Online Primary User Emulation Attacks in Cognitive Radio Networks Using Thompson Sampling

Xiang Sheng and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—Spectrum sensing is one of the main components of the cognitive radio (CR) system, based on which secondary users (SUs) can get access to spectrum holes available. On the other hand, a malicious adversary can also attack the primary user (PU) system and the legitimate CR system via spectrum sensing, which can lead to serious security issue for both systems. In this article, we study an online attacking strategy, referred to as PU emulation attacker (PEA), which transmits forged PU signals over available channels to deteriorate the spectrum sensing performance of the SUs. We propose an online learning based attacking scheme for both the single attacker and the multiple-attacker cases, and analyze the regret upper bound of the proposed algorithm. The proposed PEA strategy can work in both stationary and non-stationary CR networks where the statistical characteristics of channels and the access strategy of SUs change over time. Numerical results show that it is more efficient than others in two different performance metrics: successful accesses of SUs and effective attacks of PEAs. Our proposal raises an interesting open question on how to develop CR networks with security guarantee.**

***Index Terms*—Cognitive radio, online learning, PU emulation attack, spectrum sensing, Thompson sampling.**

I. Introduction

Cognitive radio (CR) is a promising technology to address the growing shortage of spectrum resources and has been widely studied in the past two decades [1], [2]. In a CR network, the secondary users (SUs) identify radio spectrum environment via spectrum sensing and exploit spectrum holes opportunistically. However, it is a hard task to prevent spectrum sensing mechanisms from being abused by the malicious attackers, from the aspects of both the licensed primary users (PUs) and the legitimate SUs [3]–[5]. PU emulation attacker (PEA) [6] is one of the typical malicious attackers [7], [8], which changes radio spectrum environment by transmitting forged PU signals so as to undermine the spectrum sensing process of the legitimate SUs. Note that the PEA does not damage the PU system so as to ensure its concealment and energy efficiency.

Considering a CR network being suffered from the PEA, if an SU senses the vacant channel being attacked on the spectrum sensing stage, it would mistakenly believe that a PU is using the channel and keep silent so as not to interfere with the PU. In this case, the attack is effective since it prevents the SU from exploiting the vacant channels and gives the SUs incorrect information. In other case, if no SU tries to sense the attacked channel or a PU occupies this channel, the attack will not have any impact on the SUs and is taken as an ineffective attack. Obviously, the necessary but not sufficient condition for an effective attack should focus on the spectrum holes, i.e., the channels not used by the PUs. However, the fact of the matter is that the PEA does not have any prior knowledge of the radio spectrum environment and is likely to snip those occupied channels instead of the vacant ones. Therefore, the attacking strategy of the PEA, which refers to dynamically determining which channel to attack in each time slot, significantly affects the attacking effectiveness and efficiency. In another perspective, full understanding of the attacking strategy can also help the PUs/SUs to quantify the impact of the PEA attacks on the performance of the SUs, which is essential for the assessment of the corresponding detection and defending strategies. Moreover, a smart PEA can also be modified into a new powerful jammer in the tactical networks and guide the design of security schemes in the CR systems.

The main factors determining the effectiveness of the PEA attacks include the positions of attackers, the signal powers and the mobility speeds. A few targeted methods have been proposed to fight against these attacks. In [9], a transmitter verification scheme is proposed to verify whether the received signals are transmitted by a PU or not by estimating its location with an additional wireless sensor network. In [10], an energy detection approach is proposed to classify the received signals by feeding the power features into a trained neural network. In [11], a belief propagation based defense strategy is developed, where an SU exchanges messages among his neighbors and calculates the local belief by estimating both location and mobile speed of the signal transmitters. In [12], the activity pattern of received signal, such as the ON and OFF period, is used to reconstruct the PU behavior model. By calculating the reconstruction error, it is possible to distinguish the PU from an attacker. In [13], a channel surveillance protocol is proposed to combat the selfish attacker, where an extra-sensing process is used to check whether the channel is attacked or not.

However, the aforementioned schemes have not taken the changes of PEA attacking strategy into consideration, which means that the PEA can use adaptive attacking schemes ac-

Manuscript received August 8, 2020; revised April 8, 2021; accepted June 11, 2021. This work was partially supported by the National Natural Science Foundation of China under Grants 61931023 and U1936202. The associate editor coordinating the review of this article and approving it for publication was Sudharman K. Jayaweera. *(Corresponding author: Shaowei Wang.)*

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (email: 151180109@smail.nju.edu.cn, wangsw@nju.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3091588, IEEE Transactions on Wireless Communications

cording to the behaviors of the PUs and SUs. In [14], partially observable Markov decision process has been introduced to devise the PEA attacking strategy, which depends crucially on the feedback of the attack, i.e. the PEA should know whether an SU ever attempted to sense and access the attacked channel or not. In [15], three attacking strategies including uniformly random, selectively random and maximal interception attacks, have been proposed to evaluate the performance of the proposed SU passive defend approach, where the PEA is assumed to be aware of the parameters of SU's access strategy at the beginning of each time slot. In [16], a PEA attacking strategy, referred to as play and random observe learning algorithm, is proposed for the CR network. The attacker cannot know whether its attack is effective or not because this attack happens on the spectrum sensing stage. Note that the PEA has the ability of channel measurement and identification of PU signal features so as to create forged PU signals. Therefore, the channel measurement capability of the attacker can be employed to observe the state of partial channels in each time slot.

In this paper, we study the PEA attacking strategy with unknown and changing statistical characteristics of channels, where we also take the access strategy of the legitimate SUs into consideration. If the statistical characteristics of channel states are stable, we call it a stationary CR network. If the statistical characteristics of channel states change over time, we call it a non-stationary CR network. The key challenge for the PEA is how to deal with the uncertainties of both the licensed channels and the behavior of the SUs. We formulate the decision task as a multi-armed bandit (MAB) problem, for which an online learning algorithm, named as γ-discounted Thompson sampling, is proposed to attack a set of channels and observe the communication on some channel in each time slot. The online attacking method can keep balance between the exploitation of well-performing channels and the exploration of channels. Analysis shows that its performance approaches that of attacking the best constant channel chosen in hindsight as time goes on. Numerical results confirm the effectiveness of our method in both stationary channel and non-stationary channel scenarios. The main contributions of this paper are summarized as follows:

- We analyze the impact of both stationary and non-stationary CR networks on the design of PEA strategy, and present an MAB framework to model the general optimization task.
- We develop an online attacking strategy based on Thompson sampling. Theoretical analysis and numerical results show that its performance outperforms others in stationary channel scenarios.
- We prove the regret upper bound of online attacking strategy is in the order of $\tilde{O}(\sqrt{T})$, which improves on the state-of-the-art theoretical result [16], and implies that its performance approaches the optimal over time.
- We extend the proposed online attacking strategy to the non-stationary channel scenario and the multiple-attacker case. Numerical results show it can partially track the time-variant characteristics of the CR network and effectively coordinate the multiple attackers.

The remainder of our paper is organized as follows. We illustrate our system model and formulate the optimization task in Section II. In Section III, we develop an online attacking strategy for both the single attacker and the multiple-attacker cases. Section IV provides the performance analysis for our proposed online single attacking strategy. In Section V, we evaluate the performance of the existing and proposed methods via a series of numerical experiments. Conclusions and future work are given in Section VI.

II. System Model and Problem Formulation

A. System model

Consider a network with N SUs, M PEAs and K channels, which operates in time slot mode. The activities of PUs stay the same in each time slot. In other words, the idle/busy state of all channels is determined by the PUs at the beginning of each time slot, and remains unchanged in the rest of this time slot. The PEA is essentially a malicious SU who wants to benefit from blocking the access of other legitimate SUs [17]. Assume the framework of CR network is open, so it is possible for the PEA to synchronize with the legitimate SUs. As can be seen from Fig. 1, the base station (BS) serving PUs transmits data to PU_1, PU_2 via channel f_1, f_3, respectively. Channels f_2 and f_4 are idle at this moment. SU_1 and SU_2 served by the access point (AP) of the CR network can find these two idle channels by spectrum sensing and use these channels for transmissions. However, a PEA, referred to as PEA_1, tries to prevent the SUs from accessing the idle channel f_2 by transmitting forged PU signals over channel f_2 , which makes the SUs mistake it for a PU. This example shows that the appearance of PEA attacks can severely impact on the performance of CR networks.

