# Physical Detection of Misbehavior in Relay Systems With Unreliable Channel State Information

Tiejun Lv, *Senior Member, IEEE*, Yajun Yin, Yueming Lu, Shaoshi Yang, *Member, IEEE*, Enjie Liu, and Gordon Clapworthy

***Abstract*— We study the detection of misbehavior in a Gaussian relay system, where the source transmits information to the destination with the assistance of an amplify-and-forward relay node subject to unreliable channel state information (CSI). The relay node may be potentially malicious and corrupt the network by forwarding garbled information. In this situation, misleading feedback may take place, since reliable CSI is unavailable at the source and/or the destination. By classifying the action of the relay as detectable or undetectable, we propose a novel approach that is capable of coping with any malicious attack detected and continuing to work effectively in the presence of unreliable CSI. We demonstrate that the detectable class of attacks can be successfully detected with a high probability. Meanwhile, the undetectable class of attacks does not affect the performance improvements that are achievable by cooperative diversity, even though such an attack may fool the proposed detection approach. We also extend the method to deal with the case in which there is no direct link between the source and the destination. The effectiveness of the proposed approach has been validated by numerical results.**

***Index Terms*— Physical layer security, integrity check, unreliable CSI, cooperative relay communications.**

## I. Introduction

PHYSICAL layer security (PLS) is a promising technology that provides secure wireless transmissions by smartly exploiting imperfections of the communications medium [1]. Cooperative relaying is beneficial for improving the coverage and transmission reliability of wireless systems [2], where single-antenna devices can form a virtual antenna array to provide cooperative spatial diversity [3], [4]. However, such benefits are attained only when the relays are trustworthy and always comply with cooperative protocols. In an adversarial case, some relays might maliciously alter the information sent by the source, thus degrading the performance of the relaying system significantly. The dependence of cooperative systems on the relays represents an inherent vulnerability [5]. Therefore, early detection of misbehavior is essential to maintaining the security of relaying systems and to combating malicious attacks.

Traditionally, detection methods are based on cryptography keys or authentication keys, requiring the source and the destination to share a secret key [6]–[8]. The key-based detection approach is far from ideal as it imposes a high computational cost and needs a key distribution mechanism. Alternatively, it is possible to detect malicious relays from the physical layer perspective. In particular, Mao and Wu [9] proposed a cross-layer detecting scheme, where pseudo-random tracing symbols were inserted into information bits. To identify the malicious relays, the destination measures the error probability of the observed tracing symbols, according to their *a priori* ground truth. In [10]–[12], Lo *et al.* applied a tracing-based method to non-coherent detection in various scenarios, requiring no channel state information (CSI). Note that the transmission of tracing symbols also requires the support from a key-distribution mechanism. Moreover, the performance of tracing-based schemes is highly dependent on the number of tracing symbols used, and an excessive number of them can significantly reduce the bandwidth efficiency.

To avoid the use of external assistance, many detecting schemes exploit 'clean' references stemming from the relaying system itself. A 'clean' reference contains information that has not been manipulated by the relay for sure. For example, in the orthogonal frequency-division multiplexing (OFDM) based detection scheme of [13], the source regards the transmitted information as a reference. Thus, the misbehavior of the relay is detected by examining the correlation between the reference and the information that is forwarded by the relay but overheard at the source. Detection schemes can also be implemented at the destination [14], [15]. The direct link between the source and the destination, as a 'clean' reference to the relay link, is used to compare between two different links to determine the relay behavior. However, these schemes [9]–[15] assume that each malicious relay behaves in an independent identically distributed (i.i.d.) manner of a specific form. With respect to arbitrary i.i.d. attacks, Graves and Wong [16] and Cao *et al.* [17] proposed a novel detection approach in which the relay behavior is modeled as an attack channel to check for any misbehavior. In [16], the source

Manuscript received September 15, 2017; revised February 1, 2018; accepted February 16, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671072 and in part by the National Key Research and Development Program of China under Grant 2016YFB0800302. *(Corresponding author: Tiejun Lv.)*

T. Lv and Y. Yin are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: lvtiejun@bupt.edu.cn; yinyajun@bupt.edu.cn).

Y. Lu is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: ymlu@bupt.edu.cn).

S. Yang is with the Department of Wireless Network Research, Huawei Technologies Co., Ltd., Shenzhen 518129, China (e-mail: shaoshi.yang@ieee.org).

E. Liu and G. Clapworthy are with the School of Computer Science and Technology, University of Bedfordshire, Luton LU1 3JU, U.K. (e-mail: enjie.liu@beds.ac.uk; gordon.clapworthy@beds.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2824728

extracts the estimation of an attack channel based on its own transmitted and observed information. This detection method has been extended to the scenario in which a direct 'clean' reference is unavailable [17]. However, [17] entirely depends on the source distribution. In [18], a detecting and tracing scheme for a multi-relay network was studied by partitioning the network into several sub-networks as described in [17].

The detection schemes [9], [13]–[18] above are enabled under an ideal assumption that reliable CSI is known in advance. However, reliable CSI may not be available in practice, especially when relays are malicious. For instance, malicious relays are reluctant to cooperate initially and, hence, they may deliberately manipulate the channel estimation process with ease. The whole system is then deceived into a state of unreliable CSI. In such cases, the previously mentioned schemes [9], [13]–[18] may be severely compromised. Considering a point-to-point system, Tugnait [19] proposed a scheme to detect the pilot contamination attack, which causes unreliable CSI, by superimposing a random sequence on the training sequence and using source enumeration methods.

In this paper, we consider a cooperative relaying system with a source-destination pair and a single relay employing an amplify and forward (AF) strategy [20]. The potentially malicious relay is capable of forwarding false information in an arbitrary i.i.d. manner. It can also provide unreliable CSI to degrade the system's performance. Falsified forwarding together with the unreliable CSI makes the detection of misbehavior very difficult. Our goal is to detect misbehavior based on physical-layer observations. The key difference between existing work [16]–[18] and ours is that we take into account that the channel estimation process may be compromised and hence the available CSI is unreliable. The main contribution of this paper is summarized as follows.

1) We study the misbehavior of the malicious relay under the assumption that the misbehavior arises not only from falsified forwarding, but also from dishonest feedback. According to different combinations of misbehavior and from the detection point of the view, we define two mutually exclusive attack types – *detectable* and *undetectable*. We prove that a detectable attack can be detected asymptotically by examining the distance measure between the distribution of physical-layer observations and the distribution of the calculated received symbols. The proposed detection scheme needs no extra secret keys.
2) We prove that an undetectable attack does not affect the bit error rate (BER) performance that is achievable by cooperative diversity, even though it cannot be identified. This implies that an undetectable attack hardly influences the reliability performance of the relay network, in the sense that the benefits of diversity gain are retained.
3) For relay systems having direct links, we choose the direct link as a 'clean' reference. We then extend the proposed detection scheme to relay systems having no direct link, where the source distribution is known. Furthermore, in the absence of prior information of the source, we design a 'clean' reference by introducing artificial noise (AN) to aid the proposed detection scheme.

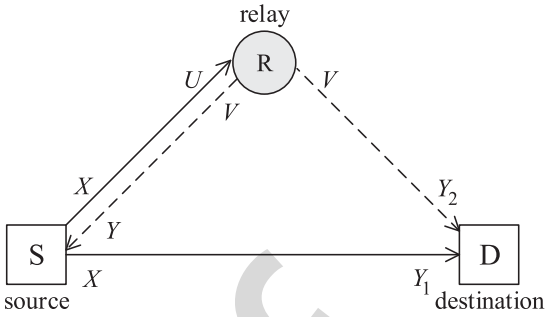



Fig. 1. A cooperative relay system consisting of a source-destination pair and a potential malicious relay with direct link.

The remainder of this paper is organized as follows. The system model and the misbehavior types are described in Section II. In Section III, we elaborate on the proposed detection scheme for detectable attacks and prove that an undetectable attack can achieve the same BER as a detectable attack. The detection scheme is extended to the scenario in which a direct link is absent in Section IV. Section V provides numerical examples and discussions, and conclusions are drawn in Section VI.

*Notation:* Upper and lower case letters denote, respectively, random variables and their realizations. Sans-serif letters denote general elements. $|\cdot|$ represents an absolute value and $\|\cdot\|$ represents the Euclidean norm. The transpose of the vector $a$ is $a^T$. For a sequence $x^N$, both $x[i]$ and $x_i$ denote the $i$-th element in $x^N$. $\mathcal{X}$ denotes the alphabet of $X$. $I(x[i]=\mathsf{x})$ is the indicator function denoting whether or not $x[i]$ is $\mathsf{x}$. $F_{X^N}(\mathsf{x}) = 1/N\sum_{i=1}^{N} I(x[i]=\mathsf{x})$ is used to denote the empirical distribution of $x^N$, and implies the relative proportion of occurrence of symbol $\mathsf{x}$ in $x^N$. For a sequence $y^N$ with consecutive values, the empirical distribution function is trivially defined as $F_{Y^N}(t) = 1/N\sum_{i=1}^{N} I(y[i] < t)$. In a similar fashion, we denote the conditional empirical distribution as

$$F_{Y^N|X^N}(t|\mathsf{x}) = \frac{\sum_{i=1}^{N} I(y[i]<t)I(x[i]=\mathsf{x})}{\sum_{i=1}^{N} I(x[i]=\mathsf{x})}.$$

## II. System Model

### A. Cooperative Transmission

We consider a typical cooperative relay network consisting of a source-destination pair and a potential malicious relay as shown in Fig. 1, where the source (S) tries to send information to the destination (D) with the aid of a relay node (R) and a direct link (S-D link). A relay system without a direct link will be considered in Section IV. Although this three-node relay network model is simple, it is fundamental for studying relay aided cooperative communications. Compared with traditional non-cooperative networks, three-node relay networks can offer several benefits, such as better connectivity, higher throughput

and greater reliability [23]–[25]. The three-node relay network model can also be extended to more complicated network topology.

In Fig. 1, the solid and dashed lines represent two transmission phases, i.e. phases 1 and 2, respectively. The wireless channels are assumed to be quasi-static in the same phase.

*1) Phase 1:* S first broadcasts an $N$-length i.i.d. sequence $X^N$ simultaneously to R and D. Let $U$ and $Y_1$ be the symbols received at R and D, respectively. In the symbol-by-symbol expression, the time index is omitted. The received symbols in Phase 1 can be expressed as

$$U = h_{sr}X + W_{sr}, \tag{1a}$$

$$Y_1 = h_{sd}X + W_{sd}. \tag{1b}$$

*2) Phase 2:* R receives $U^N$, processes it, and then forwards $V^N$ to D. Here, the symbol $V$ is a processed version of the received symbol $U$. Due to the broadcast nature of wireless communication, S can overhear the forwarded information $V^N$ at the same moment. Let $Y$ denote the received symbol overheard by S and $Y_2$ denote the received symbol at D. The received symbols in Phase 2 are given by

$$Y = h_{rs}V + W_{rs}, \tag{2a}$$

$$Y_2 = h_{rd}V + W_{rd}, \tag{2b}$$

where $h_{ij}$ is channel gain between node $i$ and node $j$ with $i, j \in \{\mathrm{S}, \mathrm{R}, \mathrm{D}\}$ and $i \neq j$. Statistically, we can model them as complex Gaussian random variables which capture the effects of pass loss and statistical fading in a wireless channel. The average transmit energy of the transmitted symbol is denoted as $E_s$. $W_{ij}$ represents additive white Gaussian noise (AWGN) with variance $\mathcal{N}_0$ received at node $j$.

CSI needs to be obtained from channel estimation. Before the transmission phases, all nodes participate in the channel estimation process. Since the malicious relay can manipulate the channel estimation process by sending incorrect pilot signals, unreliable CSI $g_{ij}$ may be provided, which is different from the reliable CSI $h_{ij}$. Let $\mathbf{g} = \{g_{sr}, g_{rs}, g_{rd}\}$ and $\mathbf{h} = \{h_{sr}, h_{rs}, h_{rd}\}$ denote the set of the potentially unreliable CSI provided and the set of the corresponding reliable CSI, respectively. Note that the channel gain of the direct link cannot be manipulated by the relay, hence $h_{sd}$ is omitted from both of the CSI sets.

### B. Misbehavior Types

The introduction of the relay opens a door to malicious attacks. Instead of complying with the cooperative strategy, a malicious relay node may exhibit misbehaviors both in the transmission phases and in the channel estimation process. Hence, potentially both the information forwarded and the CSI provided can be manipulated by the malicious relay. We identify the following two types of misbehaviors.

1) *Falsified Forwarding*: the relay receives $U^N$ in Phase 1, and then corrupts it into another sequence $V^N$ to be forwarded in Phase 2. If we assume that the malicious relay misbehaves in an arbitrary i.i.d. manner, the forwarded sequence $V^N$ will obey an arbitrary stochastic distribution conditioned on $U^N$. From the perspective of symbol-by-symbol, the relay processing behavior can be characterized by its conditional probability density function (PDF) $f_{V|U}(v|u)$. It is not difficult to derive that if the relay forwards the received symbol $U$ accurately, the conditional PDF is

$$f_{V|U}(v|u) = \delta(v - u), \tag{3}$$

where $\delta(\cdot)$ is the impulse function. This means that when $U = V$ the relay is amicable with respect to forwarding information. Otherwise, the relay is exhibiting *falsified forwarding*, also known as a *Byzantine attack*.

2) *Dishonest Feedback*: In many wireless communication protocols, the transmitter obtains the CSI estimate from the receiver's feedback. The malicious node is capable of dominating the channel estimation process deliberately. In this case the CSI provided may be unreliable. The unreliable CSI provides a malicious node with an opportunity to undermine relay selection, e.g., to select a malicious node as a qualified relay. Further, the destination node may combine the information received from the relay and the source inappropriately, due to the unreliable CSI. The CSI provided is said to be *reliable* if $\mathbf{g} = \mathbf{h}$. Otherwise, the relay node is considered to be initiating dishonest feedback that creates *unreliable* CSI. Note that imperfect CSI is usually caused by channel estimation error, which is an objective measurement error rather than a deliberate attack. Imperfect CSI does not belong to the scope of physical layer security. Thus, imperfect CSI is not considered in this paper.

Thus we can employ the parameter pair $(f_{V|U}, \mathbf{g})$ to describe the behavior of the relay. Maliciousness due to the misbehavior is defined as follows.

*Definition 1 (Maliciousness of Misbehavior): The relay is considered as cooperative if and only if the pair $(f_{V|U}, \mathbf{g})$ belongs to the set $\{f_{V|U}(v|u), \mathbf{g} | f_{V|U}(v|u) = \delta(v - u), \mathbf{g} = \mathbf{h}.\}$; otherwise, the relay is considered as malicious.*

It is obvious that neither of the above forms of misbehavior is allowed for a cooperative relay. Our goal is to use physical-layer observations to detect maliciousness if and when misbehavior occurs in the relay system.

## III. Detection Approach

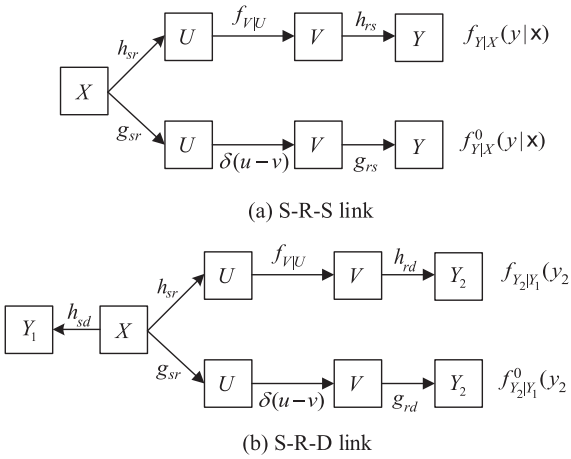
In this section, we describe the proposed approach for detecting maliciousness in a relay system with a direct link, i.e., falsified forwarding and/or dishonest feedback, but first we introduce the concept of detectability of maliciousness.

### A. Maliciousness Detectability

The source S can observe the symbol $Y$ in Phase 2 (see (2)). The symbol $Y$ goes through a real S-R-S link, which may be manipulated by a malicious relay. For S, the transmitted symbol $X$ offers a 'clean' reference.

On one hand, we use the conditional likelihood function

$$f_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; h_{sr}) \times f_{V|U}(v|u) f_{Y|V}(y|v; h_{rs})\,\mathrm{d}u\mathrm{d}v \tag{4}$$




(a) S-R-S link

(b) S-R-D link

Fig. 2. Markov chain of S-R-S link and S-R-D link.

to characterize S-R-S link, where the parameters $f_{V|U}$ and $\mathbf{h}$ are unknown for S.

On the other hand, S also tries to make use of the CSI provided, $\mathbf{g}$, even though it may be unreliable. The conditional PDF at S is computed as

$$\begin{aligned} & f^0_{Y|X}(y|\mathsf{x};\mathbf{g}) \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x};g_{sr})\delta(v-u)f_{Y|V}(y|v;g_{rs})\,\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x};g_{sr})f_{Y|V}(y|u;g_{rs})\,\mathrm{d}u, \end{aligned} \tag{5}$$

where the superscript distinguishes the conditional PDF $f^0_{Y|X}(y|\mathsf{x};\mathbf{g})$ from the conditional likelihood function $f_{Y|X}(y|\mathsf{x};f_{V|U},\mathbf{h})$. Whenever there is no ambiguity, we will employ such a notation, i.e. $f^0_{Y|X}(y|\mathsf{x})$ and $f_{Y|X}(y|\mathsf{x})$, for simplicity. It is observed that the relay is considered to faithfully forward as $f_{V|U}(v|u)=\delta(v-u)$ appears in the expression for $f^0_{Y|X}(y|\mathsf{x})$.

Since $(X,U,V,Y)$ forms a Markov chain as $X\to U\to V\to Y$, we have four cases according to different combinations of the parameter pair $(f_{V|U},\mathbf{g})$, as follows:

1) $f_{V|U}=\delta(v-u)\cap\mathbf{g}=\mathbf{h}$; full cooperative relay (no misbehavior), we have $f_{Y|X}(y|\mathsf{x})=f^0_{Y|X}(y|\mathsf{x})$.
2) $f_{V|U}\neq\delta(v-u)\cap\mathbf{g}=\mathbf{h}$; malicious relay with falsified forwarding, we have $f_{Y|X}(y|\mathsf{x})\neq f^0_{Y|X}(y|\mathsf{x})$.
3) $f_{V|U}=\delta(v-u)\cap\mathbf{g}\neq\mathbf{h}$; malicious relay with dishonest feedback, we have $f_{Y|X}(y|\mathsf{x})\neq f^0_{Y|X}(y|\mathsf{x})$.
4) $f_{V|U}\neq\delta(v-u)\cap\mathbf{g}\neq\mathbf{h}$; malicious relay with both misbehaviors. This is difficult to analyse as it is hard to determine the equivalence of $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$.

As shown in Fig. 2 (a), it is easy to check the relationship between $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$ in the four different cases. The first three are easily determined, but Case 4) is a demanding problem. From the above, based on the parameter pair $(f_{V|U},\mathbf{g})$, the inequality of $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$ is a sufficient condition to determine misbehavior.

