IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 29, 2019, accepted April 17, 2019, date of publication April 24, 2019, date of current version May 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2912996

Routing Selection With Reinforcement Learning for Energy Harvesting Multi-Hop CRN

XIAOLI HE[1,2], HONG JIANG[1], YU SONG[1,3], CHUNLIN HE[4], AND HE XIAO[1]

[1]School of Information Engineering, South West University of Science and Technology, Mianyang 621010, China
[2]School of Computer Science, Sichuan University of Science and Engineering, Zigong 643000, China
[3]Department of Network Information Management Center, Sichuan University of Science and Engineering, Zigong 643000, China
[4]School of Computer Science, China West Normal University, Nanchong 637009, China

Corresponding authors: Xiaoli He (hexiaoli_suse@hotmail.com) and Hong Jiang (jianghong_swust@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771410, in part by the Postgraduate Innovation Fund Project through the Southwest University of Science and Technology under Grant 18ycx115, in part by the Artificial Intelligence Key Laboratory of Sichuan Province under Grant 2017RYY05 and Grant 2018RYJ03, and in part by the Horizontal Project under Grant HX2017134 and Grant HX2018264.

ABSTRACT This paper considers the routing problem in the communication process of an energy harvesting (EH) multi-hop cognitive radio network (CRN). The transmitter and the relay harvest energy from the environment use it exclusively for transmitting data. In a relay on the path, a limited data buffer is used to store the received data and forward it. We consider a real-world scenario where the EH node has only local causal knowledge, i.e., at any time, each EH node only has knowledge of its own EH process, channel state, and currently received data. An EH routing algorithm based on Q learning in reinforcement learning (RL) for multi-hop CRNs (EHR-QL) is proposed. Our goal is to find an optimal routing policy that can maximize throughput and minimize energy consumption. Through continuous intelligent selection under the partially observable Markov decision process (POMDP), we use the Q learning algorithm in RL with linear function approximation to obtain the optimal path. Compared with the basic Q learning routes, the EHR-QL is superior for longer distances and higher hop counts. The algorithm produces more EH, less energy consumption, and predictable residual energy. In particular, the time complexity of the EHR-QL is analyzed and its convergence is proved. In the simulation experiments, first, we verify the EHR-QL using six EH secondary users (EH-SUs) nodes. Second, the performance (i.e., network lifetime, residual energy, and average throughput) of the EHR-QL is evaluated under the influences of different the learning rates α and discount factors γ. Finally, the experimental results show that the EHR-QL obtains a higher throughput, a longer network lifetime, less latency, and lower energy consumption than the basic Q learning routing algorithms.

INDEX TERMS Routing selection, multi-hop CRN, energy harvesting, Q learning, reinforcement learning, MDP.

I. INTRODUCTION

A multi-hop cognitive radio network (multi-hop CRN) is a wireless network formed by a plurality of network nodes, with cognitive transceivers in a self-organizing manner. It is mainly used in military, medical, environmental monitoring and disaster relief application [1]. The multi-hop CRN nodes are usually battery-powered, but the battery capacity may often be limited. When the battery is exhausted, replacing it or recharging the device through a charging system is impractical and costly. Therefore, in order to extend the lifetime of the CRN, the concept of energy harvesting (EH) is proposed, which has attracted widespread attention in the industry [2]. EH is a technology that collects energy from environmental sources (e.g., solar energy, wind, seismic energy, thermal energy and radio frequency energy (RF)). Obviously, EH technology extends the network operating lifetime and is considered as a possible alternative to address energy-constrained wireless network bottlenecks.

The key metrics of the network layer (e.g., end-to-end throughput, delay, routing effectiveness, and stability) directly impact the quality of service (QoS) of secondary

The associate editor coordinating the review of this manuscript and approving it for publication was Shashi Poddar.

2169-3536 © 2019 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

user (SU) services. Meanwhile, since the spectrum of the multi-hop CRN is dynamic in time and space, the routing protocol needs to be highly dynamic, intelligent and robust [3] and [4]. Therefore, proper routing is the key to ensuring the utility of multi-hop CRNs.

Energy harvesting multi-hop CRNs (EH Multi-hop CRNs) constitute a research field with exploratory value. On the one hand, the deployment of EH nodes in cooperative communications has been envisioned as a promising approach to energy constrained networks in fifth generation (5G) mobile communications. On the other hand, multi-hop communication is also considered an ideal solution to address the contradiction between high-speed transmission and coverage. In EH multi-hop CRN communication, the EH transmitter communicates with the receiver through many intermediate EH nodes (i.e., EH relay nodes). The EH transmitter and each EH relay node harvest energy independently and use the harvested energy for data transmission. Therefore, the capability of transmitting data from the source EH-SU node to the destination EH-SU node depends on the EH node energy harvesting process and the selection policy of the optimal path.

A. RELATED WORKS

At present, some related studies have implemented a route policy in multi-hop CRNs [5]–[9]. Khalife *et al.* [7] proposed three separate routing scheme categories. The new routing solutions were used for the static and dynamic spectrum segments, while an opportunistic forwarding scheme without pre-established routes was proposed for the main frequency band. Syed *et al.* [8] presented a platform to test the accuracy of the solutions which is composed of three routing options, software radio peripherals and GNU radios. Meanwhile, reinforcement learning (RL) and spectrum leasing (SL) were used to design the routing. Li *et al.* [9] investigated a spectrum-aware virtual coordinate (SAViC) geographic routing scheme. Therefore, geographic routing, whether it can bypass the area or not, was affected by the licensed user or bypassed with a more available spectrum. According to the different spectrum occupancy modes of the PUs, two versions of SAViC were designed on the basis of the channel utility and search time of the primary user. In the energy constrained multi-hop CRN, energy is invaluable. Thus, the above studies mainly considered the balance between routing and energy, and work on energy persistence is still needed.

Hence, researchers [10]–[12] studied routing algorithms for EH multi-hop wireless networks. Unfortunately, these routing algorithms required each node to preserve the global state of the network, which consumed a large amount of router resources while not reflecting network state changes in time. Among these studies, [12] addressed joint power allocation and routing selection to minimize the probability of an outage in an EH multi-hop CRN. The Bellman-Ford algorithm and Dijkstra's algorithm were used to select the best route path. In the previous work [13], our research focused on channel allocation and power allocation. Therefore, the purpose of this paper is to study routing in EH multi-hop CRNs.

The research of [14] and [15] is the most relevant to our work. The authors of [14] addressed the characteristics of mobile Ad hoc networks (MANETs) (i.e., dynamic topology, lack of fixed infrastructure and limited energy for mobile devices) to study the bi-objective problem of delay and energy efficient routing. It was assumed that MANETs had an EH function in which the nodes had recharging capabilities while the remaining energy levels varied randomly with the passage of time. Therefore, in order to reduce the expected long term cost function, composed of end-to-end delay and path energy costs, a bi-objective intelligent routing protocol was proposed. Specifically, the routing problem was represented as a Markov decision process (MDP), which captured the link state dynamics caused by node mobility and the rechargeable energy state dynamics. As a consequence, an algorithm based on multi-agent RL was proposed to address the optimal routing policy without any preconditions. However, for the multi-hop CRN, the state of the network is dynamic. For a single route, it is necessary to consume a large amount of resources in order to gauge the entire network in a timely manner. Therefore, it is appropriate to describe the network as a model-free partially observable Markov decision process (POMDPs).

In [15], a decode-and-forward two-hop communication scenario with EH nodes was investigated. The goal was to find power allocation strategies that could maximize the throughput of the receiver by linear function approximation RL. Each point-to-point problem was modeled as a MDP and the linear function approximation RL algorithm was applied to enhance the learning of the SARSA algorithm. In addition, for the linear function approximation, a special feature function was studied in the data receiving process of the EH node. Although RL was utilized in the EH CRN, it did not solve the problem of long distance transmission (i.e., multi-hop) and routing selection.

We will use the model-free POMDPs to establish the network model for the characteristics of the EH multi-hop CRN (e.g., dynamic topology, EH randomness and intermission). Then, Q learning with RL will be used to find a better path.

Although the EH multi-hop CRN solves the problem of energy limitation, the harvested energy is still a scarce resource. Therefore, the next focus for research in wireless networks concerns how to reduce the network energy loss and improve the EH rate and energy utilization. Q learning is based on a predictive algorithm which is applied to enhance the context awareness ability for the EH multi-hop CRN. The routing algorithm based on Q learning can establish the predictive mechanism that can predict the harvested energy, consumed energy and residual energy of the neighboring nodes. According to the energy prediction of the neighboring nodes, the possible harvested energy and energy consumption can be calculated, to select the optimal and efficient path. Ignoring the surrounding environment, Q learning algorithm starts to accumulate experience and updates the status from its own learning. Finally, it learns and seeks the optimal solution based on the accumulated experience.

B. CONTRIBUTIONS

In this paper, we focus on the path selection problem of EH multi-hop communication, i.e., routing selection. Thus, we apply RL to find the routing policy. RL is a suitable tool to design routing strategies for EH multi-hop communication scenarios because it does not require a priori information about the EH process or the channel fading process. Each routing node can be regarded as an agent, and the RL concept is used to obtain the network information through the probability that each node forwards to its neighbor without knowing the network topology. The node learns where data will be sent to from routing decisions and network responses (e.g., throughput or rate achieved). Thus, the key contributions of our paper can be summarized as follows:

- First, compared with the previous EH multi-hop CRN routing algorithm, this work not only considers the node distance and hop count, but also considers communication energy consumption and residual energy consumption in the routing selection.
- Second, we show how to apply the RL algorithm to find the best routing policy in an EH multi-hop CRN scenario. In this scenario, the EH multi-hop CRN communication problem is modeled as a POMDP, and Q learning with RL is used to find the routing selection policies aimed at maximizing the transmission rate and minimizing energy consumption.
- Third, this paper presents an energy harvesting routing model based on the Q learning with RL (EHR-QL) for multi-hop CRN. On the one hand, the model does not increase the complexity of the route. On the other hand, network congestion can be detected and avoided based on the efficiency of node EH and residual energy.
- Then, we provide numerical results to evaluate the effectiveness of our proposed routing policy. The numerical simulation results show that the EHR-QL policy we considered is superior to other routing selection algorithms, e.g., [14] and [15].
- Finally, we use an actual network topology to verify the accuracy of the EHR-QL.

