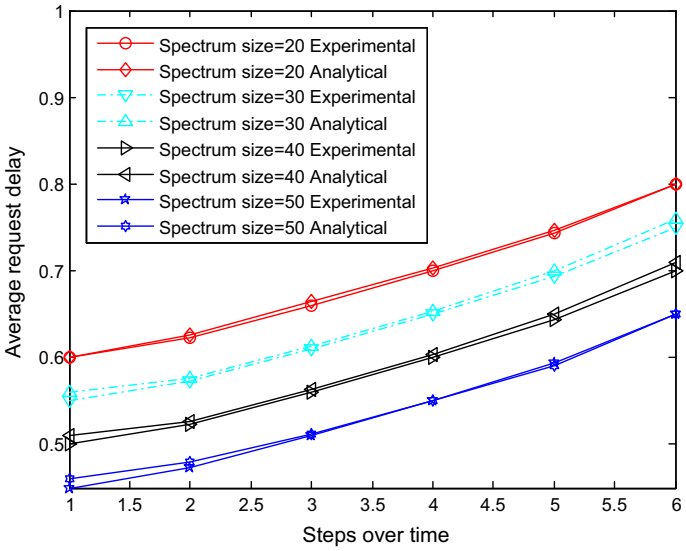

Fig. 3 Average request delay for different spectrum sizes

loss decreases significantly. We assume the average arrival and service rates are the same for all of the classes. Figure 5 shows the likelihood for reward loss for different spectrum demand. Clearly, the likelihood for losing reward decreases for higher spectrum demand. The PU generates more reward for high spectrum demand and this is clearly shown in Fig. 6.