TABLE I
THE RELATIONSHIP BETWEEN DEFINITION 1 AND DEFINITION 2

| | Detectable Class $T$ | Undetectable Class $T^c$ |
|---|---|---|
| Malicious Relay | Detectable attack | Undetectable attack |
| Cooperative Relay | $\varnothing$ | No misbehavior |

This conclusion helps to detect misbehavior in the relaying system considered. We define a set $T_1$ as:

$$T_1 := \Big\{ f_{V|U}, \{g_{sr}, g_{rs}\} \,\Big|\, f_{Y|X}(y|\mathsf{x}) \neq f^0_{Y|X}(y|\mathsf{x}) \Big\}. \tag{6}$$

If $T_1$ holds, there must be misbehavior in the S-R-S link; unfortunately we cannot jump to a conclusion of no misbehavior if $T_1$ does not hold, owing to Case 4. Thus, $T_1$ is referred to as the *detectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-S link; correspondingly, the complementary set ${T_1}^c$ of $T_1$ is called the *undetectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-S link.

In order to fully check the parameter pair $(f_{V|U},\mathbf{g})$, an S-R-D link should be included. For the S-R-D link, the set $T_2$ is defined as

$$T_2 := \Big\{ f_{V|U}, \{g_{sr}, g_{rd}\} \,\Big|\, f_{Y_2|Y_1}(y_2|y_1) \neq f^0_{Y_2|Y_1}(y_2|y_1) \Big\}, \tag{7}$$

where $f_{Y_2|Y_1}(y_2|y_1)$ and $f^0_{Y_2|Y_1}(y_2|y_1)$ are, respectively, the likelihood function and PDF of the symbol $Y_2$ received at D from the relay link conditioned on the symbol $Y_1$ received from the direct link. $T_2$ and its complementary set $T_2^c$ are referred to as, respectively, the *detectable set* and the *undetectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-D link. Fig.2 (b) helps to check the detectable set $T_2$ directly.

The parameter pair $(f_{V|U},\mathbf{g})$ is completely partitioned by combinations of $T_1$ and $T_2$. We call $T=T_1\bigcup T_2$ as the *detectable class*, in which misbehavior is inevitable. It is emphasized that the complementary set $T^c=T_1^c\bigcap T_2^c$ of $T$ implies that the behavior can be cooperative or malicious. Thus, attack types can be given by the following definition.

*Definition 2 (Attack Types): If the parameter pair $(f_{V|U},\mathbf{g})$ belongs to the detectable class $T$, misbehavior is certain, and this is called a detectable attack; if $T^c$ holds and the relay is malicious, the resulting misbehavior is called an undetectable attack.*

From Definition 2, it is seen that detectable attacks map directly to the detectable class, whereas undetectable attacks map only to a subset of the undetectable class. An undetectable attack demands that falsified forwarding and dishonest feedback occur simultaneously, but the attack is not detected by a given detection approach. The undetectable attack is a small probability event compared to the detectable attack, because the undetectable attack is required to satisfy stricter conditions. It is emphasized that the undetectable attack is still in an infinite set. Table I illustrates the relationship between Definition 1 and Definition 2, where $\varnothing$ denotes the empty set. The action of the relay, i.e., the parameter pair $(f_{V|U},\mathbf{g})$, can be fully classified by use of Definitions 1 and 2. A detectable attack results from the overlap of these two definitions, and the

identification of a detectable attack is precisely equivalent to the identification of the detectable class $T$.

### B. Identification of a Detectable Attack

As the detectable class $T$ involves both $T_1$ and $T_2$, detection is implemented at the source node and at the destination node. In order to quantify the consecutive received symbols, it is convenient to use an $n'$-length sequence $(t_1, t_2, \cdots, t_{n'})$ satisfying $a = t_1 < t_2 < t_3 \cdots < t_{n'} = b$, where the quantization range $[a, b]$ depends on $n'$. Further, we consider the quantization interval $\Delta = \frac{b-a}{n'-1}$ to be such that $\lim_{n'\to\infty} \Delta = 0$.

*1) Decision Metric at S:* The detection at S focuses on the S-R-S link, in which the source uses its transmitted symbols as a reference to check whether or not action of the relay node is in the detectable set $T_1$. We employ the empirical CDF to approximate the likelihood function $f_{Y|X}(y|\mathsf{x})$. By jointly considering the transmitted and received signal sequences $(X^N, Y^N)$, the conditional empirical CDF $F_{Y^N|X^N}(t|\mathsf{x})$ at S is written as

$$F_{Y^N|X^N}(t|\mathsf{x}) = \frac{\sum_{i=1}^{N} I(y[i] < t) I(x[i] = \mathsf{x})}{\sum_{i=1}^{N} I(x[i] = \mathsf{x})}. \tag{8}$$

Naturally, a statistical decision metric $D_1^N$ is expressed as

$$D_1^N = \frac{1}{n'} \sum_{m=1}^{n'} \left| F_{Y^N|X^N}(t_m|\mathsf{x}) - F^0_{Y|X}(t_m|\mathsf{x}) \right|, \tag{9}$$

where $F^0_{Y|X}(t_m|\mathsf{x})$ is the CDF of $f^0_{Y|X}(t_m|\mathsf{x})$ as given in (5).

*2) Decision Metric at D:* The detection at D is related to the security of the S-R-D link and takes place at the same time as the detection at S. Since D receives the signal $Y_1^N$ in Phase 1 (see (1)) and then the signal $Y_2^N$ in Phase 2 (see (2)), the likelihood function $f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h})$ characterizing the S-R-D link can be obtained as

$$\begin{aligned} &f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; h_{sr}) f_{V|U}(v|u) f_{Y_2|V}(y_2|v; h_{rd})\, \mathrm{d}u\mathrm{d}v. \end{aligned} \tag{10}$$

However, unlike S, D is inaccessible to the transmitted signal $X^N$. The received signal $Y_1^N$ in the direct link is exploited as a 'clean' reference for the detection at D. $(Y_1, X, Y_2)$ forms a Markov chain as $Y_1 \to X \to Y_2$, and $Y_1$ and $Y_2$ are conditionally independent for a given $X$, so the likelihood function conditioned on $Y_1 \le t$ can be mathematically expressed as

$$\begin{aligned} &f_{Y_2|Y_1}(y_2|t; f_{V|U}, \mathbf{h}) \\ &= \int_{-\infty}^{t} \sum_{\mathsf{x}\in\mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) \Pr(X = \mathsf{x})\, \mathrm{d}y_1 \\ &\quad \Big/ \int_{-\infty}^{t} \sum_{\mathsf{x}\in\mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) \Pr(X = \mathsf{x})\, \mathrm{d}y_1. \end{aligned} \tag{11}$$

Since the conditional PDF at D $f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g})$ is computed as

$$f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g}) = \int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; g_{sr}) f_{Y_2|V}(y_2|v; g_{rd})\, \mathrm{d}u, \tag{12}$$

the conditional PDF $f^0_{Y_2|Y_1}(y_2|t; \mathbf{g})$ can be formulated as

$$\begin{aligned} &f^0_{Y_2|Y_1}(y_2|t; \mathbf{g}) \\ &= \int_{-\infty}^{t} \sum_{\mathsf{x}\in\mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g}) \Pr(X = \mathsf{x})\, \mathrm{d}y_1 \\ &\quad \Big/ \int_{-\infty}^{t} \sum_{\mathsf{x}\in\mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) \Pr(X = \mathsf{x})\, \mathrm{d}y_1. \end{aligned} \tag{13}$$

For ease of presentation, $f_{Y_2|Y_1}(y_2|y_1; f_{V|U}, \mathbf{h})$ and $f^0_{Y_2|Y_1}(y_2|y_1; \mathbf{g})$ are written as $f_{Y_2|Y_1}(y_2|y_1)$ and $f^0_{Y_2|Y_1}(y_2|y_1)$ in the section below.

Based on the pair of received signals $(Y_1^N, Y_2^N)$, the empirical conditional CDF $F_{Y_2^N|Y_1^N}(s|t)$ can be expressed as

$$F_{Y_2^N|Y_1^N}(s|t) = \frac{\sum_{i=1}^{N} I(y_1[i] < t) I(y_2[i] < s)}{\sum_{i=1}^{N} I(y_1[i] < t)}. \tag{14}$$

By employing $F_{Y_2^N|Y_1^N}(s|t)$, the statistical decision metric $D_2^N$ for the detection at D is given by

$$D_2^N = \frac{1}{n'^2} \sum_{p=1}^{n'} \sum_{q=1}^{n'} \left| F_{Y_2^N|Y_1^N}(t_p|t_q) - F^0_{Y_2|Y_1}(t_p|t_q) \right|, \tag{15}$$

where $F^0_{Y_2|Y_1}(t_p|t_q)$ is the CDF of $f^0_{Y_2|Y_1}(t_p|t_q)$ as given in (13).

*3) Detection:* After obtaining the decision statistical metrics $D_1^N$ and $D_2^N$, we first identify whether the action of the relay falls into the detectable class $T$ or not. The following proposition will show how $D_1^N$ and $D_2^N$ identify, respectively, the detectable sets $T_1$ in the S-R-S link and $T_2$ in the S-R-D link.

*Proposition 1 (Detection at S and D): In the S-R-S link, $T_1$ can be detected by $D_1^N$ at S; in the S-R-D link, $T_2$ can be detected by $D_2^N$ at D. For $i = 1, 2$, the two decision metrics $D_1^N$ and $D_2^N$ have the following properties:*

*i)* $\lim_{N\to\infty} \Pr\left(D_i^N > \rho_1 \;\middle|\; (f_{V|U}, \mathbf{g}) \in T_i\right) = 1$, *when* $\Pr\left((f_{V|U}, \mathbf{g}) \in T_i\right) > 0$,

*ii)* $\lim_{N\to\infty} \Pr\left(D_i^N > \rho_2 \;\middle|\; (f_{V|U}, \mathbf{g}) \in T_i^c\right) = 0$, *when* $\Pr\left((f_{V|U}, \mathbf{g}) \in T_i^c\right) > 0$, *where $\rho_1$ and $\rho_2$ are strictly positive, and can be arbitrary small.*

*Proof:* See Appendix A. ■

*Remark 1: Take the detection at S for example. From (6), the detectable set $T_1$ implies that the likelihood function $f_{Y|X}(y|\mathsf{x})$ differs from the conditional PDF $f^0_{Y|X}(y|\mathsf{x})$. According to the law of large numbers, the empirical distribution $F_{Y^N|X^N}$ approaches the CDF of $f_{Y|X}(y|\mathsf{x})$ as $N \to \infty$. From the proof of Proposition 1, we can see that $D_1^N$ uses $F_{Y^N|X^N}$ as the bridge to measure the 'distance' between $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$.*

*Remark 2: Proposition 1 points out that, if the behavior of the relay follows the undetectable set $T_i^c, i = 1, 2$, then $D_i^N \to 0$. Otherwise, it is probable that the source is capable of identifying a detectable attack. In addition, the missed detection and false alarm probabilities of $D_i^N$ can be arbitrary small as $N \to \infty$.*

Combining the detection at S with the detection at D, the detectable class $T$ can be identified by the proposed Algorithm 1 below.

**Algorithm 1** The Identification Procedure for a Detectable Attack

1: Initialization: Select appropriate $N$ and $n'$, and receive the CSI set $\mathbf{g}$.
2: Calculate the decision metrics: S computes $D_1^N$ based on $(X^N, Y^N)$, and D computes $D_2^N$ based on $(Y_1^N, Y_2^N)$ simultaneously.
3: **if** $D_1^N \to 0 \cap D_2^N \to 0$ **then**
4: $(f_{V|U}, \mathbf{g}) \in T_1^c \bigcap T_2^c$, the action of the relay belongs to the undetectable class $T^c$.
5: **else**
6: $(f_{V|U}, \mathbf{g}) \in T_1 \bigcup T_2$, the action of the relay belongs to the detectable class $T$.
7: **end if**

According to Algorithm 1, if the action of the relay belongs to the detectable class, we draw a conclusion immediately that the relay is suffering from a malicious attack; if the action of the relay belongs to the undetectable class, we cannot decide whether the relay is suffering from a malicious attack or not.

### C. Signal Detection of the Undetectable Class

According to Definitions 1 and 2, we know that undetectable class consists of undetectable attacks and cooperative (or friendly) relays. In other words, if falsified forwarding and dishonest feedback occur simultaneously, it is possible that an undetectable attack has the same statistical behavior as a cooperative relay. Thus, we cannot identify whether a malicious attack is occurring by use of Algorithm 1; consequently, a malicious relay that is performing an undetectable attack can disguise itself as a cooperative one – from the signal processing point of view, the performance of an undetectable attack is the same as that of the cooperative relay. On the assumption of an i.i.d. attack, the undetectable attack can be neglected.

At D, maximum-likelihood (ML) demodulation is used, based on the CSI $\mathbf{g}$. Following (1) and (13), the symbols received from the direct link and the relay link are re-expressed as

$$\begin{cases} Y_1 = h_{sd}X + W_{sd}, \\ Y_2 = g_{sr}g_{rd}X + g_{rd}W_{sr} + W_{rd}, \end{cases}$$

which are written in vector form as $\mathbf{Y} = \mathbf{H}X + \mathbf{W}$, with $\mathbf{Y} = \left[Y_1, Y_2\right]^T$, $\mathbf{H} = \left[h_{sd},\ g_{sr}g_{rd}\right]^T$ and $\mathbf{W} = \left[W_{sd},\ g_{rd}W_{sr} + W_{rd}\right]^T$.

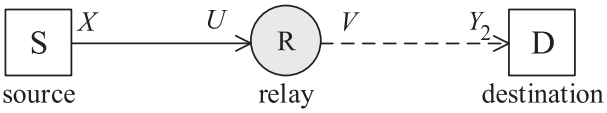



Fig. 3. A cooperative relay system consisting of a source-destination pair and a potential malicious relay without direct link.

ML detection is then performed as

$$\hat{X} = \underset{X \in \mathcal{X}}{\operatorname{argmax}} \Pr(\mathbf{Y}|X) = \underset{X \in \mathcal{X}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{H}X\|^2. \quad (16)$$

From (16), the joint PDF of $\mathbf{Y}$, $f_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; f_{V|U}, \mathbf{h})$, only effects ML detection. Then the following proposition gives a property of the undetectable class $T^c$.

*Proposition 2: If the parameter pair $(f_{V|U}, \mathbf{g})$ belongs to the undetectable class $T^c$, then there exists $f_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; \mathbf{g})$ regardless of whether there is an undetectable attack or cooperative behavior.*

*Proof:* See Appendix B. ■

*Remark 3: Essentially Proposition 2 identifies that, if the action of the relay belongs to the undetectable class $T^c$, the distributions of the received symbols from the direct link and the relay link are subject to the same joint PDF $f^0_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; \mathbf{g})$. Therefore, ML detection has the same outcome irrespective of whether it arises from an undetectable attack or from cooperative behavior.*

In terms of the signal detection performance, an undetectable attack is no worse than cooperative behavior. This implies that, for the undetectable attack, the symbols received can be properly demodulated as if they resulted from cooperative behavior. That is, although the undetectable attack cannot be identified by Algorithm 1, a relay system with an undetectable attack can still deliver the same diversity order performance as a relay system with cooperative behavior. The symbol error rate (SER) for the undetectable attack in the high signal-to-noise ratio (SNR) region is approximated as [21]

$$\Pr(e) \overset{\text{high-SNR}}{\simeq} \frac{3}{K\gamma^2}, \quad (17)$$

where $K = \frac{|g_{sr}|(|g_{sd}|+|g_{rd}|)}{|g_{sd}||g_{rd}|}$, and $\gamma = E_s/\mathcal{N}_0$ is SNR without fading. It is observed that the diversity order of the undetectable attack is 2.

An undetectable attack involves the collusion between falsified forwarding and the dishonest feedback. This escapes detection because the damage caused by the falsified forwarding is mitigated by the dishonest feedback. This intuitively explains why, for an undetectable attack, the malicious relay can still be used to maintain the cooperative diversity.

## IV. Relay System Without a Direct Link

In this section we extend our consideration from relay systems with a direct link to those without a direct link between the S and the D due to coverage, as shown in Fig. 3.

While the detection at S is unaffected as the S-R-S link is still present, in the absence of a direct link as a 'clean' reference, the approach proposed in Section III-B cannot be

applied immediately. We must develop a new detection method at D that can be used for relay systems without a direct link.

We first repeat the two-phase transmission. Here, the notation is consistent with earlier sections.

In Phase 1, S sends $X^N$ to R (solid line in Fig. 3). The symbol received at R, $U$, is written as

$$U = h_{sr}X + W_{sr}. \tag{18}$$

R processes the $U^N$ received using AF protocol, generates $V^N$ and then forwards it in Phase 2 (dashed line in Fig. 3). The symbol received at D is expressed as

$$Y_2 = h_{rd}V + W_{rd}, \tag{19}$$

where for $i, j \in \{\mathrm{S}, \mathrm{R}, \mathrm{D}\}$, $i \neq j$, $h_{ij}$ is the channel gain between node $i$ and node $j$, and $W_{ij}$ is the Gaussian noise at node $j$ with variance $\mathcal{N}_0$. Definition 1 still applies to this relay system, while Definition 2 is changed according to the following cases.

### A. Known Source Distribution

If the source distribution is known, we can use a simple extension of the previous detection approach based on a direct link. The reliable CSI set is denoted as $\mathbf{h} = \{h_{sr}, h_{rs}, h_{rd}\}$ and the CSI set provided is denoted as $\mathbf{g} = \{g_{sr}, g_{rs}, g_{rd}\}$. Since the S-R-S link remains unchanged, $T_1$ can still be checked by the detection at S. However, the detection at D will be modified based on the known source distribution.

The likelihood function is given by

$$f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) = \sum_{x \in \mathcal{X}} f_{Y_2|X}(y_2|x; f_{V|U}, \mathbf{h}) \Pr(X), \tag{20}$$

where $f_{Y_2|X}(y_2|x; f_{V|U}, \mathbf{h})$ is given in (10), and the conditional PDF is expressed as

$$f^0_{Y_2}(y; \mathbf{g}) = \textstyle\sum_{x \in \mathcal{X}} f^0_{Y_2|X}(y|x; \mathbf{g}) \Pr(X), \tag{21}$$

where $f^0_{Y_2|X}(y|x; \mathbf{g})$ is given in (12).

According to (20) and (21), $T_2$ is redefined as

$$T_2 := \left\{ f_{V|U}, \{g_{sr}, g_{rd}\} \,\middle|\, f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) \neq f^0_{Y_2}(y_2; \mathbf{g}) \right\}.$$

By observing the received sequence $Y^N$, the empirical CDF at D is given by

$$F_{Y_2^N}(t) = \frac{1}{N} \sum_{i=1}^{N} I(y_2[i] < t). \tag{22}$$

From (20), (21) and (22), the decision metric $D_2^N$ in (15) is modified to

$$D_2^N = \tfrac{1}{n'^2} \textstyle\sum_{m=1}^{n'} \left| F_{Y_2^N}(t_m) - F^0_{Y_2}(t_m) \right|, \tag{23}$$

where $F^0_{Y_2}(t)$ is the CDF of $f^0_{Y_2}(t; \mathbf{g})$ given in (21). By employing this new $D_2^N$, together with (9), Algorithm 1 can deal with the detection of misbehavior for relay systems without direct links, based on a known source distribution.

