A Byzantine Attack Defender in Cognitive Radio Networks: the Conditional Frequency Check

Xiaofan He, Huaiyu Dai, Senior Member, IEEE and Peng Ning, Senior Member, IEEE

Abstract—Security concerns are raised for collaborative spectrum sensing due to its vulnerabilities to the potential attacks from malicious secondary users. Most existing malicious user detection methods are reputation-based, which become incapable when the malicious users dominate the network. On the other hand, although Markovian models characterize the spectrum state behavior more precisely, there is a scarcity of malicious user detection methods which fully explore this feature. In this paper, a new malicious user detection method using two proposed conditional frequency check (CFC) statistics is developed under the Markovian model for the spectrum state. With the assistance of one trusted user, the proposed method can achieve high malicious user detection accuracy (∼95%) for arbitrary percentage of malicious users that may even be equipped with more advanced sensing devices, and can thus improve the collaborative spectrum sensing performance significantly. Simulation results are provided to verify the theoretical analysis and effectiveness of the proposed method.

Keywords: Cognitive radio network, collaborative spectrum sensing, malicious user detection, security.

I. INTRODUCTION

Cognitive radio is a viable and promising way of solving the spectrum scarcity problem for ever growing wireless applications. Spectrum sensing is one of the essential functions that determine the success of the cognitive radio technology [1]. Various collaborative spectrum sensing schemes have been proposed to overcome the unreliability of single user spectrum sensing due to channel uncertainties, and consequently improve the sensing performance [2, 3]. Along with all the benefits, collaborative spectrum sensing also induces security vulnerabilities [4, 5], among which the Byzantine attack [6] (a.k.a. spectrum sensing data falsification (SSDF) attack [7]) is the focus of this paper. Through the Byzantine attack, in which the malicious sensor will send falsified local spectrum inference to mislead the fusion center, the adversary can prevent honest secondary users from using the existing white space, or allure them to access the channels in use and cause excessive interference to legitimate users, thereby undermine the premise of cognitive radio technology.

Many statistical methods have been proposed to resist the Byzantine attack in literature, where the essential idea is to detect the attacker according to its abnormal behavior and then ignore its reports. In [8], a weighted sequential probability ratio test (WSPRT) is proposed for robust distributed spectrum sensing. In [9], out-lier factors are computed for secondary users to identify malicious ones. In [10], suspicious levels of all nodes are computed and the nodes with a high suspicious level are excluded from decision-making to ensure robustness. All these methods are based on the concept of reputation. One key feature of this type of methods is that lower reputations will be assigned to sensors which provide inconsistent spectrum inference to the majority’s decisions. However, the success of this type of methods is built upon the assumption that the global decision is correct, which may not be true when malicious sensors dominate the network. In fact, it was shown in [6, 11] that when the number of Byzantine attackers in the network exceeds a certain fraction, such reputation based methods become completely incapable. Non-reputation based approaches have also been developed. For example, in [12], a double-sided neighbor distance (DSND) metric and the frequency check (FC) were proposed to detect malicious users. In [13], a malicious user detection algorithm based on the non-parametric Kruskal-Wallis test is developed. In [14], a correlation filter approach, which relies on the fact that the received signal strengths among nearby secondary users are correlated, is proposed to mitigate the attack from malicious users. However, these methods still rely on the correctness of global decision and thus only investigate scenarios where a small fraction of users are malicious. When the majority are not trustworthy, global decision independent approaches are more suitable. Such type of works include the prior-probability aided method proposed in [15, 16] and the user-centric misbehavior detection presented in [17].

In literature, both non-Markovian models (e.g., the ON/OFF model in [18, 19] and the highly predictable TV white space model in [20]) and Markovian models [21–23] are investigated, and corroborated by measurements in [24, 25] and [26, 27], respectively. This paper will focus on the Markovian case. In [28, 29], it has already been shown that better spectrum sensing performance can be obtained when the Markov model is adopted. However, [28, 29] assume all secondary users are trustworthy and ignore the security aspect of the spectrum

X. He (xhe6@ncsu.edu) and H. Dai (Huaiyu_Dai@ncsu.edu) are with the Department of Electrical and Computer Engineering, North Carolina State University, NC, USA.

P. Ning (pning@ncsu.edu) is with the Department of Computer Science, North Carolina State University, NC, USA.

This work was supported in part by the National Science Foundation under Grants CNS-1016260 and ECCS-1002258.

Part of this work was presented at the IEEE International Symposium on Information Theory (ISIT), Cambridge, MA, July 2012.

1When all sensors have the same spectrum sensing ability, the fraction is 50% [6].
The spectrum has two states, i.e., \( 0 \) or \( 1 \) or focus their analysis on one time slot and ignore the correlation between the spectrum states [6, 8–10, 13, 30]. In [16], the Markov property of the spectrum is incorporated into the malicious user detection algorithm; however, it is generally difficult to obtain the required prior knowledge of the true spectrum accurately in practice.

In this paper, a global decision independent method, the Conditional Frequency Check (CFC), is proposed based on a Markov spectrum model to defend against the Byzantine attacks. As in [6, 12], it is assumed that a malicious sensor has two degrees of freedom in falsifying its local inference, i.e., flipping from 0 (idle) to 1 (occupied spectrum), or from 1 to 0. Consequently, the detection algorithm needs two testing conditions to combat these two degrees of freedom so as to confine the malicious sensors’ behaviors. From this perspective, two natural but effective CFC statistics, which explore the second order property of the Markov chain, are constructed in this paper. Corresponding analysis proves that these two statistics can be used to detect any sensor that misbehaves. In addition, two consistent histogram estimators based on the history of sensor reports are also developed for these two CFC statistics that eliminate the requirement of prior information on sensing and spectrum modeling. Furthermore, the concept of detection margin (DM) is introduced to quantify the detectability of malicious sensors. To strengthen the malicious sensor detection ability of the proposed CFC, an auxiliary hamming distance check (HDC) is applied subsequently. With the aid of one trusted sensor and a sufficient long detection window, the proposed CFC is capable of detecting any malicious sensor regardless of the proportion of malicious ones in the sensor group, without requiring prior knowledge of the true spectrum. The assumption of one available trusted sensor has been adopted in literature (e.g., [30]); for instance, the common access point or a sensor itself [17] can serve this role in distributed cooperative spectrum sensing. In the case when such a trusted sensor is not available, an auxiliary clustering procedure is developed, which can effectively detect the malicious sensors as long as honest sensors dominate the network.

The rest of this paper is organized as follows. Section II formulates the problem. The proposed malicious sensor detection method and the corresponding analytical results are presented in Section III. Threshold selection is discussed in Section IV. Extensions of the proposed methods are discussed in Section V. Simulations and corresponding discussion are presented in Section VI, and Section VII concludes the paper.

II. Problem Formulation

In this paper, the following scenario is considered: 1) The true spectrum has two states, i.e., 0 (idle) and 1 (occupied), and follows a homogeneous Markov model with state transition matrix \( A = [a_{ij}]_{2 \times 2} \) where \( a_{ij} = Pr(s_{t+1} = j | s_t = i) \) and \( s_t \) denotes the true spectrum state at time \( t \). The stationary spectrum state distribution is denoted by \( \pi = [\pi_0, \pi_1] \), which satisfies \( \pi A = \pi \). In addition, it is assumed that the Markov chain of spectrum states is in equilibrium. 2) There are two types of sensors, honest and malicious ones, and the corresponding sets are denoted by \( H \) and \( M \), respectively. One trusted honest sensor exists and is known by the fusion center. 3) The probabilities of detection and false alarm for spectrum sensing of honest sensors are assumed the same, and denoted by \( P_d \) and \( P_{fa} \), respectively. In contrast, those of malicious sensors can be arbitrarily different, denoted for the \( i \)-th malicious one by \( \gamma_{i}^{d} P_d \) and \( \gamma_{i}^{fa} P_{fa} \), respectively. (Note that \( 0 \leq \gamma_{i}^{d} P_d, \gamma_{i}^{fa} P_{fa} \leq 1 \) for any malicious sensor \( i \).) The factors \( \gamma_{i}^{d} \) and \( \gamma_{i}^{fa} \) represent the difference in spectrum sensing ability between the \( i \)-th malicious sensor and honest ones. In addition, the reporting channel between each user and the fusion center is assumed to be error-free. 4) An honest sensor will send its local spectrum sensing result directly to the fusion center. Here, it is further assumed that honest secondary users are in a proximity and make independent observations of the same spectrum. Different sensing behaviors caused by sensor location variability in large networks (e.g., [31]) can be avoided/mitigated by clustering [32, 33] and are ignored in this paper. 5) A malicious sensor, however, will tamper its local inference before reporting to the fusion center. In particular, the \( i \)-th malicious sensor will flip its local inference from 0 to 1 and from 0 to \( 1 \) with probabilities \( \varphi_{01}^{(i)} \) and \( \varphi_{10}^{(i)} \), respectively, which will be referred to as flipping attack in the following discussions.