C. ORGANIZATION

The remainder of this paper is organized as follows. In Section II, the system model is introduced. The routing selection study for maximum transmission rate and minimum energy consumption in an EH multi-hop CRN scenario is formulated in Section III. Based on RL the routing selection algorithm EHR-QL is designed in Section IV. Numerical results are presented in Section V. Finally, Section VI concludes this paper.

II. NETWORK AND SYSTEM MODEL

A. NETWORKS MODEL

In this paper, we consider an EH multi-hop CRN, in which a PU transmits its information to the base station (BS) on the licensed spectrum and allows multiple SUs to share the available spectrum of the PU. At the same time, we assume that a PU is only allocated to one licensed channel to transmit data, and a licensed channel is only allocated to one PU, so there is no spectrum contention or transmission interference between PUs. More specifically, we assume that the PU and the SUs operate in the underlay spectrum sharing mode, i.e., the PU and the SU can simultaneously transmit data on the same licensed channel, provided that the interference of the SUs to the PU does not exceed the interference temperature (IT) threshold I_{th} [13].

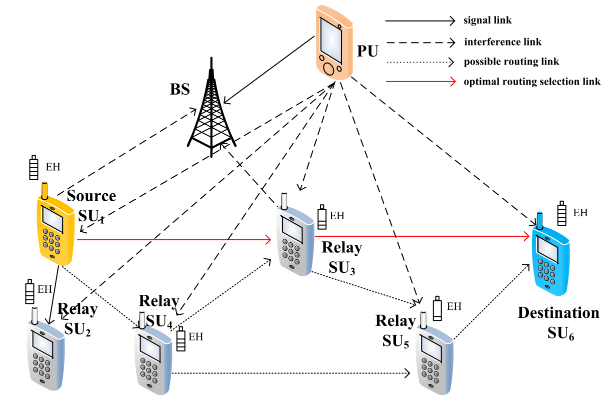


FIGURE 1. System model.

All SUs (i.e., source SU, relay SU and destination SU) have an EH function. When the transmission distance between the source SU and the destination SU exceeds a certain distance, one or more relay SUs may be used to forward the data packets to the destination SU. It is assumed that the relay SUs adopt the decode-and-forward (DF) scheme in our paper. The system model is shown in Figure 1. In our network topology diagram (see Figure 1), there are six EH-SU nodes, including four relay nodes. In addition, there are four possible routing options from the node SU$_1$ to the node SU$_6$.

The Euclidean distance can be used to represent the distance between two nodes. Then, the actual distance between node i and node j at time t is expressed as follows

$$D_{ij}^t = \sqrt{\left(x_i^t - x_j^t\right)^2 + \left(y_i^t - y_j^t\right)^2}, \quad \forall i, j \in \Phi_{SU} \tag{1}$$

$$|X| = \sqrt{(x_i^t)^2 + (y_i^t)^2} \tag{2}$$

where D_{ij}^t is the Euclidean distance between the node i and the node j. $|X|$ is the Euclidean distance from the node i to the origin, and $\Phi_{SU} = \{1, 2, \ldots, N\}$ is a set of EH-SUs. If $\left|D_{ij}^t\right| > D_{th}$, the relay node is used for multi-hop transmission, otherwise, it can be transferred directly. Therefore, in order to reduce delay due to excessive distance between nodes, there is a constraint

$$\left|D_{ij}^t\right| \le D_{th} \tag{3}$$

where D_{th} is the node transmission radius. i represents the current forwarding node, and j is any neighbor node of the

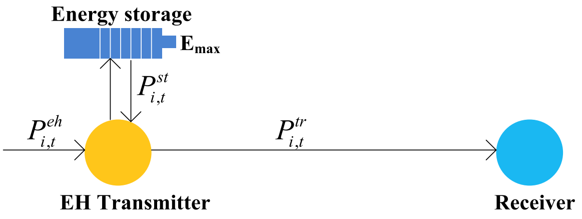


FIGURE 2. Energy harvesting model.

node i, i.e., $j \in J \in \Phi_{SU}$, $J = \{1, 2, \ldots, N-1\}$ is a set of one-hop neighbor nodes of the node i.

g_{ij} $(i, j \in \Phi_{SU} = \{1, 2, \ldots, N\})$ denotes the channel gain between the node SU_i and SU_j, which is expressed as

$$g_{ij} = D_{ij}^{-\zeta} \left| h_{ij} \right| \tag{4}$$

where $h_{ij} \sim CN(0, 1)$ is a random value generated according to the Rayleigh distribution, D_{ij} is the distance between SU_i and SU_j, and ζ is the path-loss exponent.

B. ENERGY HARVESTING MODEL

In this scenario, we assume that the EH model for EH-SU nodes follows an independent composite Poisson distribution [16]. At the same time, the EH-SUs also have many functions, such as data processing, network data packet forwarding, GPS positioning and routing selection. Because there is only one antenna in a node, we prefer the EH process in [17]. Furthermore, we assume that the battery does not leak, and that almost all of the harvested energy is stored. The EH model is shown in Figure 2. In each time slot, the node first performs EH, then stores the harvested energy, and determines if the energy is sufficient for the data transmission. If the energy is sufficient, the node will use the harvested energy to transmit its local data, otherwise it will keep EH (EH-save-judge-transmit-then EH). The variables $P_{i,t}^{eh}$, $P_{i,t}^{st}$ and $P_{i,t}^{tr}$ represent the harvested power, stored power and transmitted power of the i th EH-SU at time t, respectively.

In each time slot t, there are e energy packs that reach the EH-SUs, and the size of each energy pack is fixed at e^{fix}. The arrival time of e follows a Poisson distribution with a mean value of λ. This parameter λ also indicates the network load condition. For example, $\lambda = 0.5$ implies a low load, and $\lambda = 4$ implies a high load. Specifically, we set the entire transmission time to T, which is equally divided into h (hop count) time slots. The EH arrival times are $\{0, 1, \ldots, t, \ldots, T\}$, where $0 \le t \le T$, and $\sum_{t=0}^{T} t \le T$ is satisfied. Each EH-SU uses the harvested energy to transmit its local data. The harvested energy E_i^{eh} of the i th EH-SU at time T is expressed as

$$E_i^{eh} = \eta \int_0^T P_{i,t}^{eh} dt \tag{5}$$

where η is the energy harvesting conversion rate, $\eta \in [0, 1)$.

Similarly, the expression of the energy consumed by the transmission can be written as follows

$$E_i^{tr} = \int_0^T P_{i,t}^{tr} dt \tag{6}$$

where $P_{i,t}^{eh}$ and $P_{i,t}^{tr}$ are the harvested power and transmission power of the i th EH-SU at time slot t, respectively.

$E_{i,t}^{eh}$, $E_{i,t}^{tr}$ are used to represent the energy harvested and the energy used for transmission of the i th EH-SU at time slot t. After the node h hops, the residual energy is

$$E_{i,t}^{re} = E_{i,t}^{eh} + E_{i,t-1}^{re} - E_{i,t}^{tr} \tag{7}$$

We assume that the battery capacity is large enough and that there are no energy leaks. The harvested energy can be stored in the battery. This assumes that the battery has a very large capacity relative to the EH efficiency, which is reasonable in practical applications [14]. Therefore, we assume that the battery capacity is at most $E_{\max}$. We can write the relationship between $E_{i,t}^{re}$ and $E_{\max}$ as follows

$$0 \le E_{i,t}^{re} \le E_{\max}, \quad 0 \le t \le T \tag{8}$$

$$E_{i,t}^{re} = \min\left(E_{i,t}^{eh} + E_{i,t-1}^{re} - E_{i,t}^{tr}, E_{\max}\right), \quad 0 \le t \le T \tag{9}$$

Equation (7) represents the energy causality, and equation (8) is for the storage capacity. Let $E_{ij}^{co} = e^{\frac{hop(i,j)}{H}}$ represent the energy consumed by transmitting data packets from node i to the neighbor j within its one-hop coverage. $hop(i, j)$ represents the hop count of the next node to the destination node. H represents all hops from node i to the destination node. If the data packets are transmitted from source node 1 to destination node N, the relay node i needs multi-hop transmission, and the energy consumption can be expressed as

$$E_{all}^{co} = \sum\nolimits_{i=1, j \ne i}^{N-1} E_{ij}^{co} \tag{10}$$

For any data grouping, the optimization objective function is

$$\min E_{all}^{co} \tag{11}$$

That is, the energy consumption of each packet is minimized.

C. MARKOV DECISION PROCESS MODEL

RL is biologically inspiring, and it acquires knowledge by actively exploring its environment. It is a process of learning in "exploration-exploitation". For this reason, it is ideal for resolving distributed problems such as routing. When node i makes a routing decision, it only selects its neighbor node j as the next-hop node having the highest reward value. The self-learning mechanism will be triggered by a dynamic reward value and route policy.

