

Learning for Robust Routing Based on Stochastic Game in Cognitive Radio Networks

Wenbo Wang *Student Member, IEEE*, Andres Kwasinski *Senior Member, IEEE*,
Dusit Niyato *Senior Member, IEEE*, and Zhu Han *Fellow, IEEE*

Abstract

This paper studies the problem of robust spectrum-aware routing in a multi-hop, multi-channel Cognitive Radio Network (CRN) with the presence of malicious nodes in the secondary network. The proposed routing scheme models the interaction among the Secondary Users (SUs) as a stochastic game. By allowing the backward propagation of the path utility information from the next-hop nodes, the stochastic routing game is decomposed into a series of stage games. The best-response policies are learned through the process of smooth fictitious play, which is guaranteed to converge without flooding of the information about the local utilities and behaviors. To address the problem of mixed insider attacks with both routing-toward-primary and sink-hole attacks, the trustworthiness of the neighbor nodes is evaluated through a multi-arm bandit process for each SU. The simulation results show that the proposed routing algorithm is able to enforce the cooperation of the malicious SUs and reduce the negative impact of the attacks on the routing selection process.

Index Terms

Cognitive radio networks, spectrum-aware routing, stochastic game, two timescale learning

I. INTRODUCTION

In Cognitive Radio Networks (CRNs), Dynamic Spectrum Access (DSA) policies require Secondary Users (SUs) to opportunistically access idle channels which are temporarily unused by the Primary Users (PUs). Although being considered an efficient way of spectrum utilization

Wenbo Wang and Andres Kwasinski are with the Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA (email: wxw4213@rit.edu, axkeec@rit.edu).

Dusit Niyato is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (email: dnyato@ntu.edu.sg).

Zhu Han is with the Department of Electrical and Computer Engineering as well as the Department of Computer Science, University of Houston, TX 77004 USA (email: zhan2@uh.edu).

[1], with DSA, SUs do not have channels which are always available to access. As a result, new coupling between the PHY-MAC layers and upper layer protocols arises. At the network layer, a routing protocol is thus expected to explicitly address the impact of unstable channels on the topology and the link performance of the secondary network. Generally, routing problems in CRNs exhibits a certain level of similarity to routing problems in multi-channel ad hoc networks [2]. However, routing in CRNs faces a number of new, different challenges [2], [3]:

- (i) spectrum awareness: timely adaptation to the dynamic change of the channel availability due to DSA, and
- (ii) self-organization: proper route configuration with the limited/heterogeneous level of channel resource knowledge.

Due to the information uncertainty or locality caused by DSA, a distributed route discovery process in CRNs also tends to be more vulnerable to the insider attacks than that in conventional ad-hoc networks. As indicated by [4], [5], an attacker in CRNs exploits the following characteristics of DSA schemes: (i) vulnerabilities due to information locality with respect to sensing and reporting of spectrum states, and (ii) the imperfect knowledge of SUs about the time-varying PU channels. Since the accuracy of the channel state information directly affects the performance of the DSA schemes in CRNs, most of the identified attacks in CRNs target at channel state information distortion for attacking the PHY-MAC layer protocols [6]–[8]. Similarly, it is now possible for routing attackers to bypass the network-layer vulnerabilities used by traditional routing attack schemes and only need to distort the channel detection/access information in the DSA mechanism to disrupt the routing process.

In this paper, we study the routing mechanism in a multi-hop, multi-channel CRN and address the challenges of spectrum awareness, information locality and routing security as a joint problem. We consider limited spectrum sensing abilities of each SU in a real-world situation. We also consider the presence of malicious SUs which can apply sophisticated attacks by combining different methods of attacks including Sink-Hole (SH) and Routing-toward-Primary-User (RPU). In order to tackle the routing-under-attack problem for multiple flows, we formulate the joint channel-relay selection process of the SUs as a stochastic game. We propose a distributed, adaptive channel-relay selection scheme for SUs to learn their routing strategies with only limited amount of information exchange. To defend the SH attack with information distortion, we model

the routing performance evaluation process as a Multi-Arm Bandit (MAB) problem and use the estimated arm-selection probability as an indicator of the neighbor trustworthiness. The proposed routing scheme is featured as a self-organized strategy-learning process in a series of single-state repeated games. Neither the a-priori channel activity model nor information flooding among the SUs is required for implementing the learning scheme.

II. RELATED WORK

1) *Routing in CRNs*: The solutions to the routing problems in CRNs are usually featured by a cross-layer design that directly integrates channel sensing and MAC operations into the routing protocols. These routing protocols may vary significantly due to different assumptions on the PU activity model and DSA mechanisms. Such variation is usually reflected by differences in the selection of link metrics and routing scheme types (e.g., reactive and proactive). With respect to the different channel occupation models (e.g., underlay vs. overlay/interweaving), the link metrics may be designed in different ways. For overlay/interweaving CRNs, many studies designed the routing mechanism based on a snapshot of the channel dynamics [9]–[13]. In these studies, delay-based link quality metrics were proposed based on the collision map for the SUs over the PU channels. For underlay CRNs, the link quality metric may be designed based on the link capacity as a function of the interference to the PUs [14]. For both groups of solutions, routing schemes were usually designed in a time-slotted manner to analyze and optimize the impact of DSA mechanisms on the route performance. If complete information on the channel states and local routing decisions is assumed, the routing problem is usually formulated as an optimization programming problem (e.g., convex or integer programming) and solved with a centralized route scheduler [10]–[14].

In contrast to routing mechanisms using instantaneous collision maps, a number of works designed their link quality metric based on an a-priori probabilistic channel dynamic model [15]–[17]. Since the impact of DSA schemes on the link performance is reflected by the stochastic channel activity model, it is possible for the secondary network to treat the routing problem in CRNs as a routing problem in conventional ad-hoc networks. As a result, we can adopt existing protocols (e.g., link state routing [15], AODV [17] and RPL [18]) with little modification. The advantage of such an approach is that it provides a way of reflecting the channel dynamics in the probabilistic link metrics based on the stochastic channel activity model. Hence, the routing

protocols does not need to consider the instantaneous impact of the PHY-MAC layers. Since no collision map or route scheduler is needed, such an approach is more appropriate for designing a distributed routing mechanism. However, many of these distributed routing mechanisms only provide a heuristic routing solution. Also, it is often unrealistic to assume an a-priori channel activity model in practical scenarios and the applicability of deploying the aforementioned routing mechanisms may be limited.

In practice, the channel dynamics may exhibit heterogeneous characteristics with respect to the geolocation. In addition, SUs may have limited capability of acquiring information about the channel states and their neighbors' behaviors. Consequently, game theoretic analysis have become the focus of CRN routing protocol design, since it can efficiently solve the distributed control problems with constraints on the information exchange. Game-based routing solutions can be found in the studies on spectrum-aware, multi-flow routing [19], [20] and traffic engineering [21] in CRNs. In these studies, the SUs are assumed to be non-malicious and honest in sharing information, and the model of repeated (noncooperative) games is usually applied. The cooperation among the SUs is implicitly enforced through repeatedly playing the game and the performance of a route is ensured by the value of the game.

2) *Security Issues for Routing in CRNs*: In the literature, most of the studies on security problems in routing protocols target conventional ad-hoc networks [22], [23]. In these studies, the main focus is to prevent information distortion (e.g., with public-key distribution [24]) or to identify the attackers with limited traffic monitoring (e.g., [25], [26]). When game theoretic solutions are adopted, the interaction between the honest and malicious nodes is typically modeled as a constant/zero-sum game and solved by obtaining the minimax equilibrium strategies in the game (e.g., [27], [28]).

There are relatively few works on the secured routing protocol design in CRNs. Among them, most of the studies are confined to handling the jamming attacks or PUE attacks which distort the quality of an established link between the legitimated (normal) SUs (e.g., [29]). A more sophisticated routing attacks in CRNs recently identified is the Routing-toward-Primary-User (RPU) attack in multi-hop, overlay CRNs [30]. Unlike the PHY-MAC-layer dominated attacks, the RPU attack exploits the geographical heterogeneity of PU activities and tunnels the traffic to the SUs in the footprint of the PU transmission. An RPU attacker emulates a combined attacking mechanism of both the Sink-Hole (SH) attack [22] and the Selective-Forwarding (SF)

attack [22]. However, the RPU-caused packet drop/delay is not directly due to the network layer operation, but due to the collisions with PU transmissions on the PHY-MAC layers.

The paper is organized as follows. Section III describes the models of the PU activities and the SU behaviors. Based on these models, a spectrum-aware link quality metric is proposed to reflect the impact of the channel state dynamics on the routing process. In Section IV, the multi-flow routing process in the secondary network is formulated as a layered average-reward stochastic game and then is shown to be equivalent to a group of single-state repeated games. In Section IV-B and Section IV-C, an adaptive strategy-learning mechanism and a trustworthiness-evaluation mechanism are proposed for the normal SUs to seek the best-response routing strategies against the attackers with limited information exchange. The simulation results are provided in Section V to demonstrate the Effectiveness of the proposed routing mechanism. Section VI concludes the contribution of this paper.

III. NETWORK MODEL

We consider a multi-hop CRN that interweaves upon K orthogonal PU channels. The normal SUs abide by the interweaving DSA rule and establish links over the temporarily free PU channels. The nodes in the CRN are divided into three types: the source SUs, sink SUs and relay SUs. We consider that the relay SUs do not generate packets and only forward the received packets to their neighbors. Among the relay SUs, some malicious nodes adopt RPU-like attacks to cause delay to the traffic as much as possible.

A. Dynamic Spectrum Access Model

Based on the empirical study of the PU channel occupation time in [31], we assume that the PU activities over each channel can be modeled as an independent continuous-time Markov process with the binary states *Idle* ('0') and *Busy* ('1'). For a channel k , we assume that λ_k^{-1} and μ_k^{-1} are the mean holding times for states *Idle* and *Busy*, respectively. Then, the corresponding transition matrix is given by [31] as follows:

$$\mathbf{P}_k(t) = \frac{1}{\lambda_k + \mu_k} \begin{pmatrix} \mu_k + \lambda_k e^{-(\lambda_k + \mu_k)t} & \lambda_k - \lambda_k e^{-(\lambda_k + \mu_k)t} \\ \mu_k - \mu_k e^{-(\lambda_k + \mu_k)t} & \lambda_k + \mu_k e^{-(\lambda_k + \mu_k)t} \end{pmatrix}. \quad (1)$$

Since in practical scenarios the PU activities are usually geographically different, we assume that the CRN can be geographically divided into a set of non-overlapping, independent spectrum activity clusters according to the local PU activities. For conciseness, we consider a snapshot

of the network, during which the cluster topology remains unchanged. The cluster topology can be managed in a similar way to the CogMesh protocol [32] by trustworthy cluster heads. The cluster heads maintain the cluster formation through message exchange with neighbor nodes using a dedicated control channel.

We assume that SUs access the PU channels in a time slotted manner. Due to the practical limit on the number of radio interfaces in each SU, we assume that an SU can only sense one PU channel during one sensing slot. To reduce the detection error, SUs in the same cluster sense the PU channels following a round-robin schedule in an ascending order of the channel indices (Figure 1). The sensing results from the SUs in the same cluster are aggregated by the cluster head [1]. We assume that the detection error is negligible with aggregated sensing. For a cluster, the state of channel i , $i \in \mathcal{K} = \{0, \dots, K-1\}$, is updated only when $i = (n \bmod K)$ at slot n . For channels k , $k \neq i$, the SUs in the cluster keep the most recent sensing result at slot $\phi_k(n) = n - [(K+i-k) \bmod K]$ as their estimated state. For cluster q , let $\mathbf{o}^q(n) = [o_0^q(n), o_1^q(n), \dots, o_{K-1}^q(n)]^T$ denote the vector of estimated channel states at slot n , and $\mathbf{s}^q(n) = [s_0^q(n), s_1^q(n), \dots, s_{K-1}^q(n)]^T$ denote the real channel state vector. According to [33], the process $\mathbf{o}^j(n)$ is an irreducible, periodic, discrete-time Markov chain. Let $i' = (n+1) \bmod K$ be the channel sensed at slot $(n+1)$, then the transition probability of $\mathbf{o}^q(n)$ can be obtained based on (1) as:

$$P(\mathbf{o}_j^q(n+1) = s' | \mathbf{o}_j^q(n) = s) = \begin{cases} [\mathbf{P}_j^q(KT)]_{(s,s')}, & \text{if } j = i', \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where \mathbf{P}_j^q is the transition matrix of channel j in cluster q , T is the slot length and $[\mathbf{P}_j^q(KT)]_{(s,s')}$ is the element of \mathbf{P}_j^q transiting from s to s' .