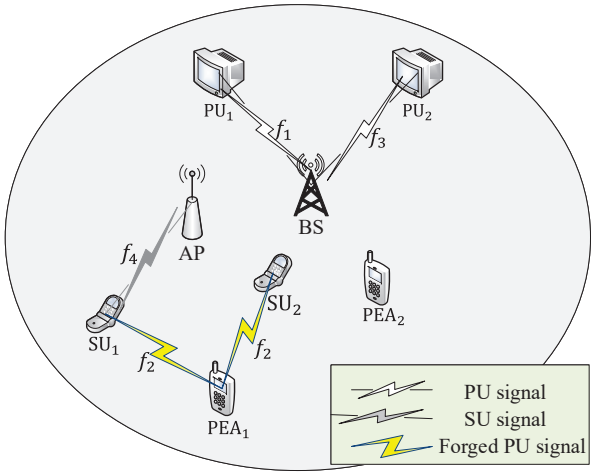

Fig. 1. Illustration of cognitive radio network.

Recall that the PUs can utilize the licensed channels at any time. In each time slot, a PU may transmit data or not and the activities of PUs determine the idle/busy state of the licensed channels. To simplify analysis, we focus on the channels and do not specify the behavioral pattern of PUs, i.e., we do not know the statistical information about the

access mode of the PUs. The set of all licensed channels is denoted by $\mathcal{K} = \{1, 2, ..., K\}$. The state of channels in time slot t has a total of 2^K possibilities and can be denoted as $\mathcal{D}_t = [D_t(1), \cdots, D_t(K)]$, where $D_t(i) \in \{0, 1\}, i = 1, 2, ..., K$, $D_t(i) = 1$ if channel i is idle; $D_t(i) = 0$ if channel i is occupied by a PU. In addition, the channel state is independent of each other. It is worth noting that the statistical characteristics of channel states are time-varying in this work.

The set of all SUs is denoted by $\mathcal{N} = \{1, 2, ..., N\}$. The SU sequentially performs spectrum sensing and data transmission in each time slot. SU i in time slot t chooses only one channel, denoted by $Q_{t,i}$ ($Q_{t,i} \in \mathcal{K}$), to sense due to its limited capability of sampling. If the SU finds the sensed channel $Q_{t,i}$ unoccupied by the PUs, it transmits data over this channel; otherwise, it keeps silent in the rest of this time slot so as not to interfere with the PUs. If two or more SUs sense and access the same channel, the transmissions will collide. We use predefined control channel to exchange information among the SUs same as discussed in [15], so that N independent SUs are equivalent to an SU who can access N channels in a time slot. The set of all channels sensed in time slot t can be denoted by $\mathcal{Q}_t = \{Q_{t,1}, Q_{t,2}, ..., Q_{t,N}\}$.

The set of all PEAs is denoted by $\mathcal{M} = \{1, 2, ..., M\}$. Each PEA is equipped with one antenna, which means that it can only receive or transmit signals on one channel alternately. The PEAs sequentially conduct attacking, observing, and learning in a given time slot. PEA j in time slot t chooses channel $I_{t,j}$ ($I_{t,j} \in \mathcal{K}$) to attack. The set of all channels attacked by the PEAs in time slot t is $\mathcal{I}_t = \{I_{t,1}, I_{t,2}, ..., I_{t,M}\}$. As discussed above, if the attacked channel is vacant and just sensed by an SU in this time slot, the attack is effective. The number of successful attacks in time slot t is denoted by

$$A(t) = \sum_{j=1}^{M} \frac{D_t(I_{t,j})}{n_{t,j}} \mathbf{1}[I_{t,j} \in \mathcal{Q}_t], \tag{1}$$

where $\mathbf{1}[x]$ is an indicator function, $\mathbf{1}[x] = 1$ if x occurs; $\mathbf{1}[x] = 0$ if x does not occur, $n_{t,j}$ is the total number of PEAs attacking channel $I_{t,j}$. Note that the PEA does not know whether its attack works or not. That is, in the attacking phase, the PEA cannot know whether an SU is sensing the attacking channel or not. Therefore, we utilize the channel measurement capability of the PEA to monitor the communication on some channels, and then obtain the environment information including channel idle states and the behavior of SUs. Specifically, the PEAs employ an observing stage, i.e., each PEA keeps silent in the rest of this time slot and observes the state of at least one channel. The number of observed channels is observation capability λ. The set of channels observed by attacker j in time slot t is denoted by $\mathcal{J}_{t,j}$, which is a subset of $\mathcal{K}$ with λ elements. Finally, the PEA adjusts the attacking strategy based on the spectrum environment information observed. We use the effective attacks of PEAs to quantify the impact of PEA attacks as follows:

$$A_T = \sum_{t=1}^{T} \sum_{j=1}^{M} \frac{D_t(I_{t,j})}{n_{t,j}} \mathbf{1}[I_{t,j} \in \mathcal{Q}_t]. \tag{2}$$

We summarize the notations in Table 1.

TABLE I
NOTATIONS

Notation	Definition
$\mathcal{K}$	Set of all licensed channels
$D_t(i)$	State of the ith channel in time slot t
$\mathcal{N}$	Set of all SUs
$Q_{t,i}$	The channel sensed by ith SU in time slot t
$\mathcal{M}$	Set of all PEAs
$\mathcal{I}_t$	Set of all channels attacked by PEAs in time slot t
$I_{t,j}$	The channel attacked by jth PEA in time slot t
$\mathcal{J}_{t,j}$	Set of channels observed by jth PEA in time slot t
λ	Observation capability
$\mathbf{1}[\cdot]$	The indicator function
A_T	Effective attacks of PEAs
G_T	Successful accesses of SUs
$r_t(k)$	Reward of kth channel in time slot t
φ	Strategy of the formulated MAB problem
Γ	Gamma function
$\mathbf{P}_k$	Transition matrix of kth channel
γ	Discount factor

B. Problem formulation

To maximize A_T, the attacking strategy needs to not only decide the attacking channel set $\mathcal{I}_t$ that helps to maximize the current number of successful attacks $A(t)$, but also decide the observing channel set $\mathcal{J}_{t,j}$ for each attacker that helps the following decision making by reducing channel uncertainty. The number of successful attacks obtained in each time slot depends on both the unknown time-varying probability of channels being idle and the unknown behaviors of the SUs, which should be learned in an online manner.

In an MAB problem, a gambler plays a slot machine with multiple arms, and tries to maximize its total expected reward in a series of trials [18]. In each round, the gambler chooses only one arm to pull and receives a random reward drawn from a fixed distribution with unknown mean. Mathematically, suppose there are K arms, and the reward of any arms is either 1 or 0. The reward of arm $k \in \{1, .., K\}$ is 1 with probability $0 \leq \theta_k \leq 1$. The probability of success for the arms $\theta = (\theta_1, ..., \theta_K)$ is unknown to the gambler but can be learned online.