We model the routing process as POMDPs. A finite MDP is made up of 5-tuple elements, $\langle S, A, P_{ss'}^a, \gamma, R \rangle$, where S is a finite set of all possible states that the agent (packet) might assume in the environment. In this paper, S is defined as a set of three sub-states, i.e., harvested energy amount, battery level, and all possible destination nodes. More specifically,

$s_{i,t} = \{E_{i,t}, B_{i,t}, N_{i,t}\}$ represents the state of the i th EH-SU at time slot t, where $E_{i,t}$ is the state of renewable energy, i.e., the amount of harvested energy, uniformly quantized into three levels (low, medium and high). $B_{i,t}$ is battery the energy level, also quantized into ten levels(less than10%, 10%-20%, 20%-30%..... 80%-90%, higher than 90%). In addition $N_{i,t}$ is the number of EH-SUs in the entire EH multi-hop CRN.

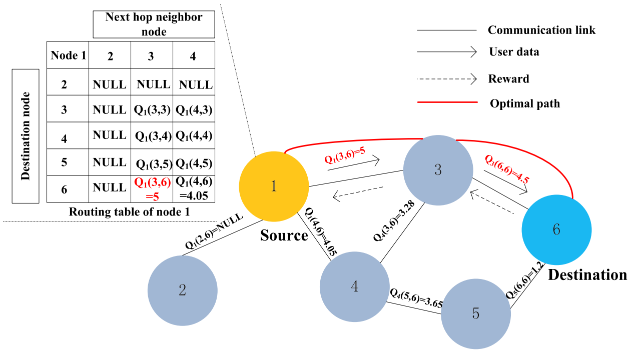
- In this paper, A is a set of all of the possible actions that all neighbor nodes may select as the next hop. $a_{i,t} \in A_i = J = \{1, 2, \ldots, N-1\}$ represents the action of the i th EH-SU at time slot t, and $J \in \Phi_{SU}$ is the number of the i th EH-SU's neighbor set. Each link in the routing may be associated with different types of action costs (e.g., queuing delay, available bandwidth, packet loss rate, energy loss level, etc.).
- $P^a_{ss'}$ is the transition probability that the system chooses the action $a \in A$ from current state $s \in S$ at time t to the next state $s' \in S$ at time $t+1$, where $P^a_{ss'} = \mathbb{P}(s_{t+1} = \{E_{i,t+1}, B_{i,t+1}, N_{i,t+1}\} = s' | s_t = \{E_{i,t}, B_{i,t}, N_{i,t}\} = s, a_t = a)$.
- $\gamma \in [0, 1]$ is the discount factor, which is used to weight the impact of future rewards on cumulative rewards. It means that the lower state is, the less it affects the reward. It is a decaying process.
- $R{:}S \times A \times S \rightarrow \mathbb{R}$ is written as $R^a_{ss'} = \mathbb{E}\left[r_{t+1} \left| s_t = s, a_t = a, s_{t+1} = s'\right.\right]$. It represents the reward function for each decision, where $R\left(s, a, s'\right)$ is the reward obtained when transiting from state $s \in S$ to the next state $s' \in S$ such that the action $a \in A$ is selected at state $s \in S$. During routing process, the reward R represents the maximum transmission rate and minimum energy consumption of the node in this state. R will vary depending on the distance between the SUs.

The routing scenario based on RL is shown in Figure 3. As seen from Figure 3, there are a total of six EH-SUs, and we will take node 1 as an example. The goal is to find the routing with the maximum transmission rate and minimum energy consumption from node 1 to node 6.Then, the state set S_1 is $S_1 = \{high; 50\% - 60\%; 1, 2, 3, 4, 5, 6\}$ and the action set A_1 is $a_{1,t} \in A_1 = \{2, 3, 4\}$. From the routing table of node 1, we can see that the value of $Q_1(3, 6) = 5$ is obtained by the algorithm, and we will discuss the maximum value later. Therefore, the routing selected from source node 1 to destination node 6 is $\{1 - 3 - 6\}$.

III. PROBLEM FORMULATION

A. REINFORCEMENT LEARNING APPROACH WITH ENERGY HARVESTING

In EH-multi-hop CRN routing environment, the entire network environment is taken as the learning object. Using a packet as an agent, each node can be treated as a state, which records its neighbor nodes as actions. Each Agent adopts a greedy ε policy. Each routing table is maintained by each node response, indicating that each node performs a possible routing policy responding to the Q-value of the next hop node. When the routing is performed, the routing table is traversed. According to the Q-value, the neighbor node with the highest Q-value is selected as the next hop node to establish an optimal path. One neighbor node is selected for data forwarding. Each node maintains a separate Q table containing all Q-values arriving at the target node. The goal of an intelligent routing optimization policy is to find a specific sequence of Q-values that maximizes the cumulative reward achieved in this sequence.



		Next hop neighbor node		
	Node 1	2	3	4
Destination node	2	NULL	NULL	NULL
	3	NULL	Q₁(3,3)	Q₁(4,3)
	4	NULL	Q₁(3,4)	Q₁(4,4)
	5	NULL	Q₁(3,5)	Q₁(4,5)
	6	NULL	Q₁(3,6) =5	Q₁(4,6) =4.05

Routing table of node 1

FIGURE 3. Routing selection scenario with RL.

Therefore, to make an optimal routing decision, a node will select the maximum Q-value across the destination column and return the neighbor's value which matches the max Q-value. To avoid the network throughput degradation caused by poor communication quality or a low transmission rate, we propose a routing rule that chooses the next hop node with higher data transmission rate as the next hop node [18]. This solution can solve the load balancing problem, improve the throughput performance of the network and enhance the link utilization of the IEEE 802.11 multi-hop network. When routing, our goal is to find the next hop node policy, which is the optimal EH policy for maximizing the average transmission rate R^{rate} of the EH transmitter within the deadline T. Then, according to Shannon theorem, the transmission rate (bits/s/Hz) of EH-SU is given by

$$R^{rate}_{i,t} = \log_2\left(1 + \frac{D^{-\zeta}_{ij}\left|h_{ij}\right| P^{tr}_{i,t}}{\underbrace{\sigma^2}_{noise\ power} + \underbrace{D^{-\zeta}_{iPU}\left|h_{iPU}\right| P^{tr}_{PU,t}}_{interferrence\ power\ of\ PU}}\right) \tag{12}$$

where $D^{-\varsigma}_{ij}$ and $D^{-\varsigma}_{iPU}$ represent the distance from node i to its neighbor node j and node i to the PU, respectively. $h_{ij} \sim CN(0, 1)$ and $h_{iPU} \sim CN(0, 1)$ are random values generated from the Rayleigh distribution. $P^{tr}_{i,t}$ and $P^{tr}_{PU,t}$ denote the power of the node i and the PU, respectively. σ^2 is the noise power, and its value is assumed to be the same for all users.

Our aim is to find the node with the highest data transmission rate (i.e., high throughput) among the neighbor nodes of the SU as the next hop. Combining equations (3), (8), (9), (11) and (12), the expression can be written as follows:

$$\begin{aligned} &\max_{i=1,2,\cdots N} R^{rate} = \max_{i=1,2,\cdots N} \sum\nolimits_{i=1}^{N} R^{rate}_{i,t} \\ &s.t.\ (3), (8), (9), (11) \end{aligned} \qquad (13)$$

Therefore, we can get the optimal node i as $i^* = \arg\max_{i=1,2,\cdots N} \sum_{i=1^N} R^{rate}_{i,t}$. Thus, the formulation of the routing policy is performed according to equation (13).

The routing algorithm can achieve load balancing of network nodes while maximizing the throughput rate and minimizing energy consumption. Therefore, when designing the reward function, the node states regarding the harvested energy, the residual energy and consumption energy are referenced. After the node obtains an instantaneous reward and forwards some data packets, it encodes the value of the residual energy, energy consumption, and the necessary Q-value to make its data packets obtain additional target rewards after reaching the target node. In this paper, the purpose of RL is to train nodes to select the neighbor node with largest reward value as the next hop node each time, and to thus select the best path. Thus, the setting of the reward value R is especially critical. We can obtain the instantaneous rewards $r_{i,t}$ as follows:

$$r_{i,t} = \begin{cases} 0 & B_{i,t} < B_{th} \quad or \quad D^t_{ij} > D_{th} \\ \beta E^{re}_{i,t} - \omega E^{co}_{ij} & B_{i,t} \geq B_{th} \quad and \quad D^t_{ij} \leq D_{th} \end{cases} \qquad (14)$$

where B_{th} is the battery level threshold. β and ω are weighting factors. If β is larger, the smaller ω is, the larger feedback value will be, and vice versa.

B. OPTIMIZATION OBJECTIVE

The proposed algorithm aims to obtain an optimal routing policy π. The node implements an action A in a certain state S, which can be represented by $\pi(s) = a$, so that an optimal reward value can be obtained. From this state $s_{i,t}$ at time t, the node i starts to perform the action $a_{i,t}$, and learns the state in the next round, and repeats the iteration until the goal of policy π is finally adopted. In summary, the cumulative reward value of the process is calculated by equation (14).

$$\begin{aligned} R^n\left(s_{i,t}\right) &= r_{i,t+1} + \gamma r_{i,t+2} + \gamma^2 r_{i,t+3} + \cdots \\ &= r_{i,t+1} + \gamma R^\pi\left(s_{i,t+2}\right) \\ &= \sum\nolimits_{k=0}^{T} \gamma^k r_{i,t+k+1} \end{aligned} \qquad (15)$$

The value of $r_{i,t+1}$ can be obtained by equation (12). There are three different scenarios for different discount factor γ:

$$\begin{cases} R^n\left(s_{i,t}\right) = r_{i,t+1} & \gamma = 0 & \text{scenario1} \\ R^n\left(s_{i,t}\right) = r_{i,t+1} + \gamma r_{i,t+2} & & \\ \quad + \gamma^2 r_{i,t+3} + \cdots & 0 < \gamma < 1 & \text{scenario2} \\ R^n\left(s_{i,t}\right) = r_{i,t+1} + r_{i,t+2} & & \\ \quad + r_{i,t+3} + \cdots & \gamma = 1 & \text{scenario3} \end{cases} \qquad (16)$$

Scenario1 indicates that only the reward value of the current action is considered. Scenario2 means considering both current and future rewards. Scenario3 shows that the current reward is the same as the future reward. In general, we choose scenario2, setting $\gamma = [0.1, 0.3, 0.5, 0.7, 0.9]$.

