IET Communications

Research Article

Pre-processing of incomplete spectrum sensing data in spectrum sensing data falsification attacks detection: a missing data imputation approach

ISSN 1751-8628
Received on 24th November 2015
Revised on 25th February 2016
Accepted on 16th March 2016
doi: 10.1049/iet-com.2015.1111
www.ietdl.org

Junnan Yao[1], *Jianjun Cao*[2] ✉, *Qibin Zheng*[1], *Jingang Ma*[1]

[1]*College of Command Information System, PLA University of Science and Technology, Nanjing 210007, People's Republic of China*
[2]*Armament Department of Nanjing Military Region, Nanjing Telecommunication Technology Institute, Nanjing 210016, People's Republic of China*
✉ *E-mail: jianjuncao@yeah.net*

Abstract: Attacks detection is an important issue in collaborative spectrum sensing (CSS) system of cognitive radio networks. Many approaches are proposed to cope with the malicious behaviour of attackers. In existing works, spectrum sensing data (SSD) received by the fusion centre is generally assumed to be integrated. However, in practical scenarios, the received SSD may be incomplete due to the imperfect reporting channel or specific CSS schemes. The performance of existing attacks detection approaches may degrade especially when the probability of missing data is large. To alleviate this challenge, the authors focus on pre-processing of incomplete SSD and propose a practical imputation algorithm, which is derived from the maximum a posteriori probability criterion, to fill in the missing values. Simulation results indicate that the proposed algorithm restores the characteristics of the SSD, and mitigate the impacts of missing value on existing attacks detection algorithm effectively.

1 Introduction

Cognitive radio networks (CRN) improve spectrum utilisation by adopting opportunistic spectrum access scheme [1]. The cognitive users (CUs) are permitted to use the specific spectrum band only when it is not occupied by the primary user (PU). Therefore, the ability of CUs to detect the state of the PU is indispensable in the CRN. The spectrum sensing approaches are discussed extensively in [2]. Due to the unfavourable channel condition of wireless transmission, such as fading and shadowing, the single-node spectrum sensing is not reliable. To deal with this problem, collaborative spectrum sensing (CSS) schemes are proposed [3, 4]. Multiple CUs observe the concerned spectrum band individually, and then they send the sensing results to the fusion centre (FC). According to a specific fusion rule, the global decision regards to the state of PU is derived.

Although CSS schemes improve the sensing performance by the diversity gain, the open access of it provides an opportunity for attackers (or malicious users) to undermine the spectrum sensing system of a CRN. Specifically, the attackers may take part in the local sensing procedure and report manipulated results to the FC to implement spectrum sensing data falsification (SSDF) attacks [5]. In [6], reputation-based algorithm is proposed to eliminate the impacts of the SSDF attacks. In a given sensing slot, the reputation of a CU increases if its sensing report supports the global decision, and it is assigned larger weight in data fusion in the following sensing slots. In [7], the spectrum sensing data (SSD) of CUs are analysed statistically, and data mining based attacks detection algorithm is studied.

In [8], the feature vectors, which reveals the dissimilarity among the CUs, are calculated, and they are utilised to identify the attackers in the CRN. In [9], the joint probability of SSD of a CU in two consecutive sensing slots is computed, and a clustering algorithm to divide the two types of CUs (i.e. the honest users and the malicious users) is proposed. Moreover, some practical malicious behaviours are considered in the existing works, such as stealthy attacks [10], massive malicious users attacks [11], and time-varying strategy attacks [12].

In most studies, including literatures mentioned above, the SSD is generally assumed to be without missing values. However, the FC cannot receive the SSD of a CU at a specific sensing slot due to the imperfect report channel [13] or the energy-saving CSS mode [14]. In this circumstance, the performance of attacks detection algorithms degrade [For instance, the joint probability of SSD in [7–9] cannot be calculated directly when there are missing values within the SSD.]. To the best of our knowledge, there is no literature that studies on the attacks detection when SSD is incomplete.

To alleviate this challenge, in this paper, we focus on pre-processing of incomplete SSD for SSDF attacks detection. The causes of missingness in SSD and the impacts of them are analysed. Based on maximum a posteriori probability (MAP) criterion, we propose the empirical joint probability imputation (EJPI) algorithm, which fills in the missing values of SSD and restores the feature vectors of CUs. Numerical experiments show that the attackers can be isolated from honest users in Euclidian space after missing data imputation [15]. Moreover, the performance of attack detection with imputed SSD improves comparing with that based on incomplete SSD without imputation.

The rest of this paper is organised as follows. In Section 2, the model of CSS system is introduced. In Section 3, the causes of missingness in SSD and the impacts of it are studied. The EJPI algorithm is proposed in Section 4. In Section 5, the performance of proposed algorithm is evaluated by numerical experiments, and the paper is concluded in Section 6.

2 System model

In this section, we introduce the typical CSS system with presence of SSDF attacks, and several attacks detection algorithms are presented as examples. Then the issue of incomplete SSD is investigated.

2.1 CSS model

We consider a CRN, in which N CUs coexist with a PU. As the PU has the priority to access to the licensed spectrum band, the CUs are permitted to use that band only when the PU is absent. Let P_0 be the prior probability that the PU is absent, and the prior probability that the PU is present is denoted by P_1. Both the PU and CUs use the time slotted system [16]. At the beginning of each time slot, the CUs sense the status of the PU in the licensed spectrum band individually. In [17], local sensing schemes have been studied extensively. Without loss of generality, in this paper, energy detection [18] is used as local sensing scheme by CUs. The local sensing result of CU$_i$ at sensing slot t is denoted by

$$u_{i,t} = \begin{cases} 0, & Y_{i,t} < \zeta_i \\ 1, & Y_{i,t} \geq \zeta_i \end{cases}, \tag{1}$$

where $u_{i,t}$ denotes the local sensing result of CU$_i$ at sensing slot t, $Y_{i,t}$ is cumulated energy of PU's signal received by the CU$_i$, ζ_i is the decision threshold of CU$_i$. For convenient denotation, the index t is omitted when the distribution of $Y_{i,t}$ is invariable with time. Therefore, local sensing performance of CU$_i$ could be evaluated by false alarm probability

$$P_{\text{fa}}^{(i)} = \Pr\left(u_i = 1|\mathcal{H}_0\right), \tag{2}$$

and missed detection probability

$$P_{\text{md}}^{(i)} = \Pr\left(u_i = 0|\mathcal{H}_1\right), \tag{3}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ are the hypotheses that PU is absent and present, respectively. When Y_i, $i \in \{1, 2, \ldots, N\}$ follows conditional independent and identical distribution [13], and all the CUs adopt the same decision threshold (the threshold can be decided by the network control centre), the local sensing performance is identical among the CUs, i.e. $P_{\text{fa}}^{(i)} = P_{\text{fa}}^{\text{H}}$ and $P_{\text{md}}^{(i)} = P_{\text{md}}^{\text{H}}$, $i \in \{1, 2, \ldots, N\}$.