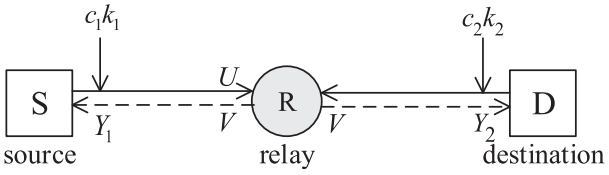



Fig. 4. A cooperative relay system with added artificial noise, where the solid and dashed lines denote Phase 1 and Phase 2, respectively.

### B. Unknown Source Distribution

If the source distribution is unknown to the relay system, the destination has no access to any 'clean' reference, since all physical layer observations tend to be manipulated. Adding auxiliary information can help to detect pilot contamination attack [19]. We employ additive AN to assist in building trustworthy reference information.

Traditionally, AN is designed to lie in the null space of the main channel [22], and it is exploited to degrade an eavesdropper's channel so that a secure capacity is guaranteed. In this paper, instead of using the traditional design of AN, we propose a different type of AN, as described below.

1) The structure of AN requires that the source is equipped with multiple antennas. Single-antenna nodes can emulate a distributed multi-antenna array. By executing a two-way communication protocol (see Fig. 4), the source and the destination simultaneously send information to the relay, thus the condition of forming AN can be satisfied.
2) The AN is defined as the product of coefficient matrix $\mathbf{C}$ and key vector $\mathbf{k}$. Then, the AN is denoted as $\mathbf{Ck}$, where $\mathbf{C} = \mathrm{diag}\{c_1,\ c_2\}$ and $\mathbf{k} = [k_1,\ k_2]^T$.
3) According to the two-way communication protocol, the AN lies in the null space of the provided CSI vector $\mathbf{g}_r = [g_{sr},\ g_{dr}]^T$ so that $\mathbf{g}_r^T\mathbf{Ck} = 0$.
4) For a given $\mathbf{C}$, when $\mathbf{g}_r$ is known and $\|\mathbf{k}\| = 1$, the AN is deterministic rather than random.
5) The AN changes with time, which takes place when the coefficient matrix $\mathbf{C}$ changes.
6) Conventionally, the wiretap channel is assumed to be uncorrelated with the main channel, which implies $\mathbf{h}_r^T\mathbf{Ck} \neq 0$. This assumption is invalid in the case considered, because $\mathbf{g}_r$ represents unreliable CSI that can be of any value. For example, the dishonest feedback can allow $\mathbf{g}_r$ to be correlated with $\mathbf{h}_r$, say, $\mathbf{g}_r = \alpha\mathbf{h}_r$ for $\alpha \neq 1$. Then, we have $\mathbf{h}_r^T\mathbf{Ck} = 0$ and AN will fail. Therefore, our analysis of the dishonest feedback covers two separate cases: $\mathbf{g}_r$ is either correlated or uncorrelated with $\mathbf{h}_r$.

In Phase 1, both S and D send AN $\mathbf{Ck}$ simultaneously. The signal received at R is expressed as

$$U = \mathbf{h}_r^T\mathbf{Ck} + W_r, \tag{24}$$

where $\mathbf{h}_r = [h_{sr},\ h_{dr}]^T$. $W_r$ is Gaussian noise at R with variance $\mathcal{N}_0$.

In Phase 2, R receives $U^N$ and then forwards a processed version, $V^N$, to S and D due to the broadcast nature of a wireless channel. The signals received at S and at D are

written as

$$Y_1 = h_{rs}V + W_{rs}, \tag{25a}$$

$$Y_2 = h_{rd}V + W_{rd}, \tag{25b}$$

where $h_{rs}$ and $h_{rd}$ are channel gains, and $W_{rs}$ and $W_{rd}$ are Gaussian noise with variance $\mathcal{N}_0$ at S and at R, respectively.

In the channel estimation process, R can know the CSI of both the S-R link and the D-R link, as S and D send pilot signals to R. Then, due to dishonest feedback, R broadcasts the potentially unreliable CSI, instead of the valid one, to S and D. When the unreliable CSI is obtained at S and D, the proposed AN-aided scheme comes into play.

Because of the symmetry of the system considered, we show the detection results from a source perspective, and the conditional likelihood function is given by

$$f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_U(u; h_{sr}) f_{V|U}(v|u) f_{Y_1|V}(y_1|v; h_{rs})\,\mathrm{d}u\mathrm{d}v, \tag{26}$$

where $\mathbf{h} = [h_{sr}, h_{dr}, h_{rs}, h_{rd}]$ is the reliable CSI set. The conditional PDF is formulated as

$$\begin{aligned} f^0_{Y_1}(y_1; \mathbf{g}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_U(u; g_{sr})\delta(v-u) f_{Y_1|V}(y_1|v; g_{rs})\,\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{+\infty} f_U(u; g_{sr}) f_{Y_1|V}(y_1|u; g_{rs})\,\mathrm{d}u, \end{aligned} \tag{27}$$

where $\mathbf{g} = [g_{sr}, g_{dr}, g_{rs}, g_{rd}]$ is the CSI set provided, with $f_{Y_2}(y_2; f_{V|U}, \mathbf{h})$ and $f^0_{Y_2}(y_2; \mathbf{g})$ being expressed in the similar way.

We discuss the four cases of the parameter pair $(f_{V|U}, \mathbf{g})$ as follows.

1) $f_{V|U} = \delta(v-u) \cap \mathbf{g} = \mathbf{h}$. It is easy to obtain $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.
2) $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} = \mathbf{h}$. First, we have $f_U(u; h_{sr}) = f_U(u; g_{sr})$ since AN is nulled out; then we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ because $f_{V|U} \neq \delta(v-u)$ and $f_{Y_1|V}(y_1|v; h_{rs}) = f_{Y_1|V}(y_1|v; g_{rs})$.
3) - $f_{V|U} = \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r \neq \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   According to (26) and (27), we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ as $\mathbf{h}_r^T\mathbf{C}\mathbf{k} \neq 0$.
   - $f_{V|U} = \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r = \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   It is observed that $\mathbf{C}\mathbf{k}$ lies in the null space of $\mathbf{h}_r$, so $\mathbf{h}_r^T\mathbf{C}\mathbf{k} = 0$; if $\mathbf{g} \neq \mathbf{h}$ but $g_{rs} = h_{rs}$, we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.
4) - $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r \neq \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   The two types of misbehavior have the potential to make $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$. By artificially operating $\mathbf{C}$, $\mathbf{h}_r^T\mathbf{C}\mathbf{k}$ changes over time and cannot be bounded by i.i.d. attack manner – $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$ may hold for some $\mathbf{C}$s with the specific pair $(f_{V|U}, \mathbf{g})$, but it does not hold when $\mathbf{C}$ changes. In general, we must have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ by using a time-varying coefficient matrix $\mathbf{C}$.
   - $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r = \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   The matrix $\mathbf{C}$ fails to change $\mathbf{h}_r^T\mathbf{C}\mathbf{k}$ as $\mathbf{C}\mathbf{k}$ lies in the null space of $\mathbf{h}_r$. It is possible to obtain $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$ with the specific pair $(f_{V|U}, \mathbf{g})$, which we will discuss later.

From the above discussion, if $\mathbf{g}_r \neq \alpha\mathbf{h}_r$ for $\alpha \neq 1$, a sufficient condition to determine misbehavior of the relay is that $(f_{V|U}, \mathbf{g})$ makes $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$. When $\mathbf{g}_r = \alpha\mathbf{h}_r$ for $\alpha \neq 1$, it is still possible that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$, because AN $\mathbf{C}\mathbf{k}$ fails to enable the distribution $Y_1$ to distinguish $f_{Y_1}(y_1; f_{V|U}, \mathbf{h})$ from $f^0_{Y_1}(y_1; \mathbf{g})$. To address this, we modify the AN $\mathbf{C}\mathbf{k}$ to $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$, where $\mathbf{g}_r^T\tilde{\mathbf{C}}\tilde{\mathbf{k}} \neq 0$. Therefore, for the second case of 3), the introduction of $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ means that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$. However, for the second case of 4), it is still possible that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.

As previously, we define

$$T_{AN1} := \left\{ f_{V|U}, \mathbf{g} \,\middle|\, f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g}) \right\},$$

and

$$T_{AN2} := \left\{ f_{V|U}, \mathbf{g} \,\middle|\, f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) \neq f^0_{Y_2}(y_2; \mathbf{g}) \right\}.$$

$T_{AN} = T_{AN1} \bigcup T_{AN2}$ is referred to as the *detectable class*, and its complement, $T^c_{AN}$, as the *undetectable class*.

1) To identify the detectable class $T_{AN}$, we need detection at both S and D. For $j = 1, 2$, based on the received sequences $Y_1^N$ and $Y_2^N$, the empirical CDFs at S and at D are given by

$$F_{Y_j^N}(t) = \frac{1}{N}\sum_{i=1}^{N} I(y_2[i] < t). \tag{28}$$

Similarly, for $j = 1, 2$, the decision metric $D_j^N$ is written as

$$D_j^N = \frac{1}{n'^2}\sum_{m=1}^{n'} \left| F_{Y_j^N}(t_m) - F^0_{Y_j}(t_m) \right|, \tag{29}$$

where $F^0_{Y_j}(t)$ is the CDF of $f^0_{Y_j}(t; \mathbf{g})$. The identification procedure of the detectable attack is elaborated in Algorithm 2.

2) We now focus on the undetectable class $T^c_{AN}$. From the expression of $f^0_{Y_2}(y_2; \mathbf{g})$, $Y_2$ is formulated as

$$Y_2 = g_{rd}(W_r + M\mathbf{g}_r^T\tilde{\mathbf{C}}\tilde{\mathbf{k}}) + W_{rd}, \tag{30}$$

where $M$ is the number of occurrences of $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ in an $N$-length block (usually taken to be $N/3$). Specifically, by setting $\tilde{\mathbf{C}} = diag\{1/\alpha M, 0\}$ and $\tilde{\mathbf{k}} = [X, 0]^T$ when $\mathbf{g}_r = \alpha\mathbf{h}_r$, (30) is rewritten as

$$Y_2 = g_{rd}(W_r + h_{sr}X) + W_{rd}, \tag{31}$$

According to the definition of $T_{AN2}$, we have $f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) = f^0_{Y_2}(y_2; \mathbf{g})$. Following the same logic as in Section III-C, the signal detection performance

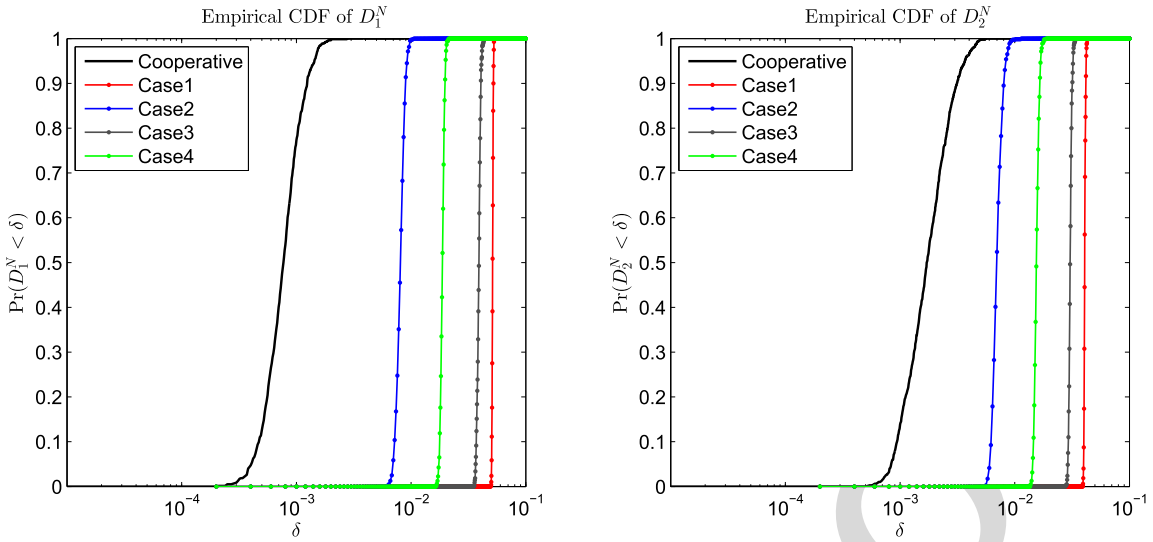



Fig. 5. The empirical CDFs of $D_1^N$ and $D_2^N$ for the four detectable attacks considered.

**Algorithm 2** The Identification Procedure for Attack Detection With Aided AN

1: Initialization: generate coefficient matrices $\mathbf{C}_1$, $\mathbf{C}_2$ and $\tilde{\mathbf{C}}$ and give the CSI set $\mathbf{g}$.
2: Calculate AN: compute $\mathbf{k}_1$, $\mathbf{k}_2$, and $\tilde{\mathbf{k}}$ based on $\mathbf{C}_1$, $\mathbf{C}_2$, and $\tilde{\mathbf{C}}$, respectively.
3: Add AN: take turns adding $\mathbf{C}_1\mathbf{k}_1$, $\mathbf{C}_2\mathbf{k}_2$ and $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ at S and D in each instant.
4: Calculate decision metric: $D_1^N$ and $D_2^N$ are computed at S and at D, respectively.
5: **if** $D_1^N \to 0 \cap D_2^N \to 0$ **then**
6: The relay action is a member of the undetectable class $T_{AN}^c$.
7: **else**
8: The relay action belongs to the detectable class $T_{AN}$ – the relay must be misbehaving.
9: **end if**

of the undetectable attack is the same as that of the cooperative scenario.

## V. Numerical Examples

### A. Relay Systems With a Direct Link

As illustration, we present here both detectable and undetectable attacks; we also evaluate the effectiveness of the proposed approach in identifying the two types of attack.

*1) Detectable Attack:* We consider a the relay system shown in Fig. 1, with S transmitting a BPSK signal with unit energy. Assume that the reliable CSI set $\mathbf{h} = [1, 1, 1]$, the AWGN variance $\mathcal{N}_0 = 0.01$, and the direct link channel gain $h_{sd} = 0.8$. The block length was selected to have $N = 1000$, and for quantization purposes $n' = 100$, $-a = b = \sqrt{n'}/2$, which implies that $\Delta = 1/\sqrt{n'}$.

To verify the effectiveness of the proposed detection schemes, the following four detectable malicious attacks were considered:

- CASE 1 - Dishonest Feedback: The relay provides an unreliable CSI with $\mathbf{g} = [0.6, 0.8, 0.7]$.
- CASE 2 - Falsified Forwarding $I$: The relay actively injects Gaussian noise distributed with $\mathcal{N}(0, 0.04)$.
- CASE 3 - Falsified Forwarding $II$: The relay intentionally adds noise with uniform distribution $\mathcal{U}(-1, +1)$.
- CASE 4 - Mixed Attack: Both dishonest feedback and falsified forwarding are considered in this case; the relay injects Gaussian noise distributed with $\mathcal{N}(0, 0.0025)$ and provides $\mathbf{g} = [0.9, 0.9, 1]$.

Fig. 5 shows the empirical CDFs of $D_1^N$ and $D_2^N$ after 800 computer simulation runs for each of the above cases. It can be observed that there is a clear separation between the undetectable class and the detectable class; this can be used as a threshold (e.g. $\delta = 0.005$ for the detection at S) for identifying the detectable class. These results further verify the effectiveness of Proposition 1.

*2) Undetectable Attack:* We assume that the reliable CSI $\mathbf{h} = [1, \sqrt{2}/2, \sqrt{2}/2]$ and the CSI provided $\mathbf{g} = [\sqrt{2}/2, 1, 1]$, and that the malicious relay performs falsified forwarding by injecting Gaussian noise distributed with $\mathcal{N}(0, 0.01)$. Fig. 6 shows the empirical CDFs of $D_1^N$ and $D_2^N$ for cooperative behavior and an undetectable attack. It is evident that the cooperative behavior and the undetectable attack are not distinguishable.

*3) BER Performance in the Presence of an Undetectable Attack:* We assume that the channel gain of the direct link $h_{sd} = 0.4$ and the injected noise power (falsified forwarding) is set at the same level as $\mathcal{N}_0$. Fig. 7 illustrates the BER performance versus SNR for different noise powers; the undetectable attack is seen to have the same BER performance as both cooperative behaviour and direct transmission from S to D. These results verify the previous claim that, even for undetectable attacks, the diversity gain is maintained.

### B. Systems Without a Direct Link

*1) Detectable Attack:* The source transmits BPSK signals and the reliable CSI is set as $\mathbf{h} = [1/2, 1/3, 1/2, 1/2]$.

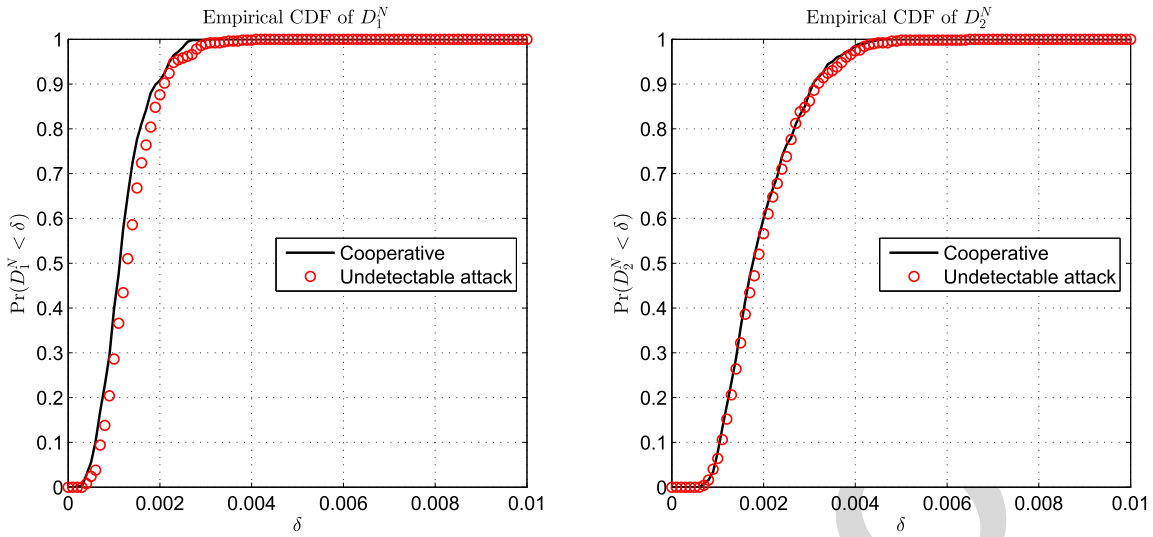



Fig. 6. The empirical CDFs of $D_1^N$ and $D_2^N$ for the undetectable attack considered.