B. Impact of Node Behavior on Link Quality

Let \mathcal{N}_i denote the set of one-hop neighbors of an SU i (including i). We assume that in slot n , SU i can freely choose its target relay SU in the neighborhood and target channel among the PU channels for data forwarding, if no constraint on SU behaviors is presented. We denote such an action by the action vector $a_i(n) = (j, k)$, where $j \in \mathcal{N}_i \setminus \{i\}$ and $k \in \mathcal{K}$. In a multi-hop CRN, it is natural to consider that the more hops used for packet forwarding, the larger total delay the path has. To enforce that packets are forwarded toward the sink SUs and no cyclic path is formed, each SU is able to exchange its geographical information with its neighbors. Using the geographical information of the neighbor SUs, we introduce the distance advancement metric

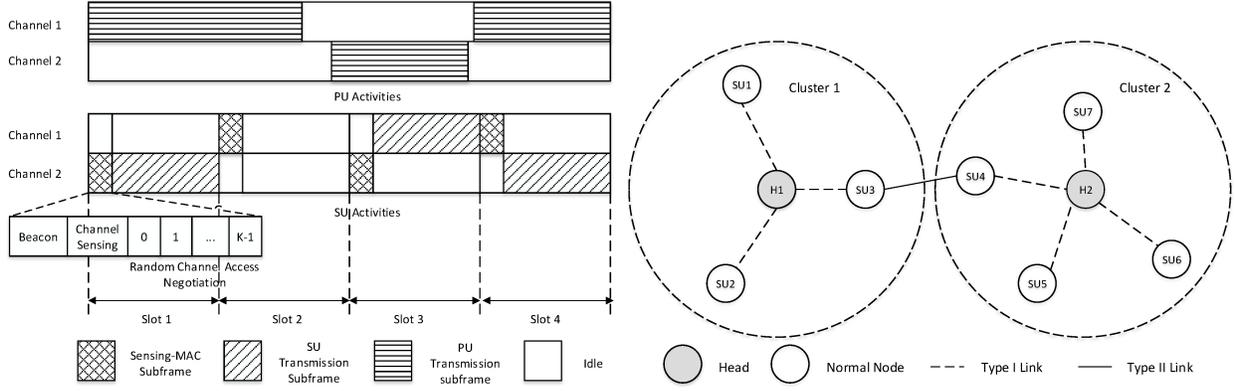


Fig. 1. Radio activities in a two-channel CRN with coordinated periodic sensing in the secondary network. Fig. 2. The SU links in a CRN of two clusters.

of a relay toward its sink [16] to help evaluate the link quality. Let L denote the sink SU, the distance advancement of SU i by choosing $a_i = (j, k)$ is defined as the reduction of distance from SU i to SU L when routing via SU j :

$$A_i(a_i) = D(i, L) - D(j, L), \quad (3)$$

where $D(i, j)$ is the Euclidean distances between SUs i and j . Based on the relay distance advancement in (3), we impose the rule that an SU is forbidden to select relays that produce negative distance advancement. Then for SU i , the set of candidate actions for channel-relay selection $a_i = (j, k)$ is defined by $\{j : j \in \mathcal{N}_i \setminus \{i\}, A_i(j) \geq 0\}$.

Let $(i, j)_k$ denote the link formed from SU i to SU j over channel k when SU i takes action $a_i = (j, k)$. According to the rule of DSA interweaving, link $(i, j)_k$ is accessible only when channel k is free for both SUs. Let $q(i)$ denote the spectrum activity cluster that SU i is in, then link $(i, j)_k$ can be classified into two types as shown in Figure 2:

- Type I: i and j are in the same cluster: $q(i) = q(j)$.
- Type II: i and j are in different clusters: $q(i) \neq q(j)$.

For Type I links, we only need to consider the channel state of one cluster, while for Type II links it is necessary to consider the joint channel state evolution of the two involved clusters.

We consider that the quality of link $(i, j)_k$ is measured based on the Effective Transmission Time (ETT) [34]. When the link is stable, the ETT over link $(i, j)_k$ can be measured as:

$$d_{(i,j)_k}^{\text{ETT}} = \frac{L}{R(1 - P_e(k))}, \quad (4)$$

where L is the packet length, R is the transmit rate and $P_e(e)$ is the packet error rate due to the physical layer error over channel k . When lacking stable channels, it is necessary to explicitly reflect in the link quality metric the impact of the DSA mechanism and MAC protocol. We note from (2) that due to the imperfect knowledge on channel states, a transmission failure may occur in the secondary network even when the current channel state vector indicates that channel k is free. Therefore, in order to determine the accessibility of link $(i, j)_k$, it is necessary to consider the conditional probability for channel k to be *Idle* during slot n given the observed state vectors of clusters $q(i)$ and $q(j)$ at the beginning of slot n . Based on (1), the probability for channel k to be *Idle* for a period τ from the beginning of slot n in cluster $q(i)$ can be calculated as:

$$\begin{aligned} P_k^{q(i)}(\tau, \mathbf{o}_k^{q(i)}(n)) &= P_k^{q(i)}\left(\mathbf{s}_k^{q(i)}(nT+\tau)=0 \mid \mathbf{o}_k^{q(i)}(n) = \mathbf{s}_k^{q(i)}(\phi_k(n)T)\right) \\ &= e^{-\lambda_k^{q(i)}\tau} \left[\mathbf{P}_k^{q(i)}((n - \phi_k(n))T) \right]_{\left(\mathbf{s}_k^{q(i)}(\phi_k(n)T), 0\right)}, \end{aligned} \quad (5)$$

where $e^{-\lambda_k^{q(i)}\tau}$ is the probability for the channel to remain idle for time τ since the beginning of slot n , and $\left[\mathbf{P}_k^{q(i)}((n - \phi_k(n))T) \right]_{\left(\mathbf{s}_k^{q(i)}(\phi_k(n)T), 0\right)}$ is obtained from (1).

The link availability probability for $(i, j)_k$ at slot n depends on the probability of channel k staying idle at both end SUs. Based on our discussion of the link type, the probability of channel k being available for link $(i, j)_k$ during slot n can be expressed as:

$$P_k^{i,j}(\mathbf{o}(n)) = P_k^{i,j}(\mathbf{o}_k^{q(i)}(n), \mathbf{o}_k^{q(j)}(n)) = \begin{cases} P_k^{q(i)}(T, \mathbf{o}_k^{q(i)}(n)), & \text{if } q(i) = q(j), \\ P_k^{q(i)}(T, \mathbf{o}_k^{q(i)}(n)) P_k^{q(j)}(T, \mathbf{o}_k^{q(j)}(n)), & \text{if } q(i) \neq q(j), \end{cases} \quad (6)$$

where \mathbf{o} is the concatenation of the observed state vectors of all the clusters. Based on (4) and (6), we can obtain the spectrum-aware link delay metric for $(i, j)_k$ at slot n as follows:

$$d_{(i,j)_k}(\mathbf{o}(n)) = T \left(1 - P_k^{i,j}(\mathbf{o}_k^{q(i)}(n), \mathbf{o}_k^{q(j)}(n)) \right) + d_{(i,j)_k}^{\text{ETT}} P_k^{i,j}(\mathbf{o}_k^{q(i)}(n), \mathbf{o}_k^{q(j)}(n)), \quad (7)$$

where $\mathbf{o}(n)$ represents the joint state of the entire secondary network at slot n .

Now, we consider the impact of the MAC protocol on the state of link availability. Let $\mathcal{A}_i = \mathcal{N}_i \setminus \{i\} \times \mathcal{K}$ denote the set of candidate actions for SU i . Due to the channel instability, it is difficult to directly adopt MAC protocols based on single-channel random access with exponential backoff in the secondary network. Instead, we consider that the contention over each channel is resolved through a reservation mechanism over the common control channel. We consider that the negotiating phase over the control channel is divided into K subslots, and the SUs compete for channel k in the corresponding subslot by sending Request-To-Send (RTS)

packets and listening to Clear-To-Send (CTS) packets from their target relay SUs (Figure 1). Since more than one RTS sent in SU i 's neighborhood over the same channel will result in collision, the channel negotiation can be considered as a random access mechanism which is similar to slotted-ALOHA. If channel k is free, the probability of SU i successfully sending the RTS packet over channel k after taking action a_i in \mathcal{N}_i can be written as:

$$P_i^k(\mathbf{a}_{\mathcal{N}_i}) = I(a_{i,2}, k) \prod_{m \in \mathcal{N}_i \setminus \{i\}} (1 - I(a_{m,2}, k)), \quad (8)$$

where $\mathbf{a}_{\mathcal{N}_i}$ is the joint SU action in \mathcal{N}_i and $I(x, y)$ is the indicator function. $I(x, y) = 0$ if $x \neq y$ and $I(x, y) = 1$ if $x = y$. Similarly, considering the existence of hidden terminals, the probability of SU j successfully receiving the RTS packet from SU i over channel k can be written as:

$$P_j^k(\mathbf{a}_{\mathcal{N}_j}) = I(a_{i,2}, k) \prod_{m \in \mathcal{N}_j \setminus \{i,j\}} (1 - I(a_{m,2}, k)). \quad (9)$$

Based on (8) and (9), we can express (7) under the joint actions $\mathbf{a}_{\mathcal{N}_i}$ and $\mathbf{a}_{\mathcal{N}_j}$ as follows:

$$d_{(i,j)_k}(\mathbf{o}(n), \mathbf{a}_{\mathcal{N}_i}, \mathbf{a}_{\mathcal{N}_j}) = T(1 - P_i^k(\mathbf{a}_{\mathcal{N}_i})P_j^k(\mathbf{a}_{\mathcal{N}_j})) + d_{(i,j)_k}(\mathbf{o}^{q(i)}(n), \mathbf{o}^{q(j)}(n))P_i^k(\mathbf{a}_{\mathcal{N}_i})P_j^k(\mathbf{a}_{\mathcal{N}_j}). \quad (10)$$

In addition to the delay caused by the SU actions following the proposed DSA-MAC, we also need to consider the delay caused by interference between multiple flows in the CRN. We assume that each SU can only respond to one randomly chosen RTS during a transmission slot. For the proposed DSA-MAC, the number of potential links that can be established to SU i is:

$$N_i(\mathbf{a}) = \sum_{k \in \mathcal{K}} \left(\sum_{m \in \mathcal{N}_i \setminus \{i\}} P_i^k(\mathbf{a}_{\mathcal{N}_i})P_m^k(\mathbf{a}_{\mathcal{N}_m}) \right), \quad (11)$$

where \mathbf{a} is the joint action of all the SUs. According to the queueing delay model based on round-robin packet processing [3], we need to adjust the expected link delay in (10) by substituting $d_{(i,j)_k}^{\text{ETT}}$ in (7) with $N_i(\mathbf{a})d_{(i,j)_k}^{\text{ETT}}$:

$$d_i(\mathbf{o}(n), \mathbf{a}) = T(1 - P_i^k(\mathbf{a}_{\mathcal{N}_i})P_j^k(\mathbf{a}_{\mathcal{N}_j})) + \left(T(1 - P_k^{i,j}(\mathbf{o}(n))) + N_i(\mathbf{a})d_{(i,j)_k}^{\text{ETT}}P_k^{i,j}(\mathbf{o}(n)) \right) P_i^k(\mathbf{a}_{\mathcal{N}_i})P_j^k(\mathbf{a}_{\mathcal{N}_j}), \quad (12)$$

C. Link Quality Metric

Let $\mathcal{P}(i_0, i_L) = \{(i_0, i_1)_{k_0}, (i_1, i_2)_{k_1}, \dots, (i_{L-1}, i_L)_{k_{L-1}}\}$ denote the path formed by a sequence of links between SU i_0 and SU i_L . According to Section III-B, the additional path delay after

including SU i into $\mathcal{P}(i_0, i_L)$ is jointly determined by the cluster states of its neighbor nodes and the joint action of its two-tier neighbor nodes, see (10) and (11). Based on (4)-(11), we can express the link added by SU i as a function of the joint action of all the SUs $\mathbf{a} = (a_1, \dots, a_{|\mathcal{N}|})$ in (12). Combining the metrics of the adjusted link delay in (12) and the relay distance advancement in (3), we can define the instant local utility of SU i as a function of the joint state $\mathbf{o}(n)$ and the joint action \mathbf{a} in the secondary network in (13):

$$u_i(\mathbf{o}(n), \mathbf{a}) = \frac{A_i(a_i)}{d_i(\mathbf{o}(n), \mathbf{a})}. \quad (13)$$

According to (3) and (12), $u_i(\mathbf{o}(n), \mathbf{a}) \geq 0$. With (13), a normal SU i_0 measures the quality of its path $\mathcal{P}(i_0, i_L)$ as the expected average of the cumulative link utility along the path as follows:

$$U_{\mathcal{P}(i_0, i_L)} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} E_{\mathbf{o}} \left[\sum_{n=0}^{\tau-1} \sum_{j \in \mathcal{P}(i_0, i_L)} u_j(\mathbf{o}(n), \mathbf{a}(n)) \middle| \mathbf{o}(0) = \mathbf{o} \right]. \quad (14)$$

Then for a normal SU, the goal of its relay-selection scheme is to maximize the value of $U_{\mathcal{P}(i_0, i_L)}$.