From the fusion center’s viewpoint, the equivalent detection and false alarm probabilities of the \( i \)-th malicious sensor with flipping probabilities \( \varphi_{01}^{(i)}, \varphi_{10}^{(i)} \) are given by

\[
P_d^{(M,i)} = (1 - \varphi_{10}^{(i)}) \gamma_{i}^{d} P_d + \varphi_{01}^{(i)} (1 - \gamma_{i}^{d} P_d),
\]

\[
P_{fa}^{(M,i)} = (1 - \varphi_{10}^{(i)}) \gamma_{i}^{fa} P_{fa} + \varphi_{01}^{(i)} (1 - \gamma_{i}^{fa} P_{fa}).
\]

If a malicious sensor \( i \) attacks, i.e., \( \{\varphi_{01}^{(i)}, \varphi_{10}^{(i)}\} \neq \{0, 0\} \), its statistical behavior will deviate from that of the honest sensor. The objective of this paper is to detect the malicious sensors by observing their statistical deviations.

III. THE PROPOSED METHOD

The proposed malicious sensor detection method consists of two parts: a conditional frequency check (CFC) and an auxiliary hamming distance check (HDC). In this section, it is assumed that the spectrum sensing capabilities of the malicious and the honest devices are identical, i.e., \( \gamma_{1}^{d} = \gamma_{0}^{d} = 1 \) for any malicious sensor \( i \). Extension to more general attackers with arbitrary \( \gamma_{1}^{d} \) and \( \gamma_{0}^{d} \) will be presented in Section V.

\(^2\)The superscript \( H \) is dropped for simplicity.
A. Conditional Frequency Check

According to the preceding modeling, a malicious sensor has two degrees of freedom, i.e., two parameters $\varphi_{01}$ and $\varphi_{10}$, in launching a flipping attack. The conventional frequency check (FC), which detects malicious sensors by computing their frequencies of reporting 1 [12], enforces only one constraint to the attacker’s behavior as indicated in Eq.(6) below. This is insufficient to prevent the malicious sensor from attacking. In contrast, our proposed CFC can enforce two constraints by exploring the correlation between consecutive spectrum states when the true spectrum states are Markovian, and consequently can identify any flipping attacker easily. In particular, the CFC consists of two statistics as defined below.

Definition 1: The two conditional frequency check statistics of a sensor are defined as $\Psi_1 \triangleq Pr(r_t = 1|r_{t-1} = 1)$, and $\Psi_0 \triangleq Pr(r_t = 0| r_{t-1} = 0)$, respectively, where $r_t$ denotes the sensor’s report at time $t$.

According to the definition, these two statistics are related to the model parameters as (see Appendix A)

$$
\Psi_1 = \frac{\pi_0 \pi_{a1} P_{fa} + (\pi_0 \pi_{a1} + \pi_1 \pi_{a0}) P_d P_{fa} + \pi_1 \pi_{a1} P_d^2}{\pi_0 P_{fa} + \pi_1 P_d},
$$

$$
\Psi_0 = \frac{\pi_0 \pi_{a0}(1 - P_{fa})^2 + (\pi_0 \pi_{a0} + \pi_1 \pi_{a1}) (1 - P_d) (1 - P_{fa})}{\pi_0 (1 - P_{fa}) + \pi_1 (1 - P_d)}.
$$

In the CFC, the fusion center will evaluate $\Psi_1$ and $\Psi_0$ for every sensor and compare the resulting values with those of the trusted sensor. If the values are different, the corresponding sensor will be identified as malicious. In the following, the effectiveness of this statistical check is demonstrated through two analytical results, followed by a practical approach to estimating these two statistics that eliminates the requirement of any prior knowledge about the sensing and spectrum models.

Proposition 1: For the Markov spectrum model considered in this paper, any sensor that survives the CFC can pass the FC.

Proof: A malicious sensor can pass the FC as long as $Pr(r^{(M)}_t = 1) = Pr(r^{(tr)}_t = 1)$, where $r^{(M)}_t$ ($r^{(tr)}_t$) denotes the malicious (trusted) sensor’s report at time $t$. However, the malicious sensor needs to satisfy both $\Psi_1^{(M)} = \Psi_1^{(tr)}$ and $\Psi_0^{(M)} = \Psi_0^{(tr)}$ to survive the CFC.

Note that $Pr(r^{(tr)}_t = i) = Pr(r^{(tr)}_{t-1} = i) (i \in \{0, 1\})$ when the true spectrum states are in equilibrium, and $Pr(r^{(tr)}_t = 1) = \Psi_1 P_t^{(tr)} + (1 - \Psi_0) P_{fa}(r^{(tr)}_{t-1} = 0)$. Consequently, for any sensor that survives the CFC, we have

$$
Pr(r^{(M)}_t = 1) = \frac{1 - \Psi_0^{(M)}}{2 - \Psi_1^{(M)} - \Psi_0^{(M)}} = \frac{1 - \Psi_0^{(tr)}}{2 - \Psi_1^{(tr)} - \Psi_0^{(tr)}} = Pr(r^{(tr)}_t = 1),
$$

which implies that this sensor can also pass the FC.

Proposition 2: If $\frac{a_{10} P_{fa} + a_{01} P_d}{a_{10} + a_{01}} \neq \frac{1}{2}$, a malicious sensor can never pass the CFC if it attacks, i.e., $\{\varphi_{01}, \varphi_{10}\} \neq \{0, 0\}$. If $\frac{a_{10} P_{fa} + a_{01} P_d}{a_{10} + a_{01}} = \frac{1}{2}$, an active malicious sensor can pass the CFC only if it sets $\{\varphi_{01}, \varphi_{10}\}$ to $\{1, 1\}$.

Proof: According to Proposition 1, passing the FC is a necessary condition for a malicious sensor to pass the CFC. Thus, $\varphi$ must satisfy $\pi_0 P_{fa} + \pi_1 P_d = Pr(r^{(M)}_t = 1) = Pr(r^{(tr)}_t = 1) = \pi_0 P_{fa} + \pi_1 P_d$ to survive the CFC. Considering (1) and (2), this implies the following linear constraint on $\varphi_{01}$ and $\varphi_{10}$:

$$
\varphi_{01}(\pi_0 (1 - P_{fa}) + \pi_1 (1 - P_d)) = \xi_1 (\pi_0 P_{fa} + \pi_1 P_d).
$$

When (6) holds, define $g_1(\varphi_{01}) \triangleq (\pi_0 P_{fa} + \pi_1 P_d) \cdot (\Psi_1^{(M)} - \Psi_1)$. After some algebra, it can be shown that

$$
g_1(\varphi_{01}) = \xi_1 \pi_0 \pi_1 (\pi_0 \pi_{a1} P_{fa} + \pi_1 a_{01} P_d) (\Psi_1^{(M)} - \Psi_1).
$$

Note that the malicious sensor can pass the CFC only if it could find a $\varphi^* = [\varphi_{01}^*, \varphi_{10}^*]$ that satisfies both $g_1(\varphi_{10}^*) = 0$ (i.e., $\Psi_1^{(M)} = \Psi_1$) and (6). Denote $\varphi_{10}^*$ as the non-zero root of $g_1(\varphi_{01}) = 0$, which can be found as:

$$
\varphi_{10}^* = -\frac{\xi_2}{\kappa_1},
$$

where $\xi_1 \triangleq \pi_0 \pi_1 (\pi_0 \pi_{a1} P_{fa} + \pi_1 a_{01} P_d)$ and $\xi_2 \triangleq 2\pi_1 a_{01} P_{fa} + (\pi_1 a_{10} + \pi_0 a_{01}) \cdot (\pi_0 P_{fa} - \pi_1 P_d)$.

According to (6) and (8), $\varphi_{01}^*$ is given by

$$
\varphi_{01}^* = -\frac{\xi_2}{\kappa_0 \xi_1},
$$

where $\kappa_0 \triangleq \pi_0 P_{fa} - P_d (\pi_0 P_{fa} + \pi_1 P_d)$.