Finishing the local spectrum sensing, the CUs send the sensing results $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)^{\text{T}}$, $u_i \in \{0, 1\}$, $i = 1, 2, \ldots, N$, to the FC. With the assumption that the channels between CUs and the FC are error free [5], the local decision from CU$_i$ could be received by the FC correctly. Specifically, '0' indicates that the CU infers the absence of the PU in the licensed band, whereas '1' indicates that the CU infers the presence of the PU. According to a predetermined fusion rule [3], the global decision of the PU's status is generated at the FC.

When all the CUs in the CRN are the honest users, CSS improves the global sensing performance. However, if there are SSDF attackers in the CSS system, the sensing performance will be degraded by their malicious behaviour. We assume that the SSD sent to the FC by CU$_i$ is $\hat{u}_i$, $i = 1, 2, \ldots, N$, then the behaviour of CU$_i$ could be denoted by the conditional probability matrix

$$Q^{(i)} = \begin{pmatrix} q_{00}^{(i)} & q_{01}^{(i)} \\ q_{10}^{(i)} & q_{11}^{(i)} \end{pmatrix}, \tag{4}$$

where $q_{kl}^{(i)} = \Pr\left(\hat{u}_i = l|u_i = k\right)$, $k, l \in \{0, 1\}$, denotes the probability that CU$_i$ sends 'l' to the FC when its sensing result is 'k'.

2.2 SSDF attacks detection

In practical scenarios, the only clue can be utilised to check the attacks is SSD from CUs, i.e. SSD matrix

$$\hat{\boldsymbol{U}} = \left(\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2, \ldots, \hat{\boldsymbol{u}}_t, \ldots, \hat{\boldsymbol{u}}_L\right), \tag{5}$$

where $\hat{\boldsymbol{u}}_t = \left(\hat{u}_{1,t}, \hat{u}_{2,t}, \ldots, \hat{u}_{i,t}, \ldots, \hat{u}_{N,t}\right)^{\text{T}}$, $t = 1, 2, \ldots, L$, is the historical SSD of CU$_i$, L is the length of SSD of CU$_i$.

In *reputation-based* attacks detection algorithms (or secure CSS algorithms), the value α_i, $i = 1, 2, \ldots, N$, is assigned to CU$_i$ as the reputation of it, and α_i updates after each time slot according to specific algorithms [6]. Usually, α_i decreases if SSD from CU$_i$ at slot t, i.e. $\hat{u}_{i,t}$, is not equal to the global decision, and it will increase when $\hat{u}_{i,t}$ equal to the global decision. Although checking local sensing results with the global decision (or a fusion result based on a subset of the CUs in the CRN) may be affected by falsified SSD, this type of schemes are efficient in most situations especially when the malicious users in the CRN are not many.

For *statistic-based* attacks detection algorithms (or data mining based algorithms) [7, 8], the global decision, which is used as a benchmark, is not needed. The detection algorithms investigate the SSD matrix $\hat{\boldsymbol{U}}$, and calculate the probabilities or joint probabilities among the SSD. After that, the algorithms identify malicious users by clustering or classifying. The statistic-based attacks detection algorithms have three *advantages* and achieve a favourable performance in checking attacks. First of all, this type of algorithms exploit SSD in both time domain and space domain with a statistical fashion [In [9], the joint probability of SSD from a single CU in two consecutive time slots is calculated, and the joint probability among the SSD of multiple CUs are studied in [8].]. Moreover, the algorithms need no fusion results, which may be generated by falsified data, as benchmarks to calibrate the CUs' behaviour. At last, they need no prior information of the CUs and the PU. Therefore, we investigate degradation of this type of detection algorithms when SSD is incomplete and try to improve the performance by pre-processing the incomplete SSD in this paper.

3 Incomplete SSD in SSDF attacks detection

In this section, we consider the issue of incomplete SSD in SSDF attacks detection. The mechanisms of missing data are investigated, and the impacts are analysed.

3.1 Causes of missing data

We consider the CSS system with N CUs introduced in Section 2. After L sensing slots, the SSD collected by the FC could be denoted by the $N \times L$ SSD matrix as follows

$$\boldsymbol{V} = \begin{pmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,L} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N,1} & v_{N,2} & \cdots & v_{N,L} \end{pmatrix}, \tag{6}$$

where $v_{i,t} \in \{0, 1\}$, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, L$, is the sensing reports of CU$_i$ at sensing slot t. In most exiting works, the matrix $\boldsymbol{V}$ is assumed to be complete and $\boldsymbol{V} = \hat{\boldsymbol{U}}$. However, this assumption may not hold in a practical CSS system. For instance, the imperfect reporting channel makes some of the SSD cannot be collected, or some CUs have not send their SSD to the FC at specific sensing slots. To depict missingness of $\boldsymbol{V}$, we define the $N \times L$ indicator matrix

$$\boldsymbol{R} = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,L} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N,1} & r_{N,2} & \cdots & r_{N,L} \end{pmatrix}, \tag{7}$$

where $r_{i,t} \in \{0, 1\}$, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, L$. $r_{i,t} = 0$ means that the value of SSD $\hat{u}_{i,t}$ is missing, while $r_{i,\ t} = 1$ indicates $\hat{u}_{i,t}$ is observed by the FC, i.e. $v_{i,t} = \hat{u}_{i,t}$.

To describe the relation between the missingness and the value of SSD, we define $\boldsymbol{V}_{\text{ob}}$ as the observed SSD and $\boldsymbol{V}_{\text{miss}}$ as the missed SSD. The probability of missingness could be denoted by $P\left(\boldsymbol{R}|\boldsymbol{V}_{\text{ob}}, \boldsymbol{V}_{\text{miss}}, \xi\right)$, where ξ is the parameter of missingness.

There are two main factors which may cause missing data of SSD:

(i) *Reporting channel condition:* The FC cannot detect the SSD sent by CU$_i$ at sensing slot t when the reporting channel condition between the FC and CU$_i$ is poor. In this circumstance, $r_{i,t} = 0$.

Moreover, $r_{i,t}$ is only dependent on the reporting channel, and it is independent both on the missed value $\boldsymbol{V}_{\text{miss}}$ and the observed value $\boldsymbol{V}_{\text{ob}}$. This type of missingness is called *missing completely at random (MCAR)* [19], which could be depicted by

$$P(\boldsymbol{R}|\boldsymbol{V}_{\text{ob}}, \boldsymbol{V}_{\text{miss}}, \xi) = P(\boldsymbol{R}|\xi), \tag{8}$$

where ξ is the missingness parameter which is related to the reporting channel condition. Then the observed values $\boldsymbol{V}_{\text{ob}}$ could be regarded as random samples of the complete data set $\hat{\boldsymbol{U}}$.

(ii) *Collaborative sensing schemes:* When the CRN adopts an energy-saving sensing scheme, the SSD may be incomplete. For instance, a CU reports its SSD to the FC only when it wants to access to the licensed band, and it keeps silent if it has no message to transmit. Obviously, the missingness of $v_{i,t}$ depends on the transmission probability of CU$_i$, and it does not depend on the values of SSD. The missingness is also MCAR.