D. Impact of Malicious SUs

We consider that in the CRN, no SU is superior to the other SUs in obtaining network information. As a result, both the normal SUs and the malicious SUs make their relay-channel selection decisions based on the same level of local information. For a malicious relay SU j in path $\mathcal{P}(i_0, i_L)$, the goal is to cause delay as much as possible by minimizing the expected cumulative utility $U_{\mathcal{P}(j, i_L)}$ while avoiding being detected as an attacker. To avoid detection, SU j disguises itself by complying with most of the routing rules in the network layer. Based on its local channel state record, SU j performs the RPU-like attacks by violating the interweaving DSA rule and forwarding the packet over the link that has the highest probability of being at state *Busy*. SU j may also attempt to send packets to the neighbor SUs which experience larger delay due to channel contention caused by flow intersection. Since normal SUs are limited by the number of the equipped radio interfaces, they are not able to passively monitor the neighbors' behaviors. Also, due to imperfect information about the instantaneous channel states, normal SUs may have difficulties in discerning the delay due to attacks from the delay due to PU activities.

Since the SUs cannot exchange the routing information (e.g., local utility and relay-selection decision) with the entire CRN, to form an efficient path they mainly rely on the information exchanged between the neighbors. From the perspective of malicious SUs, such a situation of

information locality allows them to provide fake information by distorting the announced value of the expected cumulative link utility for sub-route $\mathcal{P}(j, i_L)$ and induce the normal neighbors to forward packets to them. Malicious SUs behave similarly to the SH attackers. Since the operation of information distortion heavily depends on the routing scheme adopted by the normal SUs, we will provide more details of this type of attack in the following sections.

IV. ROBUST ROUTING BASED ON STOCHASTIC GAME

For ease of presentation, in this section we will temporarily ignore the possibility of information distortion by malicious relays and assume truthful information exchange between neighbor SUs. Following our discussion of the node behavior and the link quality metric in (13), we analyze the routing mechanism using a game theoretic model, which explicitly addresses the interaction between the normal and the malicious SUs.

A. Relay Selection as a Stochastic Game

We define the secondary network global state as the concatenation of the state vectors from all the clusters: $\mathbf{o} = \mathbf{o}^1 \parallel \dots \parallel \mathbf{o}^q \parallel \dots \parallel \mathbf{o}^Q$, where $q = 1, \dots, Q$ is the cluster index. With a slight abuse of notation, we omit the index of the sink SU i_L and denote a path starting from SU i as \mathcal{P}_i . Then, from Section III-A, the evolution of the joint states of any SU sequence retains the Markovian property, as stated in Proposition 1:

Proposition 1. *For any sequence of SUs \mathcal{P}_i , its joint observed state vector $\parallel_{q:j \in \mathcal{P}_i(j), \forall j \in \mathcal{P}_i} \mathbf{o}^q$ forms a Markov chain, of which the transition of each state element is independent of the SU actions and can be described by (2).*

The instant utility of path \mathcal{P}_i can be obtained from (14) as:

$$u_{\mathcal{P}_i}(\mathbf{o}(n), \mathbf{a}(n)) = u_i(\mathbf{o}(n), \mathbf{a}(n)) + \sum_{j \in \mathcal{P}_i \setminus \{i\}} u_j(\mathbf{o}(n), \mathbf{a}(n)). \quad (15)$$

For conciseness, we use $U_{\mathcal{P}_i}$ to represent the value of $U_{\mathcal{P}(i, i_L)}$ in (14). Since an SU only controls its own decision of choosing the next-hop and observes its local utility, the path quality evaluation by SU i will depend on the utility information provided by the next-hop SU. Then, based on the path utility in (15) and Proposition 1, we can define a stochastic routing game in the secondary network as a five-tuple multi-agent Markov Decision Process (MDP) [35]:

Definition 1 (Stochastic routing game). *The SUs in the CRN form a general-sum stochastic game in the form of a five-tuple: $\mathcal{G}_r = \langle \mathcal{N}, \mathcal{O}, \mathcal{A}, \{u_{\mathcal{P}_i}\}_{i \in \mathcal{N}}, P(\mathbf{o}' | \mathbf{o}) \rangle$, in which*

- \mathcal{N} is the set of SUs.
- \mathcal{O} is the space of the concatenated cluster state vectors.
- $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}_i$ is the set of joint actions of the SUs.
- $u_{\mathcal{P}_i} : \mathcal{O} \times \mathcal{A} \rightarrow \mathbb{R}$ is the instantaneous utility of the path starting from SU i as in (15).
- $P : \mathcal{O} \times \mathcal{O} \rightarrow [0, 1]$ is the state transition map.

Let a_{-i} denote the joint actions of all SUs except SU i , $\pi_i(\mathbf{o}) = (\pi_i(\mathbf{o}, a) : a \in \mathcal{A}_i)$ denote the mixed strategy of SU i at state \mathbf{o} , and $\pi_{-i}(\mathbf{o}) = (\pi_i(\mathbf{o}, a_{-i}) : a_{-i} \in \mathcal{A}_{-i})$ denote the mixed strategy of all SUs except SU i at state \mathbf{o} . We note that given the SUs' joint strategy, $\pi = (\pi_i(\mathbf{o}), \pi_{-i}(\mathbf{o}) : \mathbf{o} \in \mathcal{O})$, the goal of normal SU i is to maximize its expected average utility in (14), while the goal of malicious SU m is to minimize the average utility. Given π , we have:

$$U_{\mathcal{P}_i}(\mathbf{o}, \pi) = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} E_{\mathbf{o}, \pi} \left\{ \sum_{n=0}^{\tau-1} \sum_{j \in \mathcal{P}(i_0, i_L)} u_j(\mathbf{o}(n), \mathbf{a}(n)) \middle| \mathbf{o}(0) = \mathbf{o} \right\}. \quad (16)$$

With (15) and (16), we can define the Nash Equilibrium (NE) of the game as:

Definition 2 (NE). $\pi^* = (\pi_i^*, \pi_{-i}^*)$ is an NE for \mathcal{G}_r , if $\forall i \in \mathcal{N}$ and $\forall \mathbf{o} \in \mathcal{O}$ the following conditions are satisfied for any π_i :

$$\begin{cases} U_{\mathcal{P}_i}(\mathbf{o}, \pi_i^*, \pi_{-i}^*) \geq U_{\mathcal{P}_i}(\mathbf{o}, \pi_i, \pi_{-i}^*), & \text{if SU } i \text{ is normal,} \\ U_{\mathcal{P}_i}(\mathbf{o}, \pi_i^*, \pi_{-i}^*) \leq U_{\mathcal{P}_i}(\mathbf{o}, \pi_i, \pi_{-i}^*), & \text{if SU } i \text{ is malicious.} \end{cases}$$

Observing (15), we note that game \mathcal{G}_r differs from a typical stochastic game because the instantaneous individual payoff is determined by not only the local link utility, but also the utility of the sub-route starting from the next-hop SU. Therefore, to obtain the NE for \mathcal{G}_r , the SUs are required to know the sub-route utility of their next-hop nodes. In order to examine the property of the NE for \mathcal{G}_r , we introduce the concept of the bias value in a multi-agent MDP:

Definition 3 (Bias value). With initial state \mathbf{o} and policy $\pi = (\pi_i, \pi_{-i})$, the bias value of SU i is the expected accumulated difference between its instantaneous and stationary utilities:

$$h_{\mathcal{P}_i}(\mathbf{o}, \pi) = \lim_{\tau \rightarrow \infty} E \left\{ \sum_{n=0}^{\tau-1} (u_{\mathcal{P}_i}(\mathbf{o}(n), \pi) - U_{\mathcal{P}_i}(\mathbf{o}(n), \pi)) \middle| \mathbf{o}(0) = \mathbf{o} \right\}. \quad (17)$$

Based on (1) and Proposition 1, we can readily conclude that game \mathcal{G}_r in the sense of a multi-agent MDP is ergodic/recurrent [36]. Then, using the bias value in Definition 3, we introduce the representation of an average utility MDP in the form of the Bellman optimality equation:

Lemma 1. *Regardless of the initial state \mathbf{o} , the bias value of each SU in game \mathcal{G}_r is constant given any stationary policy $\boldsymbol{\pi}$ and can be expressed as:*

$$h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) = u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) - U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}). \quad (18)$$

Proof. According to Proposition 1, the transition of the observed states is independent of the SUs' actions. Therefore, according to (1), for any deterministic strategy \mathbf{a} , the underlying Markov chain converges to the same limiting distribution and thus is ergodic. Observing (13), we note that the instantaneous local link utility u_i is bounded for a finite number of relays. Then, for a sub-route \mathcal{P}_i , the expectation and summation in (16) is interchangeable. From (16) we obtain:

$$U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{n=0}^{\tau-1} E_{\mathbf{o}} \left(E_{\boldsymbol{\pi}} \left(\sum_{j \in \mathcal{P}_i} u_j(\mathbf{o}(n), \mathbf{a}(n)) \right) \middle| \mathbf{o}(0) = \mathbf{o} \right) = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{n=0}^{\tau-1} P(\mathbf{o}'|\mathbf{o}) \left(\sum_{j \in \mathcal{P}_i} u_j(\mathbf{o}', \boldsymbol{\pi}) \right). \quad (19)$$

where

$$u_j(\mathbf{o}, \boldsymbol{\pi}) = \sum_{\mathbf{a}_1 \in \mathcal{A}_1} \cdots \sum_{\mathbf{a}_{|\mathcal{N}|} \in \mathcal{A}_{|\mathcal{N}|}} \left(u_j(\mathbf{o}, \mathbf{a}) \times \boldsymbol{\pi}_1 \times \cdots \times \boldsymbol{\pi}_{|\mathcal{N}|} \right). \quad (20)$$

Therefore, with respect to the stationary joint strategy $\boldsymbol{\pi}$, each SU's state-value evolution in game \mathcal{G}_r is reduced to a finite-state, recurrent Markov reward process. Then, Lemma 1 immediately follows Theorem 8.2.6 of [36]. \square

By fixing the observed channel state as \mathbf{o} in the stochastic game, we define the stage game of \mathcal{G}_r at state \mathbf{o} as $\mathcal{G}_r(\mathbf{o}) = \langle \mathcal{N}, \mathcal{A}, \{u_{\mathcal{P}_i}(\mathbf{o})\}_{i \in \mathcal{N}} \rangle$. $\mathcal{G}_r(\mathbf{o})$ is a normal-form repeated game with normal SUs aiming at maximizing their instantaneous path utilities and malicious SUs aiming at minimizing the instantaneous path utilities at state \mathbf{o} . Based on Lemma 1, we can derive the following results on the NE points of \mathcal{G}_r :

Theorem 1. (i) $\boldsymbol{\pi}^*$ is an NE of \mathcal{G}_r , only if the following conditions are satisfied $\forall \boldsymbol{\pi}_i$:

$$h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) \geq u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) - U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}^*), \quad (21)$$

$$h_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}^*) \leq u_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}_j, \boldsymbol{\pi}_{-j}^*) - U_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}^*) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_m}(\mathbf{o}', \boldsymbol{\pi}^*), \quad (22)$$

for every normal SU i and malicious SU j .

(ii) $\boldsymbol{\pi}^*(\mathbf{o})$ is also an NE strategy of $\mathcal{G}_r(\mathbf{o})$. The NE strategies of all the stage games, $\mathcal{G}_r(\mathbf{o} : \forall \mathbf{o} \in \mathcal{O})$, constitute an NE strategy of \mathcal{G}_r .