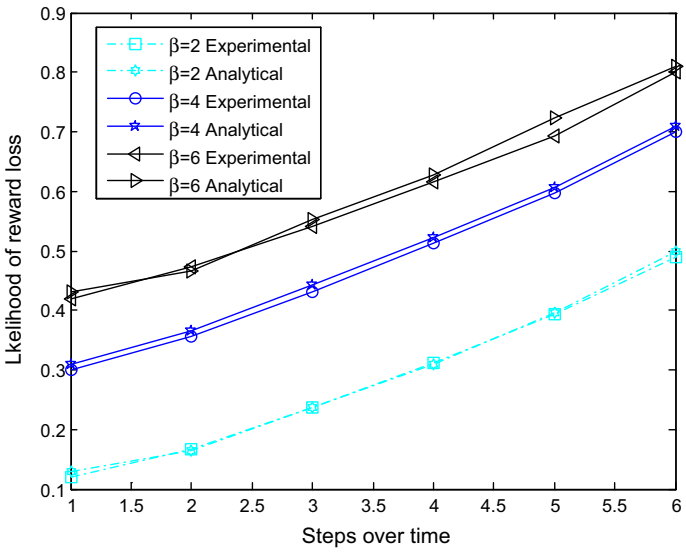

Fig. 4 The likelihood of reward loss for different spectrum cost

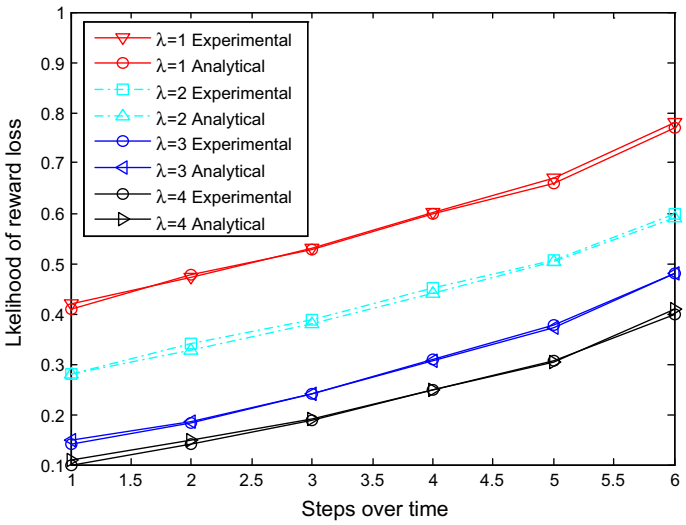

Fig. 5 The likelihood of reward loss for different spectrum demand

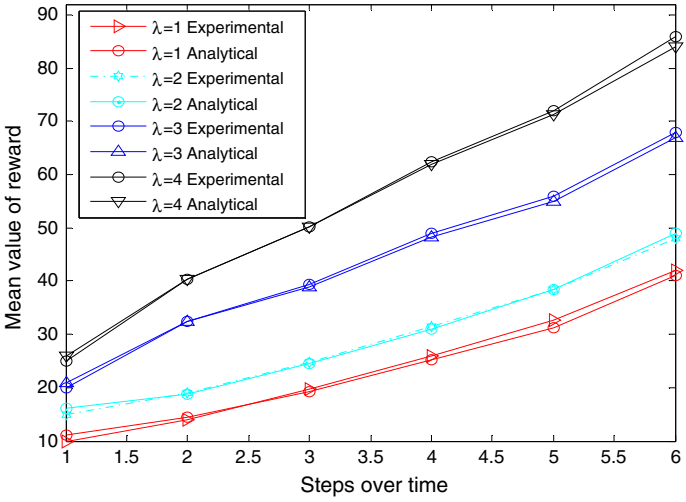

Fig. 6 The mean value of reward for PUs under different spectrum demand

Figure 7 shows the performance of RL scheme (likelihood of reward loss) as a function of the queue size. The results show that the chance of losing reward decreases as the queue size increases. However, increasing queue size leads to large queuing delay.

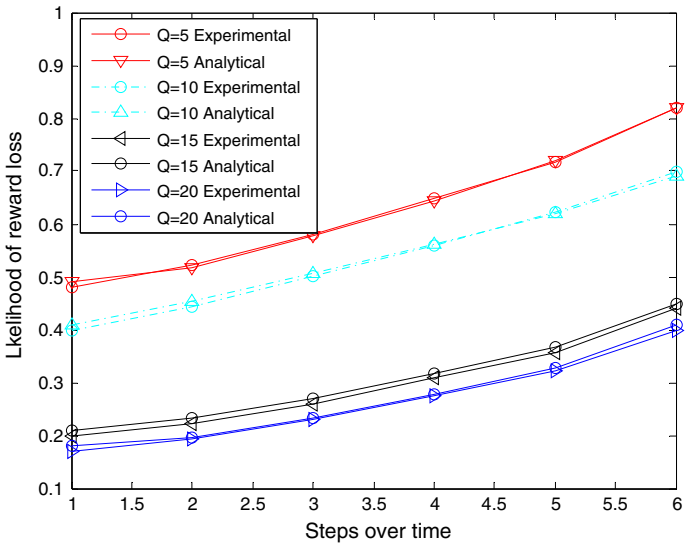

Fig. 7 The likelihood of reward loss for different queue sizes

6 Conclusion

In this paper, we formulated the spectrum leasing problem as a reward maximization problem with penalty for request delay and spectrum cost. In this formulation, each request class is characterized by its reward parameter defining the expected reward for serving a request from this class. Such a formulation allows for applying RL to solve the problem. We have presented the RL scheme to extract the optimal leasing policy for dynamic spectrum sharing in cognitive radio networks. We have considered an environment in which multiple SUs lease spare spectrum from PUs. There are two conflicting objectives to be satisfied: the first is how to select the requests that give the maximum reward; the second is how to reduce the likelihood of reward loss because of the QoS performance degradation (delay). This complex contradicting requirements is embedded in our RL model that is developed and implemented as shown in this paper. The numerical results show that our scheme is able to find an efficient trade-off between PU reward and average request delay.

The proposed model has two contributions for solving spectrum scarcity problem. From the application side, the main contribution is developing a spectrum sharing paradigm that considers different requirements such as reward for PUs, the leasing cost, and SUs requirements. All basic functions are integrated and optimized into one homogenous, theoretically based model. From the modeling side, we formulate a spectrum sharing problem as a reward maximization problem. Such a formulation allows RL to optimize the spectrum shortage problem. We are in the process of carrying similar analysis taking into account the competition among PUs of leasing the spectrum. We wish to derive the optimal solutions for PUs in an uncertain market. Furthermore, we wish to carry similar analysis on a real system.