Considering another energy-saving sensing scheme, a CU reports its SSD to the FC only when the PU's signal is detected [Generally, the prior probability that the PU is present is less than that of the PU is absent, i.e. $P_1 < P_0$. To send SSD only when PU is present is energy efficient.] and keeps silent when PU is absent (i.e. sensing result '0' will never be sent). The missingness of $v_{i,t}$ is dependent on the value of $v_{i,t}$, specifically, if $v_{i,t} = 0$, then $r_{i,t} = 0$. The missing data mechanism could be denoted by

$$P(\boldsymbol{R}|\boldsymbol{V}_{\text{ob}}, \boldsymbol{V}_{\text{miss}}, \xi) = P(\boldsymbol{R}|\boldsymbol{V}_{\text{ob}}, \boldsymbol{V}_{\text{miss}}, \xi), \tag{9}$$

or

$$P(\boldsymbol{R}|\boldsymbol{V}_{\text{ob}}, \boldsymbol{V}_{\text{miss}}, \xi) = P(\boldsymbol{R}|\boldsymbol{V}_{\text{miss}}, \xi), \tag{10}$$

which is missing not at random (MNAR).

Although MNAR is a great challenge in statistic literatures, it is not a difficult problem in aforementioned attacks detection frameworks. In this scenario, only the sensing reports '1's will be sent to the FC, and the '0's will not be sent for energy saving. In the view of the FC, values of missed sensing reports in this MNAR mechanism are known, and the FC could impute all the missed values by '0's. Therefore, we focus on pre-processing of incomplete SSD that is MCAR in this paper.

3.2 *Impacts of missing data*

In many SSDF attacks detection algorithms, missingness of SSD degrade the performance of attacks detection, and some algorithms even cannot be utilised with incomplete SSD directly. In the following of this subsection, impacts of incomplete SSD on several existing algorithms are investigated.

To analyse the behaviour of CUs, the algorithm in [9] calculates the empirical probability that sensing reports of CU$_i$ in two consecutive time slots are equal by

$$\tilde{P}(v_{i,t} = v_{i,t+1}) = \sum_{k=1}^{L-1} I(v_{i,k} = v_{i,k+1})/(L-1), \tag{11}$$

where $I(\cdot)$ is the indicator function. It is clear that if the value of $v_{i,t}$ is missing, we cannot obtain the value of indicators $I(v_{i,t} = v_{i,t+1})$ and $I(v_{i,t} = v_{i,t-1})$. An intuitive countermeasure is ignoring the two indicators above and calculating the empirical probability by

$$\tilde{P}(v_{i,t} = v_{i,t+1}) = \sum_{\substack{k=1 \\ k \neq t-1,t}}^{L-1} I(v_{i,k} = v_{i,k+1})/(L-3). \tag{12}$$

The worst case is that missing data exists in each pair of two consecutive sensing slots of a CU. Specifically, $\forall i \in \{1, 2, \ldots, N\}$ and $\forall t \in \{1, 2, \ldots, L-1\}$, the inequality $r_{i,t} + r_{i,\ t+1} \leq 1$ holds. In this condition, the empirical probability $\tilde{P}(v_{i,t} = v_{i,t+1})$ cannot be derived, and the algorithm becomes invalid.

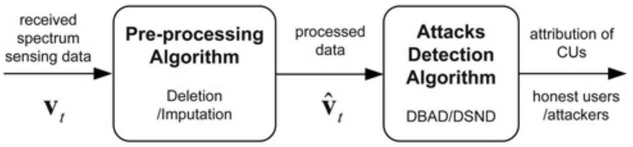


Fig. 1 *Procedure of attacks detection with incomplete SSD*

In [7], the double-side neighborhood distance (DSND) algorithm calculates the empirical probability that the SSD from two CUs in the same time slot are different, i.e.

$$\tilde{P}\left(v_i \neq v_j\right) = \sum_{t=1, i\neq j}^{L} I\left(v_{i,t} \neq v_{j,t}\right)/L. \tag{13}$$

Missingness of any two sensing reports in the same time slot makes the indicator $I(v_{i,t} \neq v_{j,t})$ invalid.

In algorithm dissimilarity based attacker detection (DBAD) [8], the dissimilarity between SSD of CU$_i$ and that of the other CUs is computed by

$$d_i = \sum_{j=1, j\neq i}^{N} I\left(v_{i,t} \neq v_{j,t}\right), \tag{14}$$

then the empirical probability mass function (pmf) of d_i is given by

$$\boldsymbol{g}_i = \left(\tilde{P}(d_i = 0), \tilde{P}(d_i = 1), \ldots, \tilde{P}(d_i = N-1)\right), \tag{15}$$

where $\tilde{P}(d_i = n)$, $n = 0, 1, \ldots, N-1$ is the empirical probability that $d_i = n$. In the clustering phase, the empirical pmfs $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_N$ are used as feature vectors, and the DBAD adopts clustering method (such as k-means) to separate the N feature vectors into two classes. One is the class of honest users, and the other is that of attackers. As the number of attackers is usually less than that of honest users [7], the DBAD regards the class with less elements as the set of attackers, then the attackers are identified.

At a given sensing slot, missingness of any SSD of the N CUs will make the indicator function invalid. Assuming the sensing data of the CUs follow the same missing data probability P_{miss}, the probability that the indicator function is invalid could be calculated by

$$P_{\text{inva}} = 1 - \left(1 - P_{\text{miss}}\right)^N, \tag{16}$$

where N is the number of the CUs in the CSS system. With increasing of P_{miss} and N, P_{inva} increases, and the performance of DBAD degrades. In a CSS system with ten CUs, P_{inva} is larger than 0.5 when $P_{\text{miss}} = 0.067$. It indicates that even a very low missing data probability could undermine the efficiency of DBAD. A countermeasure is needed. In the following of this paper, we take the DBAD algorithm as an example to investigate how to process the incomplete SSD to recover the performance of the attacks detection algorithm. The research could be also applied for other statistic-based attacks detection algorithms.

4 Pre-processing of incomplete SSD

To achieve an acceptable performance, pre-processing of SSD should be applied before attacks detection algorithms. As shown in Fig. 1, the input of pre-processing algorithm is the received SSD $\boldsymbol{v}_t$, and it outputs processed SSD $\hat{\boldsymbol{v}}_t$. Based on $\hat{\boldsymbol{v}}_t$, the attacks detection algorithms, such as DBAD, identify the attackers in the

CSS system. There are generally two sorts of pre-processing approaches, i.e. *deletion* and *imputation*.

4.1 Deletion

Deletion of incomplete SSD means that the FC deletes the SSD of a time slot if there is any missing data in it. Specifically, in SSD matrix $\boldsymbol{V}$, if $\exists i \in \{1, 2, \ldots, N\}$, $t \in \{1, 2, \ldots, L\}$, that satisfy $r_{i,t} = 0$, then the FC deletes the vector $\boldsymbol{v}_t = \left(v_{1,t}, v_{2,t}, \ldots, v_{N,t}\right)^{\mathrm{T}}$ from $\boldsymbol{V}$, and we have the processed SSD matrix $\hat{\boldsymbol{V}}$. The processed SSD could be directly utilised by conventional attacks detection algorithms, such as DBAD.