Proof. See Appendix A. \square

Remark 1. Theorem 1 establishes the equivalence between the NE strategies of \mathcal{G}_r and the group of NE strategies of its corresponding stage games. It is worth noting that Theorem 1 is based on Proposition 1. In this case, all the NE of the stochastic game are the stationary Markov perfect equilibria. In contrast, for a general-case stochastic game where the state transition is usually a function of the players' actions, the equality in (18) may not hold except for the equilibrium strategies, and the second property in Theorem 1 does not exist. \square

Due to the overhead caused by information flooding, it is unrealistic for the SUs to frequently exchange the information about their private actions and utilities with the SUs beyond the one-hop neighbors. To determine the level of information exchange in the routing game, we consider another multi-agent MDP based on each SU's local utility u_i . We define the MDP as $\mathcal{G}_l = \langle \mathcal{N}, \mathcal{O}, \mathcal{A}, \{u_i\}_{i \in \mathcal{N}}, P(\mathbf{o}'|\mathbf{o}) \rangle$. Let $h_i(\mathbf{o}, \boldsymbol{\pi})$ denote the bias value of \mathcal{G}_l and $U_i(\mathbf{o}, \boldsymbol{\pi})$ denote the average gain value of \mathcal{G}_l . Then, we can show that Lemma 1 also applies to the pair of h_i and U_i in \mathcal{G}_l . Let $\boldsymbol{\pi}_{i,1}$ denote the strategy of SU i for selecting the next hop, $\boldsymbol{\pi}_{i,2}$ denote the strategy of SU i for selecting the transmitting channel, and $\mathcal{P}_{a_{i,1}}$ denote the sub-route in \mathcal{P}_i starting from the node that is chosen by SU i with action $a_{i,1}$. Based on Lemma 1 and Theorem 1, we can show that \mathcal{G}_r can be decomposed into a layered multi-agent MDP in the following theorem:

Theorem 2. (i) With stationary joint policy $\boldsymbol{\pi}$, the relay selection process of SU i can be expressed as (23):

$$h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) + U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) = u_i(\mathbf{o}, \boldsymbol{\pi}) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_i(\mathbf{o}', \boldsymbol{\pi}) + E_{\boldsymbol{\pi}_{i,1}} \left\{ u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\}. \quad (23)$$

(ii) Strategy $\tilde{\boldsymbol{\pi}}$ is an NE point of \mathcal{G}_r when for any normal SU i and malicious SU j ,

$$\begin{cases} \tilde{\boldsymbol{\pi}}_i = \arg \max_{\boldsymbol{\pi}_i} (h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \tilde{\boldsymbol{\pi}}_{-i}) + U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \tilde{\boldsymbol{\pi}}_{-i})), \\ \tilde{\boldsymbol{\pi}}_j = \arg \min_{\boldsymbol{\pi}_j} (h_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}_j, \tilde{\boldsymbol{\pi}}_{-j}) + U_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}_j, \tilde{\boldsymbol{\pi}}_{-j})), \end{cases} \quad (24)$$

Proof. See Appendix B. \square

Remark 2. Theorem 2 shows that given a stationary joint policy $\boldsymbol{\pi}$, the relay-selecting process of an SU is composed of two value iteration processes in the form of the Bellman optimality equation. The first one is determined by the local multi-agent MDP \mathcal{G}_l , and the second one is

determined by the sub-path starting from the selected next-hop SU $a_{i,1}$ in \mathcal{G}_r . Furthermore, the second Bellman optimality equation for SU $a_{i,1}$ can be decomposed into the same two-layer form as (23) with respect to its own decision on next-hop selection. \square

According to Theorem 2, to derive its local NE strategy π_i^* , SU i needs its neighbor nodes $j \in \mathcal{N}_i \setminus \{i\}$ to truthfully provide the information on the equilibrium value of $h_{\mathcal{P}_j}(\mathbf{o}, a_{i,1} = j, \pi_{i,2}, \pi_{-i}^*) + U_{\mathcal{P}_j}(\mathbf{o}, a_{i,1} = j, \pi_{i,2}, \pi_{-i}^*)$. It requires that the NE for game \mathcal{G}_r is solved through backward induction. Observing (16) and (17), it is straightforward to show that when stochastic game \mathcal{G}_r is reduced to a stage game $\mathcal{G}_r(\mathbf{o})$ with a single state \mathbf{o} , providing the value $h_{\mathcal{P}_j}(\mathbf{o}, a_{i,1} = j, \pi_{i,2}, \pi_{-i}^*) + U_{\mathcal{P}_j}(\mathbf{o}, a_{i,1} = j, \pi_{i,2}, \pi_{-i}^*)$ is equivalent to providing the value $u_{\mathcal{P}_j}(\mathbf{o}, a_{i,1} = j, \pi_{i,2}, \pi_{-i}^*)$. Such an observation paves the way for developing a strategy-learning method based on limited information exchange between the SUs.

B. Strategy Learning with Truthful Information Exchange

According to Theorem 1, an NE for the stochastic routing game can be constructed based on the state-dependent NE strategies for each stage routing game with fixed estimated channel states. Therefore, we consider a stage routing game at state \mathbf{o} : $\mathcal{G}_r(\mathbf{o}) = \langle \mathcal{N}, \mathcal{A}, \{u_{\mathcal{P}_i}(\mathbf{o})\}_{i \in \mathcal{N}} \rangle$, where $u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}) = u_i(\mathbf{o}, \mathbf{a}) + u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, \mathbf{a})$. Based on Theorem 2, the NE for $\mathcal{G}_r(\mathbf{o})$ is achieved when each normal SU i and malicious SU j play the strategies π^* that satisfy the following conditions:

$$\begin{cases} U_{\mathcal{P}_i}^* = \max_{\pi_i} \left(u_i(\mathbf{o}, \pi_i, \pi_{-i}^*) + E_{\pi_{i,1}} \left\{ u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \pi_{i,2}, \pi_{-i}^*) \right\} \right), \\ U_{\mathcal{P}_j}^* = \min_{\pi_j} \left(u_j(\mathbf{o}, \pi_j, \pi_{-j}^*) + E_{\pi_{j,1}} \left\{ u_{\mathcal{P}_{a_{j,1}}}(\mathbf{o}, a_{j,1}, \pi_{j,2}, \pi_{-j}^*) \right\} \right). \end{cases} \quad (25)$$

To avoid information flooding, we assume that the SUs do not share with their neighbors the local action information. An SU is only able to share its value of actions by exchanging routing request (RREQ) and routing response (RREP) packets with its neighbors. In this section, we consider that malicious SUs do not provide distorted information. Since an SU cannot observe other nodes' actions, we resort to reinforcement learning to obtain the NE under the condition of incomplete information. We assume that a stationary joint strategy π is adopted by the SUs in game $\mathcal{G}_r(\mathbf{o})$. Then, we consider the following action-value learning process for SU i :

$$\tilde{u}_{\mathcal{P}_i}^{n+1}(\mathbf{o}, \mathbf{a}_i) = \tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i) + \alpha(n) I(\mathbf{a}_i(n), \mathbf{a}_i) \left(u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i(n), \mathbf{a}_{-i}(n)) - \tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i) \right), \quad (26)$$

where $\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i)$ is the expected path utility learned for action \mathbf{a}_i at slot n , and $0 < \alpha(n) < 1$ is a sequence of learning rates. According to reinforcement learning theory [37], if $u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i(n), \mathbf{a}_{-i}(n))$

is perfectly known by SU i , $\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i)$ converges almost surely to the real value of $u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i, \boldsymbol{\pi}_{-i})$, given that all the possible action combinations are visited infinitely often by the SUs and $\alpha(n)$ satisfies the conditions $\sum_n \alpha(n) = \infty$ and $\sum_n \alpha^2(n) < \infty$.

We first assume that an SU i is able to timely calculate the instantaneous accumulated utility of path \mathcal{P}_i based on its local observation of $u_i(\mathbf{o}(n), \mathbf{a}(n))$ and the instantaneous sub-path utility $u_{\mathcal{P}_{\mathbf{a}_{i,1}}}(\mathbf{o}(n), \mathbf{a}(n))$, which is fed back by its next-hop SU $j = \mathbf{a}_{i,1}$. In this case, (26) can be adopted by each SU to estimate their action value in stage game $\mathcal{G}_r(\mathbf{o})$. Then, we can adopt the algorithm of Stochastic Fictitious Play (SFP) [38] for SU i to learn $\boldsymbol{\pi}_i(\mathbf{o}, \mathbf{a}_i, \boldsymbol{\pi}_{-i})$:

$$\tilde{\boldsymbol{\pi}}_i^{n+1}(\mathbf{o}, \mathbf{a}_i) = \tilde{\boldsymbol{\pi}}_i^n(\mathbf{o}, \mathbf{a}_i) + \beta(n) (\text{BR}(\tilde{\mathbf{u}}_{\mathcal{P}_i}^n(\mathbf{o}), \mathbf{a}_i) - \tilde{\boldsymbol{\pi}}_i^n(\mathbf{o}, \mathbf{a}_i)), \quad (27)$$

where $\tilde{\mathbf{u}}_{\mathcal{P}_i}^n(\mathbf{o})$ is the vector of utility $\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i)$ for all action \mathbf{a}_i at time slot n , and $\text{BR}(\cdot)$ is the perturbed best response strategy of SU i for action \mathbf{a}_i in the form of the Logit function:

$$\text{BR}(\tilde{\mathbf{u}}_{\mathcal{P}_i}^n(\mathbf{o}), \mathbf{a}_i) = \begin{cases} \frac{\exp(\lambda_i(\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i)))}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp(\lambda_i(\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{b})))}, & i \text{ is normal,} \\ \frac{\exp(\lambda_i(\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i))^{-1})}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp(\lambda_i(\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{b}))^{-1})}, & i \text{ is malicious.} \end{cases} \quad (28)$$

The utility learning process in (26) and the SPF-based strategy learning process in (27) and (28) form a two timescale learning scheme, which has the following convergence property:

Theorem 3. *If $u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}(n))$ is known to each SU at every time slot, and the following conditions are satisfied: $\lim_{n \rightarrow \infty} \sum_n \alpha(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \alpha^2(n) < \infty$, $\lim_{n \rightarrow \infty} \sum_n \beta(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \beta^2(n) < \infty$ and $\lim_{n \rightarrow \infty} (\beta(n)/\alpha(n)) = 0$, then $\{\tilde{\boldsymbol{\pi}}_i^n(\mathbf{o}, \mathbf{a}_i)\}$ given by the learning process (26)-(28) converges almost surely to an NE for stage game $\mathcal{G}_r(\mathbf{o})$.*

Proof. See Appendix C. □

Although the learning scheme given by (27) and (28) possesses good convergence property, the assumption of perfectly knowing the instantaneous path utility is fairly strict. It requires a large amount of signaling to be performed within a single time slot. To address such a problem, we relax the requirement on information exchange by assuming that SU i only shares its locally estimated value of $u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi})$ with the neighbors. Based on the discussion of (26), it is obvious that an SU can learn its expected local link utility $u_i(\mathbf{o}, \mathbf{a}_i, \boldsymbol{\pi}_{-i})$ through an iteration which is

similar to (26), as long as all possible joint actions are visited infinitely often:

$$\tilde{u}_i^{n+1}(\mathbf{o}, \mathbf{a}_i) = \tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i) + \alpha(n) I(\mathbf{a}_i(n), \mathbf{a}_i) (u_i(\mathbf{o}, \mathbf{a}(n)) - \tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i)). \quad (29)$$

Using the value of $\tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i)$ and the value of $\tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o})$ provided by the next-hop SU $j = \mathbf{a}_{i,1}$, we introduce the learning scheme for $u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi})$:

$$\tilde{u}_{\mathcal{P}_i}^{n+1}(\mathbf{o}) = \tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}) + \gamma_i(n) \left(\sum_{\mathbf{a}_i} \tilde{\pi}_i^n(\mathbf{o}, \mathbf{a}_i) (\tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o})) - \tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}) \right). \quad (30)$$

We also modify the learning scheme of (27) and obtain:

$$\tilde{\pi}_i^{n+1}(\mathbf{o}, \mathbf{a}_i) = \tilde{\pi}_i^n(\mathbf{o}, \mathbf{a}_i) + \beta(n) \left(\text{BR}(\tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o}), \mathbf{a}_i) - \tilde{\pi}_i^n(\mathbf{o}, \mathbf{a}_i) \right), \quad (31)$$

where $\text{BR}(\cdot)$ is the modified perturbed best response strategy:

$$\text{BR}(\tilde{u}_i^n(\mathbf{o}) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o}), \mathbf{a}_i) = \begin{cases} \frac{\exp(\lambda_i(\tilde{u}_i(\mathbf{o}, \mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o})))}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp(\lambda_i(\tilde{u}_i(\mathbf{o}, \mathbf{b}) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o})))}, & i \text{ is normal,} \\ \frac{\exp(\lambda_i(\tilde{u}_i(\mathbf{o}, \mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o}))^{-1})}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp(\lambda_i(\tilde{u}_i(\mathbf{o}, \mathbf{b}) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}^n(\mathbf{o}))^{-1})}, & i \text{ is malicious.} \end{cases} \quad (32)$$

The learning scheme defined by (29)-(32) does not require SU i to immediately report the instantaneous path utility to its previous-hop SU. However, comparing (30) with (25), we note that (30) provides a biased estimation of $u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi})$, given that $\tilde{\pi}_i^n$ converges. Since the learned path utility in (30) is a biased estimation, the new learning scheme can only obtain an approximation of the NE point of the stage game. The convergence condition of the learning scheme given by (29)-(32) is provided in Theorem 4.