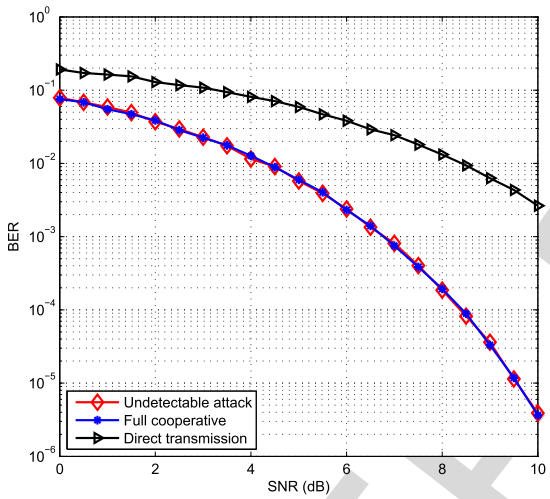



Fig. 7. BER performance comparisons among cooperative behavior, the undetectable class, and direct transmission.

The coefficient matrices are $\mathbf{C}_1 = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$, $\mathbf{C}_2 = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$ and $\tilde{\mathbf{C}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Correspondingly, $\mathbf{k}_1 = [2/\sqrt{5}, 1/\sqrt{5}]^T$, $\mathbf{k}_2 = [1/\sqrt{5}, 2/\sqrt{5}]^T$ and $\tilde{\mathbf{k}} = [1, 0]^T$. $\mathcal{N}_0 = 1/\sqrt{5}$. The block length is chosen to have $N = 1000$ and, for quantization purposes, $n' = 100$, $-a = b = \sqrt{n'}/2$, which implies that $\Delta = 1/\sqrt{n'}$. The three different cases are discussed below.

- CASE 1 - Dishonest Feedback: The relay provides the unreliable CSI $\mathbf{g} = [1/2, 1/2, 1/3, 1/3]$.
- CASE 2 - Malicious Forwarding $I$: The relay actively injects Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$.
- CASE 3 - Mixed Attack: We consider both dishonest feedback and falsified forwarding, where the relay injects Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$ and provides $\mathbf{g} = [1/3, 1/3, 1/2, 1/2]$.

Fig. 8 shows the empirical CDFs of $D_1^n$ after 800 computer simulation runs, in each of the three cases. The proposed decision metric is clearly capable of distinguishing between the detectable and undetectable classes.

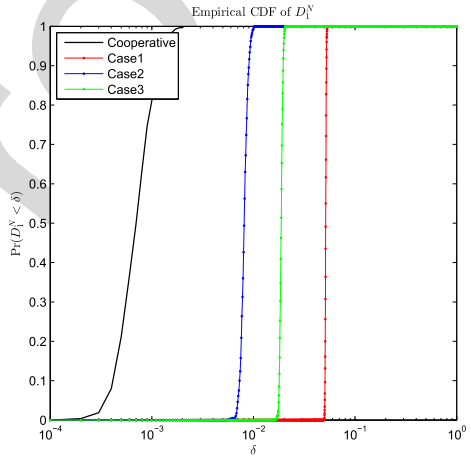



Fig. 8. The empirical CDFs of $D_1^N$ for the three detectable attacks considered.

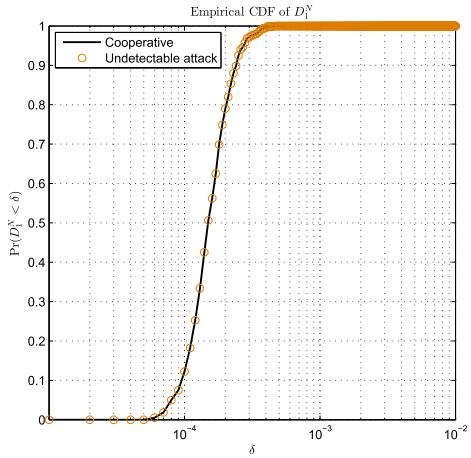



Fig. 9. The empirical CDFs of $D_1^N$ for the undetectable attack considered.

*2) Undetectable Attacks:* We consider the previously described simulation model with a different $\mathbf{g} = [1/4, 1/6, 1/4, 1/4]$, and $\mathbf{g}_r = \alpha \mathbf{h}_r$ for $\alpha = 1/2$. The malicious

relay conducts falsified forwarding by injecting Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$. Fig. 9 demonstrates that it is impossible to differentiate between cooperative behavior and an undetectable attack.

## VI. Conclusion

This paper has investigated the problem of detecting malicious attacks in a two-hop AF relay network in the presence of an unreliable CSI. In particular, we have proposed a detection approach applicable to a system with a direct link which is capable of clearly distinguishing between the detectable and undetectable classes. It has also been shown that, for the detectable class, the proposed approach detects malicious attacks with high probability. The relay system retains the benefits of diversity gain even in the presence of an undetectable attack. Further, we extended the proposed approach to a more common scenario in which no direct link is available.

## Appendix A
## Proof of Proposition 1

Without loss of generality, we firstly prove that the decision metric $D_1^N$ satisfies the two properties of Proposition 1. According to Borel's strong law of large numbers, for any arbitrary small positive $\mu$, we have

$$\lim_{N\to\infty} \Pr\left(\left|\frac{\sum_{i=1}^{N} I(y[i]<y)I(x[i]=\mathsf{x})}{\sum_{i=1}^{N} I(x[i]=\mathsf{x})} - \Pr(Y<y|X=\mathsf{x}; f_{V|U}, \mathbf{h})\right| \le \mu\right) = 1. \quad (32)$$

By defining a typical set as

$$\mathcal{A}_\mu\left(F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\right) \triangleq \left\{F \;\middle|\; \left|F - F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\right| \le \mu\right\},$$

where $F_{Y|X}(y|\mathsf{x}; \mathbf{\Psi})$ is the CDF of $f_{Y|X}(y|\mathsf{x}; \mathbf{\Psi})$, (32) can be modified as

$$\lim_{n\to\infty} \Pr\left\{F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\left(F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\right)\right\} = 1. \quad (33)$$

Under the assumption that $(f_{V|U}(v|u), \mathbf{g}) \in T$, we have $F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) \neq F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$, where $F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$ is the CDF of $f^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$.

For any sufficiently small positive $\delta$, we assume that

$$\left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta. \quad (34)$$

From (33), it follows that

$$\left|F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \in \left|\mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\}\right|,$$

which in turn implies that

$$\left|F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta - \mu. \quad (35)$$

Let us define $\rho_1 \triangleq \delta - \mu$ and assume that $\mu$ is chosen to be small enough to satisfy $\rho_1 > 0$. From the definition of $D_1^N$ in (9), (35) leads us to conclude that $D_1^N > \rho_1$.

Furthermore, according to (33) and (34), for any $\delta > 0$, we have

$$\begin{aligned}
&\Pr\left(D_1^N \ge \rho_1, (f_{V|U}(v|u), \mathbf{g}) \in T\right)\\
&= \Pr\left(D_1^N \ge \rho_1, \left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta\right)\\
&\ge \Pr\left(D_1^N \ge \rho_1, \left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta,\right.\\
&\quad \left. F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)\\
&\overset{(a)}{=} \Pr\left(\left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta,\right.\\
&\quad \left. F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)\\
&\ge \Pr\left(\left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta\right)\\
&\quad - \Pr\left(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right), \quad (36)
\end{aligned}$$

where (a) is derived by using (33), (34) and (35). From (36), we have

$$\begin{aligned}
&\Pr\left(D_1^N \ge \rho_1 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T\right)\\
&= \frac{\Pr\left(D_1^N \ge \rho_1, (f_{V|U}(v|u), \mathbf{g}) \in T\right)}{\Pr\left(\left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta\right)}\\
&\overset{(b)}{\ge} 1 - \frac{\Pr\left(F_{Y^N|X^N}(y|\mathsf{x})\mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)}{\Pr\left(\left|F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \ge \delta\right)}, \quad (37)
\end{aligned}$$

where (b) is derived by using (33).

As a result, $\lim_{n\to\infty} \Pr\left(D_1^N > \rho_1 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T\right) = 1$, which proves that $D_1^N$ satisfies the first property of Proposition 1.

We proceed now to prove that $D_1^N$ will satisfy the second property of Proposition 1. For this, assume that when $(f_{V|U}(v|u), \mathbf{g}) \in T^c$, we have $F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) = F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$. According to (33), it is also true that $F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\}$, which implies that

$$\left|F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| \in |\mathcal{A}_\mu\{0\}|, \quad (38)$$

and which yields

$$\left|F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\right| < \mu. \quad (39)$$

By defining $\rho_2 \triangleq \mu$, we have $D_1^N < \rho_2$, and thus

$$\Pr\left(D_1^N \ge \rho_2, (f_{V|U}(v|u), \mathbf{g}) \in T^c \;\middle|\; \left|F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right.\right) = 0. \quad (40)$$

where $\Pr\left(D_1^N \ge \rho_2 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T^c\right)$ is easily obtained and has been placed on top of the next page.

According to (33), this implies that $\Pr(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}) \to 0$. Finally, by means

$$\Pr\left(D_1^N \geq \rho_2 \left| (f_{V|U}(v|u), \mathbf{g}) \in T^c \right.\right)$$
$$= \frac{\Pr\left(D_1^N \geq \rho_2, (f_{V|U}(v|u), \mathbf{g}) \in T^c \left| F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right.\right)}{\Pr\left((f_{V|U}(v|u), \mathbf{g}) \in T^c\right)}$$
$$+ \frac{\Pr\left(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)}{\Pr\left((f_{V|U}(v|u), \mathbf{g}) \in T^c\right)} \quad (41)$$

of (40) and (41), as shown at the top of this page, we have $\lim_{N\to\infty} \Pr\left(D_1^N \geq \rho_2 \left| (f_{V|U}(v|u), \mathbf{g}) \in T^c \right.\right) = 0$, which proves that $D_1^N$ satisfies the second property of Proposition 1.

By a similar procedure, we can prove that $D_2^N$ also satisfies the two properties of Proposition 1, which then concludes the proof of Proposition 1. ■

## Appendix B
## Proof of Proposition 2

For the convenience of the proof, we introduce the following Lemma.

*Lemma 1: Let us consider a set of random variables $U_i$, $i = 1, 2, \cdots, 5$, in which $U_4 = U_2 + U_1$, $U_5 = U_3 + U_1$, and $U_1$ is independent of both $U_2$ and $U_3$. If there exists a PDF such that $f_{U_4|X}(u_4|\mathsf{x}) = f_{U_5|X}(u_5|\mathsf{x})$, then $f_{U_2|X}(u_2|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x})$ must hold.*

*Proof:* Since $U_4 = U_2 + U_1$, and $U_1$ and $U_2$ are independent of each other, we have

$$f_{U_4|X}(u_4|\mathsf{x}) = f_{U_2|X}(u_2|\mathsf{x}) + f_{U_1|X}(u_1|\mathsf{x}). \quad (42)$$

From (42), and by taking the characteristic function (CF) of $U_4$ conditioned on $X = \mathsf{x}$, we obtain

$$\varphi_{U_4|X}(t|\mathsf{x}) = \varphi_{U_2|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \quad (43)$$

where $\varphi_{U_2|X}(t|\mathsf{x})$ and $\varphi_{U_1|X}(t|\mathsf{x})$ are, respectively, the CFs of $U_2$ and $U_1$ conditioned on $X = \mathsf{x}$.

Similarly, since $U_5 = U_3 + U_1$ with $U_1$ and $U_3$ being independent with each other, we have

$$f_{U_5|X}(u_5|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x}) + f_{U_1|X}(u_1|\mathsf{x}). \quad (44)$$

Thus, the CF of $U_5$ conditioned on $X = \mathsf{x}$ can be expressed as

$$\varphi_{U_5|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \quad (45)$$

where $\varphi_{U_3|X}(t|\mathsf{x})$ is the CF of $U_3$ conditioned on $X = \mathsf{x}$, respectively.

Since $f_{U_4|X}(u_4|\mathsf{x}) = f_{U_5|X}(u_5|\mathsf{x})$, we have

$$\varphi_{U_4|X}(t|\mathsf{x}) = \varphi_{U_5|X}(t|\mathsf{x}). \quad (46)$$

Using (43), (45) and (46), we obtain

$$\varphi_{U_2|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \quad (47)$$

and as $\varphi_{U_1|X}(t|\mathsf{x})$ is non-zero, we have

$$\varphi_{U_2|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x}). \quad (48)$$

Since any PDF can be uniquely determined by its CF, (48) implies that

$$f_{U_2|X}(u_2|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x}). \quad (49)$$

■

We now return to the proof of Proposition 2. Since the detectable class $T = T_1 \cup T_2$, we have $T^c = T_1^c \bigcap T_2^c$. For the set $T_1^c$, $f_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})$ is identical to $f^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$. Following (1), (2) and (5), we have

$$f_{h_{rs}V+N_{rs}|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{g_{rs}(g_{sr}X+N_{sr})+N_{rs}|X}(t|\mathsf{x}; \mathbf{g}). \quad (50)$$

According to Lemma 1, it is easy to obtain that

$$f_{h_{rs}V|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{g_{rs}(g_{sr}X+N_{sr})|X}(t|\mathsf{x}; \mathbf{g}), \quad (51)$$

and if we note that $f^0_{g_{rs}(g_{sr}X+N_{sr})|X}(t|\mathsf{x}; \mathbf{g}) = \frac{1}{\pi\sigma_{sr}^2 g_{rs}^2} \exp(-\frac{\|t - g_{sr}g_{rs}\mathsf{x}\|^2}{\sigma_{sr}^2 g_{rs}^2})$, then we have

$$f_{V|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = \frac{h_{rs}^2}{\pi\sigma_{sr}^2 g_{rs}^2} \exp(-\frac{\|h_{rs}t - g_{sr}g_{rs}\mathsf{x}\|^2}{\sigma_{sr}^2 g_{rs}^2}). \quad (52)$$

Following (2b), (52) can be re-expressed as

$$f_{Y_2|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = \frac{1}{\pi(K^2\sigma_{sr}^2 + \sigma_{rd}^2)} \exp(-\frac{\|t - g_{sr}K\mathsf{x}\|^2}{(K^2\sigma_{sr}^2 + \sigma_{rd}^2)}), \quad (53)$$

where $K = g_{rs}h_{rd}/h_{rs}$ is unknown. According to (12), we have

$$f^0_{Y_2|X}(t|\mathsf{x}; \mathbf{g}) = \frac{1}{\pi\sigma_2^2} \exp(-\frac{\|t - g_{sr}g_{rd}\mathsf{x}\|^2}{\sigma_2^2}), \quad (54)$$

where $\sigma_2^2 = g_{rd}^2\sigma_{sr}^2 + \sigma_{rd}^2$.

Let us now consider $T_2^c$. For any $y_1$ and $y_2$, we obtain that

$$f_{Y_2|Y_1}(y_2|y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_2|Y_1}(y_2|y_1; \mathbf{g}). \quad (55)$$

Furthermore, since $(Y_1, X, Y_2)$ forms a Markov chain as $Y_1 \to X \to Y_2$, we have

$$\sum_{\mathsf{x}\in\mathcal{X}} \Pr(X = \mathsf{x}|Y_1 = y_1) f_{Y_2|X}(Y_2 = y_2|X = \mathsf{x}; f_{V|U}, \mathbf{h})$$
$$= \sum_{\mathsf{x}\in\mathcal{X}} \Pr(X = \mathsf{x}|Y_1 = y_1) f^0_{Y_2|X}(Y_2 = y_2|X = \mathsf{x}; \mathbf{g}). \quad (56)$$

Note that $\Pr(X=\mathsf{x}|Y_1=y_1)$ in (56) can be written as

$$\begin{aligned}&\Pr(X=\mathsf{x}|Y_1=y_1)\\&=\frac{\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}{\Pr(Y_1=y_1)}\\&=\frac{\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}{\sum_{\mathsf{x}\in\mathcal{X}}\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}\\&=\frac{1}{1+\sum_{X\neq\mathsf{x}}\exp(y_1h_{sd}(x-\mathsf{x})/\sigma_{sd}^2)}.\end{aligned}\tag{57}$$

Without loss of generality, we consider $X\in(-1,+1)$. If $\mathsf{x}=+1$, it is easy to show that $\Pr(X=\mathsf{x}|Y_1=y_1)$ becomes very small when $y_1$ is far less than 0. When $y_1\to-\infty$, we have $\lim_{y_1\to-\infty}\Pr(X=\mathsf{x}|Y_1=y_1)=0$ and $\lim_{y_1\to-\infty}\Pr(X\neq\mathsf{x}|Y_1=y_1)=1$. Therefore, (56) can be reduced to

$$\begin{aligned}&f_{Y_2|X}(Y_2=y_2|X=\mathsf{x};f_{V|U},\mathbf{h})\\&\qquad=f^0_{Y_2|X}(Y_2=y_2|X=\mathsf{x};\mathbf{g}).\end{aligned}\tag{58}$$

Substituting (53) and (54) into (58), we can obtain $K=g_{rd}$, which means that $f_{Y_2|X}(y_2|\mathsf{x};f_{V|U},\mathbf{h})$ can be expressed only by the known unreliable CSI.

In addition, since the direct link S-D and the relay link S-R-D are independent of each other, we have

$$\begin{aligned}&f_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};f_{V|U},\mathbf{h}))\\&=f_{Y_1|X}(y_1|\mathsf{x})f_{Y_2|X}(y_2|\mathsf{x};f_{V|U},\mathbf{h})\\&=\frac{1}{\pi(K^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}K\mathsf{x}\|^2}{K^2\sigma_{sr}^2+\sigma_{rd}^2}\right)\\&=\frac{1}{\pi(g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}g_{rd}\mathsf{x}\|^2}{g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2}\right).\end{aligned}\tag{59}$$

On the other hand, according to (54), we have

$$\begin{aligned}&f^0_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};\mathbf{g})\\&=f_{Y_1|X}(y_1|\mathsf{x})f^0_{Y_2|X}(y_2|\mathsf{x};\mathbf{g})\\&=\frac{1}{\pi(g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}g_{rd}\mathsf{x}\|^2}{g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2}\right).\end{aligned}\tag{60}$$

From (59) and (60), we see that $f_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};f_{V|U},\mathbf{h})=f^0_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};\mathbf{g})$, which completes the proof of Proposition 2. ■

## REFERENCES


[1] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. Di Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Trans. Commn.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[3] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[4] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler, "Fading relay channels: Performance limits and space-time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1099–1109, Aug. 2004.

[5] S. Dehnie, H. T. Sencar, and N. Memon, "Cooperative diversity in the presence of a misbehaving relay: Performance analysis," in *Proc. IEEE Sarnoff Symp.*, Princeton, NJ, USA, May 2007, pp. 1–7.

[6] P. Papadimitratos and Z. J. Haas, "Secure data communication in mobile ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 2, pp. 343–356, Feb. 2006.

[7] S. Capkun, L. Buttyan, and J. P. Hubaux, "Self-organized public-key management for mobile ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 52–64, Jan. 2003.

[8] Y.-C. Hu, A. Perrig, and D. B. Johnson, "Ariadne: A secure on-demand routing protocol for ad hoc networks," *Wireless Netw.*, vol. 11, nos. 1–2, pp. 21–38, Jan. 2005.

[9] Y. Mao and M. Wu, "Tracing malicious relays in cooperative wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 2, pp. 198–212, Jun. 2007.