We maximize the value function by selecting the appropriate policy. If policy π is taken at a certain time t, then its value function in state S can be calculated by equation (17):

$$V^\pi(s) = \mathbb{E}_\pi\left[R^n_{i,t}\,|S_t = s\right] \qquad (17)$$

$V^\pi(s)$ is a state value function, which is defined as the expected reward that state s can obtain at time t.

$$V^\pi(s) = \sum\nolimits_{a\in A} \pi(s,a)\left[R^a_{ss'} + \gamma \sum\nolimits_{s'\in S} P^a_{ss'} V^\pi\left(s'\right)\right] \qquad (18)$$

The above value function has the form of a Bellman equation, that is, a dynamic programming (DP) equation, which can be solved in several ways, such as value iteration, policy iteration, Q-learning, SARSA, etc. We use Q learning to design our routing algorithm. Please refer to **Appendix A-A** for the specific derivation process of the Bellman formula.

$V^\pi(s)$ is the state value function, we also need to design the action value function, which can be used to make actions according to the size of the action value. Q learning is to learn the quality of each action in different states.

In addition to the state value function $V^\pi(s)$, we also need to design an action value function $Q^\pi(s,a)$, which can be used to select actions based on the magnitude of the action value. Q learning is used to learn the quality of each action in different states.

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[R^n_{i,t}\,|S_t = s, A_t = a\right] \qquad (19)$$

The action value function $Q^\pi(s,a)$ is an estimate of the reward value. A low Q-value does not mean that the reward value of this state is less than the reward value of another state. Therefore, the method of selecting the optimal action when the node is in state ^{s} must incorporate various factors, and the corresponding action policy cannot be adopted by relying only on the immediate reward value.

$$\begin{aligned} Q^\pi(s,a) = R^a_{ss'} + \gamma \sum\nolimits_{s'\in S} P^a_{ss'} \\ \times \sum\nolimits_{a'\in A} \pi\left(s',a'\right) Q^\pi\left(s',a'\right) \end{aligned} \qquad (20)$$

Equation (20) is also a Bellman equation. Please also refer to **Appendix A-B** for the specific derivation process of the Bellman formula.

When routing, EHR-QL yields the optimal routing policy compared with other policies. The reward value of the EHR-QL is the maximum value, which can be obtained by the following equation:

$$\begin{aligned} V^{\pi^*}(s) &= \max\left[R^a_{ss'} + \gamma \sum\nolimits_{s'\in S} P^a_{ss'} V^{\pi^*}\left(s'\right)\right] \\ &= R^a_{ss'} + \max_{a\in A}\gamma \sum\nolimits_{s'\in S} P^a_{ss'} V^{\pi^*}\left(s'\right) \geq V^\pi(s) \end{aligned} \qquad (21)$$

$$\begin{aligned}Q^{\pi^*}(s,a) &= \max\left[R^a_{ss'} + \gamma \sum{}'_s \in SP^a_{ss'} \max Q^{\pi^*}(s',a')\right]\\ &= R^a_{ss'} + \max{}_{a\in A}\gamma \sum{}_{s'\in S}P^a_{ss'} \max Q^{\pi^*}(s',a')\\ &\geq Q^{\pi}(s,a) \qquad (22)\end{aligned}$$

Through the optimal Bellman equations of equations (20) and (21), the optimal action policy π^* can be solved as follows:

$$\pi^* = \arg_\pi \max{}_{a\in A}\gamma \sum{}_{s'\in S}P^a_{ss'}V^{\pi^*}(s') \qquad (23)$$

$$\pi^* = \arg_\pi \max Q^*(s,a),\quad \left\{R^*(s,a) \geq R^{\pi_i}(s,a),\pi_i \in \pi\right\} \qquad (24)$$

In our proposed formulation, the goal is to calculate the routing policy π_i:$S_i \rightarrow A_i$, using each routing decision a_i for each node in each state s_i at time t. Eventually, the long term cumulative reward value, namely, the difference between the residual energy and the energy consumption of the nodes is maximized. This routing policy selects the route with the maximum average rate and minimum energy consumption. More specifically, each node's optimal relay selection policy is computed based on π^*.

IV. ALGORITHM DESIGN AND IMPLEMENTATION

A. RL FOR OPTIMAL ROUTING POLICY WITH Q LEARNING

The important idea of Q learning is to use an episode as a training period, which is from the initial state to the final state. After each episode is completed, the agent moves on to the next episode to learn. Therefore, it can be seen that the outer loop of Q learning is an episode, and the inner loop is every step of the episode.

The Q learning algorithm learns the state of the network based on the Q-value, which is the value of action A when the state is S, and these Q-values are used to determine the routing policies. The Q-value update rule is the core of the Q learning algorithm. The update equation is shown:

$$Q(s_{i,t},a_{i,t}) \leftarrow (1-\alpha)\underbrace{Q(s_{i,t},a_{i,t})}_{\substack{old\\ value}} + \underbrace{\alpha}_{\substack{learning\\ rate}} \overbrace{\left[r_{i,t} + \gamma\underbrace{\max{}_{a\in A}Q(s_{i,t+1},a_{i,t+1})}_{estimate\ of\ optimal\ future\ value}\right]}^{learned\ value} \qquad (25)$$

where $0 \leq \alpha \leq 1$ is the learning rate, it controls how much of the difference between the old Q-value and the new Q -value will be taken into account. In particular, when $\alpha = 1$, the two Q-values are offset and the update is exactly the same as the Bellman equation. Here, we specify that $\alpha = 0.7$. $r_{i,t}$ is the instantaneous reward, which can be obtained from equation (12).

The Q learning algorithm for RL has been widely used in network routing. The Q-routing algorithm, which is based on the Q learning model-free RL framework, is the most well-known among them [19]. The node makes a routing decision based on the estimated time to the destination node, that is, the neighbor node with the smallest Q-value is selected as the next hop node. While not exactly the same as Q-routing, our goal is to maximize the transfer rate and minimize energy consumption, so in our EHR-QL, the Q-value is designed as the value of the maximum reward value. The specific assumptions are as follows.

Each node x in the network represents its own view of the network state through its Q table. Let the Q-value in the Q table be represented by $Q_x(y,d)$, where d is the destination node and $y \in J$ is the neighbor node of node x. Specifically, the $Q_x(y,d)$ is the best estimate of the energy that the SU source node x reaches its destination d as it travels through its next hop neighbor node y, considering the harvested energy, the residual energy, and the consumed energy. The reward value is used as the Q-value, and the neighbor SU node with the largest Q-value is selected as the next hop node.

As shown in Figure 3, when the node 1 receives the packet destined for node 6, node 1 checks its Q table to select the neighbor node with the minimum value of $Q_1(3,6)$. However, these Q-values are not accurate. Therefore, routing decisions based on Q-values may not provide the best solution. These Q-values should be updated frequently for accurate routing decisions. Instead, the Q-value will be updated whenever a node sends a packet to its neighbor. The Q-value update equation is as follows:

$$\underbrace{Q_x(y,d)}_{new} = (1-\alpha)\underbrace{Q_x(y,d)}_{old} + \alpha\left[\underbrace{r_{i,t}}_{reward} + \gamma \max_{z\in neighbors\ of\ y} Q_y(z,d)\right] \qquad (26)$$

As seen from equation (26), when α is higher, the Q-value is more dependent on $r_{i,t}$ (i.e., the current knowledge). Conversely, when α is lower, the Q-value is more dependent on the old $Q_x(y,d)$ (i.e., previous knowledge).

B. ROUTING SELECTION FOR EHR-QL

First, in the Q learning phase, the SU source node sends a packet to the destination node at a specific time and initializes the Q-value to zero. The packet encapsulates information such as energy, hop count and distance. Second, each SU relay node obtains the learned information from its neighbor SU nodes, which includes the Q-value of the neighboring nodes, the number of hops, the harvested energy, the remaining energy, and the energy consumption. Then, the routing action is selected according to a Boltzmann probability distribution. Next, information such as the hop count, reward, and residual energy is stored in the model and updates the account value iteratively. The source node periodically forwards a learned message to its neighboring nodes and each node similarly forwards the received messages and updates the model. Finally, each neighbor node continuously sends a learned message to the next node, and each node continuously updates the internal model. Through continuous iteration, the evaluation value gradually converges. The details of the routing algorithm are shown in Algorithm 1.

C. ALGORITHM ANALYSIS

1) ALGORITHM CONVERGENCE ANALYSIS

Watkins proved the convergence of the Q learning algorithm under certain conditions [20]. As shown in this paper, as long as all actions are repeatedly sampled in all states and their action values are discretely represented, Q learning will converge to the optimal action value with a probability of 1. Therefore, we also prove the convergence of our proposed EHR-QL algorithm by formulaic derivation and experimental methods.

First, our network model is a POMDP, which has two characteristics, i.e., the instant return is bounded, and each action has return information when routing. Second, a formula is used to prove the convergence of the EHR-QL algorithm. See **Appendix B** for the specific certification process. Finally, through experimental simulation, it is proved that the convergence of EHR-QL is achieved in approximately 30 episodes. See Part C of Section v for the simulation and results analysis.