References

1. Alsarhan, A., & Agarwal, A. (2011). Profit optimization in multi-service cognitive mesh network using machine learning,' EURASIP. *Journal of Wireless Communication and Networking, 36,* 1–14.
2. Alsarhan, A., & Agarwal, A. (2012). Optimizing spectrum trading in cognitive mesh network using machine learning. *Journal of Electrical and Computer Engineering, 2012*(1), 1–12.
3. Alsarhan, A., Al-Khasawneh, A., Itradat, A., & Bsoul, M. (2013). Economic model for routing and spectrum management in cognitive wireless mesh network. *International Journal of Networking and Virtual Organisations, 12*(4), 331–351.
4. Alsarhan, A., Agarwal, A., Obeidat, I., Bsoul, M., Al-Khasawneh, A., & Kilani, Y. (2013). Optimal spectrum utilisation in cognitive network using combined spectrum sharing approach: overlay, underlay and trading. *International Journal of Business Information Systems, 12*(4), 423–454.
5. Alsarhan, A., Quttoum, A., & Bsoul, M. (2015). Dynamic auction for revenue maximization in spectrum market. *Wireless Personal Communications, 83*(2), 1405–1423.
6. Bajaj, I., Lee, Y. H., & Gong, Y. (2015). A spectrum trading scheme for licensed user incentives. *IEEE Transactions on Communications, 63*(11), 4026–4036.
7. Bao, S., & Fujii, T. (2013). Learning-based p-persistent CSMA for secondary users of cognitive radio networks. *International Journal of Space-Based and Situated Computing, 3*(2), 102–112.
8. Barto, S. (1998). *Reinforcement learning: An introduction*. Cambridge: The MIT Press.
9. Bertsekas, D., & Tsitsiklis, J. (1997). *Neuro-dynamic programming*. Nashua: Athena Scientific.
10. Brik, V., Rozner, E. & Banerjee, S. (2005) DSAP: A protocol for coordinated spectrum access. In *Proceeding of IEEE symposium on new frontiers dynamic spectrum access networks (Dyspan 2005),* Maryland, USA, pp. 611–614.
11. Buddhikot, M. M., Kolody, P., Miller, S., Ryan, K. & Evans, J. (2005) DIMSUMNet: new directions in wireless networking using coordinated dynamic spectrum access. In *Proceedings of IEEE international symposium on world of wireless mobile and multimedia networks (WoWMoM 2005)*, Taormina, Italy, pp. 78–85.
12. Chieochan, S., & Hossain, E. (2013). Channel assignment for throughput optimization in multi-channel multi-radio wireless mesh networks using network coding. *IEEE Transactions on Mobile Computing, 1*(1), 118–135.
13. Cicconetti, C., Akyildiz, I. F., & Lenzini, L. (2009). FEBA: A bandwidth allocation algorithm for service differentiation in IEEE 802.16 mesh networks. *IEEE Transactions on Networking, 17*(3), 884–897.
14. Feng, X., Lin, P., & Zhang, Q. (2015). FlexAuc: Serving dynamic demands in a spectrum trading market with flexible auction. *IEEE Transactions on Wireless Communications, 14*(2), 821–830.
15. Foukalas, F., Karetsos, G., & Merakos, L. (2012). Cross-layer design in opportunistic spectrum access-based cognitive radio networks. *International Journal of Communication Networks and Distributed Systems, 8*(3/4), 230–246.
16. Gallego, G., & Ryzin, G. V. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science, 40,* 999–1020.
17. Gross, D., Shortle, J., Thompson, J., & Harris, C. (2008). *Fundamentals of queueing theory*. New York: Willey.
18. Haddadi, S., & Ghasemi, A. (2016). Pricing-based Stackelberg game for spectrum trading in self-organised heterogeneous networks. *IET Communications, 10*(11), 1374–1383.
19. He, J., Zhang, Y., Kaleshi, D., Munro, A., & McGeehan, J. (2008). Dynamic spectrum access in heterogeneous unlicensed wireless networks. *International Journal of Autonomous and Adaptive Communications Systems, 1*(1), 148–163.
20. Hossain, E., & Bhargava, V. K. (1997). *Cognitive wireless communication networks*. Berlin: Springer.
21. Hossain, E., Niyato, D., & Han, Z. (2009). *Dynamic spectrum access and management in cognitive radio networks*. Cambridge: Cambridge University Press.
22. How, K., Ma, M., & Qin, Y. (2012). Differentiated service provisioning in the MAC layer of cognitive radio mesh networks. *International Journal of Communication Networks and Distributed Systems, 8*(3/4), 213–229.
23. Huang, J., Berry, R., & Honig, M. L. (2006). Auction-based spectrum sharing. *ACM Mobile Networks and Applications, 11*(3), 405–418.
24. Ishibashi, B., Bouabdallah, N., & Boutaba, R. QoS (2008) Performance analysis of cognitive radio-based virtual wireless networks. In *Proceeding IEEE computer and communications (Infocom 2008)*, Phoenix, USA, pp. 336–340.