Deletion is a convenient approach to process missing data, but it is not efficient for attacks detection. It is because in order to achieve equivalent performance as with complete SSD $\hat{\boldsymbol{U}}$, the attacks detection algorithms should keep collecting the SSD for more sensing slots, until the number of valid sensing slots of $\boldsymbol{V}$ is equal to the length of $\hat{\boldsymbol{U}}$. For instance, when the length of $\hat{\boldsymbol{U}}$ is L, the attacks detection algorithms have to observe the matrix $\boldsymbol{V}$ for $L/(1 - P_{\text{inva}})$ sensing slots. The algorithms become impractical when P_{inva} is large. Another drawback of deletion is that the observed data in deleted $\boldsymbol{v}_t$ is also discarded. The imputation approaches, which fill in the holes in $\boldsymbol{V}$, should be considered.

4.2 Imputation

Imputation is the process that plugs in plausible values where none exists. The object of imputation is not just to find out what is missed, but to restore the important characteristics of the data set as a whole [15]. In the following of this subsection, we present several ordinary imputation approaches.

(i) *Random-value imputation (RVI):* Within L sensing slots observation, there are L'_i valid SSD of CU$_i$. The empirical distribution of $v_{i,t}$ could be calculated by

$$\tilde{P}\left(v_{i,t} = w\right) = \sum_{t=1}^{L} I\left(v_{i,t} = w\right) / L'_i, \tag{17}$$

where $w \in \{0, 1\}$. In RVI, $\hat{v}_{i,t}$ which follows the distribution $P\left(\hat{v}_i = w\right) = \tilde{P}\left(v_i = w\right)$ is assigned to the missed value of $v_{i,t}$. The *advantage* of this approach is that the mean and the variance of processed SSD of a CU are close to that of complete SSD when $L \to \infty$, and these characteristics of SSD from each individual CU are restored.

(ii) *Majority-rule imputation (MRI):* For MRI scheme, the missing data $v_{i,t}$ in time slot t will be assigned the value that is reported by most CUs in that time slot, i.e.

$$\hat{v}_{i,t} = I\left(\sum_{j \in \mathcal{C}_{\text{ob}}} v_{j,t} > \frac{\left|\mathcal{C}_{\text{ob}}\right|}{2}\right), \tag{18}$$

where $\mathcal{C}_{\text{ob}}$ is the set of CUs whose SSD are observed in the current time slot. The *advantage* of this approach is that the global sensing result at FC does not change after imputation.

4.3 MAP-based imputation

Although the aforementioned pre-processing approaches are easy to implement, the relationship among the SSD of different CUs has not been exploited properly, especially RVI treats the SSD of different CUs individually. To utilise the additional information, we consider an imputation algorithm which is based on MAP criterion in this subsection.

In the CSS system discussed in Section 2, the sensing reports of CUs in a given sensing slot are dependent with each other, because the set of SSD $\{v_{1,t}, v_{2,t}, \ldots, v_{N,t}\}$ is generated based on the same truth (i.e. the status of the PU), and it could be described as follows

$$P\left(\boldsymbol{v}_t\right) = \prod_{i=1}^{N} P\left(v_{i,t} | \mathcal{H}_0\right) P_0 + \prod_{i=1}^{N} P\left(v_{i,t} | \mathcal{H}_1\right) P_1. \tag{19}$$

When we process the missing data in set $\{v_{1,t}, v_{2,t}, \ldots, v_{N,t}\}$, the observed values in it should be considered. The MAP-based imputation could be denoted by

$$\hat{\boldsymbol{y}}_t = \arg\max_{\boldsymbol{y}_t} P\left(\boldsymbol{y}_t | \boldsymbol{x}_t\right), \tag{20}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the observed SSD and the missed SSD in sensing slot t, respectively. According to the Bayes theorem, we have

$$\hat{\boldsymbol{y}}_t = \arg\max_{\boldsymbol{y}_t} \frac{P\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right)}{P\left(\boldsymbol{x}_t\right)}. \tag{21}$$

As $P\left(\boldsymbol{x}_t\right)$ is fixed when $\boldsymbol{x}_t$ is observed SSD, then the estimation of missing data in a time slot is written by

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= \arg\max_{\boldsymbol{y}_t} P\left(\boldsymbol{y}_t, \boldsymbol{x}_t\right) \\ &= \arg\max_{\boldsymbol{y}_t} P\left(\boldsymbol{v}_t\right). \end{aligned} \tag{22}$$

Equation (22) indicates that as long as the pmf of the SSD in a time slot is known, we can derive the estimation of the missing data. As shown in (19), although $P\left(\boldsymbol{v}_t\right)$ cannot be calculated directly because the conditional probability of SSD of the attacker is usually unknown, we can derive the empirical distribution $\tilde{P}\left(\boldsymbol{v}_t\right)$ by counting the occurrence of different values of $\boldsymbol{v}_t$ from $\boldsymbol{V}'$. Here $\boldsymbol{V}'$ is the $L' \times N$ complete SSD matrix after deleting the incomplete columns from $\boldsymbol{V}$. Giving a specific value $\boldsymbol{w} = \{0, 1\}_N$, the empirical probability that $\boldsymbol{v}_t = \boldsymbol{w}$ is

$$\tilde{P}\left(\boldsymbol{v}_t = \boldsymbol{w}\right) = \sum_{t \in \mathcal{S}_{\text{com}}} I\left(\boldsymbol{v}_t = \boldsymbol{w}\right) / \left|\mathcal{S}_{\text{com}}\right|, \tag{23}$$

where $\mathcal{S}_{\text{com}}$ is the set of sensing slot indexes in which $\boldsymbol{v}_t$ is complete, and $\left|\mathcal{S}_{\text{com}}\right| = L'$. There are 2^N possible values of $\boldsymbol{w}$, and the empirical pmf of $\boldsymbol{v}_t$ is given by

$$\tilde{f}_N\left(\boldsymbol{v}_t\right) = \left(\tilde{P}\left(\boldsymbol{v}_t = \boldsymbol{w}_1\right), \tilde{P}\left(\boldsymbol{v}_t = \boldsymbol{w}_2\right), \ldots, \tilde{P}\left(\boldsymbol{v}_t = \boldsymbol{w}_{2^N}\right)\right). \tag{24}$$

We notice that calculation of $\tilde{f}_N\left(\boldsymbol{v}_t\right)$ involves counting the occurrence of $\boldsymbol{v}_t \in \{0, 1\}_N$ on all the 2^N candidates. The number of candidates increases exponentially with increasing of N, and it is difficult to implement when N is large. Moreover, calculation of $\tilde{f}_N\left(\boldsymbol{v}_t\right)$ needs complete vector $\boldsymbol{v}_t$. According to (16), it needs a long period of time to collect the complete vectors to converge $\tilde{f}_N\left(\boldsymbol{v}_t\right)$. Therefore, an efficient imputation algorithm is needed.