Consider the relation $\pi A = \pi$, (8) and (9) can be simplified as

$$
\varphi_{10}^* = -\frac{2(a_{10} P_{fa} + a_{01} P_d)}{a_{10} + a_{01}},
$$

$$
\varphi_{01}^* = \frac{2(a_{10} P_{fa} + a_{01} P_d)}{a_{10} + a_{01}}.
$$

As a direct consequence of (10) and (11), $\varphi_{10}^* + \varphi_{01}^* = 2$ must hold if the malicious sensor wants to pass the CFC. On the other hand, $0 \leq \varphi_{01}^*, \varphi_{10}^* \leq 1$ by definition. These two conditions imply that $\{\varphi_{01}^*, \varphi_{10}^*\}$ exists only if $\frac{2(a_{10} P_{fa} + a_{01} P_d)}{a_{10} + a_{01}} = 1$ and the corresponding $\{\varphi_{01}^*, \varphi_{10}^*\}$ equals (1, 1). Otherwise, there is no valid non-zero solution for both $g_1(\varphi_{10}) = 0$ and (6). That is, the malicious sensor cannot pass the CFC if it attacks.

Define the error function

$$
e(\varphi) = ||\Psi^{(tr)} - \Psi^{(M)}||_2.
$$
where \( \Psi(\text{tr}) = [\Psi_1^{(tr)}, \Psi_0^{(tr)}] \) and \( \Psi(\text{M}) = [\Psi_1^{(M)}, \Psi_0^{(M)}] \) are the CFC statistics of the trusted sensor and the malicious sensors, respectively. A typical figure of \( e(\varphi) \) when the condition \( \frac{a_{10}P_{fa} + a_{01}P_d}{a_{10} + a_{01}} = \frac{1}{2} \) holds is shown in Fig. 1. As can be seen, \([1, 1]\) is the only blind spot of the CFC. In contrast, the conventional FC only enforces a linear constraint (6) on the attacker, and thus leaves a blind line as indicated in Fig. 1. Furthermore, for a fixed \( \varphi \), a larger \( e(\varphi) \) implies a more significant statistical deviation of the malicious report as compared to the honest report and consequently easier detection. Considering this, the notion of detection margin is proposed in the following to characterize the difficulty of detecting the malicious users.

**Definition 2:** Given \( P_d \) and \( P_{fa} \) of honest sensors and the spectrum state transition matrix \( A \), the detection margin (DM) of the CFC is defined as the maximum \( e(\varphi) \) for any attacking strategy \( \varphi \) that survives the frequency check. That is,

\[
DM(P_d, P_{fa}, A) \triangleq \max_{\varphi \text{ satisfies (6)}} e(\varphi). 
\]

Remark: Seen from (7), for a given blind line (6), \( (\Psi_1^{(M)} - \Psi_1^{(tr)}) \) is a quadratic function, and so is \( (\Psi_0^{(M)} - \Psi_0^{(tr)}) \). The DM is the peak of the corresponding quadratic curve above the blind line by definition (see Fig. 1). Thus for any given blind line, a larger DM implies easier detection of malicious sensors. The following proposition relates the value of the detection margin to the probabilities of detection and false alarm of honest users.

**Proposition 3:** The detection margin \( DM \) is an increasing function of \( P_d \) and \( (1 - P_{fa}) \) when \( P_d \geq 0.5 \) and \( P_{fa} \leq 0.5 \). The minimum value of \( DM \) is zero and is achieved when \( P_d = P_{fa} = 0.5 \).

Proof: From (7), it can be seen that, for any fixed \( P_d, P_{fa} \), and \( A \), the maximum of \( e(\varphi) \) is achieved at \( \varphi_0 = \frac{1}{2}(0 + \varphi_0^*) \) and \( \varphi_0^* = 1 - 2(0 + \varphi_0^*) \) when (6) is satisfied, where \( \varphi_0^{*10} \) and \( \varphi_0^{*01} \) are given in (10) and (11) respectively. Further, it can be verified that \( \Psi_1^{(M)} = \frac{1}{2}\varphi_0^{*10} \) at \( \varphi = \{\varphi_0, \varphi_0^{*10}\} \). Consequently,

\[
\begin{align*}
|\Psi_{1}^{(tr)} - \Psi_{1}^{(M)}(\varphi)| &= |\Psi_{1}^{(tr)} - \frac{1}{2}\varphi_0^{*10}| \\
&= \frac{a_{10}a_{01}(1 - a_{01} - a_{10})(P_d - P_{fa})^2}{(a_0 + a_1)(a_0P_d + a_1P_{fa})} \\
&= \frac{a_{10}a_{01}(1 - a_{01} - a_{10})}{(a_0 + a_1)} \cdot \frac{(P_d - P_{fa})^2}{(a_0P_d + a_1P_{fa})}.
\end{align*}
\]

Taking the derivative of the second term with respect to \( P_d \) leads to

\[
\frac{\partial e}{\partial P_d} = \frac{(P_d - P_{fa})^2}{(a_0P_d + a_1P_{fa})} = \frac{1}{(a_0P_d + a_1P_{fa})}\cdot\left(\frac{2a_{10}P_{fa} + a_{01}P_d}{a_{10} + a_{01}}(P_d - P_{fa})P_{fa} \geq 0, \right.
\]

which implies that \( |\Psi_{1}^{(tr)} - \Psi_{1}^{(M)}(\varphi)| \) is an increasing function of \( P_d \). It can be shown that \( |\Psi_{1}^{(tr)} - \Psi_{1}^{(M)}(\varphi)| \) is also an increasing function of \((1 - P_{fa})\). Similarly, one can verify that \( |\Psi_{1}^{(tr)} - \Psi_{1}^{(M)}(\varphi)| \) is an increasing function of \( P_d \) and \((1 - P_{fa})\). Therefore, \( DM = ||\Psi_{1}^{(tr)} - \Psi_{1}^{(M)}(\varphi)||_2 \) is an increasing function of \( P_d \) and \((1 - P_{fa})\). The fact that the minimum value of \( DM \) is zero and is achieved at \( P_d = P_{fa} = 0.5 \) can readily be observed from (14).

Remark: Proposition 3 indicates that if honest users have better sensing quality, the detection margin is larger, which in turn implies stronger resistance to Byzantine attackers. On the other hand, a sensor network of poor quality is vulnerable to Byzantine attacks; the worst case is shown in Fig. 2 where \( P_d = P_{fa} = 0.5 \) and \( DM = 0 \). The notion of detection margin and its relation to \( P_d \) and \( P_{fa} \) will be used in Section VI to explain the behaviors of CFC for different \( P_d \) and \( P_{fa} \).

**Definition 3:** For any sensor, two histogram estimators for \( \Psi_{1} \) and \( \Psi_{0} \) are defined as

\[
\Psi_{1} \triangleq \left( \sum_{i=0}^{T-1} \delta_{t+1,i} \delta_{t+1,j} \right) / \left( \sum_{i=0}^{T-1} \delta_{t+1,i} \right),
\]

\[
\Psi_{0} \triangleq \left( \sum_{i=0}^{T-1} \delta_{t+1,i} \delta_{t+1,j} \right) / \left( \sum_{i=0}^{T-1} \delta_{t+1,i} \right),
\]

respectively, where \( \delta_{i,j} = 1 \) iff \( i = j \) and \( T \) is the detection window length.

**Proposition 4:** The two estimators \( \hat{\Psi}_1 \) and \( \hat{\Psi}_0 \) converge to \( \Psi_{1} \) and \( \Psi_{0} \), respectively, as \( T \to \infty \).

Proof: The proof is given in the Appendix B.

Remark: According to Proposition 4, the CFC statistics of all honest sensors (including the trusted one) will converge to the same value, i.e., \( \Psi(\text{tr}) \). On the other hand, the CFC statistics of any malicious sensor \( i \) will converge to some value \( \Psi(M.i) \) (depending on its \( \varphi(i) \)), which is different from \( \Psi(\text{tr}) \) according to Proposition 2. Therefore, any sensor whose CFC statistic differs from that of the trusted sensor is malicious.

Based on this rationale, the proposed CFC procedure is presented in Algorithm 1 where a threshold \( \beta_{CFC} \) is required due to the finite detection window length in practice.