25. Jia, J., Zhang, Q., Zhang, Q., & Liu, M. (2009) Revenue generation for truthful spectrum auction in dynamic spectrum access. In *Proceeding of ACM international symposium on mobile Ad Hoc networking and computing (MobiHoc 2009),* New Orleans, USA, pp. 3–12.
26. Kasbekar, G. S., & Sarkar, S. (2016). Spectrum white space trade in cognitive radio networks. *IEEE Transactions on Automatic Control, 61*(3), 585–600.
27. Kloeck, C., Jaekel, H., & Jondral, F. K. (2005) Dynamic and local combined pricing, allocation and billing system with cognitive radios. In *Proceeding of IEEE symposium on new frontiers dynamic spectrum access networks (Dyspan 2005),* Maryland, USA, pp. 73–81.
28. Kordali, A. V., & Cottis, P. G. (2016). A reinforcement-learning based cognitive scheme for opportunistic spectrum access. *Wireless Personal Communications: An International Journal, 86*(2), 751–769.
29. Mitchell, T. (1997). *Machine learning networks*. New York: McGraw-Hill.
30. Mutlu, H., Alanyali, M., & Starobinski, D. (2008) Spot pricing of secondary spectrum usage in wireless cellular networks. In *Proceeding of IEEE Computer and Communications (Infocom 2008)*, Phoenix, USA, pp. 1355–1363.
31. Niyato, D., & Hossain, E. (2008). Spectrum trading in cognitive radio networks: A market-equilibrium-based approach. *IEEE Wireless Communications, 15*(6), 71–80.
32. Niyato, D., Hossain, E., & Han, Z. (2009). Dynamics of multiple-seller and multiple-buyer spectrum trading in cognitive radio networks: A game-theoretic modeling approach. *IEEE Transaction on Mobile Computing, 8*(8), 1009–1022.
33. Pan, M., Li, P., Song, Y., Fang Y., & Lin, P. (2012). Spectrum clouds: A session based spectrum trading system for multi-hop cognitive radio networks. In *Proceeding of IEEE computer and communications (Infocom 2012)*, Orlando, USA, pp. 1557–1565.
34. Song, L., & Hatzinakos, D. (2009). Cognitive networking of large scale wireless systems. *International Journal of Communication Networks and Distributed Systems, 2*(4), 452–475.
35. Thiagarajan, M., & Srinivasan, A. (2011). M/M/c/K loss and delay interdependent queueing model with controllable arrival rates and no passing. *The Indian Journal of Statistics, 73*(2), 316–330.
36. Tijims, H. (1986). *Stochastic modeling and analysis: A computational approach*. New York: Willey.
37. Wang, F., & Cui, M. (2008) Spectrum sharing in cognitive radio networks. In *Proceeding of IEEE Computer and Communications (Infocom 2008)*, Phoenix, USA, pp. 1885–1893.
38. Wang, B., Ji, Z., & Liu, K. (2007) Primary-prioritized markov approach for dynamic spectrum access. In *Proceeding of IEEE symposium on new frontiers dynamic spectrum access networks (Dyspan 2007)*, Dublin, Ireland, pp. 507–515.
39. Wang, X., Li, Z., Xu, P., Xu, Y., Gao, X., & Chen, H. (2010). Spectrum sharing in cognitive radio networks—an auction-based approach. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics, 40*(3), 587–596.
40. Wang, J., Ding, W., Guo, Y., Zhang, C., Pan, M., & Song, J. (2016). M^3-STEP: Matching-based multi-radio multi-channel spectrum trading with evolving preferences. *IEEE Journal on Selected Areas in Communications, 34*(11), 3014–3024.
41. Yu, H., Gao, L., Wang, Z., & Hossain, E. (2010). Pricing for uplink power control in cognitive radio networks. *IEEE Transactions on Vehicular Technology, 59*(4), 1769–1778.
42. Zhang, L., & Zheng, G. (2010). Adaptive QoS-aware channel access scheme for cognitive radio networks. *International Journal of Ad Hoc and Ubiquitous Computing, 6*(3), 172–182.
43. Zheng, H., & Cao, L. (2005) Device-centric spectrum management. In *Proceeding of IEEE symposium on new frontiers dynamic spectrum access networks (Dyspan 2005),* Maryland, USA, pp. 56–65.

Dr. Ayoub Alsarhan received his Ph.D. in Electrical and Computer Engineering from Concordia University, Canada in 2011, his MSc in Computer Science from Al al-Bayt University, Jordan in 2001, and his BE in Computer Science from the Yarmouk University, Jordan in 1997. He is currently an Assistant Professor at the Computer Information System at Hashemite University, Zarqa, Jordan. His research interests include cognitive network, parallel processing, machine learning, and real time multimedia communication over internet.

Ahmad N. Quttoum received his Ph.D. degree in Electrical and Computer Engineering from University of Quebec, Canada in 2011, his M.Sc. degree in Computer Engineering from University of Sunderland, United Kingdom in 2007, and his B.E. degree in Computer Engineering from Jordan University of Science & Technology, Jordan in 2006. He is currently an Assistant Professor at the Computer Engineering Department of the Hashemite University, Zarqa, Jordan. His research interests include Data Center Networks, Cloud Computing,Virtualized Networks, and Behavior-Based Network Security.

Dr. Yousef Kilani is currently an Associate Professor in the Department of Computer Information Systems of Hashemite University (HU), Jordan. He received his Bsc in Electrical Engineering from Jordan, his Master from the New York Institute of Technology, and his Ph.D. from George Washington University, USA. His research interests lays in networking. He has experience in networking reliability. Prior to joining the Hashemite University of Jordan, he held several key positions with major international ICT and IS consultancy and solutions firms.