4.4 EJPI algorithm

In this subsection, we propose the EJPI algorithm which is more practical to complete the missed values of SSD. To some extent, the EJPI is a modification of the MAP-based algorithm. Before introducing the procedure of EJPI, we discuss how to estimate the missed value of $v_{i,t}$ based on a single observed value $v_{j,t}$. In the view of MAP criterion, observing the value of $v_{j,t}$, the most possible value of the missing data $v_{i,t}$ could be denoted by

$$\begin{aligned}\hat{v}_{i,t} &= \underset{v_{i,t}\in\{0,1\}}{\arg\max}\, P\left(v_{i,t}|v_{j,t}\right)\\ &= \underset{v_{i,t}\in\{0,1\}}{\arg\max}\, P\left(v_{i,t}, v_{j,t}\right)/P\left(v_{j,t}\right).\end{aligned} \tag{25}$$

As $v_{j,t}$ is observed at the FC, then $P(v_{j,t})$ has a determined value, and $P(v_{j,t})$ does not affect imputation of $v_{i,t}$. Therefore, we have the following equation:

$$\hat{v}_i = \underset{v_i\in\{0,1\}}{\arg\max}\, P\left(v_i, v_j\right), \tag{26}$$

where the subscript t is omitted in the following of this section for convenient notation. Equation (26) reveals how to impute a missed value of SSD v_i based on another observed v_j.

Generally, in a sensing slot, there are more than one observed sensing reports which could be used as 'evidences' to infer the missed values. How to integrate these evidences to estimate the missed value is an interesting question. A simple but effective way is combining the joint probabilities of each pair of the missed value and observed value, i.e.

$$\hat{v}_i = \underset{v_i\in\{0,1\}}{\arg\max} \sum_{j\in\mathcal{C}_{\text{ob}}} P\left(v_i, v_j\right), \tag{27}$$

where $\mathcal{C}_{\text{ob}}$ is the set of CUs whose sensing reports are observed in the current sensing slot. Although $P(v_i, v_j)$ cannot be derived directly when the behaviour of attackers is unknown, the empirical joint probability could be calculated by

$$\begin{aligned}&\tilde{P}\left(v_{i,t} = w_1, v_{j,t} = w_2\right)\\ &\quad= \sum_{t\in\mathcal{S}^{i,j}_{\text{com}}} I\left(v_{i,t} = w_1, v_{j,t} = w_2\right)/|\mathcal{S}^{i,j}_{\text{com}}|,\end{aligned} \tag{28}$$

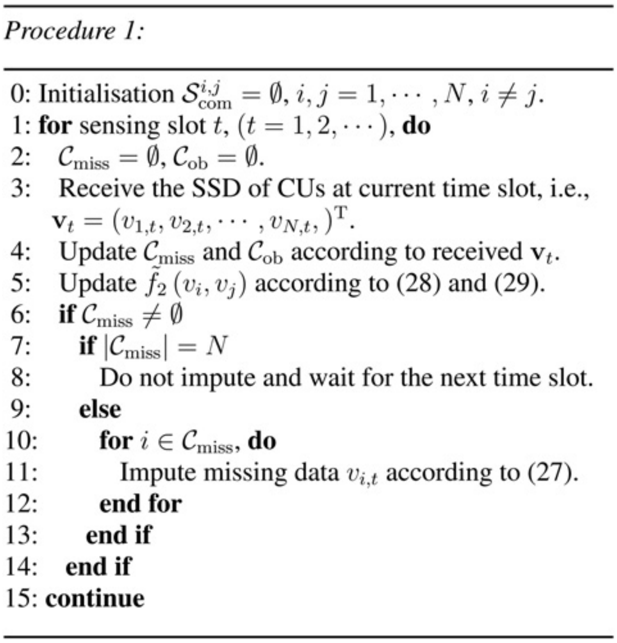
where $\mathcal{S}^{i,j}_{\text{com}}$ is the set of sensing slot indexes in which $v_{i,t}$ and $v_{j,t}$ are both observed, $|\mathcal{S}^{i,j}_{\text{com}}|$ is the cardinality of set $\mathcal{S}^{i,j}_{\text{com}}$, and w_1, $w_2 \in \{0, 1\}$. The empirical pmf $\tilde{f}_2\left(v_i, v_j\right)$ is given by

$$\begin{aligned}\tilde{f}_2\left(v_i, v_j\right) = \Big(&\tilde{P}\left(v_i = 0, v_j = 0\right), \tilde{P}\left(v_i = 0, v_j = 1\right),\\ &\tilde{P}\left(v_i = 1, v_j = 0\right), \tilde{P}\left(v_i = 1, v_j = 1\right)\Big).\end{aligned} \tag{29}$$

where the subscript '2' means that the empirical pmf involves two CUs.

On the basis of equations above, we propose the procedure of EJPI algorithm (i.e. Procedure 1 (see Fig. 2)), which imputes the missing data at each time slot according to the observed data. In the algorithm, $\mathcal{C}_{\text{ob}}$ is the set of CUs whose sensing reports are missed in the current sensing slot. When all the SSD are missed in a certain time slot and no 'evidence' could be used, the EJPI will discard that time slot without imputation.

We notice that only 2^2 statistics are involved in the calculation of $\tilde{f}_2$, which is much less than that of $\tilde{f}_N$ in MAP-based imputation where 2^N statistics are concerned. Moreover, to update $\tilde{f}_2$, $r_{i,t} = 1$, $r_{j,t} = 1$ should be satisfied, and the probability to achieve that is $(1 - P_{\text{miss}})^2$, where P_{miss} is the probability that SSD of a CU is missed in a given time slot. This probability is larger than that of updating $\tilde{f}_N$, i.e. $(1 - P_{\text{miss}})^N$. These analyses indicate that $\tilde{f}_2$ could converge in shorter length of observation, and EJPI algorithm involves *lower computational complexity* and is more practical than MAP-based imputation algorithm.



```
Procedure 1:
Initialisation $\mathcal{S}^{i,j}_{\text{com}} = \emptyset, i, j = 1, \cdots, N, i \neq j$.
for sensing slot $t$, $(t = 1, 2, \cdots)$, do
  $\mathcal{C}_{\text{miss}} = \emptyset, \mathcal{C}_{\text{ob}} = \emptyset$.
  Receive the SSD of CUs at current time slot, i.e.,
    $\mathbf{v}_t = (v_{1,t}, v_{2,t}, \cdots, v_{N,t},)^{\text{T}}$.
  Update $\mathcal{C}_{\text{miss}}$ and $\mathcal{C}_{\text{ob}}$ according to received $\mathbf{v}_t$.
  Update $\tilde{f}_2\,(v_i, v_j)$ according to (28) and (29).
  if $\mathcal{C}_{\text{miss}} \neq \emptyset$
    if $|\mathcal{C}_{\text{miss}}| = N$
      Do not impute and wait for the next time slot.
    else
      for $i \in \mathcal{C}_{\text{miss}}$, do
        Impute missing data $v_{i,t}$ according to (27).
      end for
    end if
  end if
continue
```

Fig. 2 *EJPI algorithm*

4.5 Performance evaluation of EJPI

Performance of proposed EJPI could be evaluated directly by normalised Hamming distance (NHD) and error probability of attacker detection indirectly.