[10] L.-C. Lo and W.-J. Huang, "Misbehavior detection without channel information in cooperative networks," in *Proc. IEEE 74th Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.

[11] L.-C. Lo, Z.-J. Wang, and W.-J. Huang, "Noncoherent misbehavior detection in space-time coded cooperative networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 3061–3064.

[12] L.-C. Lo, W.-J. Huang, R. Y. Chang, and W.-H. Chung, "Noncoherent detection of misbehaving relays in decode-and-forward cooperative networks," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1536–1539, Sep. 2015.

[13] W. Hou, X. Wang, and A. Refaey, "Misbehavior detection in amplify-and-forward cooperative OFDM systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 5345–5349.

[14] S. W. Kim, "Physical integrity check in cooperative relay communications," *IEEE Trans. Wireless commun.*, vol. 14, no. 11, pp. 6401–6413, Nov. 2015.

[15] S. Dehnie, H. T. Sencar, and N. Memon, "Detecting malicious behavior in cooperative diversity," in *Proc. 41st IEEE Annu. Conf. Inf. Sci. Syst.*, Baltimore, MD, USA, Mar. 2007, pp. 895–899.

[16] E. Graves and T. F. Wong, "Detection of channel degradation attack by intermediary node in linear networks," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 747–755.

[17] R. Cao, E. Graves, T. F. Wong, and T. Lv, "Detecting substitution attacks against non-colluding relays," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 1856–1861.

[18] R. Cao, S. Huang, and Y. Lu, "Detecting and tracing i.i.d. Attacks in networks with any number of relays," *IEEE Access*, vol. 4, pp. 6757–6765, Oct. 2016.

[19] J. K. Tugnait, "Self-contamination for detection of pilot contamination attack in multiple antenna systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 5, pp. 525–528, Oct. 2015.

[20] K. P. Peppas, G. C. Alexandropoulos, and P. T. Mathiopoulos, "Performance analysis of dual-hop AF relaying systems over mixed $\eta-\mu$ and $\kappa-\mu$ fading channels," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3149–3163, Sep. 2013.

[21] J. N. Laneman and G. W. Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Chicago, IL, USA, Sep. 2000, pp. 7–12.

[22] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.

[23] R. Pabst *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Wireless Commun. Mag.*, vol. 42, no. 9, pp. 80–89, Sep. 2004.

[24] B. Zafar, S. Gherekhloo, and M. Haardt, "Analysis of multihop relaying networks: Communication between range-limited and cooperative nodes," *IEEE Veh. Technol. Mag.*, vol. 7, no. 3, pp. 40–47, Sep. 2012.

[25] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 146–153, Apr. 2014.

**Tiejun Lv** (M'08–SM'12) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively. From 2001 to 2002, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, China. In 2005, he was promoted to a Full Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. From 2008 to 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He has authored over 50 published IEEE journal papers and 170 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory, and networking. He was a recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He received the Nature Science Award from the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015.

**Yajun Yin** received the B.Eng. degree in electronic and information engineering from the Harbin Institute of Technology at Weihai, China, in 2015. He is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include physical-layer security and Web anti-spam.

**Yueming Lu** received the B.S. and M.S. degrees in computer science from the Xi'an University of Architecture and Technology in 1994 and 1997, respectively, and the Ph.D. degree in computer architecture from Xi'an Jiaotong University in 2000. He was a Researcher of Lucent from 2000 to 2003. He is currently a Professor with the Beijing University of Posts and Telecommunications. His research interests include network design, network security, and distributed computing.

**Shaoshi Yang** (S'09–M'13) received the B.Eng. degree in information engineering from the Beijing University of Posts and Telecommunications, China, in 2006, and the Ph.D. degree in electronics and electrical engineering from the University of Southampton, U.K., in 2013. From 2008 to 2009, he was an Intern Research Fellow of Intel Labs China, where he was involved in the mobile WiMAX standardization. From 2013 to 2016, he was a Research Fellow with the School of Electronics and Computer Science, University of Southampton. He is currently a Principal Engineer with Huawei Technologies Co., Ltd., China. He is also a member of the Isaac Newton Institute for Mathematical Sciences, Cambridge University. His research interests include high-dimensional signal processing for communications, green radio, wireless video transmission, cross-layer system design, mathematical optimization and its applications. He was recognized by the prestigious National 1000-Young-Talent Fellowship of China and the Dean's Award for Early Career Research Excellence at the University of Southampton. He was a Guest Associate Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (http://shaoshiyang.weebly.com/).

**Enjie Liu** received the Ph.D. degree in telecommunications from the Queen Mary University of London, London, U.K., in 2002. She is currently a Reader in network applications with the School of Computer Science and Technology, University of Bedfordshire, U.K. Her current research interests include power management at protocol stack for mobile terminals, MAC and network layer protocol design, vehicular ad hoc networks, small-cell propagation modeling, and Internet of Things protocols and applications.

**Gordon Clapworthy** received the B.Sc. degree (Hons.) in mathematics from the University of London, the M.Sc. degree (Hons.) in computer science from the City, University of London, and the Ph.D. degree in aeronautical engineering from the University of London. He is currently a Professor of computer graphics with the School of Computer Science and Technology, University of Bedfordshire, U.K. He has published over 300 refereed papers and has been a Principal Investigator in 28 European projects, being a Project Coordinator in eight of them. His previous interests have included computer animation, biomechanics, space robots, transonic aerodynamics, virtual reality, surface modeling, and fundamental computer graphics algorithms, but his most recent work has focused on the technological aspects of next-generation medical systems.

# Physical Detection of Misbehavior in Relay Systems With Unreliable Channel State Information

Tiejun Lv, *Senior Member, IEEE*, Yajun Yin, Yueming Lu, Shaoshi Yang, *Member, IEEE*, Enjie Liu, and Gordon Clapworthy

***Abstract*— We study the detection of misbehavior in a Gaussian relay system, where the source transmits information to the destination with the assistance of an amplify-and-forward relay node subject to unreliable channel state information (CSI). The relay node may be potentially malicious and corrupt the network by forwarding garbled information. In this situation, misleading feedback may take place, since reliable CSI is unavailable at the source and/or the destination. By classifying the action of the relay as detectable or undetectable, we propose a novel approach that is capable of coping with any malicious attack detected and continuing to work effectively in the presence of unreliable CSI. We demonstrate that the detectable class of attacks can be successfully detected with a high probability. Meanwhile, the undetectable class of attacks does not affect the performance improvements that are achievable by cooperative diversity, even though such an attack may fool the proposed detection approach. We also extend the method to deal with the case in which there is no direct link between the source and the destination. The effectiveness of the proposed approach has been validated by numerical results.**

***Index Terms*— Physical layer security, integrity check, unreliable CSI, cooperative relay communications.**

## I. Introduction

PHYSICAL layer security (PLS) is a promising technology that provides secure wireless transmissions by smartly exploiting imperfections of the communications medium [1]. Cooperative relaying is beneficial for improving the coverage and transmission reliability of wireless systems [2], where single-antenna devices can form a virtual antenna array to provide cooperative spatial diversity [3], [4]. However, such benefits are attained only when the relays are trustworthy and always comply with cooperative protocols. In an adversarial case, some relays might maliciously alter the information sent by the source, thus degrading the performance of the relaying system significantly. The dependence of cooperative systems on the relays represents an inherent vulnerability [5]. Therefore, early detection of misbehavior is essential to maintaining the security of relaying systems and to combating malicious attacks.

Traditionally, detection methods are based on cryptography keys or authentication keys, requiring the source and the destination to share a secret key [6]–[8]. The key-based detection approach is far from ideal as it imposes a high computational cost and needs a key distribution mechanism. Alternatively, it is possible to detect malicious relays from the physical layer perspective. In particular, Mao and Wu [9] proposed a cross-layer detecting scheme, where pseudo-random tracing symbols were inserted into information bits. To identify the malicious relays, the destination measures the error probability of the observed tracing symbols, according to their *a priori* ground truth. In [10]–[12], Lo *et al.* applied a tracing-based method to non-coherent detection in various scenarios, requiring no channel state information (CSI). Note that the transmission of tracing symbols also requires the support from a key-distribution mechanism. Moreover, the performance of tracing-based schemes is highly dependent on the number of tracing symbols used, and an excessive number of them can significantly reduce the bandwidth efficiency.

To avoid the use of external assistance, many detecting schemes exploit 'clean' references stemming from the relaying system itself. A 'clean' reference contains information that has not been manipulated by the relay for sure. For example, in the orthogonal frequency-division multiplexing (OFDM) based detection scheme of [13], the source regards the transmitted information as a reference. Thus, the misbehavior of the relay is detected by examining the correlation between the reference and the information that is forwarded by the relay but overheard at the source. Detection schemes can also be implemented at the destination [14], [15]. The direct link between the source and the destination, as a 'clean' reference to the relay link, is used to compare between two different links to determine the relay behavior. However, these schemes [9]–[15] assume that each malicious relay behaves in an independent identically distributed (i.i.d.) manner of a specific form. With respect to arbitrary i.i.d. attacks, Graves and Wong [16] and Cao *et al.* [17] proposed a novel detection approach in which the relay behavior is modeled as an attack channel to check for any misbehavior. In [16], the source

Manuscript received September 15, 2017; revised February 1, 2018; accepted February 16, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61671072 and in part by the National Key Research and Development Program of China under Grant 2016YFB0800302. *(Corresponding author: Tiejun Lv.)*

T. Lv and Y. Yin are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: lvtiejun@bupt.edu.cn; yinyajun@bupt.edu.cn).

Y. Lu is with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: ymlu@bupt.edu.cn).

S. Yang is with the Department of Wireless Network Research, Huawei Technologies Co., Ltd., Shenzhen 518129, China (e-mail: shaoshi.yang@ieee.org).

E. Liu and G. Clapworthy are with the School of Computer Science and Technology, University of Bedfordshire, Luton LU1 3JU, U.K. (e-mail: enjie.liu@beds.ac.uk; gordon.clapworthy@beds.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2824728

extracts the estimation of an attack channel based on its own transmitted and observed information. This detection method has been extended to the scenario in which a direct 'clean' reference is unavailable [17]. However, [17] entirely depends on the source distribution. In [18], a detecting and tracing scheme for a multi-relay network was studied by partitioning the network into several sub-networks as described in [17].

The detection schemes [9], [13]–[18] above are enabled under an ideal assumption that reliable CSI is known in advance. However, reliable CSI may not be available in practice, especially when relays are malicious. For instance, malicious relays are reluctant to cooperate initially and, hence, they may deliberately manipulate the channel estimation process with ease. The whole system is then deceived into a state of unreliable CSI. In such cases, the previously mentioned schemes [9], [13]–[18] may be severely compromised. Considering a point-to-point system, Tugnait [19] proposed a scheme to detect the pilot contamination attack, which causes unreliable CSI, by superimposing a random sequence on the training sequence and using source enumeration methods.

In this paper, we consider a cooperative relaying system with a source-destination pair and a single relay employing an amplify and forward (AF) strategy [20]. The potentially malicious relay is capable of forwarding false information in an arbitrary i.i.d. manner. It can also provide unreliable CSI to degrade the system's performance. Falsified forwarding together with the unreliable CSI makes the detection of misbehavior very difficult. Our goal is to detect misbehavior based on physical-layer observations. The key difference between existing work [16]–[18] and ours is that we take into account that the channel estimation process may be compromised and hence the available CSI is unreliable. The main contribution of this paper is summarized as follows.

1) We study the misbehavior of the malicious relay under the assumption that the misbehavior arises not only from falsified forwarding, but also from dishonest feedback. According to different combinations of misbehavior and from the detection point of the view, we define two mutually exclusive attack types – *detectable* and *undetectable*. We prove that a detectable attack can be detected asymptotically by examining the distance measure between the distribution of physical-layer observations and the distribution of the calculated received symbols. The proposed detection scheme needs no extra secret keys.
2) We prove that an undetectable attack does not affect the bit error rate (BER) performance that is achievable by cooperative diversity, even though it cannot be identified. This implies that an undetectable attack hardly influences the reliability performance of the relay network, in the sense that the benefits of diversity gain are retained.
3) For relay systems having direct links, we choose the direct link as a 'clean' reference. We then extend the proposed detection scheme to relay systems having no direct link, where the source distribution is known. Furthermore, in the absence of prior information of the source, we design a 'clean' reference by introducing artificial noise (AN) to aid the proposed detection scheme.

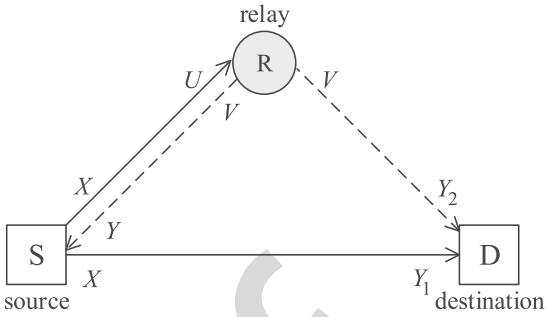



Fig. 1. A cooperative relay system consisting of a source-destination pair and a potential malicious relay with direct link.

The remainder of this paper is organized as follows. The system model and the misbehavior types are described in Section II. In Section III, we elaborate on the proposed detection scheme for detectable attacks and prove that an undetectable attack can achieve the same BER as a detectable attack. The detection scheme is extended to the scenario in which a direct link is absent in Section IV. Section V provides numerical examples and discussions, and conclusions are drawn in Section VI.

*Notation:* Upper and lower case letters denote, respectively, random variables and their realizations. Sans-serif letters denote general elements. $|\cdot|$ represents an absolute value and $\|\cdot\|$ represents the Euclidean norm. The transpose of the vector $a$ is $a^T$. For a sequence $x^N$, both $x[i]$ and $x_i$ denote the $i$-th element in $x^N$. $\mathcal{X}$ denotes the alphabet of $X$. $I(x[i]=\mathsf{x})$ is the indicator function denoting whether or not $x[i]$ is $\mathsf{x}$. $F_{X^N}(\mathsf{x}) = 1/N\sum_{i=1}^{N} I(x[i]=\mathsf{x})$ is used to denote the empirical distribution of $x^N$, and implies the relative proportion of occurrence of symbol $\mathsf{x}$ in $x^N$. For a sequence $y^N$ with consecutive values, the empirical distribution function is trivially defined as $F_{Y^N}(t) = 1/N\sum_{i=1}^{N} I(y[i] < t)$. In a similar fashion, we denote the conditional empirical distribution as

$$F_{Y^N|X^N}(t|\mathsf{x}) = \frac{\sum_{i=1}^{N} I(y[i]<t)I(x[i]=\mathsf{x})}{\sum_{i=1}^{N} I(x[i]=\mathsf{x})}.$$

## II. System Model

### A. Cooperative Transmission

We consider a typical cooperative relay network consisting of a source-destination pair and a potential malicious relay as shown in Fig. 1, where the source (S) tries to send information to the destination (D) with the aid of a relay node (R) and a direct link (S-D link). A relay system without a direct link will be considered in Section IV. Although this three-node relay network model is simple, it is fundamental for studying relay aided cooperative communications. Compared with traditional non-cooperative networks, three-node relay networks can offer several benefits, such as better connectivity, higher throughput

and greater reliability [23]–[25]. The three-node relay network model can also be extended to more complicated network topology.

In Fig. 1, the solid and dashed lines represent two transmission phases, i.e. phases 1 and 2, respectively. The wireless channels are assumed to be quasi-static in the same phase.

*1) Phase 1:* S first broadcasts an $N$-length i.i.d. sequence $X^N$ simultaneously to R and D. Let $U$ and $Y_1$ be the symbols received at R and D, respectively. In the symbol-by-symbol expression, the time index is omitted. The received symbols in Phase 1 can be expressed as

$$U = h_{sr}X + W_{sr}, \tag{1a}$$

$$Y_1 = h_{sd}X + W_{sd}. \tag{1b}$$

*2) Phase 2:* R receives $U^N$, processes it, and then forwards $V^N$ to D. Here, the symbol $V$ is a processed version of the received symbol $U$. Due to the broadcast nature of wireless communication, S can overhear the forwarded information $V^N$ at the same moment. Let $Y$ denote the received symbol overheard by S and $Y_2$ denote the received symbol at D. The received symbols in Phase 2 are given by

$$Y = h_{rs}V + W_{rs}, \tag{2a}$$

$$Y_2 = h_{rd}V + W_{rd}, \tag{2b}$$

where $h_{ij}$ is channel gain between node $i$ and node $j$ with $i, j \in \{\mathrm{S}, \mathrm{R}, \mathrm{D}\}$ and $i \neq j$. Statistically, we can model them as complex Gaussian random variables which capture the effects of pass loss and statistical fading in a wireless channel. The average transmit energy of the transmitted symbol is denoted as $E_s$. $W_{ij}$ represents additive white Gaussian noise (AWGN) with variance $\mathcal{N}_0$ received at node $j$.

CSI needs to be obtained from channel estimation. Before the transmission phases, all nodes participate in the channel estimation process. Since the malicious relay can manipulate the channel estimation process by sending incorrect pilot signals, unreliable CSI $g_{ij}$ may be provided, which is different from the reliable CSI $h_{ij}$. Let $\mathbf{g} = \{g_{sr}, g_{rs}, g_{rd}\}$ and $\mathbf{h} = \{h_{sr}, h_{rs}, h_{rd}\}$ denote the set of the potentially unreliable CSI provided and the set of the corresponding reliable CSI, respectively. Note that the channel gain of the direct link cannot be manipulated by the relay, hence $h_{sd}$ is omitted from both of the CSI sets.

### B. Misbehavior Types

The introduction of the relay opens a door to malicious attacks. Instead of complying with the cooperative strategy, a malicious relay node may exhibit misbehaviors both in the transmission phases and in the channel estimation process. Hence, potentially both the information forwarded and the CSI provided can be manipulated by the malicious relay. We identify the following two types of misbehaviors.

1) *Falsified Forwarding*: the relay receives $U^N$ in Phase 1, and then corrupts it into another sequence $V^N$ to be forwarded in Phase 2. If we assume that the malicious relay misbehaves in an arbitrary i.i.d. manner, the forwarded sequence $V^N$ will obey an arbitrary stochastic distribution conditioned on $U^N$. From the perspective of symbol-by-symbol, the relay processing behavior can be characterized by its conditional probability density function (PDF) $f_{V|U}(v|u)$. It is not difficult to derive that if the relay forwards the received symbol $U$ accurately, the conditional PDF is

$$f_{V|U}(v|u) = \delta(v - u), \tag{3}$$

where $\delta(\cdot)$ is the impulse function. This means that when $U = V$ the relay is amicable with respect to forwarding information. Otherwise, the relay is exhibiting *falsified forwarding*, also known as a *Byzantine attack*.

2) *Dishonest Feedback*: In many wireless communication protocols, the transmitter obtains the CSI estimate from the receiver's feedback. The malicious node is capable of dominating the channel estimation process deliberately. In this case the CSI provided may be unreliable. The unreliable CSI provides a malicious node with an opportunity to undermine relay selection, e.g., to select a malicious node as a qualified relay. Further, the destination node may combine the information received from the relay and the source inappropriately, due to the unreliable CSI. The CSI provided is said to be *reliable* if $\mathbf{g} = \mathbf{h}$. Otherwise, the relay node is considered to be initiating dishonest feedback that creates *unreliable* CSI. Note that imperfect CSI is usually caused by channel estimation error, which is an objective measurement error rather than a deliberate attack. Imperfect CSI does not belong to the scope of physical layer security. Thus, imperfect CSI is not considered in this paper.