2) ALGORITHM TIME COMPLEXITY ANALYSIS

Computational complexity issues are significant in all research aspects of RL mechanisms. [14] proposed that the computational complexity of the algorithm requires updating the Q-value of $\forall s_i^t \in S_i$ and for each state. Therefore, its computational complexity was $O(|A_i|\,|S_i|)$. Because $i \in \{1, 2, \cdots N\}$, we can simplify the computational complexity to $O\left(N^2\right)$. However, the influences of the number of nodes and the distance on the time complexity were not considered. The EHR-QL we proposed is applicable to multi-hop CRNs. When calculating the time complexity, it is related not only to the action and state, but also to the number of EH-SU nodes and hops. The fewer the number of nodes is, the fewer hops there will be from the source node to the destination node, the learning time will be shorter, and the time complexity of the algorithm will be lower. Hence, the time complexity of the EHR-QL algorithm is $O\left(N^2\right)$, where N is the number of EH-SUs. See Part C of Section v for simulation and results analysis.

Algorithm 1 Routing Selection for EHR-QL

```
1:  **Initialize**: $t \leftarrow 0$, $s_{i,0} = \left\{E_{i,0}, B_{i,0}, N_{i,0}\right\}, T, \alpha, \gamma,$
    $\eta,\ \beta,\ \omega,\ \lambda,\ D_{ij}^0,\ P_{i,0}^{eh},\ P_{i,0}^{tr}$,$hop\,(i,j)$,$H,\ \varepsilon,\ D_{ij}^0,\ D_{ij}^0,$
    $Q\left(s_{i,0}, a_{i,0}\right) = 0, \forall s_{i,t} \in S, \forall a_{i,t} \in A$, SU number $N$
2:  **For** $i = 1$ to $N$ do
3:  Exchange causal knowledge and observe the current
    state $s_{i,t}$ based on $E_{i,t}, B_{i,t}, N_{i,t}$
4:  **For** $j = 1, j \neq i, j \leq N$ and $EH - SU_i$ is EH do
5:  **// each** $EH - SU_i$ **EH:**
6:  Calculate the harvested energy $E_{i,t}^{eh}$ using equation (5)
7:  Calculate the consumed energy $E_{i,t}^{tr}$ and $E_i^{co}$ using equation (6) and equation (10)
8:  Calculate the residual energy $E_{i,t}^{re}$ equation (9)
9:  **//RL:**
10: If current state $s_{i,t} \in S$
11: Calculate the corresponding reward $r_{i,t}$
12: Else go back step 1
13: End if
14: If $B_{i,t} \geq B_{th}$ and $D_{ij}^t \leq D_{th}$
15:     $r_{i,t} = \beta E_i^{re} - \omega E_{ij}^{co}$
16: Else $r_{i,t} = 0$
17: End if
18: Select an action $a_{i,t}$ using $\varepsilon$-greedy
19: $\pi_i\left(s_{i,t+1}, a_t\right), \forall a_{i,t} \in A$
20: Transmit the packet with the $P_{i,0}^{tr}$
21: Observe the current reward $r_{i,t}$ and the next state $s_{i,t+1}$
22: Update the Q-value
    $Q\left(s_{i,t}, a_{i,t}\right), s_{i,t} \in S, a_{i,t} \in A$ according to equation(25)
23: Update the policy $\pi_i\left(s_{i,t}\right), s_{i,t} \in S$
24: Update the reward value $r_{i,t}$ and share with the next hop
    neighbor
25: $t \leftarrow t + 1$
26: **End for**
27: **End for**
```

V. SIMULATION AND RESULTS ANALYSIS

In this section, we use numerical simulation to evaluate the performance of our intelligent algorithm EHR-QL. As shown in Figure 1, assuming the network coverage is 100 m × 100 m, where one BS, one PU and six EH-SUs are randomly distributed. The system is deployed in a Rayleigh fading environment and the channel state information (CSI) is perfect, i.e., the channel state estimation is equal to the actual channel gain value without an estimation error. Let the distance D_{ij} between the node i and the neighbor node j vary from 1 meter to 30 meters. Referring to the literature [8], [14], [15], the simulation parameters used in this paper are shown in TABLE 1, and all of the simulation models and algorithms are coded in the MATLAB 2015b.

To evaluate the performance of our proposed EHR-QL routing algorithm, an example and algorithm comparison analysis method is used. First, a path selection example with only six EH-SUs nodes is given. According to the Q learning mechanism in the RL algorithm, the node updates its own Q-value table and selects as the next hop node the node with the maximum Q-value table to select the path. Then, the main routing information, such as the end-to-end delay, throughput and jitter delay, are simulated and compared with other routing algorithms.

A. EXAMPLE OF EHR-QL WITH SIX EH-SUs NODES

In this section, we use an example to verify the accuracy of EHR-QL. The EH multi-hop CRN routing scheme based

TABLE 1. Simulation input parameters.

Parameters	Value
Spectrum bandwidth	35kHz
Number of EH-SU N	6 to 20
Noise power σ^2	-130 dBm/Hz
Interference temperature I_{th}	5 dBm
Learning rate α	0.1-0.9
Discount factor γ	0.1-0.9
Greedy ε	0.1-0.9
EH conversion rate η	0.7
Weighting factor β	0.8
Weighting factor ω	0.1
Battery capacity	100J
Battery threshold B_{th}	30% of the total battery capacity
Distance threshold D_{th}	30m
Path-loss exponent ζ	2
Harvested power $P_{i,t}^{eh}$	4W
Stored power $P_{i,t}^{st}$	2W
Transmitted power $P_{i,t}^{tr}$	10W
Number of iterations	300
Number of episodes	60
λ	0.5

on EHR-QL RL are shown in Figure 3, there are six EH-SUs: source node 1, intermediate nodes 2, 3, 4, 5 and destination node 6. When node 1 receives the packet for the destination node 6, node 1 checks its Q table (It is similar to the routing table, but the Q table has Q-values, which are related to the learned feedback value), and selects next hop node based on the maximum Q-value of the neighbor nodes. Note that this is not exactly same as Q routing, our Q-value is the reward value rather than the delay time. Therefore, the node with the maximum Q-value is selected as the next hop node. However, these Q-values are not accurate and may not provide the best solution. These Q-values should be updated in a frequently for accurate routing decisions. In other words, the Q-value is updated whenever a node sends a packet to its neighbor.

In Figure 3, there are four possible routes to destination node 6, i.e., $\{1-3-6\}$, $\{1-3-4-5-6\}$, $\{1-4-5-6\}$ and $\{1-4-3-6\}$. According to equation (24), after 300 iterations using MATLAB, the Q-value of node 1 can be calculated. The neighbor nodes of node 1 are node 2, node 3 and node 4. However, node 2 does not have a route to destination node 6, so the Q-value is null. $Q_1(3,6) = 5$ is larger than $Q_1(4,6) = 4.05$, so node 1 selects node 3 as the next hop node. The possible paths are now $\{1-3-6\}$ and $\{1-3-4-5-6\}$. Applying Dijkstra's algorithm, the longest path $\{1-3-4-5-6\}$ is excluded from the candidate routes. Therefore, the optimal route from node 1 to node 6 is $\{1-3-6\}$. Different routing results can be obtained quickly by changing the number of EH-SUs and modifying the source and destination nodes.

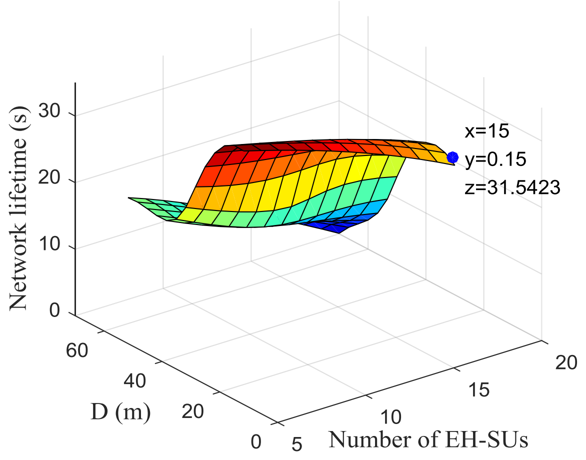


FIGURE 4. Impact of communication distance and number of EH-SUs on network lifetime.

B. ALGORITHM PERFORMANCE EVALUATION

In this section, we will simulate the effectiveness and characteristics of EHR-QL. When routing, there are many factors that affect the network lifetime, such as the network topology (number of nodes), transmission rate, transmission range (communication distance) and residual energy. As shown in Figure 4 (three-dimensional surface map), the network lifetime of an EH multi-hop CRN is related to the distances and the number of EH-SUs. It can be seen from Figure 4 that the network lifetime decreases gradually with the increase of the communication distance but increases with an increase in the number of nodes. We analyze the impact of the communication distance and the number of EH-SUs in the network lifetime. The specific analysis process is as follows:

- The loss of signal strength is related to the transmission distance. As shown in Figure 4, the network lifetime does not decrease sharply with distance but rather changes slowly. That is because the EH and multi-hop network for close-range communication is considered in a CRN. Therefore, its disadvantage that, the network lifetime will be shortened with an increase in communication distance will be compensated by EH and multi-hop. However, once the transmission distance from the source node to the destination node exceeds the distance threshold $D_{th} = 30$ meters, the transmission energy consumption will be very large. As a result, the network lifetime also changes rapidly. The simulation shows that the EHR-QL multi-hop routing algorithm still performs well in a large coverage area, which effectively expands the coverage area of the network and prolongs the network lifetime.
- Considering the complexity of the algorithm, the number of EH-SUs is set to 20. Since each SU has EH ability, the network lifetime is longer than that of an energy-limited network. Thus, the problem of dynamic topology can be solved better with a multi-hop CRN. Figure 4 illustrates how the lifetime of the entire network changes with the number of nodes. The blue dot in Figure 4 indicates that when the communication distance is 0.15 meters and the number of

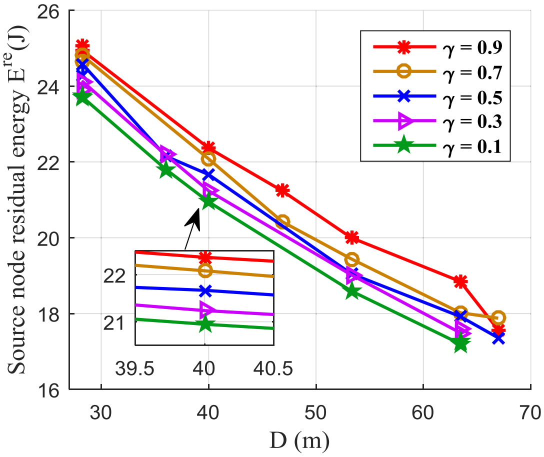


FIGURE 5. Residual energy value varies with distance for different discount factors where $\alpha = 0.7, \varepsilon = 0.3$.