**B. The Hamming Distance Check**

As shown in Fig. 1, the CFC will fail to detect a malicious sensor employing \( \varphi = \{1, 1\} \) when \( \frac{a_{10}P_{fa} + a_{01}P_d}{a_{10} + a_{01}} = \frac{1}{2} \). This may happen, for example, when \( a_{10} = a_{01} \) and \( P_d + P_{fa} = 1 \).
A. Selection of the asymptotic independence property which has already been the accurate evaluation of the standard deviations difficult. In the Markovian property of the true spectrum state renders the case that the trusted sensor is not available is presented. This section extends and a trusted sensor is always available. This section extends well by simulation results in Section VI.

\[ \beta_{CFC} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{t_i}, \]
and \( E[X_{t_i}] = \Psi_1 \). Also, it can be verified that \( \text{Var}[X_{t_i}] = \Psi_1(1 - \Psi_1) \) due to its Bernoulli distribution. Similar observations apply to \( \Psi_0 \). Therefore, we can approximate the standard deviation of \( \Psi_1 \) and \( \Psi_0 \) as \( \sqrt{\frac{\Psi_1 - \Psi}{n_1}} \) and \( \sqrt{\frac{\Psi_0 - \Psi}{n_0}} \), respectively, by assuming that \( X_{t_i} \)'s are independent. Also, \( \pi_0 T \) and \( \pi_1 T \) can be used as estimates of \( n_0 \) and \( n_1 \) (defined in the Appendix B), respectively. Further, although \( \Psi_0 \) and \( \Psi_1 \) are correlated in general, we treat them as independent to avoid tedious computation, and use \( \max \{ \sqrt{\frac{\Psi_1 (1 - \Psi_1)}{\pi_1 T}}, \sqrt{\frac{\Psi_0 (1 - \Psi_0)}{\pi_0 T}} \} \) to approximate the standard deviation of \( ||\Psi_1 - \Psi||_2 \), which equals the standard deviation of \( ||\Psi - \Psi||_2 \). In Claim 1, \( \beta_{CFC} \) is chosen as the 3\( \sigma \) confidence interval based on this approximate standard deviation, where the true values of \( \Psi \) and \( \pi \) are replaced by \( \hat{\Psi}^{(tr)} \) and 0.5, respectively.

B. Selection of \( \beta_{HDC} \)

Claim 2: An approximation of \( \beta_{HDC} \) is given by \( P_{fa}(1 - P_{fa}) + P_d(1 - P_d) + 3\sqrt{f(P_{fa}) + f(P_d)} \), where \( f(x) = x(1 - x)(1 - 2x + 2x^2) \).

Remark: As shown in the Appendix B, \( \Psi_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{t_i} \), and \( E[X_{t_i}] = \Psi_1 \). Also, it can be verified that \( \text{Var}[X_{t_i}] = \Psi_1(1 - \Psi_1) \) due to its Bernoulli distribution. Similar observations apply to \( \Psi_0 \). Therefore, we can approximate the standard deviation of \( \Psi_1 \) and \( \Psi_0 \) as \( \sqrt{\frac{\Psi_1 - \Psi}{n_1}} \) and \( \sqrt{\frac{\Psi_0 - \Psi}{n_0}} \), respectively, by assuming that \( X_{t_i} \)'s are independent. Also, \( \pi_0 T \) and \( \pi_1 T \) can be used as estimates of \( n_0 \) and \( n_1 \) (defined in the Appendix B), respectively. Further, although \( \Psi_0 \) and \( \Psi_1 \) are correlated in general, we treat them as independent to avoid tedious computation, and use \( \max \{ \sqrt{\frac{\Psi_1 (1 - \Psi_1)}{\pi_1 T}}, \sqrt{\frac{\Psi_0 (1 - \Psi_0)}{\pi_0 T}} \} \) to approximate the standard deviation of \( ||\Psi_1 - \Psi||_2 \), which equals the standard deviation of \( ||\Psi - \Psi||_2 \). In Claim 1, \( \beta_{CFC} \) is chosen as the 3\( \sigma \) confidence interval based on this approximate standard deviation, where the true values of \( \Psi \) and \( \pi \) are replaced by \( \hat{\Psi}^{(tr)} \) and 0.5, respectively.

\[ \beta_{HDC} = \frac{1}{T} \sum_{t=1}^{T} \delta(h_{i,t}) = \frac{1}{T} \sum_{t=1}^{T} \delta(h_{i,t}), \]

is expected between the reported sequences from a malicious sensor \( i \) and the trusted sensor, because of the high flipping probability \( \phi \). Based on this observation, sensor \( i \) will be identified as malicious if \( dh(i, tr) \) is greater than a pre-specified threshold \( \beta_{HDC} \).

IV. SELECTION OF THRESHOLDS

This section provides guidance on the selection of the two thresholds \( \beta_{CFC} \) and \( \beta_{HDC} \). Specifically, since both \( \Psi \) and \( dh \) are random, \( \beta_{CFC} \) and \( \beta_{HDC} \) may in principle be determined by their corresponding \( 3 \sigma \) confidence regions [34]. However, the Markovian property of the true spectrum state renders the accurate evaluation of the standard deviations difficult. In the following, some approximate results are obtained through the asymptotic independence property which has already been exploited in the proof of Proposition 4 (see the Appendix B).

A. Selection of \( \beta_{CFC} \)

Claim 1: An approximation of \( \beta_{CFC} \) is given by \( \max \left\{ 3 \sqrt{\frac{\hat{\Psi}^{(tr)}(1 - \Psi^{(tr)})}{0.5 T}}, 3 \sqrt{\frac{\hat{\Psi}^{(tr)}(1 - \Psi^{(tr)})}{0.5 T}} \right\} \).

Remark: As shown in the Appendix B, \( \hat{\Psi}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{t_i}, \)
and \( E[X_{t_i}] = \Psi_1 \). Also, it can be verified that \( \text{Var}[X_{t_i}] = \Psi_1(1 - \Psi_1) \) due to its Bernoulli distribution. Similar observations apply to \( \hat{\Psi}_0 \). Therefore, we can approximate the standard deviation of \( \hat{\Psi}_1 \) and \( \hat{\Psi}_0 \) as \( \sqrt{\frac{\hat{\Psi}_1 - \Psi}{n_1}} \) and \( \sqrt{\frac{\hat{\Psi}_0 - \Psi}{n_0}} \), respectively, by assuming that \( X_{t_i} \)'s are independent. Also, \( \pi_0 T \) and \( \pi_1 T \) can be used as estimates of \( n_0 \) and \( n_1 \) (defined in the Appendix B), respectively. Further, although \( \hat{\Psi}_0 \) and \( \hat{\Psi}_1 \) are correlated in general, we treat them as independent to avoid tedious computation, and use \( \max \{ \sqrt{\frac{\hat{\Psi}_1 (1 - \hat{\Psi}_1)}{\pi_1 T}}, \sqrt{\frac{\hat{\Psi}_0 (1 - \hat{\Psi}_0)}{\pi_0 T}} \} \) to approximate the standard deviation of \( ||\hat{\Psi}_1 - \Psi||_2 \), which equals the standard deviation of \( ||\hat{\Psi} - \Psi||_2 \). In Claim 1, \( \beta_{CFC} \) is chosen as the 3\( \sigma \) confidence interval based on this approximate standard deviation, where the true values of \( \Psi \) and \( \pi \) are replaced by \( \hat{\Psi}^{(tr)} \) and 0.5, respectively.

B. Selection of \( \beta_{HDC} \)

Claim 2: An approximation of \( \beta_{HDC} \) is given by \( P_{fa}(1 - P_{fa}) + P_d(1 - P_d) + 3\sqrt{f(P_{fa}) + f(P_d)} \), where \( f(x) = x(1 - x)(1 - 2x + 2x^2) \).

Remark: The probabilities that two honest sensors report different observations are \( 2P_{fa}(1 - P_{fa}) \) and \( 2P_d(1 - P_d) \)
when the true spectrum states are 0 and 1, respectively. Therefore, the expectation of \( dh \), the normalized hamming distance between two honest sensors, is given by \( E(dh) = 2\pi_0 P_{fa}(1 - P_{fa}) + 2\pi_1 P_d(1 - P_d) \). By a similar argument as in the derivation of Claim 1 above, we approximate the variance of \( dh \) by \( [2\pi_0 P_{fa}(1 - P_{fa})(1 - 2P_{fa} + 2P_{fa}^2) + 2\pi_1 P_d(1 - P_d)(1 - 2P_d + 2P_d^2)]/T \). In Claim 2, \( \beta_{HDC} \) is chosen as the right 3\( \sigma \) deviation bound of \( dh \) (i.e., \( E(dh) + 3\sigma_{dh} \)) based on this approximate variance, where the true values of \( \pi_0 \) and \( \pi_1 \) are replaced by 0.5. In practice, nominal values of \( P_d \) and \( P_{fa} \) can be used to compute \( \beta_{HDC} \).