The PEA attacking strategy problem can be formulated as an MAB problem, where K licensed channels can be regard as K arms. The number of successful attack obtained by selecting channel k follows a fixed distribution with unknown parameter θ_k. Specifically, in time slot t, the reward of channel $k \in \mathcal{K}$ is defined as the corresponding times of successful attack, which can be denoted by

$$r_t(k) = \begin{cases} 1 & \text{if } k \in Q_t \text{ and } D_t(k) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The strategy of the formulated MAB problem can be denoted as $\varphi = \{\varphi(t, j)\}_{t \geq 1, j \in \mathcal{M}}$, in which each element $\varphi(t, j) : \mathcal{K} \rightarrow (I_{t,j}, \mathcal{J}_{t,j})$ is only based on the previous observed rewards $\{r_m(n)\}_{t > m \geq 1, n \in \mathcal{J}_{m,j}}$. Formally, we target at finding the optimal strategy φ^*:

$$\varphi^* := \arg\max_{\varphi} \sum_{t=1}^{T} \sum_{j=1}^{M} \frac{r_t\left(I_{t,j}\right)}{n_{t,j}}. \tag{4}$$

III. Online attacking strategy

Thompson sampling (TS) is an efficient algorithm that can handle the MAB problems [19], [20]. It has been widely applied primarily due to its ability to deal with the tradeoff between the exploitation of the high-performance arms and the exploration of the uncertain arms. The key idea of TS is to track the probability of success for each arm in an online manner, and then randomly choosing each arm based on its posterior probability of being optimal. Given a prior distribution $P(\theta)$ on parameter θ, the posterior probability of θ under the history of past experiments $\mathcal{B}$ is updated by the Bayes rule,

$$P(\theta|\mathcal{B}) \propto \prod P\left(r_t(I_t)|I_t,\theta\right) P(\theta). \tag{5}$$

where $\mathcal{B}$ is made of tuples $(I_t, r_t(I_t))$ and $r_t(I_t)$ is the reward obtained by pulling arm I_t.

After estimating the posterior probability of θ, the gambler can calculate the probability of each arm being optimal. Specifically, the probability of arm k being optimal given $P(\theta|\mathcal{B})$ is:

$$\int \mathbf{1}\left[\mathbb{E}(r_t(k)|k,\theta) = \max_{k'} \mathbb{E}\left(r_t(k')|k',\theta\right)\right] P(\theta|\mathcal{B}) d\theta \tag{6}$$

Note that it is unnecessary to compute explicitly the mentioned probability: it suffices to sample θ in each round and selects the arm with the largest expected reward as explained in **Algorithm 1**, where the sampling result of θ in round t is denoted by $\theta(t)$. Both regret analysis and empirical results of TS show that it is highly competitive in general cases [21], [22].

Algorithm 1 Thompson sampling

Initialization: $\mathcal{B} = \emptyset$.
for $t = 1, 2, ..., T$ **do**
  Draw $\theta(t) \sim P(\theta|\mathcal{B})$.
  Select arm $I_t := \arg\max_k \mathbb{E}(r_t(k)|k, \theta(t))$.
  Observe reward r_t
  $\mathcal{B} = \mathcal{B} \cup (I_t, r_t(I_t))$.
end for

A. PEA attacking for single attacker case

For the single attacker case, the PEA acts as the gambler in the MAB problem and the CR system with K channels for sensing is a slot machine with K arms. The objective of the PEA is to choose the optimal channel which yields the greatest reward. For each channel, its potential reward is a random variable obeying an unknown distribution. Specifically, for channel k, $r_t(k)$ can be approximated by sampling from a random variable obeying *Bernoulli* distribution, in which the probability distribution is given by:

$$P(r_t(k)|\theta_k) = \begin{cases} \theta_k & r_t(k) = 1, \\ 1 - \theta_k & r_t(k) = 0. \end{cases} \tag{7}$$

Note that the parameters $(\theta_1, ..., \theta_K)$ are unknown to the PEA but can be learned online. In order to track the parameters $(\theta_1, ..., \theta_K)$, the PEA gives a prior distribution on those parameters. Specifically, the value of θ_k obeys *Beta* distribution *Beta*(S_k, F_k), the probability distribution of which is given by:

$$P(\theta_k) = \frac{\Gamma(S_k + F_k)}{\Gamma(S_k)\,\Gamma(F_k)} \theta_k^{S_k - 1} (1 - \theta_k)^{F_k - 1}, \tag{8}$$

where Γ denotes the gamma function, S_k and F_k are the parameters of *Beta* distribution. The conjugacy properties of *Beta* distribution make its posterior distribution also *Beta* distribution whose parameters can be simply calculated. Either one of S_k and F_k can be increased by 1 to show its posterior distribution. Note that *Beta* distribution represents a family of continuous probability distribution defined on the interval $[0, 1]$. The mean of θ_k is $S_k/(S_k + F_k)$ and its variance decreases with $S_k + F_k$. Therefore, *Beta* distribution is very convenient to be the prior distribution for Bernoulli reward case where the rewards are either 0 or 1. Our proposed online attacking strategy, called as γ-discounted TS (γ-DTS) attacking strategy, consists of two phases: attacking and observing, which is illustrated in **Algorithm 2**.

Algorithm 2 γ-discounted TS attacking strategy

Parameters: $\gamma \in (0, 1]$; $\overline{S_k}$, $\overline{F_k}$, $\forall k \in \mathcal{K}$.
Initialization: $S_k = \overline{S_k}, F_k = \overline{F_k}, \forall k \in \mathcal{K}$.
for $t = 1, 2, ..., T$ **do**
  for $k = 1, 2, ..., K$ **do**
    $\theta_k(t) \sim$ *Beta* (S_k, F_k).
  end for
  Attack channel $I_t := \arg\max_k \theta_k(t)$.
  Choose λ channels other than the attacked one uniformly at random and denote by $\mathcal{J}_t$.
  Observe reward $r_t(k)$ for each $k \in \mathcal{J}_t$.
  For $k \in \mathcal{J}_t$, update S_k and F_k using (13).
  For $k \notin \mathcal{J}_t$, update S_k and F_k using (14).
end for

In the attacking phase, the PEA determines which channel to attack. Given $P(\theta)$, the probability of channel k being optimal is given by:

$$\int \mathbf{1}[\mathbb{E}(r_t(k)|\theta) = \max_{k'} \mathbb{E}\left(r_t(k')|\theta\right)] P(\theta) d\theta \tag{9}$$

The PEA randomly selects each channel according to its probability of being optimal so as to keep balance between exploitation and exploration. Specifically, the PEA samples θ and selects the channel with the largest expected reward. The sampling result of θ_k in time slot t is denoted by $\theta_k(t)$, and the channel with the best expected reward under current samples can be given by:

$$I_t = \arg\max_k \ \mathbb{E}\left(r_t(k)|\theta_k(t)\right) = \arg\max_k \ \theta_k(t). \tag{10}$$

In the observing phase, the PEA needs to update the parameters of $P(\theta_k)$ based on the observed rewards. In time slot t, if the PEA observes channel k, it will find that this channel is occupied by a PU or is exploited by an SU or is vacant. If the channel is occupied by an SU, it means that $r_t(k)$ is 1; otherwise, $r_t(k)$ is 0. Note that the attacked channel I_t will not be used by an SU for the PEA has transmitted

forged PU signals in this channel. Therefore, the PEA chooses λ channels other than the attacked one uniformly at random to observe. We utilize the observed reward values to update the parameters of θ_k according to (5), in which the posterior distribution of θ_k after observing $r_t(k)$ is given by:

$$P(\theta_k|r_t(k)) \propto P\left(r_t(k)|\theta_k\right) P(\theta_k). \tag{11}$$

After substituting (7) and (8) into (11), S_k and F_k are updated by:

$$\begin{aligned} S_k &\leftarrow S_k + r_t(k), \\ F_k &\leftarrow F_k + 1 - r_t(k). \end{aligned} \tag{12}$$

Specifically, for channel k, if $r_t(k)$ is 1, S_k increases by 1 while F_k remains unchanged; otherwise, F_k increases by 1 while S_k remains unchanged. With the increase of S_k and F_k, the variance of θ_k decreases and the mean of θ_k approaches the real expected reward of channel k.