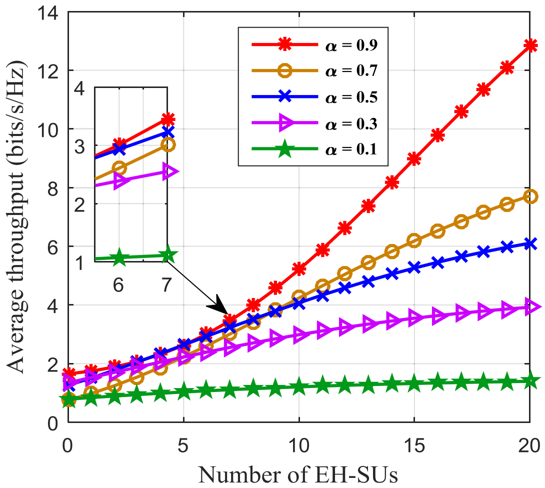


FIGURE 6. Average throughput varies with the number of EH-SUs for different learning rates where $\gamma = 0.9, \varepsilon = 0.5$.

nodes is 15, the maximum lifetime of EH multi-hop CRN is 31.5432 seconds.

Figure 5 illustrates that for different values of γ, the residual energy of the source node decreases as the communication distance increases, where $\alpha = 0.7, \varepsilon = 0.3$. Once the communication distance exceeds the threshold of 30 meters, the node energy consumption will increases. As a result, the residual energy also drops sharply. When the distance reaches 63.35 meters and $\gamma = 0.1$, the residual energy will be at least 17.21 J. At the same time, it can be seen that in the case of the same communication distance, a larger value of γ leads to more attention paid to the experience and the energy-saving routing path is more likely to be selected, and the residual energy will be larger.

Figure 6 provides the average throughput of the EH multi-hop CRN that is obtained through simulation. As the number of EH-SUs increases, the throughput increases and the routing energy consumption decreases. It can be explained that as the number of EH-SUs node increases, the energy harvested in the network increases, and when the source EH-SU node sends data to the destination EH-SU node, the selected routing probability increases. By constantly updating the return value, the optimal path is selected, and the network energy is balanced while the network throughput is improved. It can also be seen from equation (25), that the larger the learning rate α is, the less the training is before retention. As seen from Figure 6, EHR-QL can achieve the highest network average throughput when the learning factor α is 0.9. The higher the learning factor α is, the more balanced the energy utilization is.

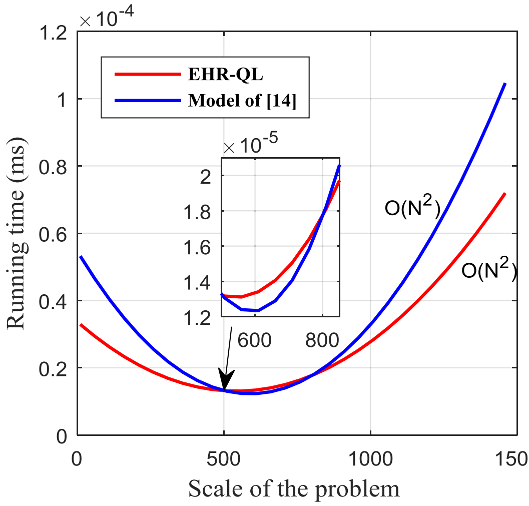


FIGURE 7. Comparison of algorithmic time complexity relative to the model of [14].

C. ALGORITHM PERFORMANCE COMPARISON

To evaluate our proposed algorithm, we analyze it relative to other similar algorithms. From the aspects of algorithmic time complexity, convergence, end-to-end delay, system throughput and jitter delay, we compare EHR-QL with algorithms in [14] and [15].

Figure 7 shows how the running time of the EHR-QL algorithm changes as the scale of the problem increases. It can be seen that by increasing the training time, the agent Q table is trained better, and the agent is more likely to find the optimal path of the target state. In this way, the algorithm running time is slowly reduced. When the problem scale reaches 500, the algorithm running time is slowly increasing with the number of EH-SUs. Compared with the model of literature [14], the EHR-QL running is slightly less, although the time complexity both is $O(N^2)$. The main reason for this is that in the state setting, we consider the harvesting energy, the battery residual energy and the number of EH-SUs, while the model of literature [14] only considers energy and distance.

Figure 8 depicts the number of iterations required for the three algorithms at different episodes. We can also clearly see the convergence speed of the three algorithms from the graph. When the convergence target value is 0.15, the convergence average iteration number is 10 and the episode

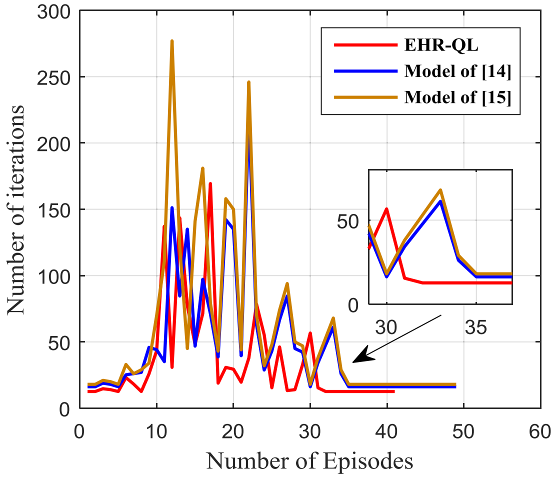


FIGURE 8. Comparison of algorithm convergence with models of [14] and [15].

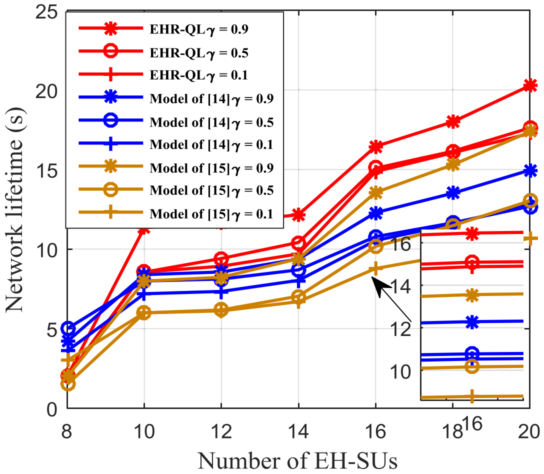


FIGURE 9. The network lifetime VS. number of EH-SUs with different discount factors.

is approximately 31.28, EHR-QL converges. The models of [14] and [15] need to continue iterating and training until the number of episodes is approximately 35.15. It can be concluded that the convergence speed of EHR-QL is the fastest (in other words, the iterations and episodes of EHR-QL are the lowest), followed by the models of [14] and [15].

In Figure 9, as the number of EH-SU nodes increases, the number of available routes also increases, as does the network lifetime. When the number of EH-SU is 20 and $\gamma = 0.9$, the network lifetime under the EHR-QL algorithm is the largest, and at a value of approximately 20.56 seconds. All three algorithms consider the energy persistence problem of the SU node, but the EHR-QL algorithm also considers the residual energy maximization problem, so its lifetime is longer than the other two algorithms. At the same time, we also noticed that when $\gamma = 0.9$, the lifetime of the model of [15] is higher than the model of [14]. The main reason is that in the model of [15] pairs of each point-to-point problem are modeled as a Markov decision process and the RL algorithm, SARSA, is combined with linear function approximation.

Figure 10 demonstrates the process of increasing the average throughput as the number of episodes changes. When the γ values are the same, the average throughput of EHR-QL is slightly higher than that of the other two algorithms. There are several reasons for this. First, the EHR-QL algorithm balances the load on the network. Second, it considers maximizing the throughput rate and minimizing energy consumption. Finally, when setting the reward function, EHR-QL also considers energy harvesting, residual energy and energy consumption.

In addition, γ also changes the average throughput. The larger the γ is, the larger the average throughput is. From Figure 9 and Figure 10 we can know that the γ value will affect the length of the network lifetime and the size of the average throughput. However, it is worth noting that the optimal path is constant regardless of the γ value.

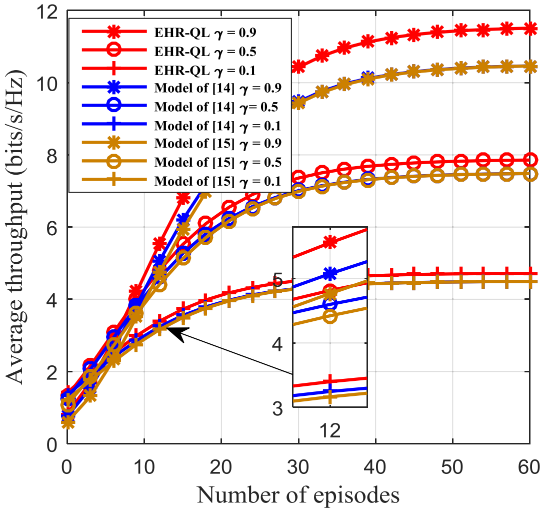


FIGURE 10. The average throughput VS. number of episodes with different discount factors.