V. EXTENSIONS OF THE CFC

The previous analysis demonstrates the effectiveness of the proposed CFC under the assumptions that both honest and malicious users are equipped with the same spectrum sensing devices (i.e., \( \gamma_0 = \gamma_1 = 1 \) for any malicious sensor \( i \)) and a trusted sensor is always available. This section extends the study to more general circumstances. In particular, the effectiveness of the CFC to more general attackers are shown analytically, and a complementary approach of the CFC for the case that the trusted sensor is not available is presented.

A. More General Attackers

In practice, malicious users may carry devices that have different spectrum sensing accuracy as compared to the honest

4Note that \( X_{t_i} \)'s are actually dependent as argued in Appendix B. The independence assumption here is intended for simplifying the computation and obtaining an approximate threshold, which is shown to work reasonably well by simulation results in Section VI.
sensors. For this reason, more general attackers are considered in this subsection. In particular, the probabilities of detection and false alarm of the $i$-th malicious sensors are $\gamma_1(i)P_d$ and $\gamma_0(i)P_{fa}$, respectively, where $\gamma_1(i)$ and $\gamma_0(i)$ are arbitrary so long as $0 \leq \gamma_1(i)P_d + \gamma_0(i)P_{fa} \leq 1$. The following proposition shows that the proposed CFC is still effective even in the presence of such more general attackers.

**Proposition 5:** For any attacker with spectrum sensing ability $\gamma_1P_d$ and $\gamma_0P_{fa}$, the blind spot of the CFC $\{\varphi_{01}, \varphi_{01}^*\}$ is given by

$$\varphi_{01}^* = 2\frac{a_{01}P_d + a_{10}P_{fa}}{a_{01} + a_{10}} - \frac{\gamma_1 - \gamma_0}{\gamma_1P_d - \gamma_0P_{fa}}, \quad (18)$$

$$\varphi_{01}^* = 1 - \varphi_{01}^* + \frac{P_d - P_{fa}}{\gamma_1P_d - \gamma_0P_{fa}}, \quad (19)$$

**Proof:** Define the equivalent flipping indices $\varphi_{10}'$ and $\varphi_{01}'$ such that

$$P_d^{(M)} = (1 - \varphi_{10} - \varphi_{01}')\gamma_1P_d + \varphi_{01}' = (1 - \varphi_{10} - \varphi_{01})P_d + \varphi_{01}', \quad (20)$$

$$P_{fa}^{(M)} = (1 - \varphi_{10} - \varphi_{01}')(\gamma_0P_{fa} + \varphi_{01}') = (1 - \varphi_{10} - \varphi_{01})P_{fa} + \varphi_{01}', \quad (21)$$

where $\varphi_{10}'$ and $\varphi_{01}'$ are not probabilities and hence could be greater than 1. Taking the difference of (20) and (21) results in

$$(1 - \varphi_{10}' - \varphi_{01}') = (1 - \varphi_{10} - \varphi_{01})\frac{\gamma_1P_d - \gamma_0P_{fa}}{P_d - P_{fa}}. \quad (22)$$

Substituting the preceding equation back into (21), it has

$$\varphi_{01}' = \varphi_{01} + (1 - \varphi_{10} - \varphi_{01})P_dP_{fa}\frac{\gamma_0 - \gamma_1}{P_d - P_{fa}}. \quad (23)$$

By noticing the similarity between the structures of (1), (2) and (20), (21), the same approach as in the proof of Proposition 2 can be applied here to find the blind spot of the CFC. In particular, the following two analogues of (10) and (11) can be observed

$$\varphi_{10}' = \frac{2\left(a_{10}P_{fa} + a_{01}P_d\right)}{a_{10} + a_{01}}, \quad (24)$$

$$\varphi_{01}' = \frac{2\left(a_{10}P_{fa} + a_{01}P_d\right)}{a_{10} + a_{01}}. \quad (25)$$

Substituting (24) and (25) into (22) and (23) leads to (18) and (19).

**Remark:** In fact, Proposition 2 is a special case of Proposition 5 when $\gamma_0 = \gamma_1 = 1$. Further, it is worth noting that another choice for the malicious sensor to pass the CFC is to set $\varphi_{01} = \varphi_{01}^* = \frac{P_d - P_{fa}}{\gamma_1P_d - \gamma_0P_{fa}}$ and $\varphi_{10} = 1 - \varphi_{01} + \frac{P_d - P_{fa}}{\gamma_1P_d - \gamma_0P_{fa}}$, which corresponds to a flipping indices pair $\{\varphi_{01}^*, \varphi_{10}'\} = \{0, 0\}$. But it can be verified that $P_d^{(M)} = P_d$ and $P_{fa}^{(M)} = P_{fa}$ in this case. That is, the malicious sensor behaves statistically the same as the honest sensor.

**Corollary 1:** For a malicious sensor with weaker spectrum sensing ability than the honest ones, the CFC has no blind spot. (Specifically, $\gamma_1$ and $\gamma_0$ of a weaker malicious sensor admit $\gamma_1 < 1$, $\gamma_0 \geq 1$, or $\gamma_1 = 1$, $\gamma_0 > 1$.)

**Proof:** In this case, it can be verified that surviving the CFC requires $\varphi_{01}^* + \varphi_{01}^* > 2$ according to (18) and (19), which implies invalid $\varphi_{01}^*$ and $\varphi_{10}'$.

**Remark:** For a malicious sensor more powerful than the honest ones, the existence of the blind spot of the CFC depends on the specific values of corresponding sensing ability and true spectrum parameters.

**Proposition 6:** The expectation of the normalized hamming distance between a CFC-surviving malicious sensor $m$ and the trusted sensor $E[d_h(m, tr)]$ is no less than that between a honest sensor $h$ and the trusted sensor $E[d_h(h, tr)]$. The expectation of the distance gap $E[d_h(m, tr)] - E[d_h(h, tr)]$ is irrespective of $\gamma_1^{(m)}$ and $\gamma_0^{(m)}$, and is lower bounded by $(P_d - P_{fa}) \cdot \min(2P_d - 1, 1 - 2P_{fa})$, which is non-trivial when $P_d > 0.5$ and $P_{fa} < 0.5$.

**Proof:** Note that given any two sensors $i$ and $j$ with equivalent spectrum sensing capabilities $P_d^{(i)}$, $P_{fa}^{(i)}$ and $P_d^{(j)}$, $P_{fa}^{(j)}$, the expected normalized hamming distance is given by

$$E[d_h(i, j)] = \sum_{k=0}^{1} \pi_k \left[P_h(i^{(k)}) = 1, r_t^{(j)} = 0 | s_t = k\right]$$

$$+ \sum_{k=0}^{1} \pi_k \left[P_h(i^{(k)}) = 0, r_t^{(j)} = 1 | s_t = k\right]$$

$$= \pi_0 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right]$$

$$+ \pi_1 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right]. \quad (26)$$

Thus,

$$E[d_h(m, tr)] - E[d_h(h, tr)] = \pi_0 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right]$$

$$+ \pi_1 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right]$$

$$= \pi_0 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right]$$

$$- \pi_0 \left[P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right] - \pi_1 \left(P_h^{(j)}(1 - 2P_{fa}^{(j)}) + P_{fa}^{(j)}\right).$$

Since the malicious sensor survives the CFC, one can substitute (20), (21), (24) and (25) into the preceding equation, which leads to

$$E[d_h(tr, m)] - E[d_h(tr, h)]$$

$$= 2(P_d - P_{fa}) a_{10}^2(2P_d - 1) + a_{01}^2(1 - 2P_{fa})$$

$$\geq (P_d - P_{fa}) a_{10}^2(2P_d - 1) + a_{01}^2(1 - 2P_{fa})$$

$$\geq (P_d - P_{fa}) \cdot \min(2P_d - 1, 1 - 2P_{fa}). \quad (28)$$

**Remark:** Although the CFC may have a blind spot to non-weaker malicious sensors and fail to detect them, Proposition 6 implies that the HDC can be employed to detect such malicious sensors effectively as long as $(P_d - P_{fa}) \cdot \min(2P_d - 1, 1 - 2P_{fa}) > 0$, which is generally true in practice.
B. In the Absence of Trusted Sensor

When there is no trusted user in the network and the fusion center is distant from the sensing spot (and thus cannot serve as a sensor), a heuristic clustering method shown in Algorithm 2 can be used to detect the malicious sensors when honest users dominate the network (accounting for more than 50% of the total number of nodes). The intuitive idea behind Algorithm 2 is that two users $i$ and $j$ will fall into the same cluster if their corresponding CFC statistics satisfy $|\Psi(i) - \Psi(j)| \leq \beta_{CFC}$ and $\forall j \in C$. The condition $|\Psi(j) - \Psi(i)| \leq \beta_{CFC}$ ($\forall j \in C$) ensures that any two users within the same cluster will not deviate from each other dramatically. As a result, $\Psi$’s of honest sensors will form a cluster with approximate radius $\beta_{CFC}/2$. In addition, due to the honest user domination assumption, the cluster formed by the honest sensors always has larger cardinality than any other cluster formed by malicious sensors.