The time-varying characteristics of both the probability of channels being idle and the behavior of the SUs cause the previous observed reward values outdated, which implies that the learned attacking strategy may become ineffective as time slot goes on. We introduce a discount factor $\gamma \in (0, 1]$ to gradually reduce the effect of previous observations, the new iterative strategy is given by: in time slot t, for each $k \in \mathcal{J}_t$, update S_k and F_k by

$$\begin{aligned} S_k &\leftarrow \gamma S_k + (1-\gamma)\overline{S_k} + r_t(k), \\ F_k &\leftarrow \gamma F_k + (1-\gamma)\overline{F_k} + 1 - r_t(k); \end{aligned} \tag{13}$$

and for each $k \notin \mathcal{J}_t$, update S_k and F_k by

$$\begin{aligned} S_k &\leftarrow \gamma S_k + (1-\gamma)\overline{S_k}, \\ F_k &\leftarrow \gamma F_k + (1-\gamma)\overline{F_k}. \end{aligned} \tag{14}$$

where $\overline{S_k}$ and $\overline{F_k}$ are prior *Beta* distribution parameters of channel k. The PEA has no prior knowledge of the CR network and then sets $\overline{S_k} = \overline{F_k} = 1$ which makes the prior *Beta* distribution uniform over $[0, 1]$. Obviously, the discount factor γ injects the uncertainty into the distribution $P(\theta_k)$ to gradually reduce the effects of the previous observation.

B. PEA attacking for multiple-attacker case

In the multiple-attacker case, two or more PEAs may select the same channel to attack, which significantly reduce the attacking effectiveness and efficiency. We apply time-division fair sharing technique, which is to let M distributed players alternately use the most promising M arms [23], to ensure the attacking effectiveness. In each round, we let M distributed attackers alternately use the most promising M channels, i.e., each PEA has a different attacking priority as shown in Fig. 2. Specifically, the attacking priority of PEA j in time slot t is $(j + t - 1) \bmod M$, which means that it chooses the $((j + t - 1) \bmod M)$-th most promising channel to attack. The assignment of the attacking priorities can be achieved by each PEA broadcasting signal to other PEAs on arrival and departure. For example, in the case of $M = 2$, the attacking priority of PEA_1 and PEA_2 are 1 and 2 respectively. If a new PEA_3 comes, it broadcasts arrival signal and gets attacking priority 1. Correspondingly, the attacking priority of PEA_1 and PEA_2 are reduced to 2 and 3. For PEA j, we let the *Beta* distribution $\mathit{Beta}\left(S_{j,k}, F_{j,k}\right)$ be the distribution of $\theta_{j,k}$. The distributed γ-DTS, takes the γ-DTS as the main framework, as given in **Algorithm 3**.

Algorithm 3 Distributed γ-discounted TS attacking strategy

Parameters: j; $\gamma_j \in (0, 1]$; $\overline{S_{j,k}}$, $\overline{F_{j,k}}$, $\forall k \in \mathcal{K}$.
Initialization: $S_{j,k} = \overline{S_{j,k}}, F_{j,k} = \overline{F_{j,k}}, \forall k \in \mathcal{K}$.
for $t = 1, 2, ..., T$ **do**
  for $k = 1, 2, ..., K$ **do**
    $\theta_{j,k}(t) \sim \mathit{Beta}\left(S_{j,k}, F_{j,k}\right)$.
  end for
  Attack channel $I_{t,j}$ with $((j + t - 1) \bmod M)$-th largest sample $\theta_{j,k}(t)$.
  Choose λ channels other than the attacked one uniformly at random and denote by $\mathcal{J}_{t,j}$.
  Observe reward $r_t(k)$ for each $k \in \mathcal{J}_{t,j}$.
  For $k \in \mathcal{J}_{t,j}$, update $S_{j,k}$ and $F_{j,k}$ using (15).
  For $k \notin \mathcal{J}_{t,j}$, update $S_{j,k}$ and $F_{j,k}$ using (16).
end for

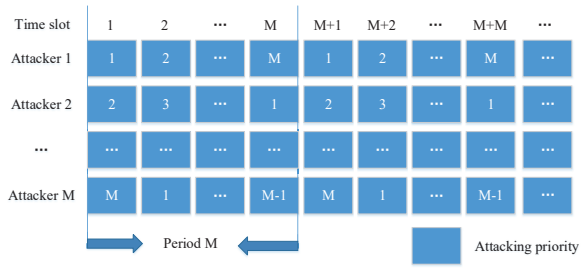

Fig. 2. The attacking priority of each attacker in multiple-attacker case

In the attacking phase, PEA j needs to select the $((j + t - 1) \bmod M)$-th most promising channel to attack. Specifically, the PEA randomly selects each channel according to its probability of being the $((j+t-1) \bmod M)$-th most promising channel. In time slot t, PEA j samples $\theta_{j,k}$ for each $k \in \mathcal{K}$ and selects the channel with the $((j+t-1) \bmod M)$-th largest sampling result.

In the observing phase, PEA j chooses λ channels other than the attacked one uniformly at random. The iterative strategy of parameters $S_{j,k}$ and $F_{j,k}$ under discount factor $\gamma_j \in (0, 1]$ is given by: in time slot t, for each $k \in \mathcal{J}_{t,j}$, update $S_{j,k}$ and $F_{j,k}$ by

$$\begin{aligned} S_{j,k} &\leftarrow \gamma_j S_{j,k} + (1-\gamma_j)\overline{S_{j,k}} + r_t(k), \\ F_{j,k} &\leftarrow \gamma_j F_{j,k} + (1-\gamma_j)\overline{F_{j,k}} + 1 - r_t(k); \end{aligned} \tag{15}$$

and for each $k \notin \mathcal{J}_{t,j}$, update $S_{j,k}$ and $F_{j,k}$ by

$$\begin{aligned} S_{j,k} &\leftarrow \gamma_j S_{j,k} + (1-\gamma_j)\overline{S_{j,k}}, \\ F_{j,k} &\leftarrow \gamma_j F_{j,k} + (1-\gamma_j)\overline{F_{j,k}}. \end{aligned} \tag{16}$$

where $\overline{S_{j,k}}$ and $\overline{F_{j,k}}$ are prior *Beta* distribution parameters of channel k for PEA j.

IV. Performance Analysis

We give the performance analysis of the proposed online attacking strategy γ-DTS in stationary environment. For an

MAB problem with K arms, the best constant arm chosen in hindsight is denoted by I^*,

$$I^* = \arg\max_{a\in\mathcal{K}} \sum_{t=1}^{T} r_t(a). \tag{17}$$

We define *regret* as the gap between A_T and the cumulative reward of constantly choosing I^*,

$$\text{Regret}(T) := \sum_{t=1}^{T} \left[r_t\left(I^*\right) - r_t\left(I_t\right)\right] = \sum_{t=1}^{T} \triangle_t, \tag{18}$$

where $\triangle_t = r_t\left(I^*\right) - r_t\left(I_t\right)$.

We define the Γ_t as the ratio of the square of expected *regret* in time slot t to information gain in time slot t [24]:

$$\Gamma_t := \frac{\mathbb{E}\left[\triangle_t\right]^2}{\mathbb{I}\left(I^*; \Phi_t\left(I_t\right)\right)}, \tag{19}$$

where $\mathbb{E}$ and $\mathbb{I}$ represent expectation and mutual information, and $\Phi_t\left(I_t\right)$ denote the observations in time t. If Γ_t is bounded, the *regret* upper bound exists and can be given as follows [25],

Proposition 1. *If there exists a constant $\overline{\Gamma} \geq \Gamma_t$ almost surely for all $t \in \{1, .., T\}$. Then, $\mathbb{E}\left[\text{Regret}\left(T\right)\right] \leq \sqrt{\overline{\Gamma} T \log K}$, where K is the total number of arms.*

Note that the *regret* bound derived from **Proposition 1** is tight when all channel rewards are generated from Bernoulli distribution and independent of any history of observations. **Proposition 1** shows that we only need to focus on the bound of Γ_t in the PEA attacking problem to simplify the analysis. The PEA observes the rewards associated with λ channels except the attacked channel I_t in time slot t. The observations of the PEA in time slot t can be expressed by $\Phi_t^{\lambda}(I_t)$. In the full observing case where λ is $K-1$, $\Phi_t^{K-1}(I_t)$ is $\{r_t(1), \ldots, r_t(I_t-1), r_t(I_t+1), \ldots, r_t(K)\}$. The Γ_t of the full observing case can be given by:

Lemma 1. *In the full observing case, $\Gamma_t \leq 2$ almost surely for all $t \in \{1, .., T\}$.*

The proof of **Lemma 1** can be referred to Appendix B. The full observing case is a special case of our problem. Then, we give the bound of the Γ_t in a more general case where λ can be any number between 1 and $K-1$.