Theorem 4. *Assume that the following are satisfied: $\lim_{n \rightarrow \infty} \sum_n \alpha(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \alpha^2(n) < \infty$, $\lim_{n \rightarrow \infty} \sum_n \gamma_i(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \gamma_i^2(n) < \infty$, $\lim_{n \rightarrow \infty} \sum_n \beta(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \beta^2(n) < \infty$, $\lim_{n \rightarrow \infty} (\gamma_i(n)/\alpha(n)) = 0$, $\lim_{n \rightarrow \infty} (\beta(n)/\gamma_i(n)) = 0$, and $\lim_{n \rightarrow \infty} (\gamma_i(n)/\gamma_j(n)) = 0$, if SU i is closer to the sink SU than SU j in terms of distance. Then, $\{\tilde{\pi}_i^n(\mathbf{o}, \mathbf{a}_i)\}$ obtained through the learning process defined by (29)-(32) converges almost surely.*

Proof. See Appendix D. □

C. Strategy Learning with Truth-telling Enforcement

Now, we consider the situation when the malicious SUs also perform the SH attacks. In this case, the malicious SUs may distort the value of $\tilde{u}_{\mathcal{P}_i}(\mathbf{o})$ and report an estimated utility which is much larger than the real value to their neighbor SUs. As a result, a normal neighbor with strategy learning scheme in (31) will choose the malicious SU as its relay with a higher probability. Then, the malicious SU will induce the neighbor SUs to forward more packets to them. To address this situation, we introduce a feedback mechanism for a relay SU to measure the real delay of the path that it chooses toward a sink SU. We consider that a normal relay SU i is able to insert a Request-ACK packet into the flows that it serves in random time intervals. SU i records the time stamp for sending the Request-ACK packet. When receiving the Request-ACK packet, the corresponding sink SU L replies the Response packet to SU i by including the time stamp for reception in the packet. We assume that the data in the Response packet is protected by a pair of keys and is always reliable. With the two time stamps, SU i is able to calculate the total delay time of the Request-ACK packet over the sub-path through the next-hop SU $j = \mathbf{a}_{i,1}$ that it chooses with action \mathbf{a}_i . Let $c_i(\mathbf{a}_i)$ denote such a delay measured by SU i . Then, SU i needs to evaluate the trustworthiness of its next-hop SU j based on sequence $\{c_i^{\hat{n}}(\mathbf{a}_i(\hat{n}))\}$, in which \hat{n} is a time slot for SU i to send a Request-ACK packet.

We consider that with a certain termination condition, the learning scheme given in (29)-(32) can always reach a stationary policy π . Meanwhile, a malicious SU m shares a fixed distorted value of $\tilde{u}_{\mathcal{P}_m}(\mathbf{o})$ with the neighbors. From the perspective of a normal SU i , when sending a Request-ACK packet, its relay selection can be considered as a Multi-Arm Bandit (MAB) [39] process, since at slot \hat{n} SU i can choose only one neighbor as its relay according to $\mathbf{a}_{i,1}(\hat{n})$, and only the real path delay through relay node $\mathbf{a}_{i,1}(\hat{n})$ can be confirmed. We note that the real path delay (i.e., the cost of each arm) is a stochastic function determined by stationary distribution π , while the arm selection sequence is generated by the local strategy π_i . Formally, we can define the MAB for trustworthiness evaluation as follows:

Definition 4. For each normal SU i , the MAB for trustworthiness evaluation in state \mathbf{o} can be defined by a 4-tuple: $\mathcal{B}_i = \langle \mathcal{A}_i, \{c_i^{\hat{n}}(\mathbf{o}, \mathbf{a}_i(\hat{n}))\}_{\hat{n}}, \{x(\hat{n})\}_{\hat{n}}, \{\hat{n}\} \rangle$, in which

- \mathcal{A}_i is the set of the single-bandit processes and corresponds to the set of actions of SU i .
- $\{c_i^{\hat{n}}(\mathbf{o}, \mathbf{a}_i(\hat{n}))\}_{\hat{n}}$ is the sequence of cost.

- $\{x(\hat{n}) = \mathbf{a}_i(\hat{n})\}_{\hat{n}}$ is the sequence of relay (i.e., arm) selection decision.

It is worth noting that the MAB given in Definition 4 differs from a typical MAB in that the sequence of arm selection $\{x(\hat{n})\}_{\hat{n}}$ is generated following a given policy π_i . Therefore, we can consider the MAB process up to time slot n as a utility exploration phase with a given sampling distribution π_i . With the MAB given in Definition 4, SU i is able to calculate the accumulated delay for the Request-ACK packets that it sends with action $\mathbf{a}_i(n) = \mathbf{a}$:

$$C_i^n(\mathbf{o}, \mathbf{a}) = \begin{cases} c_i^n(\mathbf{o}, \mathbf{a}_i(n)) + C_i^{n-1}(\mathbf{o}, \mathbf{a}) & \text{if } n \in \{\hat{n}\}, \mathbf{a}_i(n) = \mathbf{a} \\ C_i^{n-1}(\mathbf{o}, \mathbf{a}) & \text{otherwise,} \end{cases} \quad (33)$$

and the sampled frequency of each action is:

$$Z^n(\mathbf{o}, \mathbf{a}) = \frac{1}{n}(I(\mathbf{a}_i(n), \mathbf{a}) + (n-1)Z^{n-1}(\mathbf{o}, \mathbf{a})). \quad (34)$$

With (33) and (34), we can obtain the sampled average path delay of SU i at action $\mathbf{a}(n) = \mathbf{a}$ as $R_i^n(\mathbf{o}, \mathbf{a}) = C_i^n(\mathbf{o}, \mathbf{a})/Z_i^n(\mathbf{o}, \mathbf{a})$. Then according to [39], a greedy, sub-optimal mixed strategy for arm allocation to minimize the average path delay can be obtained using the Logit function:

$$\tilde{\sigma}_i^n(\mathbf{o}, \mathbf{a}) = \frac{\exp(\lambda_i (R_i^n(\mathbf{o}, \mathbf{a}))^{-1})}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp(\lambda_i (R_i^n(\mathbf{o}, \mathbf{b}))^{-1})}, \quad (35)$$

which is in a similar form to (28). $\tilde{\sigma}_i^n(\mathbf{a})$ does not have to be consistent with the learned equilibrium policy when every SU is honest. However, it can represent the ranking value of the trustworthiness of the relay associated with action \mathbf{a} . During the estimation of the perturbed best response, a normal SU will consider the contribution of the reported sub-path utility by its neighbors in proportion to the trustworthiness credit that it assigns to each neighbor. According to (53), a normal SU i modifies its smooth best response objective as follows:

$$\overline{\text{BR}}(\pi_{-i}) = \arg \max_{\pi_i} \left(\sum_{\mathbf{a}_i} \pi_i(\mathbf{a}_i) \left(u_i(\mathbf{o}, \pi(\mathbf{a}_i), \pi_{-i}) + \tilde{\sigma}_i^n(\mathbf{o}, \mathbf{a}_i) u_{\mathcal{P}_{a_i,1}}(\mathbf{o}, \pi) \right) - \lambda_i \sum_{\mathbf{a}_i} \pi_i(\mathbf{a}_i) \log \pi_i(\mathbf{a}_i) \right). \quad (36)$$

Then, its perturbed best response strategy can be adjusted as:

$$\text{BR} \left(\tilde{\mathbf{u}}_i^n(\mathbf{o}) + \tilde{u}_{\mathcal{P}_{a_i,1}}^n(\mathbf{o}), \mathbf{a}_i \right) = \frac{\exp \left(\lambda_i \left(\tilde{u}_i(\mathbf{o}, \mathbf{a}_i) + \tilde{\sigma}_i^n(\mathbf{o}, \mathbf{a}_i) \tilde{u}_{\mathcal{P}_{a_i,1}}^n(\mathbf{o}) \right) \right)}{\sum_{\mathbf{b} \in \mathcal{A}_i} \exp \left(\lambda_i \left(\tilde{u}_i(\mathbf{o}, \mathbf{b}) + \tilde{\sigma}_i^n(\mathbf{o}, \mathbf{b}) \tilde{u}_{\mathcal{P}_{a_i,1}}^n(\mathbf{o}) \right) \right)}, \quad (37)$$

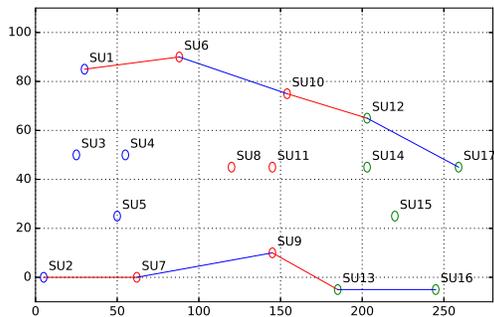


Fig. 3. An attacker-free CRN over 2 PU channels. Red lines represent packet-forwarding over channel 1 with a higher probability. Blue lines represent packet-forwarding over channel 0 with a higher probability.

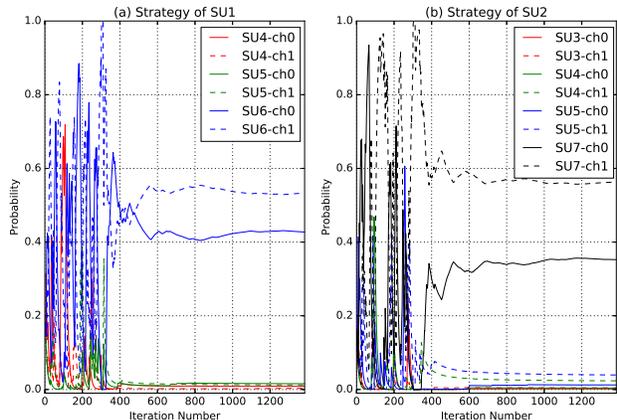


Fig. 4. Strategy evolution: channel-relay selection probability vs. iteration number.

V. SIMULATION RESULTS

Firstly, we demonstrate the convergence property of the proposed path selection mechanism given by (29)-(32). Without loss of generality, we assume that the state transition maps are identical for all PU channels. We set the parameters for channel state transition as $\lambda^{-1} = 0.2s$, $\mu^{-1} = 0.42s$, $T = 0.5s$, and for a valid link $d^{ETT} = 0.01s$. For convenience of visualization, we examine a randomly generated 2-channel, 3-cluster CRN with 2 flows in Figure 3. In Figure 3, SUs 1 and 2 are the source nodes and SUs 16 and 17 are the sink nodes. The strategy evolution for the source SUs is shown in Figure 4. According to our discussion about channel contention on (8)-(11), any source selecting SUs 3, 4 or 5 as its relay will result in a higher probability of conflict with the other source. Therefore, SUs 1 and 2 are expected to geographically separate their next hop as much as possible. As shown in Figure 4, with the learning scheme given by (29)-(32), SUs 1 and 2 separate the two flows by choosing SUs 6 and 7 as their relays with non-zero probabilities. The strategies of relaying through SUs 3, 4 and 5 finally converge to near 0. A mixed-strategy NE is reached and SUs 1 and 2 select between the two channels for transmission with non-zero probability. The highest-probability result of joint relay-channel selection for each SU at the NE is shown by the colored lines in Figure 3.