Algorithm 2 User classification by clustering

Start with the set of all sensors $U = \{1, 2, \ldots, N\}$.

while $U \neq \emptyset$ do

Pick a random $k \in U$.

Create a new cluster $C = \{k\}$.

$U = U \setminus \{k\}$.

for $i \in U$ do

if $\exists j \in C$ such that $|\Psi(j) - \Psi(i)| \leq \beta_{CFC}$ and $\forall j \in C$,

$C = C \cup \{i\}$, $U = U \setminus \{i\}$.

end if

end for

end while

Sensors in the cluster of the largest cardinality are honest.

Since Algorithm 2 may fail to filter out some malicious sensors due the existence of the blind spot, a modified HDC is presented in Algorithm 3 to catch these malicious sensors. Particularly, the modified HDC divides the sensors that survive Algorithm 2 into two groups $U_1$ and $U_2$ by the corresponding normalized hamming distances. By the assumption that the honest sensors dominate the network, the one of $U_1$ and $U_2$ with larger cardinality will be identified as the honest sensor set. In particular, in Algorithm 3, $H$ and $M$ denote the sets of honest sensors and malicious sensors identified by Algorithm 2, respectively. In addition, $|H|$ denotes the cardinality of the set $H$. Note that $|M| > |H|$ may happen, which implies that although the effectiveness of Algorithm 3 is ensured only when $|M| < |H|$, it may work for certain cases with $|M| > |H|$. Specifically, such case happens when the CFC statistics of malicious sensors form multiple clusters with each having a cardinality less than $|H|$.

VI. SIMULATIONS

Throughout the simulations, each malicious sensor randomly selects its own $\{\varphi_0, \varphi_1\}$ uniformly over $(0, 1)^2$. The total number of honest and malicious sensors are denoted by $n_H$ and $n_M$, respectively. In addition, it is assumed that the first honest sensor is the trusted one, which is known by the fusion center. The thresholds $\beta_{CFC}$ and $\beta_{HDC}$ are selected according to Claim 1 and Claim 2, respectively. At the fusion center, AND rule, OR rule and majority voting are all valid options for data fusion. Among them, the AND rule tends to exploit the primary channel aggressively with the risk of introducing excessive interference to the primary users while OR rule is the opposite; majority voting lies in between. As the proposed malicious user detection method is invariant to the specific fusion rule, any one of these data fusion rules can be used. Specifically, majority voting is adopted in the following simulations. Fig. 3 depicts the block diagram of the proposed algorithm. As shown, the most recent $T$ spectrum sensing history of all users $r_{t-T+1:t}$’s are stored for malicious user detection, and the fusion center will determine the current spectrum state estimate $\hat{s}_t$ based on current sensing reports $r_t$’s and the detected user types. (The probabilities of detection and false alarm at the fusion center will be denoted as $P_d'$ and $P_{fa}'$, respectively, in the following.)

A. Basic Examples

Two different cases are simulated in this subsection to demonstrate how the proposed method effectively detects the malicious sensors and enhances the spectrum sensing results accordingly. In both cases, it is assumed that all malicious users are equipped with the same sensing devices as honest ones (i.e., $\lambda(0) = \lambda(1) = 1$ for any malicious sensor $i$),

Algorithm 3 Modified HDC

Start with $H = \{h_1, h_2, \ldots\}$.
Pick a random $k$ (1 $\leq k \leq |H|$).
Set $U_1 = H$ and $U_2 = \emptyset$.
for $i = 1 : |H|$ do

if $d_h(h_k, h_i) \geq \beta_{HDC}$ then

Move $h_i$ from $U_1$ to $U_2$.
end if
end for

if $|U_1| \geq |U_2|$ then

Malicious sensors: $M = U_2 \cup M$.
Honess sensors: $H = U_1$.
else

Malicious sensors: $M = U_1 \cup M$.
Honess sensors: $H = U_2$.
end if

Current sensing reports: $r_t$

Data fusion

Spectrum state estimate: $s_t$

Buffer

History: $r_{t-T+1:t}$

Detected user types

Malicious user detection

Fig. 3. The block diagram of the proposed algorithm.
and specifically $P_d = 0.9$ and $P_{fa} = 0.1$. In the first case, $A = [0.8, 0.2]$; while in the second case, $A = [0.8, 0.2]$. Thus, the condition $\frac{a_{10}P_{fa} + a_{01}P_d}{a_{10} + a_{01}} = \frac{1}{2}$ is satisfied in the first case but not in the second one. There are $n_H = 8$ honest sensors and $n_M = 13$ malicious sensors, i.e., the malicious sensors dominate the network. The detection window length is $T = 100$ (time slot).

Simulation results of a typical run of the first case are shown in Fig. 4–Fig. 6. By comparing Fig. 4 and Fig. 5, it can be seen that two malicious sensors whose flipping probabilities $\varphi_{01}$ and $\varphi_{10}$ are close to 1 successfully pass the CFC. However, these two malicious sensors fail to pass the succeeding HDC. Also, it can be seen by comparing Fig. 5 and Fig. 6 that there is one malicious user surviving both CFC and HDC. Further examination reveals that the flipping probabilities of this malicious user are low: $\varphi_{01} \approx 0$ and $\varphi_{10} \approx 0.1$. Although this malicious sensor is not detected, its negative influence on the spectrum sensing result of the fusion center is negligible.

Fig. 7–Fig. 8 show results of a typical run of the second case where $\frac{a_{10}P_{fa} + a_{01}P_d}{a_{10} + a_{01}} \neq \frac{1}{2}$. As expected, the CFC alone successfully detects the severe attacker without activating the HDC, because the CFC has no blind point in this case and can detect any active attacker according to Proposition 2.

Table I summarizes the simulation results over 1000 Monte Carlo runs for these two cases. As can be seen, in both cases, the proposed algorithm (using both CFC and HDC) provides high malicious sensor detection accuracy ($\eta > 95\%$) with a detection window of length 100 time slots. Accordingly, it also achieves nearly perfect spectrum sensing results in both cases, i.e., $P_d = 0.995$ and $P_{fa} = 0.001$ in the first case, and $P_d = 0.996$ and $P_{fa} = 0.001$ in the second case, which are significantly better than the performance of using only the trusted sensor and that of using all sensors without malicious user detection.

B. Further Simulations

This subsection will demonstrate how user classification accuracy and spectrum sensing performance of the proposed method are affected by 1) the detection window length $T$, 2) the spectrum sensing ability of the honest sensors, i.e., $P_d$
Table I

Average Performance Comparison over 1000 runs.

<table>
<thead>
<tr>
<th></th>
<th>No detection</th>
<th>Trusted only</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{d}^* ) (case one)</td>
<td>0.921</td>
<td>0.900</td>
<td>0.995</td>
</tr>
<tr>
<td>( P_{fa} ) (case one)</td>
<td>0.076</td>
<td>0.102</td>
<td>0.001</td>
</tr>
<tr>
<td>( \eta ) (case one)</td>
<td>( \eta ) (case one)</td>
<td>( \eta ) (case one)</td>
<td>( \eta ) (case one)</td>
</tr>
<tr>
<td>( P_{d}^* ) (case two)</td>
<td>0.928</td>
<td>0.900</td>
<td>0.996</td>
</tr>
<tr>
<td>( P_{fa} ) (case two)</td>
<td>0.077</td>
<td>0.099</td>
<td>0.001</td>
</tr>
<tr>
<td>( \eta ) (case two)</td>
<td>( \eta ) (case two)</td>
<td>( \eta ) (case two)</td>
<td>( \eta ) (case two)</td>
</tr>
</tbody>
</table>

![Image](image-url)

Fig. 9. Spectrum sensing performance at the fusion center in scenario 1.

![Image](image-url)

Fig. 10. Spectrum sensing performance at the fusion center in scenario 2.

\[
P_{d}^* = Pr(\text{More sensors reporting } 1|s_k = 1) = \sum_{i=\lceil \frac{n_H}{2} \rceil + 1}^{n_H} \left( \frac{n_H}{i} \right) (1 - \rho)^{n_H-i}, \tag{29}
\]

As can be seen from Fig. 9–Fig. 12, increasing the detection window length \( T \) improves both the spectrum sensing performance and user classification accuracy of the fusion center. When \( T \) is sufficiently large, the classification accuracy is high (\( \geq 95\% \)), and thus the corresponding spectrum sensing performances are very close to the theoretical optimal ones given in (29) and (30).