Lemma 2. *In PEA attacking problem, $\Gamma_t \leq \frac{2(K-1)}{\lambda}$ almost surely for all $t \in \{1, .., T\}$.*

The proof of **Lemma 2** can be referred to Appendix C. Applying **Proposition 1** and **Lemma 2**, we get a *regret* upper bound for proposed online attacking strategy in our problem.

Theorem 1. *For proposed online attacking strategy in PEA attacking problem, we have $\mathbb{E}\left[\text{Regret}\left(T\right)\right] \leq \sqrt{\frac{2(K-1)}{\lambda} T \log K}$.*

Our *regret* upper bound for the PEA attacking problem improves on the result in [16] that $\mathbb{E}\left[\text{Regret}\left(T\right)\right] \leq 4\sqrt{(e-2)}\sqrt{T\frac{K-1}{\lambda}\log K}$. If observation capability $\lambda = 0$, the *regret* is unbounded. By adding λ from 0 to 1, the *regret* upper bound becomes $\sqrt{2(K-1)T\log K}$, which significantly improves the attacking effectiveness and efficiency. The *regret* upper bound decreases with observation capability λ and is proportional to $\sqrt{\frac{1}{\lambda}}$. Note that the *regret* bound is a convex function of λ. Therefore, a relatively small λ is sufficient to get a good attacking performance. The *regret* upper bound increases with the number of licensed channels K and is proportional to $\sqrt{(K-1)\log K}$, which shows that it can effectively handle situations with large K. The performance analysis of γ-DTS in non-stationary environment remains to be an open question since it is difficult to quantitatively analyze the variation of both the probability of channels being idle and the access strategy of the SUs.

V. Performance evaluation

We validate the proposed online attacking strategy in both stationary channel and non-stationary channel scenarios. The idle/busy state of each channel is subject to an independent two-state Markov chain, as discussed in [16], [26]. The transition matrix of channel k is denoted as

$$\mathbf{P}_k = \begin{bmatrix} p_{00}(k) & p_{01}(k) \\ p_{10}(k) & p_{11}(k) \end{bmatrix}, \tag{20}$$

where $p_{ij}(k)$ is the transition probability from state i to state j and $p_{i0}(k) + p_{i1}(k) = 1$. In the stationary channel scenario, we randomly generate K independent $\mathbf{P}_k$ at the beginning of each experiment and keep each $\mathbf{P}_k$ unchanged, while in the non-stationary scenario, we randomly generate K independent $\mathbf{P}_k$ at the beginning of each experiment and regenerate each $\mathbf{P}_k$ independently every ΔT time slots.

We introduce the successful accesses of SUs as a performance metric of the attacking strategy. The total number of successful accesses by the SUs in total time slots T is

$$G_T = \sum_{t=1}^{T}\sum_{i=1}^{N} D_t(Q_{t,i})\mathbf{1}[Q_{t,i} \notin \mathcal{I}_t]. \tag{21}$$

The SUs employ an adversarial bandit based passive defend approach proposed in [15], which use the experience of spectrum access to adapt the sensing policy to the current attacking strategy. All results are averaged by 10,000 Monte Carlo simulations.

A. Single Attacker Case

For single attacker case, we compare the proposed γ-DTS strategy with two conventional attacking strategies: play and random observe learning algorithm (PROLA) [16] and uniformly random attack (URA) [15]. In the PROLA, the attacker applies an exponential weighting distribution based on the observed rewards to determine which channel to attack. In the URA, the probability of each channel being attacked is equal in each time slot. Consider a network with $K = 10$, $N = 1$, $M = 1$, $\lambda = 1$ and $\Delta T = 200$. The hyperparameters in both stationary and non-stationary channel scenarios are given by $\overline{S_k} = 1$, $\overline{F_k} = 1$, $\gamma = 0.99$. The parameter γ is determined by a series of numerical experiments in both stationary and non-stationary channel scenarios as shown in Fig. 3.

Fig. 4 shows the *regret* as a function of time slots. As we can see from Fig. 4, the difference of regret between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3091588, IEEE Transactions on Wireless Communications

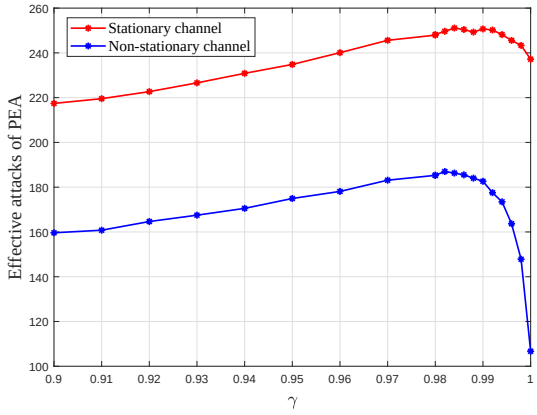

Fig. 3. Attacking performance as a function of γ for single attacker case.

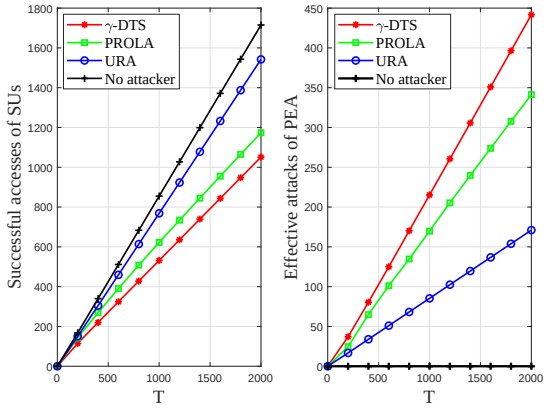

Fig. 5. Attacking performance for single attacker case in stationary channel scenarios.

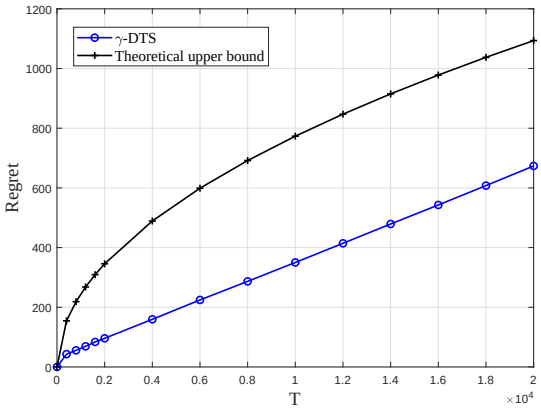

Fig. 4. *Regret* as a function of time slots.

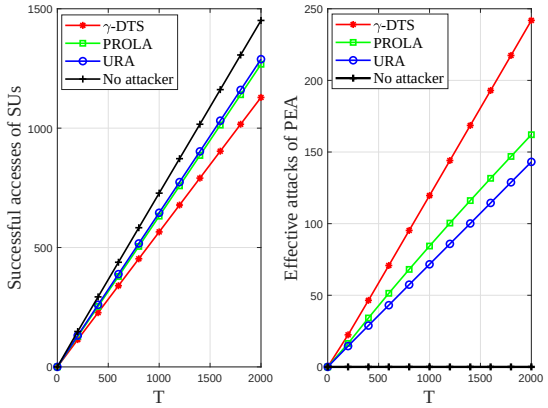

Fig. 6. Attacking performance for single attacker case in non-stationary channel scenarios.

the γ-DTS and the upper bound converges towards stability, indicating that the proposed algorithm can learn the statistical characteristics of channel states and the behaviors of SUs.