Thus we can employ the parameter pair $(f_{V|U}, \mathbf{g})$ to describe the behavior of the relay. Maliciousness due to the misbehavior is defined as follows.

*Definition 1 (Maliciousness of Misbehavior): The relay is considered as cooperative if and only if the pair $(f_{V|U}, \mathbf{g})$ belongs to the set $\{f_{V|U}(v|u), \mathbf{g} | f_{V|U}(v|u) = \delta(v - u), \mathbf{g} = \mathbf{h}.\}$; otherwise, the relay is considered as malicious.*

It is obvious that neither of the above forms of misbehavior is allowed for a cooperative relay. Our goal is to use physical-layer observations to detect maliciousness if and when misbehavior occurs in the relay system.

## III. Detection Approach

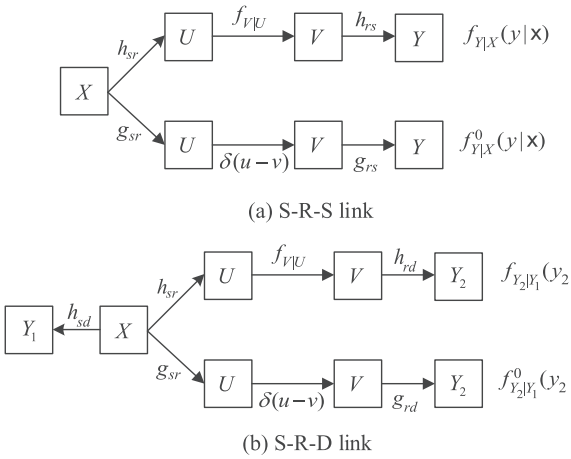
In this section, we describe the proposed approach for detecting maliciousness in a relay system with a direct link, i.e., falsified forwarding and/or dishonest feedback, but first we introduce the concept of detectability of maliciousness.

### A. Maliciousness Detectability

The source S can observe the symbol $Y$ in Phase 2 (see (2)). The symbol $Y$ goes through a real S-R-S link, which may be manipulated by a malicious relay. For S, the transmitted symbol $X$ offers a 'clean' reference.

On one hand, we use the conditional likelihood function

$$f_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; h_{sr}) \times f_{V|U}(v|u) f_{Y|V}(y|v; h_{rs})\,\mathrm{d}u\mathrm{d}v \tag{4}$$




(a) S-R-S link

(b) S-R-D link

Fig. 2. Markov chain of S-R-S link and S-R-D link.

to characterize S-R-S link, where the parameters $f_{V|U}$ and $\mathbf{h}$ are unknown for S.

On the other hand, S also tries to make use of the CSI provided, $\mathbf{g}$, even though it may be unreliable. The conditional PDF at S is computed as

$$\begin{aligned} & f^0_{Y|X}(y|\mathsf{x};\mathbf{g}) \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x};g_{sr})\delta(v-u)f_{Y|V}(y|v;g_{rs})\,\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x};g_{sr})f_{Y|V}(y|u;g_{rs})\,\mathrm{d}u, \end{aligned} \tag{5}$$

where the superscript distinguishes the conditional PDF $f^0_{Y|X}(y|\mathsf{x};\mathbf{g})$ from the conditional likelihood function $f_{Y|X}(y|\mathsf{x};f_{V|U},\mathbf{h})$. Whenever there is no ambiguity, we will employ such a notation, i.e. $f^0_{Y|X}(y|\mathsf{x})$ and $f_{Y|X}(y|\mathsf{x})$, for simplicity. It is observed that the relay is considered to faithfully forward as $f_{V|U}(v|u)=\delta(v-u)$ appears in the expression for $f^0_{Y|X}(y|\mathsf{x})$.

Since $(X,U,V,Y)$ forms a Markov chain as $X\to U\to V\to Y$, we have four cases according to different combinations of the parameter pair $(f_{V|U},\mathbf{g})$, as follows:

1) $f_{V|U}=\delta(v-u)\cap\mathbf{g}=\mathbf{h}$; full cooperative relay (no misbehavior), we have $f_{Y|X}(y|\mathsf{x})=f^0_{Y|X}(y|\mathsf{x})$.
2) $f_{V|U}\neq\delta(v-u)\cap\mathbf{g}=\mathbf{h}$; malicious relay with falsified forwarding, we have $f_{Y|X}(y|\mathsf{x})\neq f^0_{Y|X}(y|\mathsf{x})$.
3) $f_{V|U}=\delta(v-u)\cap\mathbf{g}\neq\mathbf{h}$; malicious relay with dishonest feedback, we have $f_{Y|X}(y|\mathsf{x})\neq f^0_{Y|X}(y|\mathsf{x})$.
4) $f_{V|U}\neq\delta(v-u)\cap\mathbf{g}\neq\mathbf{h}$; malicious relay with both misbehaviors. This is difficult to analyse as it is hard to determine the equivalence of $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$.

As shown in Fig. 2 (a), it is easy to check the relationship between $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$ in the four different cases. The first three are easily determined, but Case 4) is a demanding problem. From the above, based on the parameter pair $(f_{V|U},\mathbf{g})$, the inequality of $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$ is a sufficient condition to determine misbehavior.

TABLE I
THE RELATIONSHIP BETWEEN DEFINITION 1 AND DEFINITION 2

| | Detectable Class $T$ | Undetectable Class $T^c$ |
|---|---|---|
| Malicious Relay | Detectable attack | Undetectable attack |
| Cooperative Relay | $\varnothing$ | No misbehavior |

This conclusion helps to detect misbehavior in the relaying system considered. We define a set $T_1$ as:

$$T_1 := \Big\{ f_{V|U}, \{g_{sr}, g_{rs}\} \,\Big|\, f_{Y|X}(y|\mathsf{x}) \neq f^0_{Y|X}(y|\mathsf{x}) \Big\}. \tag{6}$$

If $T_1$ holds, there must be misbehavior in the S-R-S link; unfortunately we cannot jump to a conclusion of no misbehavior if $T_1$ does not hold, owing to Case 4. Thus, $T_1$ is referred to as the *detectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-S link; correspondingly, the complementary set ${T_1}^c$ of $T_1$ is called the *undetectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-S link.

In order to fully check the parameter pair $(f_{V|U},\mathbf{g})$, an S-R-D link should be included. For the S-R-D link, the set $T_2$ is defined as

$$T_2 := \Big\{ f_{V|U}, \{g_{sr}, g_{rd}\} \,\Big|\, f_{Y_2|Y_1}(y_2|y_1) \neq f^0_{Y_2|Y_1}(y_2|y_1) \Big\}, \tag{7}$$

where $f_{Y_2|Y_1}(y_2|y_1)$ and $f^0_{Y_2|Y_1}(y_2|y_1)$ are, respectively, the likelihood function and PDF of the symbol $Y_2$ received at D from the relay link conditioned on the symbol $Y_1$ received from the direct link. $T_2$ and its complementary set $T_2^c$ are referred to as, respectively, the *detectable set* and the *undetectable set* of the parameter pair $(f_{V|U},\mathbf{g})$ in the S-R-D link. Fig.2 (b) helps to check the detectable set $T_2$ directly.

The parameter pair $(f_{V|U},\mathbf{g})$ is completely partitioned by combinations of $T_1$ and $T_2$. We call $T=T_1\bigcup T_2$ as the *detectable class*, in which misbehavior is inevitable. It is emphasized that the complementary set $T^c=T_1^c\bigcap T_2^c$ of $T$ implies that the behavior can be cooperative or malicious. Thus, attack types can be given by the following definition.

*Definition 2 (Attack Types): If the parameter pair $(f_{V|U},\mathbf{g})$ belongs to the detectable class $T$, misbehavior is certain, and this is called a detectable attack; if $T^c$ holds and the relay is malicious, the resulting misbehavior is called an undetectable attack.*

From Definition 2, it is seen that detectable attacks map directly to the detectable class, whereas undetectable attacks map only to a subset of the undetectable class. An undetectable attack demands that falsified forwarding and dishonest feedback occur simultaneously, but the attack is not detected by a given detection approach. The undetectable attack is a small probability event compared to the detectable attack, because the undetectable attack is required to satisfy stricter conditions. It is emphasized that the undetectable attack is still in an infinite set. Table I illustrates the relationship between Definition 1 and Definition 2, where $\varnothing$ denotes the empty set. The action of the relay, i.e., the parameter pair $(f_{V|U},\mathbf{g})$, can be fully classified by use of Definitions 1 and 2. A detectable attack results from the overlap of these two definitions, and the

identification of a detectable attack is precisely equivalent to the identification of the detectable class $T$.

### B. Identification of a Detectable Attack

As the detectable class $T$ involves both $T_1$ and $T_2$, detection is implemented at the source node and at the destination node. In order to quantify the consecutive received symbols, it is convenient to use an $n'$-length sequence $(t_1, t_2, \cdots, t_{n'})$ satisfying $a = t_1 < t_2 < t_3 \cdots < t_{n'} = b$, where the quantization range $[a, b]$ depends on $n'$. Further, we consider the quantization interval $\Delta = \frac{b-a}{n'-1}$ to be such that $\lim_{n' \to \infty} \Delta = 0$.

*1) Decision Metric at S:* The detection at S focuses on the S-R-S link, in which the source uses its transmitted symbols as a reference to check whether or not action of the relay node is in the detectable set $T_1$. We employ the empirical CDF to approximate the likelihood function $f_{Y|X}(y|\mathsf{x})$. By jointly considering the transmitted and received signal sequences $(X^N, Y^N)$, the conditional empirical CDF $F_{Y^N|X^N}(t|\mathsf{x})$ at S is written as

$$F_{Y^N|X^N}(t|\mathsf{x}) = \frac{\sum_{i=1}^{N} I(y[i] < t) I(x[i] = \mathsf{x})}{\sum_{i=1}^{N} I(x[i] = \mathsf{x})}. \tag{8}$$

Naturally, a statistical decision metric $D_1^N$ is expressed as

$$D_1^N = \frac{1}{n'} \sum_{m=1}^{n'} \left| F_{Y^N|X^N}(t_m|\mathsf{x}) - F^0_{Y|X}(t_m|\mathsf{x}) \right|, \tag{9}$$

where $F^0_{Y|X}(t_m|\mathsf{x})$ is the CDF of $f^0_{Y|X}(t_m|\mathsf{x})$ as given in (5).

*2) Decision Metric at D:* The detection at D is related to the security of the S-R-D link and takes place at the same time as the detection at S. Since D receives the signal $Y_1^N$ in Phase 1 (see (1)) and then the signal $Y_2^N$ in Phase 2 (see (2)), the likelihood function $f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h})$ characterizing the S-R-D link can be obtained as

$$f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; h_{sr}) f_{V|U}(v|u) f_{Y_2|V}(y_2|v; h_{rd}) \, \mathrm{d}u \mathrm{d}v. \tag{10}$$

However, unlike S, D is inaccessible to the transmitted signal $X^N$. The received signal $Y_1^N$ in the direct link is exploited as a 'clean' reference for the detection at D. $(Y_1, X, Y_2)$ forms a Markov chain as $Y_1 \to X \to Y_2$, and $Y_1$ and $Y_2$ are conditionally independent for a given $X$, so the likelihood function conditioned on $Y_1 \leq t$ can be mathematically expressed as

$$f_{Y_2|Y_1}(y_2|t; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{t} \sum_{\mathsf{x} \in \mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) f_{Y_2|X}(y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) \Pr(X = \mathsf{x}) \, \mathrm{d}y_1 \Big/ \int_{-\infty}^{t} \sum_{\mathsf{x} \in \mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) \Pr(X = \mathsf{x}) \, \mathrm{d}y_1. \tag{11}$$

Since the conditional PDF at D $f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g})$ is computed as

$$f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g}) = \int_{-\infty}^{+\infty} f_{U|X}(u|\mathsf{x}; g_{sr}) f_{Y_2|V}(y_2|v; g_{rd}) \, \mathrm{d}u, \tag{12}$$

the conditional PDF $f^0_{Y_2|Y_1}(y_2|t; \mathbf{g})$ can be formulated as

$$f^0_{Y_2|Y_1}(y_2|t; \mathbf{g}) = \int_{-\infty}^{t} \sum_{\mathsf{x} \in \mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) f^0_{Y_2|X}(y_2|\mathsf{x}; \mathbf{g}) \Pr(X = \mathsf{x}) \, \mathrm{d}y_1 \Big/ \int_{-\infty}^{t} \sum_{\mathsf{x} \in \mathcal{X}} f_{Y_1|X}(y_1|\mathsf{x}) \Pr(X = \mathsf{x}) \, \mathrm{d}y_1. \tag{13}$$

For ease of presentation, $f_{Y_2|Y_1}(y_2|y_1; f_{V|U}, \mathbf{h})$ and $f^0_{Y_2|Y_1}(y_2|y_1; \mathbf{g})$ are written as $f_{Y_2|Y_1}(y_2|y_1)$ and $f^0_{Y_2|Y_1}(y_2|y_1)$ in the section below.

Based on the pair of received signals $(Y_1^N, Y_2^N)$, the empirical conditional CDF $F_{Y_2^N|Y_1^N}(s|t)$ can be expressed as

$$F_{Y_2^N|Y_1^N}(s|t) = \frac{\sum_{i=1}^{N} I(y_1[i] < t) I(y_2[i] < s)}{\sum_{i=1}^{N} I(y_1[i] < t)}. \tag{14}$$

By employing $F_{Y_2^N|Y_1^N}(s|t)$, the statistical decision metric $D_2^N$ for the detection at D is given by

$$D_2^N = \frac{1}{n'^2} \sum_{p=1}^{n'} \sum_{q=1}^{n'} \left| F_{Y_2^N|Y_1^N}(t_p|t_q) - F^0_{Y_2|Y_1}(t_p|t_q) \right|, \tag{15}$$

where $F^0_{Y_2|Y_1}(t_p|t_q)$ is the CDF of $f^0_{Y_2|Y_1}(t_p|t_q)$ as given in (13).

*3) Detection:* After obtaining the decision statistical metrics $D_1^N$ and $D_2^N$, we first identify whether the action of the relay falls into the detectable class $T$ or not. The following proposition will show how $D_1^N$ and $D_2^N$ identify, respectively, the detectable sets $T_1$ in the S-R-S link and $T_2$ in the S-R-D link.

*Proposition 1 (Detection at S and D): In the S-R-S link, $T_1$ can be detected by $D_1^N$ at S; in the S-R-D link, $T_2$ can be detected by $D_2^N$ at D. For $i = 1, 2$, the two decision metrics $D_1^N$ and $D_2^N$ have the following properties:*

*i)* $\lim_{N \to \infty} \Pr\left( D_i^N > \rho_1 \;\middle|\; (f_{V|U}, \mathbf{g}) \in T_i \right) = 1$, *when* $\Pr\left( (f_{V|U}, \mathbf{g}) \in T_i \right) > 0$,

*ii)* $\lim_{N \to \infty} \Pr\left( D_i^N > \rho_2 \;\middle|\; (f_{V|U}, \mathbf{g}) \in T_i^c \right) = 0$, *when* $\Pr\left( (f_{V|U}, \mathbf{g}) \in T_i^c \right) > 0$, *where $\rho_1$ and $\rho_2$ are strictly positive, and can be arbitrary small.*

*Proof:* See Appendix A. ■

*Remark 1: Take the detection at S for example. From (6), the detectable set $T_1$ implies that the likelihood function $f_{Y|X}(y|\mathsf{x})$ differs from the conditional PDF $f^0_{Y|X}(y|\mathsf{x})$. According to the law of large numbers, the empirical distribution $F_{Y^N|X^N}$ approaches the CDF of $f_{Y|X}(y|\mathsf{x})$ as $N \to \infty$. From the proof of Proposition 1, we can see that $D_1^N$ uses $F_{Y^N|X^N}$ as the bridge to measure the 'distance' between $f_{Y|X}(y|\mathsf{x})$ and $f^0_{Y|X}(y|\mathsf{x})$.*

*Remark 2: Proposition 1 points out that, if the behavior of the relay follows the undetectable set $T_i^c, i = 1, 2$, then $D_i^N \to 0$. Otherwise, it is probable that the source is capable of identifying a detectable attack. In addition, the missed detection and false alarm probabilities of $D_i^N$ can be arbitrary small as $N \to \infty$.*

Combining the detection at S with the detection at D, the detectable class $T$ can be identified by the proposed Algorithm 1 below.

**Algorithm 1** The Identification Procedure for a Detectable Attack

1: Initialization: Select appropriate $N$ and $n'$, and receive the CSI set $\mathbf{g}$.
2: Calculate the decision metrics: S computes $D_1^N$ based on $(X^N, Y^N)$, and D computes $D_2^N$ based on $(Y_1^N, Y_2^N)$ simultaneously.
3: **if** $D_1^N \to 0 \cap D_2^N \to 0$ **then**
4: $(f_{V|U}, \mathbf{g}) \in T_1^c \bigcap T_2^c$, the action of the relay belongs to the undetectable class $T^c$.
5: **else**
6: $(f_{V|U}, \mathbf{g}) \in T_1 \bigcup T_2$, the action of the relay belongs to the detectable class $T$.
7: **end if**

According to Algorithm 1, if the action of the relay belongs to the detectable class, we draw a conclusion immediately that the relay is suffering from a malicious attack; if the action of the relay belongs to the undetectable class, we cannot decide whether the relay is suffering from a malicious attack or not.

### C. Signal Detection of the Undetectable Class

According to Definitions 1 and 2, we know that undetectable class consists of undetectable attacks and cooperative (or friendly) relays. In other words, if falsified forwarding and dishonest feedback occur simultaneously, it is possible that an undetectable attack has the same statistical behavior as a cooperative relay. Thus, we cannot identify whether a malicious attack is occurring by use of Algorithm 1; consequently, a malicious relay that is performing an undetectable attack can disguise itself as a cooperative one – from the signal processing point of view, the performance of an undetectable attack is the same as that of the cooperative relay. On the assumption of an i.i.d. attack, the undetectable attack can be neglected.

At D, maximum-likelihood (ML) demodulation is used, based on the CSI $\mathbf{g}$. Following (1) and (13), the symbols received from the direct link and the relay link are re-expressed as

$$\begin{cases} Y_1 = h_{sd}X + W_{sd}, \\ Y_2 = g_{sr}g_{rd}X + g_{rd}W_{sr} + W_{rd}, \end{cases}$$

which are written in vector form as $\mathbf{Y} = \mathbf{H}X + \mathbf{W}$, with $\mathbf{Y} = \left[Y_1, Y_2\right]^T$, $\mathbf{H} = \left[h_{sd},\ g_{sr}g_{rd}\right]^T$ and $\mathbf{W} = \left[W_{sd},\ g_{rd}W_{sr} + W_{rd}\right]^T$.