(i) *NHD:* We define the SSD reported by CU$_i$ at sensing slot t by $v_{i,t}$, and $\hat{v}_{i,t}$ is the imputed SSD. If $v_{i,t}$ is observed, then $\hat{v}_{i,t} = v_{i,t}$. The metric NHD tries to reveal the dissimilarity between $v_{i,t}$ and $\hat{v}_{i,t}$ in observed T sensing slots, i.e.

$$D_{\text{NH}}(T) = \sum_{t=1}^{T}\sum_{i=1}^{N}\left(v_{i,t} \oplus \hat{v}_{i,t}\right)/NT, \tag{30}$$

where '⊕' is the XOR operator. $D_{\text{NH}}(T)$ could be regarded as the probability that $v_{i,t}$ is different from $\hat{v}_{i,t}$, which depicts the accuracy of imputation algorithms.

(ii) *Error probability of attacker detection:* The ultimate purpose of imputation is not just filling in the missing values but providing the qualified SSD to the attacks detection algorithm. Error probability of attacker detection, i.e. P_{e}, could be used to infer the effectiveness of imputation approaches. In the simulation part, we adopt the 'imputation+detection' model to compare the performance of several imputation algorithms.

5 Simulation results

In this section, we analyse the impacts of missing data and performance of proposed imputation algorithm by extensive numerical experiments. The performance of deletion, RVI, and MRI are also demonstrated.

Table 1 Statistics in existing algorithms

Statistics	Algorithm
$\tilde{P}(v_{i,t} = v_{i,t+1})$, (11)	FCM-clustering [9]
$\tilde{P}(v_i \neq v_j)$, (13)	DSND [7]
$\tilde{P}(d_i = n)$, (14)	DBAD [8]

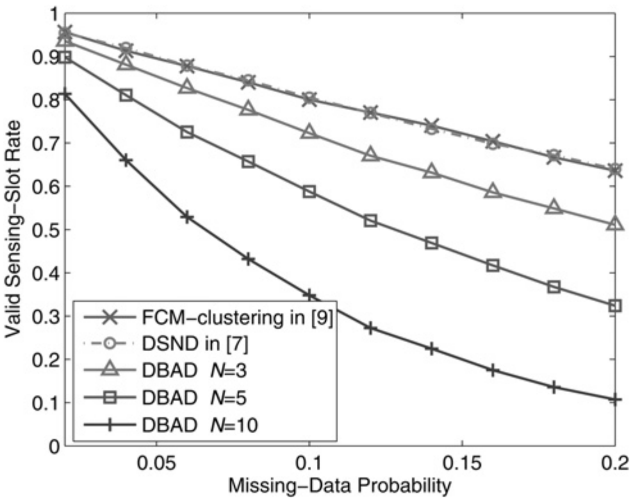


Fig. 3 *Valid sensing slot rate under different missing data probabilities*

5.1 Impacts of missing data

In many attacks detection algorithms, specific statistics are used to depict CUs' behaviours as shown in Table 1. To update the statistic at current sensing slot, the SSD involved in calculation should be received by the FC. In this circumstance, the sensing slot is regarded as a *valid* sensing slot, and the valid sensing slot rate is presented by

$$R_{\text{valid}} = \frac{|\mathcal{T}_{\text{valid}}|}{T}, \tag{31}$$

where $|\mathcal{T}_{\text{valid}}|$ is the number of valid slots in T sensing slots.

In Fig. 3, impacts of incomplete SSD are illustrated via the valid sensing slot rate R_{valid} under different values of missing data probability. As the value R_{valid} depends on missingness of the SSD and independent on the attribute of the CUs (honest users or attackers), we assume that there is no attackers in the CRN for computational simplicity. Moreover, because R_{valid} of Fuzzy C-Means (FCM)-clustering algorithm and DSND algorithm are also independent on the number of CUs, N is set to be 10 as an example. The same results of these two algorithms could be derived when N chooses other values. In the experiment, different values of missing data probability from 0.02 to 0.2 are tested, and the length of sensing slots is $T = 100$.

The simulation results show that R_{valid} of presented algorithms decrease with increasing of missing data probability of SSD. It indicates that the larger P_{miss} brings about lower R_{valid}, and the attack detection algorithms have to accumulate more SSD (wait for longer sensing slots) to achieve the same performance as they do when the SSD is complete. Moreover, the DBAD is more sensitive to the missing data because the statistics it applied is calculated based on the sensing reports from all the N CUs in a sensing slot. Missingness of any one of these reports could invalidate the current sensing slot.

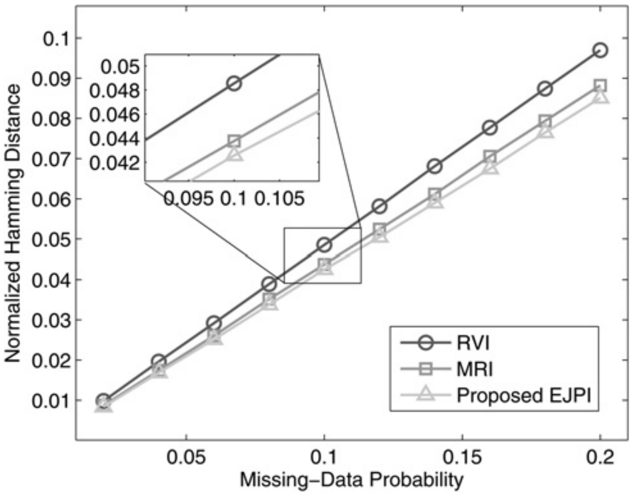


Fig. 4 *NHD under different missing data probabilities*

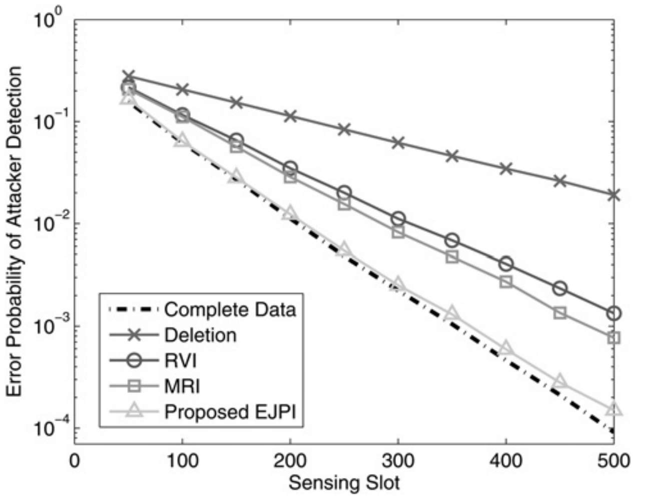


Fig. 5 *Attacker detection performance with elapsing of sensing slots*

5.2 Performance of imputation algorithms

In Fig. 4, the NHD between complete SSD and imputed SSD is utilised to evaluate the performance of different imputation algorithms. In this experiment, the simulation parameters are set as follows: $P_0 = 0.7$, $P_{\text{fa}} = P_{\text{md}} = 0.3$, $q_{01} = q_{10} = 1$, $N = 10$, $M = 3$, and the length of observation is $T = 100$. As shown in the figure, the NHD increases linearly with increasing of missing data probability, and proposed EJPI achieves the lowest NHD among the presented imputation schemes.