Fig. 5 shows the attacking performance for single attacker case in stationary channel scenarios. From Fig. 5 we can see, the γ-DTS strategy outperforms the PROLA and the URA by 100 and 271 in effective attacks of PEA, respectively. As for the reductions of successful accesses of SUs compared to no attacker case, the γ-DTS strategy outperforms the PROLA and the URA strategies by 112 and 491, respectively. The reason for the improvement is that the γ-DTS strategy can find the channel with the largest reward faster than the PROLA and the URA without requiring the prior knowledge about the channels and the behaviors of the SUs. Note that the reduction of successful access of SUs is more than the increment of effective attacks of PEA because a successful attack can impact the access of the SUs in two ways: preventing the SUs from accessing the attacked channel explicitly or misleading the SUs the true spectrum environment. For the latter, the SUs may not attempt to access this attacked channel.

Fig. 6 shows the attacking performance for single attacker case in non-stationary channel scenarios. Compared to stationary channel scenarios, the performance of the PROLA and the γ-DTS deteriorates gradually since non-stationary channels introduce more uncertainty into the CR network. We can see that the effective attacks of the PROLA is only 13% higher than that of the URA, but the PROLA needs more computation power and hardware capabilities. However, the effective attacks of the γ-DTS is 68% higher than that of the URA and only needs the same computation power and hardware capabilities as the PROLA. It indicates that the γ-DTS strategy can track the changes of the probability of channels being idle fairly quickly.

As we can see from Fig. 5 and Fig. 6, the proposed γ-DTS algorithm shows a better attacking performance in both stationary and non-stationary channel scenarios. Moreover, its computational complexity is at least $\mathcal{O}(TK)$, which is the same as the computational complexity of the PROLA algorithm. Although the computational complexity of the URA algorithm is at least $\mathcal{O}(T)$, the proposed γ-DTS algorithm produces much higher attacking performance and its regret converges in a logarithmic order.

Fig. 7 shows the attacking performance as a function of observation capability λ for single attacker case, where $K = 40$, $N = 4$, $M = 1$ and $\lambda \in [1, 37]$. As we can see from Fig. 7, the effective attacks of PEA increases with the increasing of observation capability λ. The reason is that the increasing of λ

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3091588, IEEE Transactions on Wireless Communications

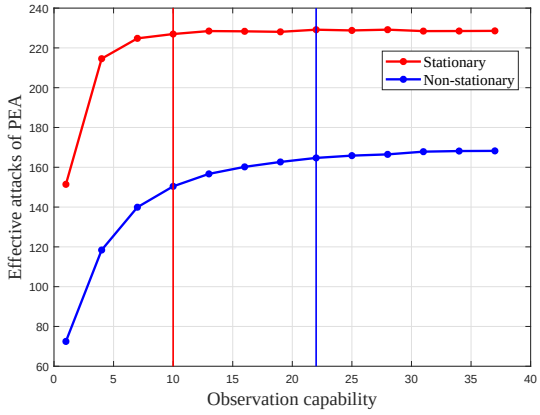

Fig. 7. Attacking performance as a function of observation capability λ for single attacker case.

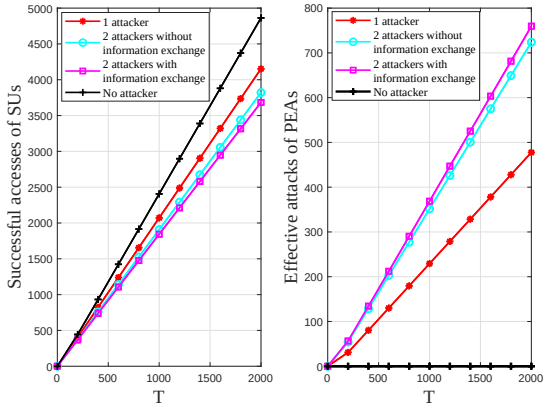

Fig. 8. Attacking performance for multiple-attacker case in stationary channel scenarios.

makes the PEA get more information about both the channel states and the behaviors of the SUs. As a result, the parameters of *Beta* distribution can be updated faster to the promising values. Notice that the hardware cost of the PEA will increase if we improve the observation capability λ. Therefore, it is reasonable to find a tradeoff between the attacking effectiveness and the cost. In the beginning, by adding a little observation capability (from $\lambda = 1$ to $\lambda = 4$), the effective attacks of PEA increases dramatically. The increment becomes marginal as the observation capability λ is sufficiently large. This implies that, to achieve a good tradeoff between performance and cost, the observation capability of PEA is not necessary great. In stationary channel scenarios, the effective attacks of PEA is near optimal when the observation capability is 10, while in non-stationary channel scenarios, $\lambda = 22$ yields almost the optimal.

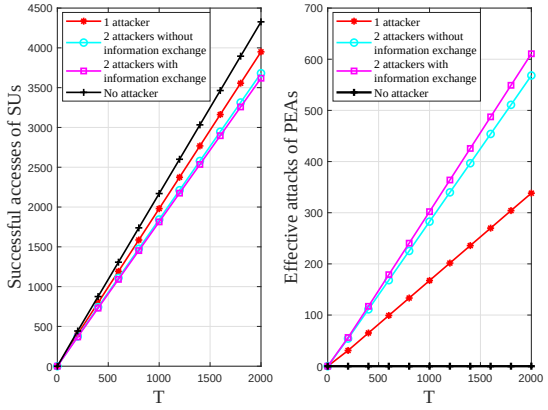

Fig. 9. Attacking performance for multiple-attacker case in non-stationary channel scenarios.

B. multiple-attacker Case

For multiple-attacker case, we employ the proposed strategy distributed γ-DTS. If the PEAs can exchange information about the attacked channels, the attacking efficiency should be higher than that without information exchange. Specifically, M PEAs with the ability of information exchange can be taken as a single PEA who can attack M channels in a time slot. Consider a network with $K = 20$, $N = 3$, $M = 2$, $\lambda = 1$ and $\Delta T = 200$. The hyperparameters are given by $\overline{S_{j,k}} = 1$, $\overline{F_{j,k}} = 1$, $\gamma_j = 0.99$.

Fig. 8 shows the attacking performance for multiple-attacker case in stationary channel scenarios. From Fig. 8 we can see that two attackers with information exchange yield 759 effective attacks in 2000 time slots, making the SUs less accesses 1180 times, while two attackers without information exchange produce 723 effective attacks and the SUs fail to access 1042 times. The difference of effective attacks of the two schemes is not significant, indicating that the distributed γ-DTS is a promising strategy for practical applications in which exchanging information is usually difficult.

Fig. 9 shows the attacking performance for multiple-attacker case in non-stationary channel scenarios. We can see that two attackers with information exchange yield 610 effective attacks in 2000 time slots and make the SUs less accesses 708 times. For two attackers without information exchange, effective attacks are 568, which makes the SUs less accesses 645 times. Compared to the stationary channel scenarios, effective attacks are cut down for both the cases of exchanging information or not in non-stationary channel scenarios. The reason is that the variations of the statistical characteristics of channel states boost that difficulties of learning. Again, the performance difference between the two schemes is slight, which means that our proposed distributed γ-DTS is still effective in non-stationary channel scenarios.

Fig. 10 shows the attacking performance as a function of number of PEAs M for multiple-attacker case, where $K = 40$, $N = 4$, $\lambda = 1$ and $M \in [1, 8]$. As can be seen from Fig. 9, though the effective attacks of PEAs increases with the number of PEAs, the average effective attacks of each PEA decreases gradually. It can be explained as follows: For a given network, the probabilities of channels being idle and the behaviors of the SUs are also determined. Increasing PEAs can result in more ineffective attacks.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2021.3091588, IEEE Transactions on Wireless Communications

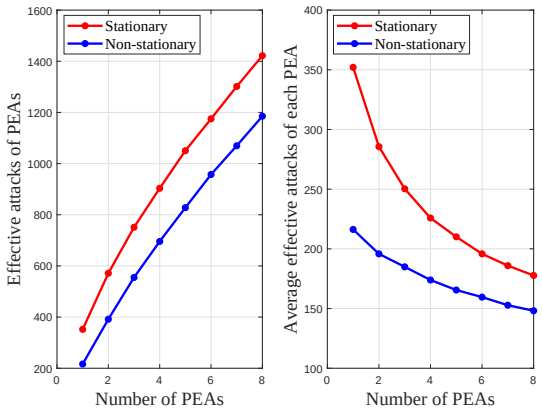

Fig. 10. Attacking performance as a function of number of PEAs M for multiple-attacker case.