Comparing Fig. 9 and Fig. 10, it can be seen that the required detection window length \( T \) for similar performance is much smaller in the scenario 2, as compared to scenario 1. The same observation can be made in Fig. 12. In fact, it can be verified that the corresponding blind lines are the same in these two scenarios while the detection margin in scenario 2 is higher as a consequence of larger \( P_d \) and smaller \( P_{fa} \) (according to Proposition 3). That is, for the same attacking strategy \( \phi \), the statistical deviations \( e(\phi) \) of malicious sensors will be more significant in scenario 2, and thus a shorter detection window will suffice.

 Naturally the adversary may attempt to enhance the attack through increasing the number of malicious sensors. Considering this, the percentage of malicious sensors is increased from \( \rho = 62\% \) in scenario 1 to \( \rho = 90\% \) in scenario 3. As can be seen from Fig. 11 and Fig. 12, the proposed algorithm can still provide appealing classification accuracy and spectrum sensing performances by slightly increasing the detection window length, even though the adversary dramatically increases the amount of malicious sensors. Consequently, the adversary’s attempt to launch more severe flipping attack by increasing the number of malicious sensors is ineffective for our scheme.

C. Malicious Users with More/Less Powerful Sensors

In practice, malicious users may be equipped with different spectrum sensing devices hoping to further disrupt the fusion center. For this reason, this subsection explores the performance of the proposed method when more general malicious sensors exist. In particular, assume that \( \gamma_1 = 1.1, \gamma_0 = 0 \) for strong attackers and \( \gamma_1 = 0.8, \gamma_0 = 1.1 \) for weak attackers. Other parameters remains the same as the first case in Section VI-A. The corresponding results are summarized in Table II.

It can be seen that the proposed algorithm still can detect malicious sensor detection are given by:\(^7\)

\[
P_{fa}^* = Pr(\text{More sensors reporting } 0|s_k = 0) = \sum_{i=\lceil \frac{n_H}{2} \rceil + 1}^{n_H} \left( \frac{n_H}{i} \right) (1 - P_d)^{n_H-i}, \tag{30}
\]

In this paper, only hard-decision malicious user detection is considered. That is, a secondary user is identified either as an honest one or as a malicious one, and the reports from malicious users will be discarded. The analysis for the soft-decision case remains a future work.
malicious sensors with high accuracy and provide satisfactory sensing performance, which justifies the analytical results in Section V-A.

D. Removing the Trusted Sensor

When the trusted sensor is not available, the fusion center can adopt Algorithm 2 and Algorithm 3 to detect the malicious sensors so as to ensure the robustness of collaborative spectrum sensing. Table III summarizes the results over 1000 runs for such a circumstance where \( P_d = 0.8 \) and \( P_{fa} = 0.2 \), \( \gamma_0(i) = \gamma_1(i) = 1 \) for all malicious sensor \( i \), and \( A = [0.4 \ 0.2 \ 0.6] \). There are 5 severe attackers with \( \varphi_{01} = \varphi_{10} = 1 \) while other attackers select their \( \varphi \)'s uniformly over \((0,1)^2\). As can be seen, even when \( \rho = 45\% \) sensors in the network (of 21 sensors) are malicious, the proposed algorithm still can detect them satisfactorily (\( \eta = 92.5\% \)) and improve the spectrum sensing performance from \( P^*_d = 0.864 \), \( P^*_fa = 0.142 \) to \( P_d = 0.975 \), \( P_{fa} = 0.012 \).

Table IV summarizes the results of an even worse case, where the trusted sensor is not available and the malicious users are equipped with more advanced sensing devices \((\gamma_0(i) = 0 \text{ and } \gamma_1(i) = 1.2 \text{ for all malicious sensor } i)\). However, as can be seen from Table IV, the proposed algorithm can still detect the malicious user accurately and improve the spectrum sensing performance significantly.

The similar performances shown in Table III and Table IV justify the robustness of the proposed algorithm to more powerful malicious sensors even when the trusted sensor is not available.

VII. CONCLUSIONS

With Markovian modeling of spectrum states, a new malicious cognitive radio user detection method, which consists of two natural but effective CFC statistics and an auxiliary HDC, is proposed in this paper against the flipping attack. In addition, two consistent histogram estimators of these two statistics are developed so that no prior information on sensing and spectrum models is required in the proposed method. Both theoretical analysis and simulation results show that the proposed method is effective even when the malicious users are equipped with more advanced sensing devices. With the assistance of one trusted sensor, the proposed method can provide high detection accuracy and achieve near optimal collaborative spectrum sensing performance for arbitrary percentage of malicious sensors. In the case when the trusted sensor is not available, the extended algorithms can still maintain satisfactory detection and enhance the sensing performance significantly when the honest sensors dominate the network.

### Table II

<table>
<thead>
<tr>
<th>( P_d ) (strong attacker)</th>
<th>No detection</th>
<th>Trusted only</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_{fa} ) (strong attacker)</td>
<td>0.080</td>
<td>0.099</td>
<td>0.001</td>
</tr>
<tr>
<td>( \eta ) (strong attacker)</td>
<td>( \gamma_0 = 0 )</td>
<td>( \gamma_0 = 0 )</td>
<td>( \eta = 95.0% )</td>
</tr>
<tr>
<td>( P_\eta ) (weak attacker)</td>
<td>0.956</td>
<td>0.900</td>
<td>0.994</td>
</tr>
<tr>
<td>( P_{\eta d} ) (weak attacker)</td>
<td>0.054</td>
<td>0.099</td>
<td>0.001</td>
</tr>
<tr>
<td>( \eta ) (weak attacker)</td>
<td>( \gamma_0 = 0 )</td>
<td>( \gamma_0 = 0 )</td>
<td>98.3%</td>
</tr>
</tbody>
</table>

### Table III

| Performance of the proposed method without the trusted sensor over 1000 runs. |
|-----------------------------|--------------|--------------|----------|
| \( \rho \) | 25% | 30% | 35% | 40% | 45% |
| \( \eta \) | 94.9% | 94.3% | 93.8% | 92.8% | 91.5% |
| \( P_d \) (proposed) | 0.993 | 0.993 | 0.988 | 0.984 | 0.974 |
| \( P_{fa} \) (proposed) | 0.928 | 0.901 | 0.864 | 0.832 | 0.784 |
| \( P_\eta \) (proposed) | 0.004 | 0.005 | 0.005 | 0.008 | 0.011 |
| \( P_{\eta d} \) (proposed) | 0.084 | 0.115 | 0.146 | 0.182 | 0.247 |

### Table IV

| Performance of the proposed method without the trusted sensor over 1000 runs when malicious sensors are more powerful. |
|-----------------------------|--------------|--------------|----------|
| \( \rho \) | 25% | 30% | 35% | 40% | 45% |
| \( \eta \) | 94.9% | 94.3% | 93.8% | 92.8% | 91.5% |
| \( P_d \) (proposed) | 0.993 | 0.993 | 0.988 | 0.984 | 0.974 |
| \( P_{fa} \) (proposed) | 0.928 | 0.901 | 0.864 | 0.832 | 0.784 |
| \( P_\eta \) (proposed) | 0.004 | 0.005 | 0.005 | 0.008 | 0.011 |
| \( P_{\eta d} \) (proposed) | 0.084 | 0.115 | 0.146 | 0.182 | 0.247 |
More advanced attackers with adaptive time-varying flipping probabilities or that can intelligently construct falsified reported sequences satisfying the statistical characteristics of the CFC may exist. Defence against such types of attackers is a more challenging and open problem, which remains a future work.

**APPENDIX A**

**DERIVATIONS OF (3) AND (4)**

By the Bayesian formula and the total probability theorem, we have

\[ \Psi_1 = Pr(r_t = 1|r_{t-1} = 1) = \frac{Pr(r_t = 1, r_{t-1} = 1)}{Pr(r_{t-1} = 1)} \]

(31)

\[ = \sum_{i,j=0}^{1} Pr(s_t = j, s_{t-1} = i)Pr(r_t = 1, r_{t-1} = 1|s_t = j, s_{t-1} = i) \]

\[ = \frac{1}{\pi_0 P_{a0} + P_{a1}^2 + (\pi_0 a_{10} + \pi_1 a_{10}) P_{d} P_{a0} + \pi_1 a_{11} P_{d}^2}, \]

where \( Pr(r_t = 1, r_{t-1} = 1|s_t = j, s_{t-1} = i) = Pr(r_t = 1|s_t = j)Pr(r_{t-1} = 1|s_{t-1} = i) \) is applied. Similar steps verify (4).