In this section, we will describe the effect of the learning rate α on average throughput and end-to-end delay. α is an important parameter that affects the learning efficiency of Q learning. In the Q learning of RL, α is closely related to the dynamic level of the environment. Specifically, a higher (or lower) α value is required if the environmental dynamics are high (or low).

This is illustrated in Figure 11, where a higher α value indicates a higher performance enhancement. Higher throughput is also obtained compared to lower α values. At the same time, as the number of EH-SUs increases in network, the average throughput also increases. When the number of EH-SUs nodes is 20, the average throughput reaches a maximum of approximately 8.56 bits/s/Hz. From Figure 12, it can be seen that the average throughput obtained by EHR-QL is slightly higher than that of the other two algorithms. However, the average throughput of the model of [14] is very close to that model of [15].

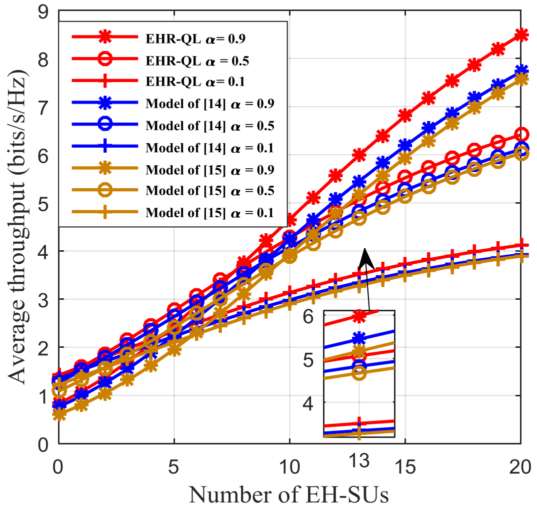


FIGURE 11. The average throughput VS. number of EH-SUs with different learning rates.

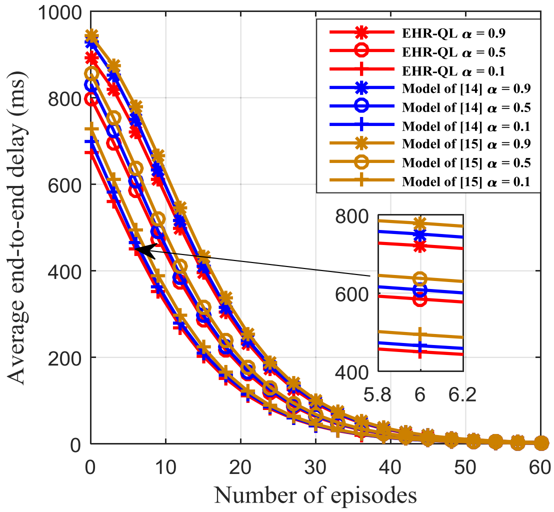


FIGURE 12. The average end-to-end delay VS. number of episodes with different learning rates.

In a multi-hop network, each node sends data through a relay in a routing protocol. Considering that the random scene test cannot accurately reflect the performance of the routing algorithm, multiple experiments with time delay are simulated according to the training times, and thus, the performance of EHR-QL is described more objectively.

Similarly, Figure 12 shows that as the learning ratios are changed from $\alpha = 0.1$ to $\alpha = 0.5$ and then to $\alpha = 0.9$, the average end-to-end delay decreases as the number of episodes increases. In the case of the learning rate $\alpha = 0.9$, episode [0, 10], the EHR-QL method provides the best network performance. The average end-to-end delay of EHR-QL is minimal. With the increase of episodes, when the episode is at [50, 60], the delays obtained by the three algorithms all reach the same value of 0. Through RL, the three algorithms select the optimal path for signal transmission through continuous learning. Therefore, the average end-to-end delay also decreases with the increase in trainings time.

VI. CONCLUSION

We have studied the routing problem in the EH multi-hop CRN communication scenario, where only the EH procedure is assumed at the transmitter and relay of the SU. Different from other researchers, we assumed that the battery does not leak, and considered the factors affecting routing, such as the distance of the node, the number of hops, the communication energy consumption and the residual energy consumption. The EH multi-hop CRN communication problem is modeled as a POMDP and the Q-learning RL algorithm is used to find a routing strategy aimed at maximizing the transmission rate and minimizing energy consumption. Combining energy harvesting and throughput maximization, we propose the EHR-QL algorithm. In addition, we provide an analysis and a proof of the time complexity and convergence for the proposed algorithm. The effectiveness of our proposed routing strategy is evaluated through experimental and numerical results. The numerical simulation results show that the EHR-QL performance is superior to other routing algorithms in terms of extending the network lifetime, saving residual energy, increasing the average throughput and decreasing the average end-to-end delay.

In the future, we intend to design cross-layer routing to optimize the performance of the physical layer, data link layer and network layer. Meanwhile, the experimental process of EH is more refined. The algorithm in this paper is applied to the further expansion of RL, i.e., multi-agent methods and deep reinforcement learning. At the same time, with the research findings of this paper, SU network performance can be improved in much more complex and more realistic scenarios, such as with power allocation, channel access, spectrum sensing and so on. This paper has important implications for the widespread use of CRNs and green communications. In addition, it also provides new research methods, research ideas and a research basis for the theory of CRN transmission technology.

APPENDIX A

A. DERIVATION OF BELLMAN EQUATION I

$$
\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\left[R^{n}_{i,t}\,|S_t = s\right] \\
V^{\pi}(s) &= \mathbb{E}_{\pi}\left[r_{i,t+1} + \gamma r_{i,t+2}\gamma^{2} r_{i,t+3} + \cdots |S_t = s\right] \\
&= \mathbb{E}_{\pi}\Big[r_{i,t+1} + \gamma\Big(r_{i,t+2} + \gamma r_{i,t+3} \\
&\qquad + \gamma^{2} r_{i,t+4} + \cdots\Big)|S_t = s\Big] \\
&= \mathbb{E}_{\pi}\left[r_{i,t+1} + \gamma R^{n}_{i,t+1}\left(s_{i,t+1}\right)|S_t = s\right] \\
&= \mathbb{E}_{\pi}\Bigg[\underbrace{r_{i,t+1}}_{immediate\ reward} \\
&\qquad + \underbrace{\gamma V^{\pi}\left(s_{i,t+1}\right)}_{discount\ value\ of\ the\ next\ state\ value\ function\ value}|S_t = s\Bigg] \\
&= \sum\nolimits_{a\in A}\pi(s,a)\left[R^{a}_{ss'} + \gamma\sum\nolimits_{s'\in S}P^{a}_{ss'}V^{\pi}\left(s'\right)\right] \quad (27)
\end{aligned}
$$

The value function can be divided into two parts, i.e., $r_{i,t+1}$ is the immediate reward, and $\gamma V^{\pi}\left(s_{i,t+1}\right)$ is the discount value of the next state value function value.

$$V^{\pi}(s) = \sum\nolimits_{a\in A} \pi(s,a)\left[R^{a}_{ss'} + \gamma \sum\nolimits_{s'\in S} P^{a}_{ss'} V^{\pi}\left(s'\right)\right] \tag{28}$$

Equation (2) represents the expected cumulative reward value obtained by the system according to equation (13) after implementing action policy π.

B. DERIVATION OF BELLMAN EQUATION II

$$\begin{aligned}
Q^{\pi}(s,a) &= \mathbb{E}_{\pi}\left[R^{n}_{i,t}\,|S_t = s, A_t = a\right] \\
Q^{\pi}(s,a) &= \mathbb{E}_{\pi}\left[r_{i,t+1} + \gamma r_{i,t+2}\right. \\
&\qquad \left. + \gamma^2 r_{i,t+3} + \cdots |S_t = s, A_t = a\right] \\
&= \mathbb{E}_{\pi}\left[r_{i,t+1} + \gamma\left(r_{i,t+2} + \gamma r_{i,t+3}\right.\right. \\
&\qquad \left.\left. + \gamma^2 r_{i,t+4} + \cdots\right) |S_t = s, A_t = a\right] \\
&= \mathbb{E}_{\pi}\left[r_{i,t+1}\right. \\
&\qquad \left. + \gamma \sum\nolimits_{k=0}^{\infty} \gamma^k r_{t+k+2}\,|S_t = s, A_t = a\right] \\
&= \sum\nolimits_{s'\in S} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma E_{\pi}\right. \\
&\qquad \left. \times\left(\sum\nolimits_{k=1}^{\infty} \gamma^k r_{i,t+k+2}\left|s_{t+1} = s'\right.\right)\right] \\
&= \sum\nolimits_{s'\in S} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma E_{\pi}\right. \\
&\qquad \left. \times\left(\sum\nolimits_{k=1}^{\infty} \gamma^k r_{i,t+k+2}\left|s_{t+1} = s', A_t = a'\right.\right)\right] \\
&= \sum\nolimits_{s'\in S} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma \sum\nolimits_{a'\in A} \pi\left(s',a'\right) Q^{\pi}\left(s',a'\right)\right] \\
&= R^{a}_{ss'} + \gamma \sum\nolimits_{s'\in S} P^{a}_{ss'} \sum\nolimits_{a'\in A} \pi\left(s',a'\right) Q^{\pi}\left(s',a'\right)
\end{aligned} \tag{29}$$

$$\begin{aligned}
V^{\pi}(s) &= \mathbb{E}_{\pi}\left[R^{n}_{i,t}\,|S_t = s\right] \\
&= \sum\nolimits_{a\in A} \mathbb{P}\left[A_t = a\,|S_t = s\right] \\
&\qquad * \mathbb{E}_{\pi}\left[R^{n}_{i,t}\,|S_t = s\ , A_t = a\right] && (30) \\
&= \pi(s,a)\, Q^{\pi}(s,a) \\
Q^{\pi}(s,a) &= R^{a}_{ss'} + \gamma \sum\nolimits_{s'\in S} P^{a}_{ss'} \sum\nolimits_{a'\in A} V^{\pi}\left(s',a'\right) && (31)
\end{aligned}$$