In Figure 5, we compare the performance of the algorithms given by (26)-(28) and (29)-(32) with that of a reference algorithm based on Opportunistic Cognitive Routing (OCR) with Cognitive Transport Throughput (CTT) as the link performance metric [16]. The original OCR-CTT algorithm was designed as a heuristic joint channel-relay searching method for efficient

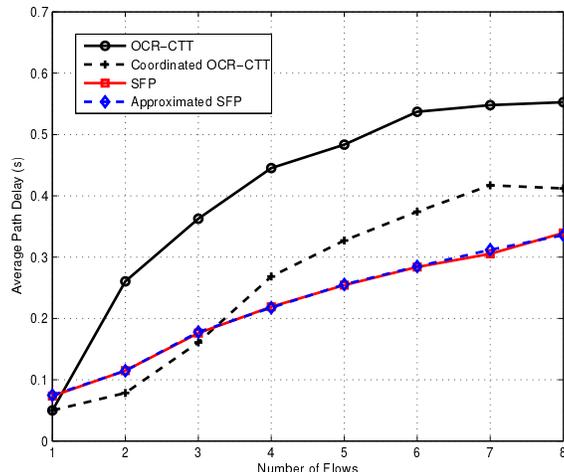


Fig. 5. Average path delay vs. number of flows for different algorithms.

single-flow routing in CRNs. To address the bottleneck effect with multiple flows, we modify the original OCR-CTT algorithm by introducing a centralized, greedy channel assignment mechanism. The simulation is set in a $250\text{m} \times 250\text{m}$ area with 100 relays randomly deployed in a 2-channel, 3-cluster CRN. The coverage radius of each SU is set to 35m. As shown in Figure 5, the proposed algorithms (SFP and Approximated SFP) with mixed-strategies have slightly larger delay than that of the deterministic OCR-CTT algorithms when the number of flows is small and the active SUs are sparse in the network. However, as the network becomes more congested with a larger number of flows, the proposed algorithms are able to better avoid channel conflicts and reduce the average path delay by 30% compared with the coordinated OCR-CTT algorithm.

In Figure 6, we evaluate the performance of the proposed strategy learning algorithm when malicious SUs exist. The simulation is conducted in the same randomly generated network for the simulation in Figure 5. We investigate the “aggressiveness” of an attacker by varying the scale of information distortion by the malicious SUs based on the real value of the sub-path utility. The larger the scale that an attacker uses for information distortion, the more aggressive the attacker is. There are 4 flows in the CRN and for each source node there is one malicious SU randomly placed in its one-hop neighborhood. Comparing the average path delay at 4 flows in Figure 5 with the average path delay at scale 1 in Figure 6, we note that the routing performance is not affected by the presence of attackers when malicious SUs do not adopt SH schemes. Intuitively, this is because with the proposed learning mechanisms, an SU is able to switch to alternative normal relays when performance deterioration from the attackers is detected and the network is

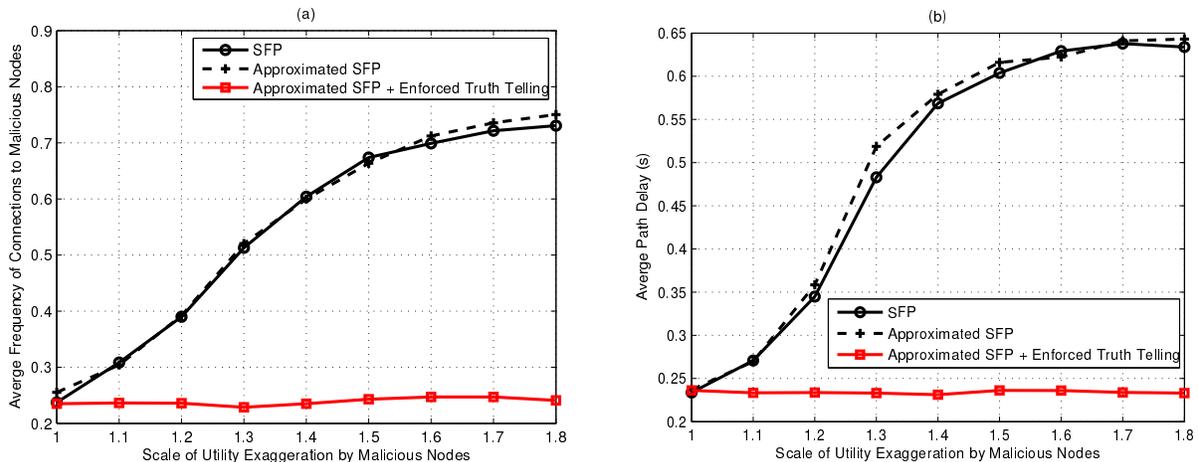


Fig. 6. (a) Frequency of connections to malicious nodes vs. scale for utility. (b) Average path delay vs. scale of exaggerated utility by malicious nodes.

not congested. However, when truth-telling enforcement is not enabled, the malicious SUs are able to quickly attract the nearby flows by exaggerating their reported value of sub-path utility (see Figure 6a). Consequently, a steep increase in average path delay can be observed in Figure 6b. In contrast, when truth-telling enforcement is enabled, the performance of multi-flow routing remains in the same level of the case of no attackers. As can be observed in (35), given sufficient time for delay-evaluation based on the proposed feedback mechanism, the exponential operator in (35) is able to reduce the weight of non-optimal relays in (37) to near-zero. Therefore, as long as the network is not congested, the source node can only get connected to the malicious nodes if the routing performance through the malicious nodes is no worse than the performance through any other neighbor nodes.

VI. CONCLUSION

In this paper, we have proposed a stochastic learning scheme for spectrum-aware, joint relay-channel selection in a multi-channel, multi-hop CRN. To address the potential vulnerabilities due to the combined Routing-toward-Primary-User (RPU) and Sink-Hole (SH) attack, we have formulated the distributed routing process as a stochastic game. By showing that the stochastic routing game can be decomposed into a group of single-state repeated games, we have proposed a Stochastic Fictitious Play (SFP) based relay selection algorithm based on limited information back propagation. We have also introduced a Multi-Arm Bandit (MAB) based truth-telling enforcement procedure for normal SUs to evaluate the trustworthiness of their candidate relays. With numerical simulations, we have demonstrated that the proposed routing algorithm is able to

reduce the average path delay by more than 30% compared to conventional routing mechanisms. Moreover, we have demonstrated that with the proposed learning algorithm, it is guaranteed that the routing performance is not affected by the inside attackers.

APPENDIX

A. Proof of Theorem 1

Let $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}_i, \{r_i\}_{i \in \mathcal{N}} P(s'|s, \mathbf{a}) \rangle$ represents a general-case average-reward recurrent stochastic game, where $r_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ($\mathcal{A} = \times \mathcal{A}_i$) is bounded and state transition probability $P(\cdot)$ is a function of all the players' joint action \mathbf{a} . Let R_i denote the expected average gain of player i given in (16) and g_i denote its expected bias value as in (17). Then, for game \mathcal{G} we have

Lemma 2 (Theorem 2.6 of [35]). *The joint strategy $\boldsymbol{\pi}^*$ is an average NE point iff the pair of $R_i(s, \boldsymbol{\pi}^*)$ and $g_i(s, \boldsymbol{\pi}^*)$ solves the following optimality equations for each play i :*

$$R_i(s, \boldsymbol{\pi}^*) = \max_{\boldsymbol{\pi}_i} \left\{ \sum_{s'} P(s'|s, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) R_i(s', \boldsymbol{\pi}^*) \right\}, \quad (38)$$

$$g_i(s, \boldsymbol{\pi}^*) = \max_{\boldsymbol{\pi}_i} \left\{ u_i(s, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) - R_i(s, \boldsymbol{\pi}^*) + \sum_{s'} P(s'|s, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) g_i(s', \boldsymbol{\pi}^*) \right\}. \quad (39)$$

According to Proposition 1, the state transition in game \mathcal{G}_r is independent of SU actions. Then, we readily obtain the two inequalities in (i) of Theorem 1 according to (39).

To prove (ii) in Theorem 1, we first consider the case of a normal SU. Based on Lemma 1, we can substitute $h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*)$ in (21) with (18) and obtain $\forall \boldsymbol{\pi}_i$:

$$u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) - U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}^*) \geq u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) - U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}^*). \quad (40)$$

From (40) we obtain $u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) \geq u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*)$, $\forall \boldsymbol{\pi}_i$, which is exactly the same as the condition equation for an NE in game $\mathcal{G}_r(\mathbf{o})$. For a malicious SU j , we can show $u_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}^*) \leq u_{\mathcal{P}_j}(\mathbf{o}, \boldsymbol{\pi}_j, \boldsymbol{\pi}_{-j}^*)$ similarly with the help of (22) in Theorem 1.

To show that the NE strategies for the stage game group $\mathcal{G}_r(\mathbf{o} : \forall \mathbf{o} \in \mathcal{O})$ constitute an NE strategy for \mathcal{G}_r , we rewrite (19) as follows:

$$U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \left(\sum_{n=0}^{\tau-1} P(\mathbf{o}'|\mathbf{o}) u_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}) \right). \quad (41)$$

Consider the case that $\boldsymbol{\pi}$ comprises of the NE strategies of the stage game groups, $\boldsymbol{\pi} = (\boldsymbol{\pi}^*(\mathbf{o}) : \forall \mathbf{o} \in \mathcal{O})$. If $\boldsymbol{\pi}$ is not an NE strategy of game \mathcal{G}_r , according to Definition 2, we can find at least

one SU i (assume that SU i is normal), satisfying the following inequality:

$$U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) - U_{\mathcal{P}_i}(\mathbf{o}, \tilde{\boldsymbol{\pi}}_i, \boldsymbol{\pi}_{-i}) < 0, \exists \tilde{\boldsymbol{\pi}}_i. \quad (42)$$

Then, after substituting $U_{\mathcal{P}_i}$ in (42) with (41), we have:

$$\lim_{\tau \rightarrow \infty} \frac{1}{\tau} \left(\sum_{n=0}^{\tau-1} P(\mathbf{o}'|\mathbf{o}) \times \left(u_{\mathcal{P}_i}(\mathbf{o}', \boldsymbol{\pi}^*(\mathbf{o}')) - u_{\mathcal{P}_i}(\mathbf{o}', \tilde{\boldsymbol{\pi}}_i(\mathbf{o}'), \boldsymbol{\pi}_{-i}^*(\mathbf{o}')) \right) \right) < 0, \quad (43)$$

which contradicts the fact that $\boldsymbol{\pi}^*(\mathbf{o})$ is the NE strategy of stage game $\mathcal{G}_r(\mathbf{o})$. Therefore, property (ii) of Theorem 1 holds.

B. Proof of Theorem 2

After exchanging the order of expectation and summation, we can expand (16) as:

$$\begin{aligned} U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) &= \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{n=0}^{\tau-1} E_{\mathbf{o}} \left(u_i(\mathbf{o}(n), \boldsymbol{\pi}) \mid \mathbf{o}(0) = \mathbf{o} \right) + \\ E_{\boldsymbol{\pi}_{i,1}} \left\{ \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \sum_{n=0}^{\tau-1} E_{\mathbf{o}} \left(u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}(n), a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \mid \mathbf{o}(0) = \mathbf{o} \right) \right\} &= U_i(\mathbf{o}, \boldsymbol{\pi}) + E_{\boldsymbol{\pi}_{i,1}} \left\{ U_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\}, \end{aligned} \quad (44)$$

where $\boldsymbol{\pi}_{i,1}$ is SU i 's strategy for choosing the next-hop SU. From (17) and (44), we obtain:

$$\begin{aligned} h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) &= \lim_{\tau \rightarrow \infty} \sum_{n=0}^{\tau-1} E_{\mathbf{o}} \left\{ u_i(\mathbf{o}(n), \boldsymbol{\pi}) + E_{\boldsymbol{\pi}_{i,1}} \left\{ u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}(n), a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\} \right. \\ &\quad \left. - U_i(\mathbf{o}, \boldsymbol{\pi}) - E_{\boldsymbol{\pi}_{i,1}} \left\{ U_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\} \mid \mathbf{o}(0) = \mathbf{o} \right\} \\ &= \lim_{\tau \rightarrow \infty} \sum_{n=0}^{\tau-1} E_{\mathbf{o}} \left\{ u_i(\mathbf{o}(n), \boldsymbol{\pi}) - U_i(\mathbf{o}, \boldsymbol{\pi}) \mid \mathbf{o}(0) = \mathbf{o} \right\} + \lim_{\tau \rightarrow \infty} \sum_{n=0}^{\tau-1} E_{\mathbf{o}, \boldsymbol{\pi}_{i,1}} \left\{ u_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}(n), a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right. \\ &\quad \left. - U_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \mid \mathbf{o}(0) = \mathbf{o} \right\} = h_i(\mathbf{o}, \boldsymbol{\pi}) + E_{\boldsymbol{\pi}_{i,1}} \left\{ h_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\}. \end{aligned} \quad (45)$$

Adding (44) and (45), we obtain:

$$\begin{aligned} h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) + U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) &= h_i(\mathbf{o}, \boldsymbol{\pi}) + U_i(\mathbf{o}, \boldsymbol{\pi}) + \\ E_{\boldsymbol{\pi}_{i,1}} \left\{ h_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) + U_{\mathcal{P}_{a_{i,1}}}(\mathbf{o}, a_{i,1}, \boldsymbol{\pi}_{i,2}, \boldsymbol{\pi}_{-i}) \right\}. \end{aligned} \quad (46)$$

After applying Lemma 1 to (46), (23) is obtained.