**APPENDIX B**

**PROOF OF PROPOSITION 4**

Proof: It can be seen that \( \Psi_1 = \frac{1}{n_i} \sum_{i=1}^{n_i} X_{i, i} \), in which \( X_{i, i} \) is defined as

\[ X_{i, i} = \begin{cases} 1, & \text{if } r_{i+1} = 1, \text{ given } r_i = 1, \\ 0, & \text{if } r_{i+1} = 0, \text{ given } r_i = 1, \end{cases} \]

(32)

where \( t_i \) is the time slot for the \( i \)-th reported 1 of the sensor, and \( n_i \) is the number of pairs \( \{r_{i+1}, r_t \} \) with \( r_i = 1 \). To prove the convergence of \( \Psi_1 \), we need to prove 1) \( E(\hat{\Psi}_1) = \Psi_1 \), which is simple to show by noticing that \( E(X_t) = Pr(r_{t+1} = 1|r_t = 1) = \Psi_1 \); 2) \( \lim_{T \to \infty} \text{Var}(\hat{\Psi}_1) = 0 \). In general, \( X_t \)'s are not independent due to the correlation between the consecutive true spectrum states in the Markov model. Thus, the central limit theorem cannot be applied. However, we will show the second fact is true by first proving that the correlation between \( X_i \) and \( X_j \) (\( i > j \)) vanishes as \( (i-j) \) approaches infinity. That is,

\[ \lim_{(i-j) \to \infty} E(X_i X_j) = E(X_i)E(X_j). \]

(33)

Note that

\[ E(X_i X_j) = Pr(r_{i+1} = 1, r_{j+1} = 1|r_j = 1, r_j = 1) = Pr(r_{j+1} = 1|r_j = 1)Pr(r_{i+1} = 1|r_i = 1, r_j = 1) \]

\[ = Pr(r_{j+1} = 1|r_j = 1)|Pr(r_{i+1} = 1|r_i = 1, r_j = 1)Pr_d \]

\[ + Pr(s_{i+1} = 0|r_i = 1, r_j = 1)Pr_f]. \]

Comparing the two preceding equations, it can be seen that, to prove (33), it is sufficient to prove

\[ \lim_{(i-j) \to \infty} Pr(s_{i+1} = 1|r_i = 1, r_j = 1) = Pr(s_{i+1} = 1|r_i = 1). \]

Note that \( Pr(s_{i+1} = 1|r_i = 1) \) is given by

\[ = P_d Pr(s_{i+1} = 1, s_i = 1, s_j = 1) \]

\[ = \frac{\pi_1 P_{a11} + \pi_0 P_{fa} a_{01}}{\pi_1 P_d + \pi_0 P_{fa}}. \]

(34)

and \( Pr(s_{i+1} = 1|r_i = 1, r_j = 1) \) is given by

\[ = \frac{P_d Pr(s_{i+1} = 1, s_i = 1, s_j = 1) \]

\[ + P_d P_{fa} Pr(s_{i+1} = 1, s_i = 0, s_j = 1) \]

\[ + ... + P_d P_{fa} Pr(s_{i+1} = 1, s_i = 1, s_j = 0) \]

\[ + P_d P_{fa} Pr(s_i = 1, s_j = 0) \]

\[ + ... + P_d P_{fa} Pr(s_i = 0, s_j = 0) \]

\[ + P_d P_{fa} P_{a11} + P_d P_{fa} (1 - P_{a11} a_{01}) \]

\[ + \pi_1 (P_d P_{a11} P_{a11} + P_d P_{fa} (1 - P_{a11} a_{01})) \]

\[ + ... + \pi_0 (P_d P_{fa} P_{a11} + P_d P_{fa} (1 - P_{a11} a_{01})), \]

(35)

where \( p_{n1} = Pr(s_{n+1} = 1|s_j = 1) \) and \( p_{n0} = Pr(s_{n+1} = 0|s_j = 0) \). According to the definition, the following recursive relation holds for \( p_{n1} \),

\[ p_{n1} = Pr(s_{j+n} = 1|s_j = 1) \]

\[ = Pr(s_{j+n} = 1, s_{j+n-1} = 1|s_j = 1) \]

\[ + Pr(s_{j+n} = 1, s_{j+n-1} = 0|s_j = 1) \]

\[ = a_{11} p_{n1} + a_{01} (1 - a_{11} p_{n1}). \]

(36)

Consequently, \( p_{n1} = \frac{a_{11}}{1 - a_{11} + a_{11} a_{01}}. \) Similarly, we have \( p_{n0} = \frac{a_{01}}{1 - a_{00} + a_{00} a_{10}}. \) Substituting these two expressions into (35), it can be verified that \( Pr(s_{i+1} = 1|r_i = 1, r_j = 1) = \frac{\pi_1 P_{a11} + \pi_0 P_{fa} a_{01}}{\pi_1 P_d + \pi_0 P_{fa}} = Pr(s_{i+1}|r_i = 1) \) as \( i-j \) approaches infinity. Therefore (33) holds.

Now, we will use (33) to prove that \( \lim_{(i-j) \to \infty} \text{Var}(\hat{\Psi}_1) = 0 \). For any positive \( \delta \), \( \exists K_\delta \) such that \( |\text{Cov}(\hat{\Psi}_1, \hat{\Psi}_j)| < \delta/2 \) when

\[ \frac{\delta - \frac{\delta}{2}}{\frac{\delta}{2}}. \]
Therefore, $\hat{\Psi}_1$ converges to $\Psi_1$. Following the same approach, it can be shown that $\Psi_0$ converges to $\Psi_0$.

REFERENCES

Huaiyu Dai (M’03, SM’09) received the B.E. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2002.

He was with Bell Labs, Lucent Technologies, Holmdel, NJ, during summer 2000, and with AT&T Labs-Research, Middletown, NJ, during summer 2001. Currently he is an Associate Professor of Electrical and Computer Engineering at NC State University, Raleigh. His research interests are in the general areas of communication systems and networks, advanced signal processing for digital communications, and communication theory and information theory. His current research focuses on networked information processing and crosslayer design in wireless networks, cognitive radio networks, wireless security, and associated information-theoretic and computation-theoretic analysis.

He has served as editor of IEEE Transactions on Communications, Signal Processing, and Wireless Communications. He co-edited two special issues for EURASIP journals on distributed signal processing techniques for wireless sensor networks, and on multiuser information theory and related applications, respectively. He co-chairs the Signal Processing for Communications Symposium of IEEE Globecom 2013, the Communications Theory Symposium of IEEE ICC 2014, and the Wireless Communications Symposium of IEEE Globecom 2014.

Peng Ning (M’01, SM’12) received the B.S. degree in information sciences from the University of Science and Technology of China (USTC), Hefei, China, in 1994, the M.E. degree in communications and electronics systems from USTC, Graduate School in Beijing, Beijing, China, in 1997, and the Ph.D. degree in information technology from George Mason University, Fairfax, VA, in 2001.

He is a Professor of Computer Science at NC State University, where he also serves as the Technical Director for Secure Open Systems Initiative (SOSI). He is a recipient of National Science Foundation (NSF) CAREER Award in 2005. He is currently the Secretary/Treasurer of the ACM Special Interest Group on Security, Auditing, and Control (SIGSAC), and is on the Executive Committee of ACM SIGSAC. He is an editor for Springer Briefs in Computer Science, responsible for Briefs on information security. He has served or is serving on the editorial boards of several international journals, including ACM Transactions on Sensor Networks, Journal of Computer Security, Ad-Hoc Networks, Ad-Hoc & Sensor Networks: an International Journal, International Journal of Security and Networks, and IET Proceedings Information Security. He also served as the Program Chair or Co-Chair for ACM SASN ’05, ICICS ’06 and ESORICS ’09, ICDCS-SPCC ’10, and NDSS ’13, the General Chair of ACM CCS ’07 & ’08, and Program Vice Chair for ICDCS ’09 & ’10 – Security and Privacy Track. He served on the Steering Committee of ACM CCS from 2007 to 2011, and is a founding Steering Committee member of ACM WiSec and ICDCS SPCC. His research has been supported by NSF, Army Research Office (ARO), the Advanced Research and Development Activity (ARDA), IBM Research, SRI International, and the NCSU/Duke Center for Advanced Computing and Communication (CACC). Peng Ning is a senior member of the ACM, the ACM SIGSAC, and a senior member of the IEEE. http://discovery.csc.ncsu.edu/~pning/