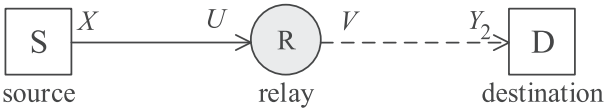



Fig. 3. A cooperative relay system consisting of a source-destination pair and a potential malicious relay without direct link.

ML detection is then performed as

$$\hat{X} = \underset{X \in \mathcal{X}}{\operatorname{argmax}} \Pr(\mathbf{Y}|X) = \underset{X \in \mathcal{X}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{H}X\|^2. \quad (16)$$

From (16), the joint PDF of $\mathbf{Y}$, $f_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; f_{V|U}, \mathbf{h})$, only effects ML detection. Then the following proposition gives a property of the undetectable class $T^c$.

*Proposition 2: If the parameter pair $(f_{V|U}, \mathbf{g})$ belongs to the undetectable class $T^c$, then there exists $f_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; \mathbf{g})$ regardless of whether there is an undetectable attack or cooperative behavior.*

*Proof:* See Appendix B. ■

*Remark 3: Essentially Proposition 2 identifies that, if the action of the relay belongs to the undetectable class $T^c$, the distributions of the received symbols from the direct link and the relay link are subject to the same joint PDF $f^0_{Y_1,Y_2|X}(y_1, y_2|\mathsf{x}; \mathbf{g})$. Therefore, ML detection has the same outcome irrespective of whether it arises from an undetectable attack or from cooperative behavior.*

In terms of the signal detection performance, an undetectable attack is no worse than cooperative behavior. This implies that, for the undetectable attack, the symbols received can be properly demodulated as if they resulted from cooperative behavior. That is, although the undetectable attack cannot be identified by Algorithm 1, a relay system with an undetectable attack can still deliver the same diversity order performance as a relay system with cooperative behavior. The symbol error rate (SER) for the undetectable attack in the high signal-to-noise ratio (SNR) region is approximated as [21]

$$\Pr(e) \overset{\text{high-SNR}}{\simeq} \frac{3}{K\gamma^2}, \quad (17)$$

where $K = \frac{|g_{sr}|(|g_{sd}|+|g_{rd}|)}{|g_{sd}||g_{rd}|}$, and $\gamma = E_s/\mathcal{N}_0$ is SNR without fading. It is observed that the diversity order of the undetectable attack is 2.

An undetectable attack involves the collusion between falsified forwarding and the dishonest feedback. This escapes detection because the damage caused by the falsified forwarding is mitigated by the dishonest feedback. This intuitively explains why, for an undetectable attack, the malicious relay can still be used to maintain the cooperative diversity.

## IV. Relay System Without a Direct Link

In this section we extend our consideration from relay systems with a direct link to those without a direct link between the S and the D due to coverage, as shown in Fig. 3.

While the detection at S is unaffected as the S-R-S link is still present, in the absence of a direct link as a 'clean' reference, the approach proposed in Section III-B cannot be

applied immediately. We must develop a new detection method at D that can be used for relay systems without a direct link.

We first repeat the two-phase transmission. Here, the notation is consistent with earlier sections.

In Phase 1, S sends $X^N$ to R (solid line in Fig. 3). The symbol received at R, $U$, is written as

$$U = h_{sr}X + W_{sr}. \tag{18}$$

R processes the $U^N$ received using AF protocol, generates $V^N$ and then forwards it in Phase 2 (dashed line in Fig. 3). The symbol received at D is expressed as

$$Y_2 = h_{rd}V + W_{rd}, \tag{19}$$

where for $i, j \in \{\mathrm{S}, \mathrm{R}, \mathrm{D}\}$, $i \neq j$, $h_{ij}$ is the channel gain between node $i$ and node $j$, and $W_{ij}$ is the Gaussian noise at node $j$ with variance $\mathcal{N}_0$. Definition 1 still applies to this relay system, while Definition 2 is changed according to the following cases.

### A. Known Source Distribution

If the source distribution is known, we can use a simple extension of the previous detection approach based on a direct link. The reliable CSI set is denoted as $\mathbf{h} = \{h_{sr}, h_{rs}, h_{rd}\}$ and the CSI set provided is denoted as $\mathbf{g} = \{g_{sr}, g_{rs}, g_{rd}\}$. Since the S-R-S link remains unchanged, $T_1$ can still be checked by the detection at S. However, the detection at D will be modified based on the known source distribution.

The likelihood function is given by

$$f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) = \sum_{x \in \mathcal{X}} f_{Y_2|X}(y_2|x; f_{V|U}, \mathbf{h}) \Pr(X), \tag{20}$$

where $f_{Y_2|X}(y_2|x; f_{V|U}, \mathbf{h})$ is given in (10), and the conditional PDF is expressed as

$$f^0_{Y_2}(y; \mathbf{g}) = \textstyle\sum_{x \in \mathcal{X}} f^0_{Y_2|X}(y|x; \mathbf{g}) \Pr(X), \tag{21}$$

where $f^0_{Y_2|X}(y|x; \mathbf{g})$ is given in (12).

According to (20) and (21), $T_2$ is redefined as

$$T_2 := \left\{ f_{V|U}, \{g_{sr}, g_{rd}\} \,\middle|\, f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) \neq f^0_{Y_2}(y_2; \mathbf{g}) \right\}.$$

By observing the received sequence $Y^N$, the empirical CDF at D is given by

$$F_{Y_2^N}(t) = \frac{1}{N} \sum_{i=1}^{N} I(y_2[i] < t). \tag{22}$$

From (20), (21) and (22), the decision metric $D_2^N$ in (15) is modified to

$$D_2^N = \tfrac{1}{n'^2} \textstyle\sum_{m=1}^{n'} \left| F_{Y_2^N}(t_m) - F^0_{Y_2}(t_m) \right|, \tag{23}$$

where $F^0_{Y_2}(t)$ is the CDF of $f^0_{Y_2}(t; \mathbf{g})$ given in (21). By employing this new $D_2^N$, together with (9), Algorithm 1 can deal with the detection of misbehavior for relay systems without direct links, based on a known source distribution.

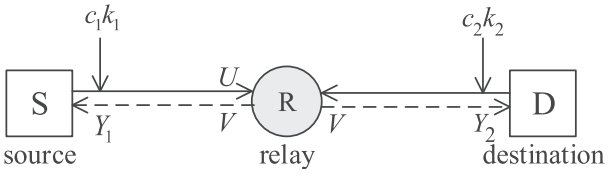



Fig. 4. A cooperative relay system with added artificial noise, where the solid and dashed lines denote Phase 1 and Phase 2, respectively.

### B. Unknown Source Distribution

If the source distribution is unknown to the relay system, the destination has no access to any 'clean' reference, since all physical layer observations tend to be manipulated. Adding auxiliary information can help to detect pilot contamination attack [19]. We employ additive AN to assist in building trustworthy reference information.

Traditionally, AN is designed to lie in the null space of the main channel [22], and it is exploited to degrade an eavesdropper's channel so that a secure capacity is guaranteed. In this paper, instead of using the traditional design of AN, we propose a different type of AN, as described below.

1) The structure of AN requires that the source is equipped with multiple antennas. Single-antenna nodes can emulate a distributed multi-antenna array. By executing a two-way communication protocol (see Fig. 4), the source and the destination simultaneously send information to the relay, thus the condition of forming AN can be satisfied.
2) The AN is defined as the product of coefficient matrix $\mathbf{C}$ and key vector $\mathbf{k}$. Then, the AN is denoted as $\mathbf{Ck}$, where $\mathbf{C} = \mathrm{diag}\{c_1,\ c_2\}$ and $\mathbf{k} = [k_1,\ k_2]^T$.
3) According to the two-way communication protocol, the AN lies in the null space of the provided CSI vector $\mathbf{g}_r = [g_{sr},\ g_{dr}]^T$ so that $\mathbf{g}_r^T\mathbf{Ck} = 0$.
4) For a given $\mathbf{C}$, when $\mathbf{g}_r$ is known and $\|\mathbf{k}\| = 1$, the AN is deterministic rather than random.
5) The AN changes with time, which takes place when the coefficient matrix $\mathbf{C}$ changes.
6) Conventionally, the wiretap channel is assumed to be uncorrelated with the main channel, which implies $\mathbf{h}_r^T\mathbf{Ck} \neq 0$. This assumption is invalid in the case considered, because $\mathbf{g}_r$ represents unreliable CSI that can be of any value. For example, the dishonest feedback can allow $\mathbf{g}_r$ to be correlated with $\mathbf{h}_r$, say, $\mathbf{g}_r = \alpha\mathbf{h}_r$ for $\alpha \neq 1$. Then, we have $\mathbf{h}_r^T\mathbf{Ck} = 0$ and AN will fail. Therefore, our analysis of the dishonest feedback covers two separate cases: $\mathbf{g}_r$ is either correlated or uncorrelated with $\mathbf{h}_r$.

In Phase 1, both S and D send AN $\mathbf{Ck}$ simultaneously. The signal received at R is expressed as

$$U = \mathbf{h}_r^T\mathbf{Ck} + W_r, \tag{24}$$

where $\mathbf{h}_r = [h_{sr},\ h_{dr}]^T$. $W_r$ is Gaussian noise at R with variance $\mathcal{N}_0$.

In Phase 2, R receives $U^N$ and then forwards a processed version, $V^N$, to S and D due to the broadcast nature of a wireless channel. The signals received at S and at D are

written as

$$Y_1 = h_{rs}V + W_{rs}, \tag{25a}$$

$$Y_2 = h_{rd}V + W_{rd}, \tag{25b}$$

where $h_{rs}$ and $h_{rd}$ are channel gains, and $W_{rs}$ and $W_{rd}$ are Gaussian noise with variance $\mathcal{N}_0$ at S and at R, respectively.

In the channel estimation process, R can know the CSI of both the S-R link and the D-R link, as S and D send pilot signals to R. Then, due to dishonest feedback, R broadcasts the potentially unreliable CSI, instead of the valid one, to S and D. When the unreliable CSI is obtained at S and D, the proposed AN-aided scheme comes into play.

Because of the symmetry of the system considered, we show the detection results from a source perspective, and the conditional likelihood function is given by

$$f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_U(u; h_{sr}) f_{V|U}(v|u) f_{Y_1|V}(y_1|v; h_{rs})\,\mathrm{d}u\mathrm{d}v, \tag{26}$$

where $\mathbf{h} = [h_{sr}, h_{dr}, h_{rs}, h_{rd}]$ is the reliable CSI set. The conditional PDF is formulated as

$$\begin{aligned} f^0_{Y_1}(y_1; \mathbf{g}) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_U(u; g_{sr})\delta(v-u) f_{Y_1|V}(y_1|v; g_{rs})\,\mathrm{d}u\mathrm{d}v \\ &= \int_{-\infty}^{+\infty} f_U(u; g_{sr}) f_{Y_1|V}(y_1|u; g_{rs})\,\mathrm{d}u, \end{aligned} \tag{27}$$

where $\mathbf{g} = [g_{sr}, g_{dr}, g_{rs}, g_{rd}]$ is the CSI set provided, with $f_{Y_2}(y_2; f_{V|U}, \mathbf{h})$ and $f^0_{Y_2}(y_2; \mathbf{g})$ being expressed in the similar way.

We discuss the four cases of the parameter pair $(f_{V|U}, \mathbf{g})$ as follows.

1) $f_{V|U} = \delta(v-u) \cap \mathbf{g} = \mathbf{h}$. It is easy to obtain $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.
2) $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} = \mathbf{h}$. First, we have $f_U(u; h_{sr}) = f_U(u; g_{sr})$ since AN is nulled out; then we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ because $f_{V|U} \neq \delta(v-u)$ and $f_{Y_1|V}(y_1|v; h_{rs}) = f_{Y_1|V}(y_1|v; g_{rs})$.
3) - $f_{V|U} = \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r \neq \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   According to (26) and (27), we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ as $\mathbf{h}_r^T\mathbf{C}\mathbf{k} \neq 0$.
   - $f_{V|U} = \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r = \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   It is observed that $\mathbf{C}\mathbf{k}$ lies in the null space of $\mathbf{h}_r$, so $\mathbf{h}_r^T\mathbf{C}\mathbf{k} = 0$; if $\mathbf{g} \neq \mathbf{h}$ but $g_{rs} = h_{rs}$, we have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.
4) - $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r \neq \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   The two types of misbehavior have the potential to make $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$. By artificially operating $\mathbf{C}$, $\mathbf{h}_r^T\mathbf{C}\mathbf{k}$ changes over time and cannot be bounded by i.i.d. attack manner – $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$ may hold for some $\mathbf{C}$s with the specific pair $(f_{V|U}, \mathbf{g})$, but it does not hold when $\mathbf{C}$ changes. In general, we must have $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$ by using a time-varying coefficient matrix $\mathbf{C}$.
   - $f_{V|U} \neq \delta(v-u) \cap \mathbf{g} \neq \mathbf{h} \cap \mathbf{g}_r = \alpha\mathbf{h}_r$, for $\alpha \neq 1$.
   The matrix $\mathbf{C}$ fails to change $\mathbf{h}_r^T\mathbf{C}\mathbf{k}$ as $\mathbf{C}\mathbf{k}$ lies in the null space of $\mathbf{h}_r$. It is possible to obtain $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$ with the specific pair $(f_{V|U}, \mathbf{g})$, which we will discuss later.

From the above discussion, if $\mathbf{g}_r \neq \alpha\mathbf{h}_r$ for $\alpha \neq 1$, a sufficient condition to determine misbehavior of the relay is that $(f_{V|U}, \mathbf{g})$ makes $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$. When $\mathbf{g}_r = \alpha\mathbf{h}_r$ for $\alpha \neq 1$, it is still possible that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$, because AN $\mathbf{C}\mathbf{k}$ fails to enable the distribution $Y_1$ to distinguish $f_{Y_1}(y_1; f_{V|U}, \mathbf{h})$ from $f^0_{Y_1}(y_1; \mathbf{g})$. To address this, we modify the AN $\mathbf{C}\mathbf{k}$ to $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$, where $\mathbf{g}_r^T\tilde{\mathbf{C}}\tilde{\mathbf{k}} \neq 0$. Therefore, for the second case of 3), the introduction of $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ means that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g})$. However, for the second case of 4), it is still possible that $f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_1}(y_1; \mathbf{g})$.

As previously, we define

$$T_{AN1} := \left\{ f_{V|U}, \mathbf{g} \,\middle|\, f_{Y_1}(y_1; f_{V|U}, \mathbf{h}) \neq f^0_{Y_1}(y_1; \mathbf{g}) \right\},$$

and

$$T_{AN2} := \left\{ f_{V|U}, \mathbf{g} \,\middle|\, f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) \neq f^0_{Y_2}(y_2; \mathbf{g}) \right\}.$$

$T_{AN} = T_{AN1} \bigcup T_{AN2}$ is referred to as the *detectable class*, and its complement, $T^c_{AN}$, as the *undetectable class*.

1) To identify the detectable class $T_{AN}$, we need detection at both S and D. For $j = 1, 2$, based on the received sequences $Y_1^N$ and $Y_2^N$, the empirical CDFs at S and at D are given by

$$F_{Y_j^N}(t) = \frac{1}{N}\sum_{i=1}^{N} I(y_2[i] < t). \tag{28}$$

Similarly, for $j = 1, 2$, the decision metric $D_j^N$ is written as

$$D_j^N = \frac{1}{n'^2}\sum_{m=1}^{n'} \left| F_{Y_j^N}(t_m) - F^0_{Y_j}(t_m) \right|, \tag{29}$$

where $F^0_{Y_j}(t)$ is the CDF of $f^0_{Y_j}(t; \mathbf{g})$. The identification procedure of the detectable attack is elaborated in Algorithm 2.

2) We now focus on the undetectable class $T^c_{AN}$. From the expression of $f^0_{Y_2}(y_2; \mathbf{g})$, $Y_2$ is formulated as

$$Y_2 = g_{rd}(W_r + M\mathbf{g}_r^T\tilde{\mathbf{C}}\tilde{\mathbf{k}}) + W_{rd}, \tag{30}$$

where $M$ is the number of occurrences of $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ in an $N$-length block (usually taken to be $N/3$). Specifically, by setting $\tilde{\mathbf{C}} = diag\{1/\alpha M, 0\}$ and $\tilde{\mathbf{k}} = [X, 0]^T$ when $\mathbf{g}_r = \alpha\mathbf{h}_r$, (30) is rewritten as

$$Y_2 = g_{rd}(W_r + h_{sr}X) + W_{rd}, \tag{31}$$

According to the definition of $T_{AN2}$, we have $f_{Y_2}(y_2; f_{V|U}, \mathbf{h}) = f^0_{Y_2}(y_2; \mathbf{g})$. Following the same logic as in Section III-C, the signal detection performance

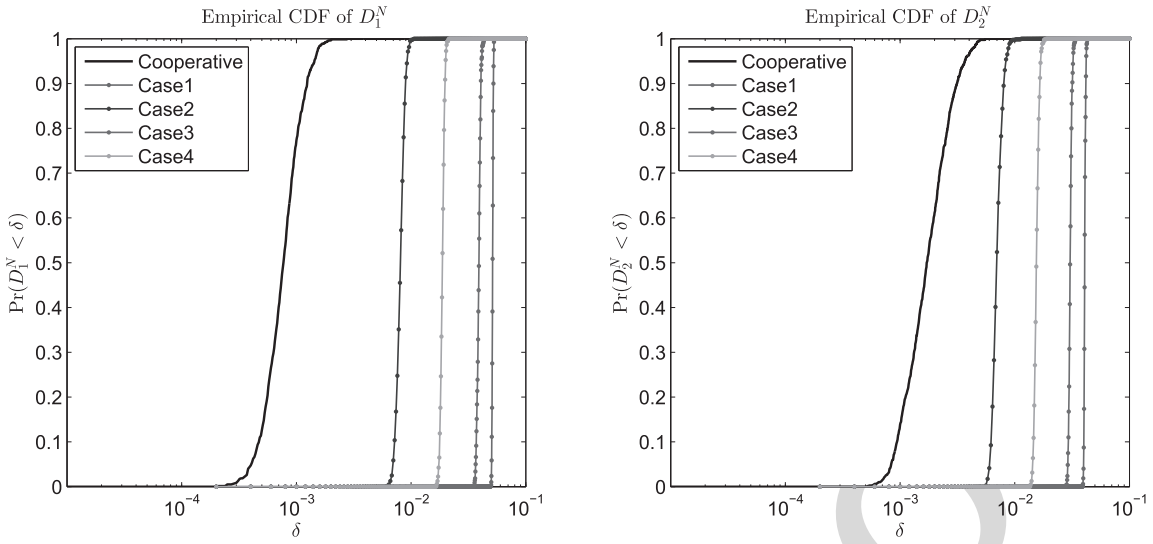



Fig. 5. The empirical CDFs of $D_1^N$ and $D_2^N$ for the four detectable attacks considered.