5.3 Attacks detection with incomplete SSD

In this subsection, the performance of DBAD with supporting of different imputation algorithms is tested based on 'imputation+DBAD' simulation

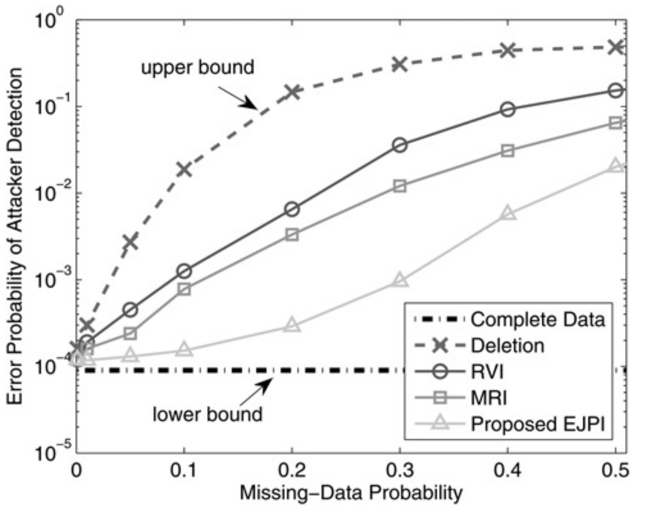


Fig. 6 *Attacker detection performance under different missing data probabilities*

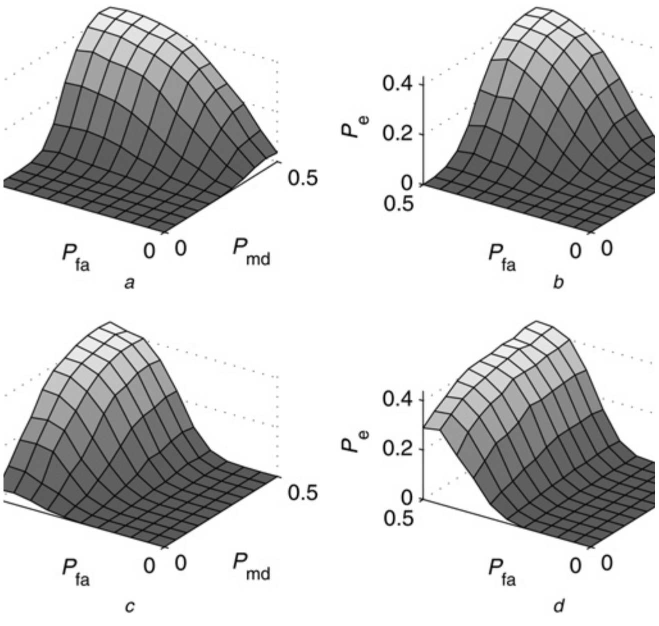


Fig. 7 *Effects of P_{fa} and P_{md} on error detection probability of CUs*

a $P_0 = 0.25$
b $P_0 = 0.5$
c $P_0 = 0.75$
d $P_0 = 0.9$

model. In the experiment, we consider a CRN with $N = 10$ CUs, in which there are $M = 3$ attackers.

In Fig. 5, the performance curves of DBAD based on different pre-processing approaches are presented. In the experiment, the prior probability that PU is absent is $P_0 = 0.7$, and $P_{fa} = P_{md} = 0.3$. The behaviour of attackers is set to be $q_{01} = q_{10} = 1$. The missing data probability P_{miss} is 0.1. As shown in the figure, P_e decreases with elapsing of sensing slots exponentially, and the curve of deletion approach (dashed line with '×') has the largest value among the tested approaches. It is because that all the sensing slots with incomplete SSD are discarded, and the observed SSD in the discarded sensing slots is not utilised. According to (16), when $P_{miss} = 0.1$ and $N = 10$, we have $P_{inva} = 0.6513$. In this circumstance, the DBAD needs about three times of observations to achieve the performance as it does based on complete SSD. In the figure, the propose EJPI (the line with triangles) has the best performance among the tested imputation algorithms, and it is close to that of complete SSD. It means that the proposed EJPI almost compensates the performance loss of DBAD when P_{miss} is small.

Missing data probability P_{miss} is a key factor that affects performance of attacks detection. In Fig. 6, the P_e of DBAD based on different values of P_{miss} is computed. The length of sensing slots is set to be $T = 500$, and other parameters are the *same* as those in experiment of Fig. 5. In this case, P_e based on complete SSD has an unchangeable value, which is about $C = 9.1 \times 10^{-5}$, and it could be regarded as the lower bound for the imputation algorithms. Especially, when $P_{miss} \to 0$, we have $P_e \to C$. The curves of imputation algorithms (including RVI, MRI, and EJPI) increase with increasing of P_{miss}. Furthermore, 'EJPI+DBAD' has the lowest error probability of attacker detection among the tested imputation algorithms, and the error probability is <0.02, even when $P_{miss} = 0.5$. The simulation results indicate that the proposed EJPI is an effective approach to restore the characteristics of the incomplete SSD for SSDF attacks detection.

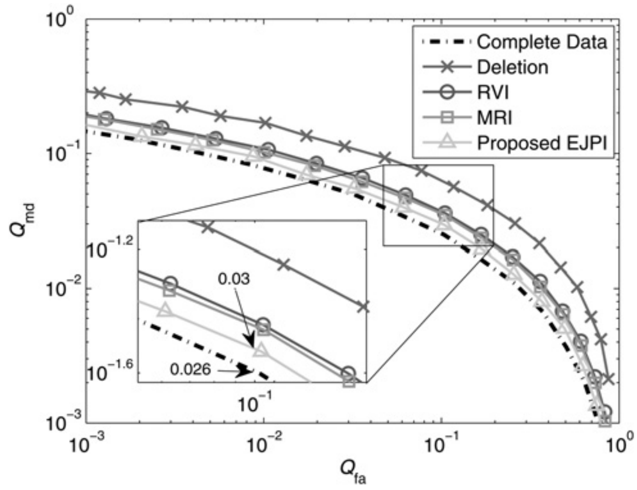


Fig. 8 *ROC curves of CSS with different SSD pre-processing schemes*

To investigate the effects of parameters P_0, P_{fa}, and P_{md} on error detection probability of CUs P_e, we provide the experiment as shown in Fig. 7. In this experiment, P_e is computed when P_0 is set to be different values (from Figs. 7*a–d*, the value of P_0 is 0.25, 0.5, 0.75, 0.9, respectively), and P_{fa}, P_{md} vary from 0 to 0.5. In the simulation, $N = 10$, $M = 3$, $q_{01} = q_{10} = 1$, $P_{miss} = 0.1$, and $T = 40$.