Consider a normal SU $i \in \mathcal{N}$. From (18), we can show that the best response of SU i to the joint strategy $\tilde{\boldsymbol{\pi}}_{-i}$ with respect to the sum of its bias value and gain value is obtained when

$$h_{\mathcal{P}_i}(\mathbf{o}, \tilde{\boldsymbol{\pi}}) + U_{\mathcal{P}_i}(\mathbf{o}, \tilde{\boldsymbol{\pi}}) = \max_{\boldsymbol{\pi}_i} \left(u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \tilde{\boldsymbol{\pi}}_{-i}) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}', \tilde{\boldsymbol{\pi}}) \right), \quad (47)$$

where $\tilde{\pi}$ is the solution to the right-hand side of (47). From (38) and (39) in Lemma 2, we have

$$h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) + \max_{\boldsymbol{\delta}} U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\delta}, \boldsymbol{\pi}_{-i}^*) = \max_{\boldsymbol{\pi}_i} \left(u_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*) + \sum_{\mathbf{o}'} P(\mathbf{o}'|\mathbf{o}) h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}^*) \right), \quad (48)$$

and $\boldsymbol{\pi}^* = (\boldsymbol{\delta}^*, \boldsymbol{\pi}_{-i}^*)$ is the NE strategy. For the malicious SUs, a similar pair of equations to (47) and (48) can be obtained by substituting operator $\max(\cdot)$ with $\min(\cdot)$ in (47) and (48). Comparing the right-hand side of (47) and (48), it is straightforward to show that the best response with respect to $h_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi}) + U_{\mathcal{P}_i}(\mathbf{o}, \boldsymbol{\pi})$ is also the NE strategy of the game.

C. Proof of Theorem 3

From [37], we introduce Lemma 3 in regard to the two timescale learning process in (26)-(28):

Lemma 3 (Theorem 5 of [37]). *Consider that in the following stochastic approximation processes*

$$\begin{cases} \theta_1^{n+1} = \theta_1^n + \gamma_1^n (F_1(\theta_1^n, \theta_2^n) + M_1^{n+1}), \\ \theta_2^{n+1} = \theta_2^n + \gamma_2^n (F_2(\theta_1^n, \theta_2^n) + M_2^{n+1}), \end{cases} \quad (49)$$

for each i , θ_i^n is bounded, $\sum_{n \rightarrow \infty} \gamma_i^n = \infty$, $\sum_{n \rightarrow \infty} (\gamma_i^n)^2 < \infty$, F_i is globally Lipschitz continuous, $\{\sum_{n=1}^k \gamma_i^n M_i^n\}_k$ converges almost surely, and $\lim_{n \rightarrow \infty} \gamma_1^n / \gamma_2^n = 0$. Suppose that for each θ_1 the Ordinary Differential Equation (ODE)

$$\frac{dY}{dt} = F_2(\theta_1, Y),$$

has a unique globally asymptotically stable equilibrium point $\xi(\theta_1)$ such that ξ is Lipschitz continuous. Then almost surely,

$$\lim_{n \rightarrow \infty} \|\theta_2^n - \xi(\theta_1^n)\| = 0,$$

and a suitable interpolation of the process $\{\theta_1^n\}$ is an asymptotic pseudo-trajectory of the flow defined by the ODE

$$\frac{dX}{dt} = F_1(X, \xi(X)).$$

Let $\{\tilde{u}_{\mathcal{P}_i}^n(\mathbf{o}, \mathbf{a}_i)\}$ in (26) be $\{\theta_2^n\}$ in (49) and $\{\tilde{\pi}_i^n(\mathbf{o})\}$ in (27) be $\{\theta_1^n\}$ in (49), then we define the following two ODEs:

$$\frac{d\tilde{u}_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i)}{dt} = F_2(\tilde{u}_{\mathcal{P}_i}, \tilde{\pi}_i) = u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i) - \tilde{u}_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i), \quad (50)$$

$$\frac{d\tilde{\pi}_i(\mathbf{o}, \mathbf{a}_i)}{dt} = F_1(\mathbf{u}_{\mathcal{P}_i}, \tilde{\pi}_i) = \text{BR}(\mathbf{u}_{\mathcal{P}_i}(\mathbf{o})) - \tilde{\pi}_i(\mathbf{o}, \mathbf{a}_i). \quad (51)$$

According to our discussion on (26), $\tilde{u}_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i)$ almost surely converges to $u_{\mathcal{P}_i}(\mathbf{o}, \mathbf{a}_i, \boldsymbol{\pi}_{-i})$. Then, by Lemma 3, a suitable interpolation of $\{\tilde{\pi}_i^n(\mathbf{o})\}$ is an asymptotic pseudo-trajectory of the flow defined by the ODE in (51). It is well known [37], [38] that (51) is equivalent to (52):

$$\frac{d\boldsymbol{\pi}_i(\mathbf{o})}{dt} = \overline{\text{BR}}(\boldsymbol{\pi}_{-i}(\mathbf{o})) - \boldsymbol{\pi}_i(\mathbf{o}). \quad (52)$$

where for a normal SU i (we omit the state indicator \mathbf{o} for simplicity)

$$\overline{\text{BR}}(\boldsymbol{\pi}_{-i}) = \arg \max_{\boldsymbol{\pi}_i} \left(u_{\mathcal{P}_i}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}) - \lambda_i \sum_{\mathbf{a}_i} \pi_i(\mathbf{a}_i) \log \pi_i(\mathbf{a}_i) \right), \quad (53)$$

and for a malicious SU j

$$\overline{\text{BR}}(\boldsymbol{\pi}_{-j}) = \arg \max_{\boldsymbol{\pi}_j} \left(u_{\mathcal{P}_j}^{-1}(\boldsymbol{\pi}_j, \boldsymbol{\pi}_{-j}) - \lambda_j \sum_{\mathbf{a}_j} \pi_j(\mathbf{a}_j) \log \pi_j(\mathbf{a}_j) \right), \quad (54)$$

because (28) provides the solutions to (53) and (54) [38]. In (53) and (54), the entropy function $v_i(\boldsymbol{\pi}_i) = -\sum_{\mathbf{a}_i} \pi_i(\mathbf{a}_i) \log \pi_i(\mathbf{a}_i)$ is called the perturbation in SFP. According to [38], we have

Lemma 4 (Proposition 3.1 of [38]). *Consider a general, normal-form repeated game $\mathcal{G} = \langle \mathcal{N}, \times_{i \in \mathcal{N}} \mathcal{A}_i, \{u_i\}_{i \in \mathcal{N}} \rangle$. Let $\hat{\pi}_i^n$ be the fixed point of the SFP dynamic given by (52) with respect to a perturbation vector $\mathbf{v}^n = (v_1^n, \dots, v_{|\mathcal{N}|}^n)$. If the perturbation sequence $\{\mathbf{v}^n\}$ converges weakly, and the sequence $\{\hat{\pi}_i^n\}$ converges to π_i^* , then π_i^* is the NE for \mathcal{G} .*

By Lemma 4, when the solution to the ODE in (52) converges to a fixed point, it converges to the NE of game $\mathcal{G}_r(\mathbf{o})$. Then, based on the discussion following Lemma 3, proving the convergence of the learning process given by (26)-(28) to the NE is equivalent to proving that the solution trajectories to the SFP dynamic in (51) converge to the set of fixed points from any initial condition. Observing the structure of $\mathcal{G}_r(\mathbf{o})$, the proof can be developed using the following properties of a repeated game:

Lemma 5 (Corollary 5.5 of [38]). *If a generic repeated game \mathcal{G} is a supermodular game, then the solutions to the smooth best response dynamic in the form of (52) for \mathcal{G} converges almost surely to its rest point set from any initial condition. The remaining nonconvergent initial conditions are contained in a finite or countable union $\cup_i M_i$, of invariant manifolds of codimension 1, and hence have measure zero.*

Lemma 6 (Supermodular game [40]). *A continuous normal-form game $\mathcal{G} = \langle \mathcal{N}, \{\boldsymbol{\Pi}_i\}_{i \in \mathcal{N}}, \{u_i(\boldsymbol{\pi}_i)\}_{i \in \mathcal{N}} \rangle$ is a supermodular game if for any player $i \in \mathcal{N}$,*

- i) the strategy space Π_i is a compact subset of \mathbb{R}^K .
- ii) the payoff function u_i is upper semi-continuous in $\pi_i = (\pi_i, \pi_{-i})$.
- iii) $\frac{\partial^2 u_i(\boldsymbol{\pi})}{\partial \pi_{i,k} \partial \pi_{j,l}} \geq 0 \quad \forall j \neq i, k, l$, where $\pi_{i,k}$ is the k -th element of vector π_i .

With Lemma 6, we can check the supermodularity of game $\mathcal{G}_r(\mathbf{o})$ with respect to strategy π_i . According to (13), we have $u_i \geq 0 \quad \forall \mathbf{o}, \mathbf{a}$. Then, according to (20), $\forall i \neq j$

$$\frac{\partial^2 (u_{\mathcal{P}_i} = \sum_{i \in \mathcal{P}_i} u_i(\boldsymbol{\pi}))}{\partial \pi_i(\mathbf{a}_i) \partial \pi_j(\mathbf{a}_j)} \geq 0, \forall \mathbf{a}_i, \mathbf{a}_j. \quad (55)$$

Therefore, game $\mathcal{G}_r(\mathbf{o})$ in the form of continuous game¹ with strategy $\boldsymbol{\pi}$ is a supermodular game. By Lemma 5, the smooth best response dynamic converges almost surely. By Lemma 4, Theorem 3 is proved.

D. Proof of Theorem 4

The proof of Theorem 4 can be achieved by applying Lemma 3 repeatedly to the learning scheme given by (29) and (30), then to the learning scheme given by (30) and (31). According to our discussion on (29), $\tilde{u}_i^n(\mathbf{o}, \mathbf{a}_i)$ has a unique globally asymptotically stable equilibrium $u_i(\mathbf{o}, \mathbf{a}_i, \boldsymbol{\pi}_{-i})$ if $\boldsymbol{\pi}$ is fixed. Then, it is sufficient to prove that the following ODE:

$$\frac{d\tilde{u}_{\mathcal{P}_i}(\mathbf{o})}{dt} = \left(\sum_{\mathbf{a}_i} \tilde{\pi}_i(\mathbf{o}, \mathbf{a}_i) (\tilde{u}_i(\mathbf{o}, \mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}(\mathbf{o})) - \tilde{u}_{\mathcal{P}_i}(\mathbf{o}) \right), \quad (56)$$

is globally asymptotically stable to show that the learning process given by (29)-(31) produces an asymptotic pseudo-trajectory of the SFP flow. Omitting state indicator \mathbf{o} for convenience, we denote $\hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) = \tilde{u}_i(\mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}$, $\xi_i = \frac{d\tilde{u}_{\mathcal{P}_i}}{dt}$ and $\epsilon(\mathbf{a}_i) = \frac{d\hat{u}_{\mathcal{P}_i}(\mathbf{a}_i)}{dt}$, and define a Lyapunov function:

$$V_i(t) = \left(\sum_{\mathbf{a}_i} \tilde{\pi}_i(\mathbf{a}_i) (\tilde{u}_i(\mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i,1}}) - \tilde{u}_{\mathcal{P}_i} \right)^2. \quad (57)$$

We sort the SUs in path \mathcal{P}_i according to their distance in hop count to sink L in an ascending order as $\{L-1, L-2, \dots, i\}$. Then, the two-timescale stochastic approximation process in Lemma 3 can be extended to multiple-timescale with the same form of function F_i as in (49):

$$\begin{cases} F_1(\tilde{u}_j(\mathbf{a}_j), \tilde{\boldsymbol{\pi}}) = u_j(\mathbf{a}_j(n)) - \tilde{u}_j^n(\mathbf{a}_j), \\ F_2^j(\tilde{u}_j(\mathbf{a}_j), \tilde{u}_{\mathcal{P}_j}, \tilde{u}_{\mathcal{P}_{j+1}}) = \sum_{\mathbf{a}_i} \tilde{\pi}_i(\mathbf{a}_i) (\tilde{u}_i(\mathbf{a}_i) + \tilde{u}_{\mathcal{P}_{\mathbf{a}_i}}) - \tilde{u}_{\mathcal{P}_i}. \end{cases} \quad (58)$$

¹Such property also holds for malicious SUs as long as their strategy learning scheme complies with SFP given by (28).