VI. Conclusions and Future Work

In this paper, we proposed a primary user emulation attacking scheme based on Thompson sampling method in CR networks, where the unknown time-varying statistical characteristics of channels and the behaviors of SUs are learned in an online manner. We prove that the *regret* upper bound of online attacking algorithm is sublinear with time slots in stationary scenarios. Numerical results have demonstrated that the proposed γ-DTS can track the variations of the probability of channels being idle and the access strategy of SUs fairly quickly in both the stationary and non-stationary channel scenarios, and is effective for practical applications where exchanging information is usually difficult. The behaviors of the primary users are not considered in this work, which could be an interesting problem for future work. On the other hand, the *regret* upper bound of γ-DTS in non-stationary scenarios remains to be an open question. Note that an efficient attacking algorithm should keep good balance between sample efficiency and expressive power. The proposed γ-DTS focuses on high sample efficiency to track dynamic environment. However, expressive power can be more important in complex CR networks, where model-based reinforcement learning algorithms would be promising methods to tackle time-varying statistical characteristics at the cost of sample efficiency.

Appendix A
Facts and Preparatory Lemmas

We first give notations and conventions used in the following proof. The relative entropy between probability measure μ and ω is $\mathrm{D}(\omega\|\mu) = \int \log\left(\frac{d\omega}{d\mu}\right) d\omega$ if $\omega \ll \mu$ and $\mathrm{D}(\omega\|\mu) = \infty$ otherwise, where log is the natural logarithm. For a measure $\mathbb{P}$ and two jointly distributed random elements X and Z, we use $\mathbb{P}_{X|Z}$ to represent the conditional probability of X given Z, where $\mathbb{P}_{X|Z}(\cdot) = \mathbb{P}(X \in \cdot|Z)$. When Z is discrete, we use $\mathbb{P}_{X|Z=i}$ to denote $\mathbb{P}(X \in \cdot|Z = i)$. The mutual information between X and Z is $\mathbb{I}(X;Z) = \mathbb{E}[\mathrm{D}(\mathbb{P}_{X|Z}|\mathbb{P}_X)]$. For any $t \le T$, the probability of the actions satisfies

$$\mathbb{P}_t = \mathbb{P}_t\left(I_t \in \cdot\right) = \varphi\left(I_1, \Phi_1\left(I_1\right), \ldots, I_{t-1}, \Phi_{t-1}\left(I_{t-1}\right)\right). \tag{22}$$

where φ is a measurable policy. With this notation, let $P_{ta} = \mathbb{P}_t(I_t = a)$, let $C_t = \arg\max_{a\in\mathcal{K}} P_{ta}$, where a is a action.

Fact 1 (Chain rule of mutual information in [24]). $\mathbb{I}\left(X;(Z_1,\ldots Z_T)\right) = \mathbb{I}\left(X;Z_1\right) + \mathbb{I}\left(X;Z_2|Z_1\right) + \ldots + \mathbb{I}\left(X;Z_T|Z_1,\ldots,Z_{T-1}\right)$.

Fact 2 (Pinskers inequality). *When a random variable $X \in [a,b]$ almost surely,*

$$\int X(d\mu - d\nu) \le (b-a)\sqrt{\frac{1}{2}\mathrm{D}(\mu|\nu)}. \tag{23}$$

Fact 3 (Data processing inequality in [27]). *For $b \ne C_t$, $r_t(C_t)$ is is a deterministic function of $\Phi_t(b)$.*

Lemma 3 (Lemma 15 in [27]). *Let $P^*_{ta} = \mathbb{P}_t\left(I^* = a\right)$,*

$$\mathbb{I}\left(I^*;\Phi_t\left(I_t\right)\right) = \sum_{a=1}^{k} P^*_{ta} \sum_{b=1}^{k} P_{tb}\mathbb{E}_t\left[\mathrm{D}\left(\mathbb{P}_{t,\Phi_t(b)|I^*=a}\|\mathbb{P}_{t,\Phi_t(b)}\right)\right]. \tag{24}$$

Appendix B

*Proof of **Lemma 1**:* By adding and subtracting $r_t(C_t)$, we expand the definition:

$$\begin{aligned}
\mathbb{E}\left[\Delta_t\right] &= \sum_{a\ne C_t} P_{ta}\mathbb{E}\left[r_t(C_t) - r_t\left(a\right)\right] \\
&\quad + \sum_{a\ne C_t} P_{ta}\mathbb{E}\left[r_t\left(a\right) - r_t(C_t)|I^* = a\right] \\
&\le \sum_{a\ne C_t} P_{ta}\Big(\sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_{t,r_t(C_t)|I^*=a}\|\mathbb{P}_{t,r_t(C_t))}\right)} \\
&\quad + \sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_{t,r_t(a)|I^*=a}\|\mathbb{P}_{t,r_t(a)}\right)}\Big) \\
&\le \sqrt{(A)} + \sqrt{(B)}.
\end{aligned} \tag{25}$$

where the first inequality follows from **Fact 2**, the second inequality follows from Cauchy-Schwarz inequality and the definitions,

$$\begin{aligned}
(\mathrm{A}) &= \frac{1-P_{tC_t}}{2}\sum_{a\ne C_t} P_{ta}\mathrm{D}\left(\mathbb{P}_{t,r_t(C_t)|I^*=a}\|\mathbb{P}_{t,r_t(C_t))}\right) \\
(\mathrm{B}) &= \frac{1-P_{tC_t}}{2}\sum_{a\ne C_t} P_{ta}\mathrm{D}\left(\mathbb{P}_{t,r_t(a)|I^*=a}\|\mathbb{P}_{t,r_t(a)}\right).
\end{aligned} \tag{26}$$

Since $1 - P_{tC_t} = \sum_{b\ne C_t} P_{tb}$, the bound of term (A) is given by:

$$\begin{aligned}
(\mathrm{A}) &= \frac{1-P_{tC_t}}{2}\sum_{a\ne C_t} P_{ta}\mathrm{D}\left(\mathbb{P}_{t,r_t(C_t)|I^*=a}\|\mathbb{P}_{t,r_t(C_t)}\right) \\
&= \frac{1}{2}\sum_{a\ne C_t} P_{ta}\sum_{b\ne C_t} P_{tb}\mathrm{D}\left(\mathbb{P}_{t,r_t(C_t)|I^*=a}\|\mathbb{P}_{t,r_t(C_t)}\right) \\
&\le \frac{1}{2}\sum_{a\ne C_t} P_{ta}\sum_{b\ne C_t} P_{tb}\mathrm{D}\left(\mathbb{P}_{t,\Phi_t^{K-1}(b)|I^*=a}\|\mathbb{P}_{t,\Phi_t^{K-1}(b)}\right) \\
&\le \frac{1}{2}\mathbb{I}\left(I^*;\Phi_t^{K-1}\left(I_t\right)\right).
\end{aligned} \tag{27}$$

where the first inequality follows from **Fact 2** and the second inequality follows from **Lemma 3** in Appendix A.

The bound of term (B) is given by:

$$\begin{aligned}(\mathrm{B}) &= \frac{1-P_{tC_t}}{2}\sum_{a\neq C_t} P_{ta}\mathrm{D}\left(\mathbb{P}_{t,r_t(a)|I^*=a}\|\mathbb{P}_{t,r_t(a)}\right)\\ &\leq \frac{1}{2}\sum_{a\neq C_t}(1-P_{ta})P_{ta}\mathrm{D}\left(\mathbb{P}_{t,r_t(a)|I^*=a}\|\mathbb{P}_{t,r_t(a)}\right)\\ &= \frac{1}{2}\sum_{a\neq C_t} P_{ta}\sum_{b\neq a} P_{tb}\mathrm{D}\left(\mathbb{P}_{t,r_t(a)|I^*=a}\|\mathbb{P}_{t,r_t(a)}\right)\\ &\leq \frac{1}{2}\sum_{a\neq C_t} P_{ta}\sum_{b\neq a} P_{tb}\mathrm{D}\left(\mathbb{P}_{t,\Phi_t^{K-1}(b)|I^*=a}\|\mathbb{P}_{t,\Phi_t^{K-1}(b)}\right)\\ &\leq \frac{1}{2}\mathbb{I}\left(I^*;\Phi_t^{K-1}(I_t)\right).\end{aligned} \tag{28}$$

where the first inequality follows from the fact that $1-P_{tC_t}\leq 1-P_{ta}$, the second inequality follows from **Fact 2** and the third inequality follows from **Lemma 3** in Appendix A.