The results show that P_e has a low value when P_{fa} and P_{md} approach to 0. It is because when P_{fa} and P_{md} are small, the CUs could detect the PU correctly. The distribution of falsified SSD is significantly different from that of honest users, and the attackers could be distinguished easily based on their abnormal SSD. Moreover, the parameter P_0 also influences the simulation results. When P_0 is large (see Fig. 7*d*, $P_0 = 0.9$), P_{fa} has more impacts on P_e than P_{md} does. It is because large value of P_0 means that the PU is absent in most of sensing slots, the parameter P_{fa} affects the SSD of CUs significantly. In addition, when $P_0 = 0.5$, the parameters P_{fa} and P_{md} have the same impacts on the error detection probability of CUs as shown in Fig. 7*b*.

5.4 *Secure CSS with incomplete SSD*

In Fig. 8, we provide the receiver operating characteristic (ROC) curves of secure CSS mechanisms. The procedure of secure CSS is described as follows: (i) *pre-process* the incomplete SSD, (ii) *detect* the attackers in the CRN based on the processed SSD, and (iii) *isolate* the detected attackers and *generate* the global spectrum sensing result. In this simulation, different SSD pre-processing methods are applied, the FC uses the DBAD to detect the attackers, and it adopts the 'majority' fusion rule to generate global spectrum sensing result. In the simulation, the sensing performance of CUs is calculated according to equation (3)(4) in [20], and other parameters are the *same* as those applied in Fig. 5.

As shown in Fig. 8, CSS with the proposed EJPI algorithm achieves the best performance among the tested pre-processing schemes. Specifically, when the false alarm probability of CSS is $Q_{fa} = 0.1$, the missed detection probability of CSS with EJPI is about $Q_{md\_EJPI} = 0.03$, which is close to that based on complete sensing data, i.e. $Q_{md\_Comp} = 0.026$.

6 Conclusions

Incomplete SSD is a real challenge in SSDF attacks detection. In this paper, the mechanism and the impacts of missing data in SSDF attacks detection are analysed, and a practical SSD pre-processing algorithm, i.e. EJPI, is proposed. The proposed algorithm fills in the missing values in the SSD without any prior information of the PU or the CUs, and it supports the attacks detection effectively. Simulation results show that the EJPI restores the characteristics of the SSD and improves the performance of existing attacks detection algorithm DBAD when the SSD is incomplete.

7 Acknowledgment

This work was supported by the National Science Foundation for Post-Doctoral Scientists of China (201003797), the National

Natural Science Foundation of China (61371196 and 61172062), and the Natural Science Foundation of Jiangsu, China (BK2011116).

8 References

1 Mitola, J.: 'Software radio architecture' (John Wiley & Sons, 2000)

2 Haykin, S.: 'Cognitive radio: brain-empowered wireless communications', *IEEE J. Sel. Areas Commun.*, 2005, **23**, pp. 201–220

3 Ghasemi, A., Sousa, E.: 'Collaborative spectrum sensing for opportunistic access in fading environments'. Proc. of IEEE DySPAN, 2005

4 Mishra, S.M., Sahai, A., Brodersen, R.W.: 'Cooperative sensing among cognitive radios'. Proc. of IEEE ICC, 2006

5 Chen, R., Park, J.M., Bian, K.: 'Robust distributed spectrum sensing in cognitive radio networks'. Proc. of IEEE INFOCOM, 2008, pp. 1876–1884

6 Zeng, K., Pawełczak, P., Čabrić, D.: 'Reputation-based cooperative spectrum sensing with trusted nodes assistance', *IEEE Lett. Commun.*, 2010, **14**, pp. 226–228

7 Li, H., Han, Z.: 'Catch me if you can: an abnormality detection approach for collaborative spectrum sensing in cognitive radio networks', *IEEE Trans. Wirel. Commun.*, 2010, **9**, pp. 3554–3565

8 Yao, J., Wu, Q., Wang, J.: 'Attacker detection based on dissimilarity of local sensing reports in collaborative spectrum sensing', *IEICE Trans. Commun.*, 2012, **E95-B**, pp. 3024–3027

9 Li, L., Chigan, C.: 'Fuzzy C-means clustering based secure fusion strategy in collaborative spectrum sensing'. Proc. of IEEE ICC, 2014, pp. 1355–1360

10 Wang, J., Yao, J., Wu, Q.: 'Stealthy-attacker detection with a multidimensional feature vector for collaborative spectrum sensing', *IEEE Trans. Veh. Technol.*, 2013, **62**, pp. 3996–4009

11 Yao, J., Zhu, H., Liu, Y., *et al.*: 'Secure spectrum data fusion with presence of massive malicious users', *Frequenz*, 2015, **69**, pp. 271–280

12 Yao, J., Wu, Q., Feng, S., *et al.*: 'Online malicious behavior detection in collaborative spectrum sensing: a change detection approach', *Radioengineering*, 2013, **22**, pp. 536–543

13 Ma, J., Zhao, G., Li, Y.: 'Soft combination and detection for cooperative spectrum sensing in cognitive radio networks', *IEEE Trans. Wirel. Commun.*, 2008, **7**, pp. 4502–4507

14 Maleki, S., Pandharipande, A., Leus, G.: 'Energy-efficient distributed spectrum sensing for cognitive sensor networks', *IEEE Sens. J.*, 2011, **11**, pp. 565–573

15 Donders, A.R.T., Heijden, G.J.M.G., Stijnen, T., *et al.*: 'Review: a gentle introduction to imputation of missing values', *J. Clin. Epidemiol.*, 2006, **59**, pp. 1087–1091

16 Khan, Z., Lehtomaki, J., Umebayashi, K., *et al.*: 'On the selection of the best detection performance sensors for cognitive radio networks', *IEEE Signal Process. Lett.*, 2010, **17**, pp. 359–362

17 Haykin, S., Thomson, D., Reed, J.: 'Spectrum sensing for cognitive radio', *Proc. IEEE*, 2009, **97**, pp. 849–877

18 Digham, F.F., Alouini, M.S., Simon, M.K.: 'On the energy detection of unknown signals over fading channels', *IEEE Trans. Commun.*, 2007, **55**, pp. 21–24

19 Graham, J.W.: 'Missing data analysis: making it work in the real world', *Annu. Rev. Psychol.*, 2009, **60**, pp. 549–576

20 Letaief, K.B., Zhang, W.: 'Cooperative communications for cognitive radio networks', *Proc. IEEE*, 2009, **97**, pp. 878–893