APPENDIX B
PROOF OF CONVERGENCE

Let $Q^*\left(s_{i,t}, a_{i,t}\right)$ denote the Q-value of the state $s_{i,t}$ after the $t+1\,th$ update, and the expression is as follows

$$\begin{aligned}
Q^*\left(s_{i,t}, a_{i,t}\right) &= (1-\alpha)\, Q\left(s_{i,t}, a_{i,t}\right) \\
&\qquad + \alpha\left[r_{i,t} + \gamma \max Q\left(s_{i,t+1}, a_{i,t+1}\right)\right]
\end{aligned} \tag{32}$$

Additionally, let ΔQ denote the maximum error of all of the entries in the Q-value table, as shown below

$$\Delta Q = \max\left|Q\left(s_{i,t+1}, a_{i,t+1}\right) - Q\left(s_{i,t}, a_{i,t}\right)\right| \tag{33}$$

where $Q\left(s_{i,t}, a_{i,t}\right)$ represents the value before the update.

$$\begin{aligned}
&Q^*\left(s_{i,t}, a_{i,t}\right) - Q\left(s_{i,t}, a_{i,t}\right) \\
&\quad = \max\Big|(1-\alpha)Q\left(s_{i,t}, a_{i,t}\right) \\
&\qquad + \alpha\left[r_{i,t} + \gamma \max Q\left(s_{i,t+1}, a_{i,t+1}\right)\right] \\
&\qquad - (1-\alpha)Q\left(s_{i,t}, a_{i,t}\right) \\
&\qquad + \alpha\left[r_{i,t} + \gamma \max Q\left(s_{i,t}, a_{i,t}\right)\right]\Big| \\
&\quad = \Big|\alpha\left[r_{i,t} + \gamma \max Q\left(s_{i,t+1}, a_{i,t+1}\right)\right] \\
&\qquad - \alpha\left[r_{i,t} + \gamma \max Q\left(s_{i,t}, a_{i,t}\right)\right]\Big| \\
&\quad = \left|\alpha\left[r \max Q\left(s_{i,t+1}, a_{i,t+1}\right) - \gamma \max Q\left(s_{i,t}, a_{i,t}\right)\right]\right| \\
&\quad = \left|\alpha\gamma\left[\max Q\left(s_{i,t+1}, a_{i,t+1}\right) - \max Q\left(s_{i,t}, a_{i,t}\right)\right]\right| \\
&\quad = \left|\alpha\gamma \max\left[Q\left(s_{i,t+1}, a_{i,t+1}\right) - Q\left(s_{i,t}, a_{i,t}\right)\right]\right| \\
&\quad = \alpha\gamma \max\left|\left[Q\left(s_{i,t+1}, a_{i,t+1}\right) - Q\left(s_{i,t}, a_{i,t}\right)\right]\right| \\
&\quad = \alpha\gamma\Delta Q
\end{aligned} \tag{34}$$

where $0 \leq \alpha \leq 1$ and $0 \leq \gamma \leq 1$ are bounded. Therefore, for any $s_{i,t}$ and $a_{i,t}$, the updated $Q^*\left(s_{i,t}, a_{i,t}\right)$ is at most $\alpha\gamma$ times the maximum error ΔQ in the Q-values table. After k stages, since each state and action is frequently accessed infinitely, the error is at most $(\alpha\gamma)^k\,\Delta Q$. The number of such intervals is infinite, so when $k \to \infty$, $(\alpha\gamma)^k\,\Delta Q \to 0$, and $Q\left(s_{i,t}, a_{i,t}\right)$ converges to $Q^*\left(s_{i,t}, a_{i,t}\right)$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their selfless reviews and valuable comments, which have improved the quality of our original manuscript.

REFERENCES

[1] *Spectrum Policy Task Force Report*, ET Docket 02-155, FCC, Nov. 2002.

[2] M. E. Haque and U. Baroudi, "Energy efficient routing scheme using leader election in ambient energy harvesting wireless ad-hoc and sensor networks," in *Proc. IEEE Sensors*, Busan, South Korea, Nov. 2015, pp. 1–4.

[3] A. S. Cacciapuoti, M. Caleffi, F. Marino, and L. Paura, "On the route priority for cognitive radio networks," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3103–3117, Sep. 2015.

[4] J. Wang, H. Yue, L. Hai, and Y. Fang, "Spectrum-aware anypath routing in multi-hop cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 4, pp. 1176–1187, Apr. 2017.

[5] Q. Chen, L. Wang, Y. Gao, R. Chai, and X. Huang, "Energy efficient constrained shortest path first-based joint resource allocation and route selection for multi-hop CRNs," *China Commun.*, vol. 14, no. 12, pp. 72–86, Dec. 2017.

[6] H. K. Boddapati, M. R. Bhatnagar, and S. Prakriya, "Ad-hoc relay selection protocols for multi-hop underlay cognitive radio networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[7] H. Khalife, N. Malouch, and S. Fdida, "Multihop cognitive radio networks: To route or not to route," *IEEE Netw.*, vol. 23, no. 4, pp. 20–25, Jul. 2009.

[8] A. R. Syed, K.-L. A. Yau, J. Qadir, H. Mohamad, N. Ramli, and S. L. Keoh, "Route selection for multi-hop cognitive radio networks using reinforcement learning: An experimental study," *IEEE Access*, vol. 4, pp. 6304–6324, 2016.

[9] D. Li, Z. Lin, M. Stoffers, and J. Gross, "Spectrum aware virtual coordinates assignment and routing in multihop cognitive networks," in *Proc. 14th IFIP Int. Conf. Netw.*, Toulouse, France, May 2015, pp. 20–22.

[10] K.-W. Chin, L. Wang, and S. Soh, "Joint routing and links scheduling in two-tier multi-hop RF-energy harvesting networks," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1864–1867, Sep. 2016.

[11] P.-D. Thanh, H. Vu-Van, and I. Koo, "Efficient channel selection and routing algorithm for multihop, multichannel cognitive radio networks with energy harvesting under jamming attacks," *Secur. Commun. Netw.*, vol. 2018, Mar. 2018, Art. no. 7543212. doi: 10.1155/2018/7543212.

[12] A. Banerjee, A. Paul, and S. P. Maity, "Joint power allocation and route selection for outage minimization in multihop cognitive radio networks with energy harvesting," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 1, pp. 82–92, Mar. 2018.

[13] X. He, H. Jiang, Y. Song, Y. Luo, and Q. Zhang, "Joint optimization of channel allocation and power control for cognitive radio networks with multiple constraints," *Wireless Netw.*, pp. 1–20, Jul. 2018. doi: 10.1007/s11276-018-1785-1.

[14] M. Maleki, V. Hakami, and M. Dehghan, "A model-based reinforcement learning algorithm for routing in energy harvesting mobile ad-hoc networks," *Wireless Pers. Commun.*, vol. 95, no. 3, pp. 3119–3139, Feb. 2017.

[15] A. Ortiz, H. Al-Shatri, X. Li, T. Weber, and A. Klein, "Reinforcement learning for energy harvesting decode-and-forward two-hop communications," *IEEE Trans. Green Commun. Netw.*, vol. 1, no. 3, pp. 309–319, Sep. 2017.

[16] X. Liu, M. Jia, Z. Na, W. Lu, and F. Li, "Multi-modal cooperative spectrum sensing based on Dempster-Shafer fusion in 5G-based cognitive radio," *IEEE Access*, vol. 6, pp. 199–208, 2017.

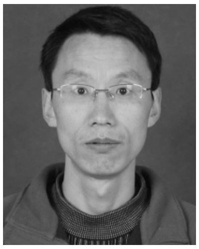
[17] K. Tutuncuoglu, A. Yener, and S. Ulukus, "Optimum policies for an energy harvesting transmitter under energy storage losses," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 467–481, Mar. 2015.

[18] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high-throughput path metric for multi-hop wireless routing," *Wireless Netw.*, vol. 11, no. 4, pp. 419–434, Jul. 2005.

[19] J. A. Boyan and M. L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," in *Proc. Int. Conf. Neural Inf. Process. Syst.* San Mateo, CA, USA: Morgan Kaufmann, 1993, pp. 671–678.

[20] C. J. C. H. Watkins and P. Dayan, "Technical note: Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

XIAOLI HE received the M.S. degree in computer application technology from the China West Normal University of China, in 2008. She is currently pursuing the Ph.D. degree with the South West University of Science and Technology. She was an Associate Professor with the Sichuan University of Science and Engineering. Her research interests include network communication and information systems, cognitive radio networks, and computer application technology. She has published several research papers in scholarly journals in the above research areas and has participated in several books.



HONG JIANG received the Ph.D. degree from the School of Communication and Information Engineering, University of Electronic Science and Technology of China. He is currently a Full Professor with the South West University of Science and Technology of China. His current interests include the cross layer Qos support in ad hoc networks and intelligent learning in cognitive radio networks.

YU SONG received the M.S. degree in computer application technology from the China West Normal University of China, in 2009. He is currently pursuing the Ph.D. degree with the South West University of Science and Technology. He is currently an Engineer with the Sichuan University of Science and Engineering. His research interests include computer application technology and computer network security.

CHUNLIN HE received the M.S. degree from the China West Normal University of China, in 2006. He is currently a Professor with the School of Computer Science, China West Normal University. His research interests include big data analysis and image processing.

HE XIAO received the M.S. degree in computer application technology from the University of Electronic Science and Technology of China, in 2008. He is currently pursuing the Ph.D. degree with the South West University of Science and Technology. His research interest includes cognitive radio networks.

• • •