Combining the bound of two terms (A) and (B), the Γ_t of the full observing case can be given by:

$$\Gamma_t = \frac{\mathbb{E}\left[\triangle_t\right]^2}{\mathbb{I}\left(I^*;\Phi_t^{K-1}(I_t)\right)} \leq 2. \tag{29}$$

■

APPENDIX C

*Proof of **Lemma 2**:* Since the arms are independent of each other, by Fact 1, we show:

$$\begin{aligned}\mathbb{I}(I^*;\Phi_t^{K-1}(I_t)) = \mathbb{I}(I^*;r_t(1)) + \ldots + \mathbb{I}(I^*;r_t(I_t-1))\\ +\mathbb{I}(I^*;r_t(I_t+1)) + \ldots + \mathbb{I}(I^*;r_t(K)).\end{aligned} \tag{30}$$

As we can see, $\Phi_t^{\lambda}(I_t)$ is the λ-subset of $\Phi_t^{K-1}(I_t)$ and the probability that each element of $\Phi_t^{K-1}(I_t)$ belongs to $\Phi_t^{\lambda}(I_t)$ is $\frac{\lambda}{K-1}$. Then, we show:

$$\begin{aligned}\mathbb{I}(I^*;\Phi_t^{\lambda}(I_t)) &= \frac{\lambda}{K-1}(\mathbb{I}(I^*;r_t(1)) + \ldots + \mathbb{I}(I^*;r_t(I_t-1))\\ &\quad + \mathbb{I}(I^*;r_t(I_t+1)) + \ldots + \mathbb{I}(I^*;r_t(K)))\\ &= \frac{\lambda}{K-1}\mathbb{I}(I^*;\Phi_t^{K-1}(I_t)).\end{aligned} \tag{31}$$

Combing (31) with **Lemma 1**, we show:

$$\begin{aligned}\Gamma_t &= \frac{\mathbb{E}\left[\triangle_t\right]^2}{\mathbb{I}\left(I^*;\Phi_t^{K-1}(I_t)\right)} * \frac{\mathbb{I}\left(I^*;\Phi_t^{K-1}(I_t)\right)}{\mathbb{I}\left(I^*;\Phi_t^{\lambda}(I_t)\right)}\\ &\leq \frac{2(K-1)}{\lambda}.\end{aligned} \tag{32}$$

■

REFERENCES

[1] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[2] Y. Zhao, S. Mao, J. O. Neel, and J. H. Reed, "Performance evaluation of cognitive radios: Metrics, utility functions, and methodology," *Proc. IEEE*, vol. 97, no. 4, pp. 642–659, Apr. 2009.

[3] S. Wang, Z.-H. Zhou, M. Ge, and C. Wang, "Resource allocation for heterogeneous multiuser-OFDM based cognitive radio networks with imperfect spectrum sensing," in *Proc. IEEE INFOCOM'12*, Mar. 2012, pp. 2264–2272.

[4] Q. Yan, M. Li, T. Jiang, W. Lou, and Y. T. Hou, "Vulnerability and protection for distributed consensus-based spectrum sensing in cognitive radio networks," in *Proc. IEEE INFOCOM'12*, Mar. 2012, pp. 900–908.

[5] S. Wang, Z.-H. Zhou, M. Ge, and C. Wang, "Resource allocation for heterogeneous cognitive radio networks with imperfect spectrum sensing," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 3, pp. 464–475, Mar. 2013.

[6] R. Yu, Y. Zhang, Y. Liu, S. Gjessing, and M. Guizani, "Securing cognitive radio networks against primary user emulation attacks," *IEEE Netw.*, vol. 29, no. 4, pp. 68–74, Nov. 2015.

[7] J. Esch, "A survey of security challenges in cognitive radio networks: Solutions and future research directions," *Proc. IEEE*, vol. 100, no. 12, pp. 3170–3171, Dec. 2012.

[8] I. Bisio, C. Garibotto, F. Lavagetto, A. Sciarrone, and S. Zappatore, "Improving WiFi statistical fingerprint-based detection techniques against UAV stealth attacks," in *Proc. IEEE GLOBECOM'18*, Dec. 2018, pp. 1–6.

[9] R. Chen, J.-M. Park, and J. H. 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.

[10] D. Pu, Y. Shi, A. V. Ilyashenko, and A. M. Wyglinski, "Detecting primary user emulation attack in cognitive radio networks," in *Proc. IEEE GLOBECOM'11*, Dec. 2011, pp. 1–5.

[11] Z. Yuan, D. Niyato, H. Li, J. B. Song, and Z. Han, "Defeating primary user emulation attacks using belief propagation in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 10, pp. 1850–1860, Nov. 2012.

[12] C. Xin and M. Song, "Detection of PUE attacks in cognitive radio networks based on signal activity pattern," *IEEE Trans. Mobile Comput.*, vol. 13, no. 5, pp. 1022–1034, May 2014.

[13] N. Nguyen-Thanh, P. Ciblat, A. T. Pham, and V.-T. Nguyen, "Surveillance strategies against primary user emulation attack in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4981–4993, Sept. 2015.

[14] H. Li and Z. Han, "Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part I: Known channel statistics," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3566–3577, Nov. 2010.

[15] H. Li and Z. Han, "Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part II: Unknown channel statistics," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 274–283, Jan. 2010.

[16] M. Dabaghchian, A. Alipour-Fanid, K. Zeng, Q. Wang, and P. Auer, "Online learning with randomized feedback graphs for optimal PUE attacks in cognitive radio networks," *IEEE/ACM Trans. Netw.*, vol. 26, no. 5, pp. 2268–2281, Oct. 2018.

[17] D.-T. Ta, N. Nguyen-Thanh, P. Maillé, and V.-T. Nguyen, "Strategic surveillance against primary user emulation attacks in cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 582–596, Sept. 2018.

[18] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. ICML'13*, May 2013, pp. 151–159.

[19] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, Dec. 1933.

[20] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *Proc. ICML'13*, May 2013, pp. 127–135.

[21] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," in *Proc. NeurIPS'11*, Dec. 2011, pp. 2249–2257.

[22] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. COLT'12*, Jun. 2012, pp. 39.1–39.26.

[23] J. Zhao, H. Zheng, and G.-H. Yang, "Distributed coordination in dynamic spectrum allocation networks," in *Proc. DySPAN'05*, Nov. 2005, pp. 259–268.

[24] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2442–2471, Apr. 2016.

[25] S. Dong and B. Van Roy, "An information-theoretic analysis for Thompson sampling with many actions," in *Proc. NeurIPS'18*, Dec. 2018, pp. 4157–4165.

[26] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in Ad Hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[27] T. Lattimore and C. Szepesvári, "An information-theoretic approach to minimax regret in partial monitoring," in *Proc. COLT'19*, Jun. 2019, pp. 2111–2139.

Xiang Sheng received the BS degree from Nanjing University, Nanjing, China, in 2019, where he is currently pursuing the MS degree at the School of Electronic Science and Engineering. His research interests include wireless communications and machine learning.

Shaowei Wang (S'06-M'07-SM'13) received the PhD degree from Wuhan University, Wuhan, China, in 2006, and joined the School of Electronic Science and Engineering at Nanjing University, Nanjing, China, as a faculty member in the same year, where he is currently a Full Professor. From 2012 to 2013, he was a Visiting Scholar/Professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include communications and networking, operations research and machine learning.