Since the learning process in (29) is globally asymptotically convergent, then at the stable point of \tilde{u}_i , $\frac{d\tilde{u}_i}{dt}=0$ and $\epsilon_i(\mathbf{a}_i)=\frac{d\tilde{u}_{\mathcal{P}_i+1}}{dt}$, where $\mathbf{a}_{i,1}=i+1$. We now examine V_i and obtain

$$\begin{aligned} \frac{1}{2} \frac{dV_i}{dt} &= \left(\sum_{\mathbf{a}_i} \tilde{\pi}_i(\mathbf{a}_i) \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) - \tilde{u}_{\mathcal{P}_i} \right) \times \left(\frac{d}{dt} \sum_{\mathbf{a}_i} \frac{e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i)}}{\sum_{\mathbf{b}} e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{b})}} \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) - \frac{d\tilde{u}_{\mathcal{P}_i}}{dt} \right), \\ &= \xi_i \left(\sum_{\mathbf{a}_i} \left(\sum_{\mathbf{b}} \frac{\lambda_i e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) \hat{u}_{\mathcal{P}_i}(\mathbf{b})}}{\left(\sum_{\mathbf{b}} e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{b})} \right)^2} (\epsilon_i(\mathbf{a}_i) - \epsilon_i(\mathbf{b})) \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) + \frac{e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i)}}{\sum_{\mathbf{b}} e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{b})}} \epsilon_i(\mathbf{a}_i) \right) - \xi_i \right), \\ &= \lambda_i \sum_{\mathbf{a}_i} \sum_{\mathbf{b}} \frac{e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i)}}{\sum_{\mathbf{b}} e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{b})}} (\epsilon_i(\mathbf{a}_i) - \epsilon_i(\mathbf{b})) \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i) \xi_i + \sum_{\mathbf{a}_i} \frac{e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{a}_i)}}{\sum_{\mathbf{b}} e^{\lambda_i \hat{u}_{\mathcal{P}_i}(\mathbf{b})}} \epsilon_i(\mathbf{a}_i) \xi_i - \xi_i^2. \end{aligned}$$

We start examining the property of $\frac{dV_i}{dt}$ in the way of backward propagation from SU $L-1$. Since $\tilde{u}_{\mathcal{P}_L}=0$, we have $\epsilon_{L-1}(\mathbf{a}_{L-1})=0$, hence $\frac{1}{2} \frac{dV_{L-1}}{dt} = -\xi_{L-1}^2 \leq 0$ at the stable point of the approximation process represented by $F_2^{L-1}(\tilde{u}_{L-1}, \tilde{\pi})$. Therefore, the ODE for SU $L-1$ in the form of (56) is globally asymptotically convergent. Then, we can apply Lemma 3 to the two-timescale learning process featured by F_2^{L-1} and F_2^{L-2} , and show that a suitable interpolation of the process $\{\tilde{u}_{\mathcal{P}_{L-2}}^n\}$ is an asymptotic pseudo-trajectory of the flow defined by the ODE $\frac{d\tilde{u}_{\mathcal{P}_{L-2}}}{dt}$ given in (56). At the stable point of $\tilde{u}_{\mathcal{P}_{L-1}}$, we have $\frac{d\tilde{u}_{\mathcal{P}_{L-1}}}{dt}=0$, so $\epsilon_{L-2}(\mathbf{a}_{L-2})=0$. With the similar way to analyzing $\frac{dV_{L-1}}{dt}$, we have $\frac{1}{2} \frac{dV_{L-2}}{dt} = -\xi_{L-2}^2 \leq 0$. By repeatedly applying the same analysis to the sequence of the learning processes featured by $\{(F_2^{L-1}, F_2^{L-2}), (F_2^{L-2}, F_2^{L-3}), \dots, (F_2^{i+1}, F_2^i)\}$, we obtain Lemma 7:

Lemma 7. *The learning process given in (29) and (30) is globally asymptotically convergent, provided that the following are satisfied: $\lim_{n \rightarrow \infty} \sum_n \alpha(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \alpha^2(n) < \infty$, $\lim_{n \rightarrow \infty} \sum_n \gamma_i(n) = \infty$, $\lim_{n \rightarrow \infty} \sum_n \gamma_i^2(n) < \infty$, $\lim_{n \rightarrow \infty} (\gamma_i(n)/\alpha(n)) = 0$ and $\lim_{n \rightarrow \infty} (\gamma_i(n)/\gamma_j(n)) = 0$, if SU i is closer to the sink SU than SU j in terms of hop count.*

If $\lim_{n \rightarrow \infty} (\beta(n)/\gamma_i(n)) = 0$, we can further conclude that the learning process given by (31) and (32) yields an asymptotic pseudo-trajectory of the flow defined by the SPF-based ODE. We note that Lemma 5 and Lemma 6 still hold for a new game with the utility of each player being the convergent biased value estimation. Then, Theorem 4 is proved.

REFERENCES

- [1] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 1, pp. 116–130, First Quarter 2009.
- [2] M. Cesana, F. Cuomo, and E. Ekici, "Routing in cognitive radio networks: Challenges and solutions," *Ad Hoc Networks.*, vol. 9, no. 3, pp. 228–248, May 2011.

- [3] M. Youssef, M. Ibrahim, M. Abdelatif, L. Chen, and A. Vasilakos, "Routing metrics of cognitive radio networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 92–109, First Quarter 2014.
- [4] G. Baldini, T. Sturman, A. Biswas, R. Leschhorn, G. Godor, and M. Street, "Security aspects in software defined radio and cognitive radio networks: A survey and a way ahead," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 2, pp. 355–379, Second Quarter 2012.
- [5] K. J. R. Liu and B. Wang, *Cognitive radio networking and security: A game-theoretic view*. Cambridge University Press, 2010.
- [6] R. Chen, J.-M. Park, and J. Reed, "Defense against primary user emulation attacks in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 25–37, Jan. 2008.
- [7] N. Hu, Y.-D. Yao, and J. Mitola, "Most active band (mab) attack and countermeasures in a cognitive radio network," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 898–902, Mar. 2012.
- [8] B. Wang, Y. Wu, K. J. R. Liu, and T. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 877–889, Apr. 2011.
- [9] H.-P. Shiang and M. van der Schaar, "Distributed resource management in multihop cognitive radio networks for delay-sensitive transmission," *IEEE Trans. Veh. Technol.*, vol. 58, no. 2, pp. 941–953, Feb. 2009.
- [10] D. Xue and E. Ekici, "Guaranteed opportunistic scheduling in multi-hop cognitive radio networks," in *2011 IEEE Proceedings INFOCOM*, Shanghai, China, Apr. 2011, pp. 2984–2992.
- [11] Y. Hou, Y. Shi, and H. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 146–155, Jan. 2008.
- [12] M. Pan, H. Yue, C. Zhang, and Y. Fang, "Path selection under budget constraints in multihop cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 6, pp. 1133–1145, Jun. 2013.
- [13] I. Filippini, E. Ekici, and M. Cesana, "A new outlook on routing in cognitive radio networks: Minimum-maintenance-cost routing," *IEEE/ACM Trans. Netw.*, vol. 21, no. 5, pp. 1484–1498, Oct. 2013.
- [14] L. Ding, T. Melodia, S. Batalama, J. Matyjas, and M. Medley, "Cross-layer routing and dynamic spectrum allocation in cognitive radio ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1969–1979, May 2010.
- [15] M. Caleffi, I. Akyildiz, and L. Paura, "Opera: Optimal routing metric for cognitive radio ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2884–2894, Aug. 2012.
- [16] Y. Liu, L. Cai, and X. Shen, "Spectrum-aware opportunistic routing in multi-hop cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1958–1968, Nov. 2012.
- [17] Z. Yang, G. Cheng, W. Liu, W. Yuan, and W. Cheng, "Local coordination based routing and spectrum assignment in multi-hop cognitive radio networks," *Mob. Netw. Appl.*, vol. 13, no. 1-2, pp. 67–81, Apr. 2008.
- [18] A. Aijaz, H. Su, and A.-H. Aghvami, "Corpl: A routing protocol for cognitive radio enabled ami networks," *IEEE Transactions on Smart Grid*, vol. 6, no. 1, pp. 477–485, Jan. 2015.
- [19] C. Pandana, Z. Han, and K. J. R. Liu, "Cooperation enforcement and learning for optimizing packet forwarding in autonomous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 8, pp. 3150–3163, Aug. 2008.
- [20] Q. Zhu, Z. Yuan, J. B. Song, Z. Han, and T. Basar, "Interference aware routing game for cognitive radio multi-hop networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 2006–2015, Nov. 2012.
- [21] Y. Song, C. Zhang, and Y. Fang, "Stochastic traffic engineering in multihop cognitive wireless mesh networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 3, pp. 305–316, Mar. 2010.

- [22] B. Kannhavong, H. Nakayama, Y. Nemoto, N. Kato, and A. Jamalipour, "A survey of routing attacks in mobile ad hoc networks," *IEEE Wireless Commun. Mag.*, vol. 14, no. 5, pp. 85–91, Oct. 2007.
- [23] B. Wu, J. Chen, J. Wu, and M. Cardei, "A survey of attacks and countermeasures in mobile ad hoc networks," in *Wireless Network Security*, ser. Signals and Communication Technology, Y. Xiao, X. Shen, and D.-Z. Du, Eds. Springer US, 2007, pp. 103–135.
- [24] H. Yih-Chun and A. Perrig, "A survey of secure wireless ad hoc routing," *IEEE Security Privacy*, vol. 2, no. 3, pp. 28–39, May 2004.
- [25] S. Marti, T. J. Giuli, K. Lai, and M. Baker, "Mitigating routing misbehavior in mobile ad hoc networks," in *Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, Aug. 2000, pp. 255–265.
- [26] B. Cui and S. Yang, "Nre: Suppress selective forwarding attacks in wireless sensor networks," in *IEEE Conference on Communications and Network Security*, San Francisco, CA, Oct. 2014, pp. 229–237.
- [27] M. Kodialam and T. V. Lakshman, "Detecting network intrusions via sampling: a game theoretic approach," in *Annual Joint Conference of the IEEE Computer and Communications*, vol. 3, San Francisco, CA, Apr. 2003, pp. 1880–1889 vol.3.
- [28] S. Bohacek, J. Hespanha, J. Lee, C. Lim, and K. Obraczka, "Game theoretic stochastic routing for fault tolerance and security in computer networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 9, pp. 1227–1240, Sep. 2007.
- [29] Q. Zhu, J. B. Song, and T. Basar, "Dynamic secure routing game in distributed cognitive radio networks," in *IEEE Global Telecommunications Conference*, Houston, Texas, Dec. 2011.
- [30] Z. Yuan, Z. Han, Y. Sun, H. Li, and J. B. Song, "Routing-toward-primary-user attack and belief propagation-based defense in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 9, pp. 1750–1760, Sep. 2013.
- [31] S. Geirhofer, L. Tong, and B. Sadler, "Cognitive medium access: Constraining interference based on experimental models," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 95–105, Jan. 2008.
- [32] T. Chen, H. Zhang, G. Maggio, and I. Chlamtac, "Cogmesh: A cluster-based cognitive radio network," in *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, Dublin, Ireland, Apr. 2007, pp. 168–178.
- [33] Q. Zhao, S. Geirhofer, L. Tong, and B. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 785–796, Feb. 2008.
- [34] R. Draves, J. Padhye, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in *Proceedings of the International Conference on Mobile Computing and Networking*, New York, NY, USA, Sep. 2004, pp. 114–128.
- [35] E. Altman, A. Hordijk, and F. M. Spieksma, "Contraction conditions for average and α -discount optimality in countable state markov games with unbounded rewards," *Math. Oper. Rre.*, vol. 22, no. 3, pp. 588–618, Aug. 1997.
- [36] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley and Sons, Inc, 2008.
- [37] D. S. Leslie and E. Collins, "Convergent multiple-timescales reinforcement learning algorithms in normal form games," *Ann. Appl. Probab.*, vol. 13, no. 4, pp. 1231–1251, Feb. 2003.
- [38] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, Nov. 2002.
- [39] S. L. Scott, "A modern bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, Nov. 2010.
- [40] Z. Han, D. Niyato, W. Saad, T. Basar, and A. Hjørungnes, *Game theory in wireless and communication networks*. Cambridge University Press, 2012.