**Algorithm 2** The Identification Procedure for Attack Detection With Aided AN

1: Initialization: generate coefficient matrices $\mathbf{C}_1$, $\mathbf{C}_2$ and $\tilde{\mathbf{C}}$ and give the CSI set $\mathbf{g}$.
2: Calculate AN: compute $\mathbf{k}_1$, $\mathbf{k}_2$, and $\tilde{\mathbf{k}}$ based on $\mathbf{C}_1$, $\mathbf{C}_2$, and $\tilde{\mathbf{C}}$, respectively.
3: Add AN: take turns adding $\mathbf{C}_1\mathbf{k}_1$, $\mathbf{C}_2\mathbf{k}_2$ and $\tilde{\mathbf{C}}\tilde{\mathbf{k}}$ at S and D in each instant.
4: Calculate decision metric: $D_1^N$ and $D_2^N$ are computed at S and at D, respectively.
5: **if** $D_1^N \to 0 \cap D_2^N \to 0$ **then**
6: The relay action is a member of the undetectable class $T_{AN}^c$.
7: **else**
8: The relay action belongs to the detectable class $T_{AN}$ – the relay must be misbehaving.
9: **end if**

of the undetectable attack is the same as that of the cooperative scenario.

## V. Numerical Examples

### A. Relay Systems With a Direct Link

As illustration, we present here both detectable and undetectable attacks; we also evaluate the effectiveness of the proposed approach in identifying the two types of attack.

*1) Detectable Attack:* We consider a the relay system shown in Fig. 1, with S transmitting a BPSK signal with unit energy. Assume that the reliable CSI set $\mathbf{h} = [1, 1, 1]$, the AWGN variance $\mathcal{N}_0 = 0.01$, and the direct link channel gain $h_{sd} = 0.8$. The block length was selected to have $N = 1000$, and for quantization purposes $n' = 100$, $-a = b = \sqrt{n'}/2$, which implies that $\Delta = 1/\sqrt{n'}$.

To verify the effectiveness of the proposed detection schemes, the following four detectable malicious attacks were considered:

- CASE 1 - Dishonest Feedback: The relay provides an unreliable CSI with $\mathbf{g} = [0.6, 0.8, 0.7]$.
- CASE 2 - Falsified Forwarding $I$: The relay actively injects Gaussian noise distributed with $\mathcal{N}(0, 0.04)$.
- CASE 3 - Falsified Forwarding $II$: The relay intentionally adds noise with uniform distribution $\mathcal{U}(-1, +1)$.
- CASE 4 - Mixed Attack: Both dishonest feedback and falsified forwarding are considered in this case; the relay injects Gaussian noise distributed with $\mathcal{N}(0, 0.0025)$ and provides $\mathbf{g} = [0.9, 0.9, 1]$.

Fig. 5 shows the empirical CDFs of $D_1^N$ and $D_2^N$ after 800 computer simulation runs for each of the above cases. It can be observed that there is a clear separation between the undetectable class and the detectable class; this can be used as a threshold (e.g. $\delta = 0.005$ for the detection at S) for identifying the detectable class. These results further verify the effectiveness of Proposition 1.

*2) Undetectable Attack:* We assume that the reliable CSI $\mathbf{h} = [1, \sqrt{2}/2, \sqrt{2}/2]$ and the CSI provided $\mathbf{g} = [\sqrt{2}/2, 1, 1]$, and that the malicious relay performs falsified forwarding by injecting Gaussian noise distributed with $\mathcal{N}(0, 0.01)$. Fig. 6 shows the empirical CDFs of $D_1^N$ and $D_2^N$ for cooperative behavior and an undetectable attack. It is evident that the cooperative behavior and the undetectable attack are not distinguishable.

*3) BER Performance in the Presence of an Undetectable Attack:* We assume that the channel gain of the direct link $h_{sd} = 0.4$ and the injected noise power (falsified forwarding) is set at the same level as $\mathcal{N}_0$. Fig. 7 illustrates the BER performance versus SNR for different noise powers; the undetectable attack is seen to have the same BER performance as both cooperative behaviour and direct transmission from S to D. These results verify the previous claim that, even for undetectable attacks, the diversity gain is maintained.

### B. Systems Without a Direct Link

*1) Detectable Attack:* The source transmits BPSK signals and the reliable CSI is set as $\mathbf{h} = [1/2, 1/3, 1/2, 1/2]$.

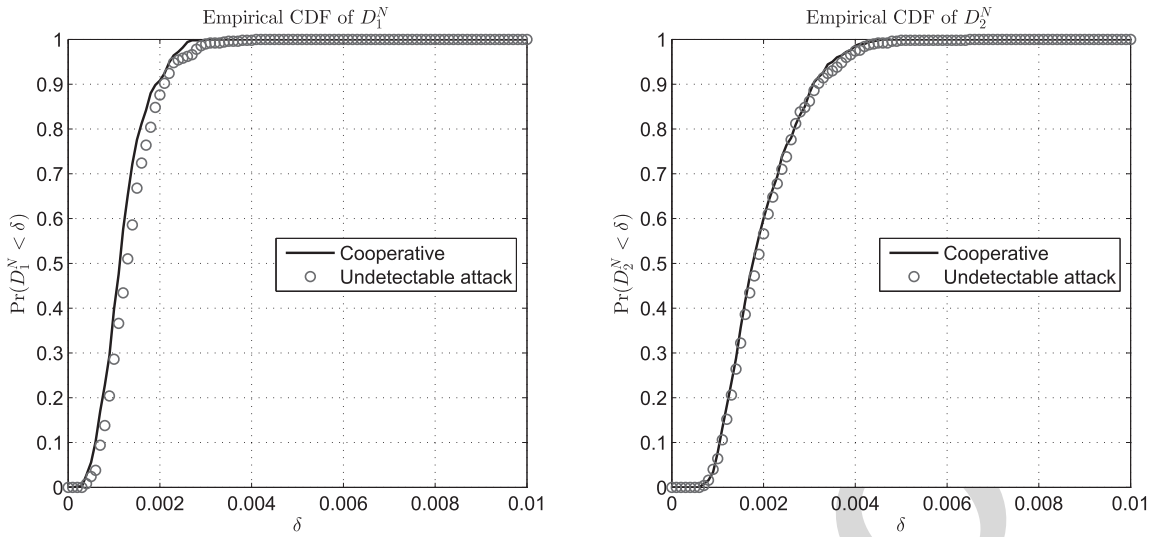



Fig. 6. The empirical CDFs of $D_1^N$ and $D_2^N$ for the undetectable attack considered.

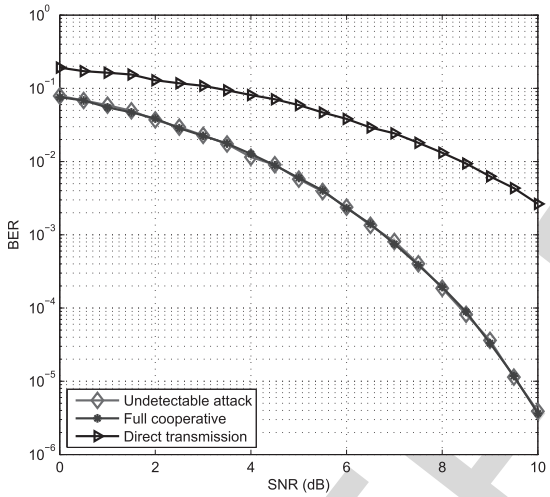



Fig. 7. BER performance comparisons among cooperative behavior, the undetectable class, and direct transmission.

The coefficient matrices are $\mathbf{C}_1 = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$, $\mathbf{C}_2 = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}$ and $\tilde{\mathbf{C}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Correspondingly, $\mathbf{k}_1 = [2/\sqrt{5}, 1/\sqrt{5}]^T$, $\mathbf{k}_2 = [1/\sqrt{5}, 2/\sqrt{5}]^T$ and $\tilde{\mathbf{k}} = [1, 0]^T$. $\mathcal{N}_0 = 1/\sqrt{5}$. The block length is chosen to have $N = 1000$ and, for quantization purposes, $n' = 100$, $-a = b = \sqrt{n'}/2$, which implies that $\Delta = 1/\sqrt{n'}$. The three different cases are discussed below.

- CASE 1 - Dishonest Feedback: The relay provides the unreliable CSI $\mathbf{g} = [1/2, 1/2, 1/3, 1/3]$.
- CASE 2 - Malicious Forwarding $I$: The relay actively injects Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$.
- CASE 3 - Mixed Attack: We consider both dishonest feedback and falsified forwarding, where the relay injects Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$ and provides $\mathbf{g} = [1/3, 1/3, 1/2, 1/2]$.

Fig. 8 shows the empirical CDFs of $D_1^n$ after 800 computer simulation runs, in each of the three cases. The proposed decision metric is clearly capable of distinguishing between the detectable and undetectable classes.

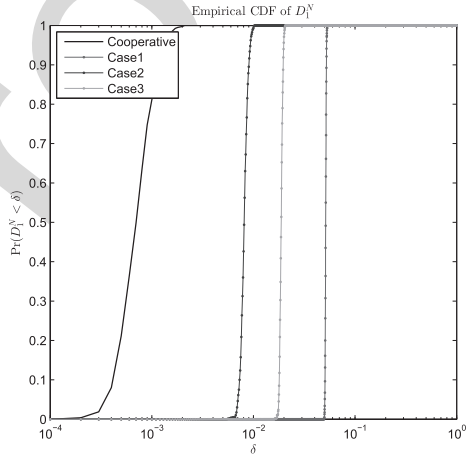



Fig. 8. The empirical CDFs of $D_1^N$ for the three detectable attacks considered.

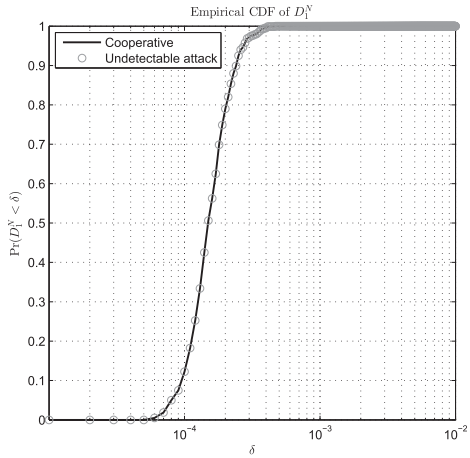



Fig. 9. The empirical CDFs of $D_1^N$ for the undetectable attack considered.

*2) Undetectable Attacks:* We consider the previously described simulation model with a different $\mathbf{g} = [1/4, 1/6, 1/4, 1/4]$, and $\mathbf{g}_r = \alpha \mathbf{h}_r$ for $\alpha = 1/2$. The malicious

relay conducts falsified forwarding by injecting Gaussian noise distributed with $\mathcal{N}(0, 1/\sqrt{5})$. Fig. 9 demonstrates that it is impossible to differentiate between cooperative behavior and an undetectable attack.

## VI. Conclusion

This paper has investigated the problem of detecting malicious attacks in a two-hop AF relay network in the presence of an unreliable CSI. In particular, we have proposed a detection approach applicable to a system with a direct link which is capable of clearly distinguishing between the detectable and undetectable classes. It has also been shown that, for the detectable class, the proposed approach detects malicious attacks with high probability. The relay system retains the benefits of diversity gain even in the presence of an undetectable attack. Further, we extended the proposed approach to a more common scenario in which no direct link is available.

## Appendix A
## Proof of Proposition 1

Without loss of generality, we firstly prove that the decision metric $D_1^N$ satisfies the two properties of Proposition 1. According to Borel's strong law of large numbers, for any arbitrary small positive $\mu$, we have

$$\lim_{N\to\infty} \Pr\left(\left| \frac{\sum_{i=1}^{N} I(y[i] < y) I(x[i] = \mathsf{x})}{\sum_{i=1}^{N} I(x[i] = \mathsf{x})} - \Pr(Y < y | X = \mathsf{x}; f_{V|U}, \mathbf{h}) \right| \le \mu \right) = 1. \quad (32)$$

By defining a typical set as

$$\mathcal{A}_\mu \left( F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) \right) \triangleq \left\{ F \;\middle|\; \left| F - F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) \right| \le \mu \right\},$$

where $F_{Y|X}(y|\mathsf{x}; \mathbf{\Psi})$ is the CDF of $f_{Y|X}(y|\mathsf{x}; \mathbf{\Psi})$, (32) can be modified as

$$\lim_{n\to\infty} \Pr\left\{ F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu \left( F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) \right) \right\} = 1. \quad (33)$$

Under the assumption that $(f_{V|U}(v|u), \mathbf{g}) \in T$, we have $F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) \neq F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$, where $F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$ is the CDF of $f^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$.

For any sufficiently small positive $\delta$, we assume that

$$\left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta. \quad (34)$$

From (33), it follows that

$$\left| F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \in \left| \mathcal{A}_\mu \{ F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \} \right|,$$

which in turn implies that

$$\left| F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta - \mu. \quad (35)$$

Let us define $\rho_1 \triangleq \delta - \mu$ and assume that $\mu$ is chosen to be small enough to satisfy $\rho_1 > 0$. From the definition of $D_1^N$ in (9), (35) leads us to conclude that $D_1^N > \rho_1$.

Furthermore, according to (33) and (34), for any $\delta > 0$, we have

$$\begin{aligned}
&\Pr\left(D_1^N \ge \rho_1, (f_{V|U}(v|u), \mathbf{g}) \in T\right) \\
&= \Pr\left(D_1^N \ge \rho_1, \left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta\right) \\
&\ge \Pr\left(D_1^N \ge \rho_1, \left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta, \right. \\
&\quad \left. F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right) \\
&\overset{(a)}{=} \Pr\left(\left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta, \right. \\
&\quad \left. F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right) \\
&\ge \Pr\left(\left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta\right) \\
&\quad - \Pr\left(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right), \quad (36)
\end{aligned}$$

where (a) is derived by using (33), (34) and (35). From (36), we have

$$\begin{aligned}
&\Pr\left(D_1^N \ge \rho_1 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T\right) \\
&= \frac{\Pr\left(D_1^N \ge \rho_1, (f_{V|U}(v|u), \mathbf{g}) \in T\right)}{\Pr\left(\left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta\right)} \\
&\overset{(b)}{\ge} 1 - \frac{\Pr\left(F_{Y^N|X^N}(y|\mathsf{x}) \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)}{\Pr\left(\left| F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \ge \delta\right)}, \quad (37)
\end{aligned}$$

where (b) is derived by using (33).

As a result, $\lim_{n\to\infty} \Pr\left(D_1^N > \rho_1 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T\right) = 1$, which proves that $D_1^N$ satisfies the first property of Proposition 1.

We proceed now to prove that $D_1^N$ will satisfy the second property of Proposition 1. For this, assume that when $(f_{V|U}(v|u), \mathbf{g}) \in T^c$, we have $F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h}) = F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$. According to (33), it is also true that $F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F^0_{Y|X}(y|\mathsf{x}; \mathbf{g})\}$, which implies that

$$\left| F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| \in |\mathcal{A}_\mu\{0\}|, \quad (38)$$

and which yields

$$\left| F_{Y^N|X^N}(y|\mathsf{x}) - F^0_{Y|X}(y|\mathsf{x}; \mathbf{g}) \right| < \mu. \quad (39)$$

By defining $\rho_2 \triangleq \mu$, we have $D_1^N < \rho_2$, and thus

$$\Pr\left(D_1^N \ge \rho_2, (f_{V|U}(v|u), \mathbf{g}) \in T^c \right. \left. \left| F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right) = 0. \quad (40)$$

where $\Pr\left(D_1^N \ge \rho_2 \;\middle|\; (f_{V|U}(v|u), \mathbf{g}) \in T^c\right)$ is easily obtained and has been placed on top of the next page.

According to (33), this implies that $\Pr(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}) \to 0$. Finally, by means

$$\Pr\left(D_1^N \geq \rho_2 \left| (f_{V|U}(v|u), \mathbf{g}) \in T^c \right.\right) = \frac{\Pr\left(D_1^N \geq \rho_2, (f_{V|U}(v|u), \mathbf{g}) \in T^c \left| F_{Y^N|X^N}(y|\mathsf{x}) \in \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right.\right)}{\Pr\left((f_{V|U}(v|u), \mathbf{g}) \in T^c\right)} + \frac{\Pr\left(F_{Y^N|X^N}(y|\mathsf{x}) \notin \mathcal{A}_\mu\{F_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})\}\right)}{\Pr\left((f_{V|U}(v|u), \mathbf{g}) \in T^c\right)} \tag{41}$$

of (40) and (41), as shown at the top of this page, we have $\lim_{N\to\infty} \Pr\left(D_1^N \geq \rho_2 \left| (f_{V|U}(v|u), \mathbf{g}) \in T^c \right.\right) = 0$, which proves that $D_1^N$ satisfies the second property of Proposition 1.

By a similar procedure, we can prove that $D_2^N$ also satisfies the two properties of Proposition 1, which then concludes the proof of Proposition 1. ■

## APPENDIX B
## PROOF OF PROPOSITION 2

For the convenience of the proof, we introduce the following Lemma.

*Lemma 1: Let us consider a set of random variables $U_i$, $i = 1, 2, \cdots, 5$, in which $U_4 = U_2 + U_1$, $U_5 = U_3 + U_1$, and $U_1$ is independent of both $U_2$ and $U_3$. If there exists a PDF such that $f_{U_4|X}(u_4|\mathsf{x}) = f_{U_5|X}(u_5|\mathsf{x})$, then $f_{U_2|X}(u_2|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x})$ must hold.*

*Proof:* Since $U_4 = U_2 + U_1$, and $U_1$ and $U_2$ are independent of each other, we have

$$f_{U_4|X}(u_4|\mathsf{x}) = f_{U_2|X}(u_2|\mathsf{x}) + f_{U_1|X}(u_1|\mathsf{x}). \tag{42}$$

From (42), and by taking the characteristic function (CF) of $U_4$ conditioned on $X = \mathsf{x}$, we obtain

$$\varphi_{U_4|X}(t|\mathsf{x}) = \varphi_{U_2|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \tag{43}$$

where $\varphi_{U_2|X}(t|\mathsf{x})$ and $\varphi_{U_1|X}(t|\mathsf{x})$ are, respectively, the CFs of $U_2$ and $U_1$ conditioned on $X = \mathsf{x}$.

Similarly, since $U_5 = U_3 + U_1$ with $U_1$ and $U_3$ being independent with each other, we have

$$f_{U_5|X}(u_5|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x}) + f_{U_1|X}(u_1|\mathsf{x}). \tag{44}$$

Thus, the CF of $U_5$ conditioned on $X = \mathsf{x}$ can be expressed as

$$\varphi_{U_5|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \tag{45}$$

where $\varphi_{U_3|X}(t|\mathsf{x})$ is the CF of $U_3$ conditioned on $X = \mathsf{x}$, respectively.

Since $f_{U_4|X}(u_4|\mathsf{x}) = f_{U_5|X}(u_5|\mathsf{x})$, we have

$$\varphi_{U_4|X}(t|\mathsf{x}) = \varphi_{U_5|X}(t|\mathsf{x}). \tag{46}$$

Using (43), (45) and (46), we obtain

$$\varphi_{U_2|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x})\varphi_{U_1|X}(t|\mathsf{x}), \tag{47}$$

and as $\varphi_{U_1|X}(t|\mathsf{x})$ is non-zero, we have

$$\varphi_{U_2|X}(t|\mathsf{x}) = \varphi_{U_3|X}(t|\mathsf{x}). \tag{48}$$

Since any PDF can be uniquely determined by its CF, (48) implies that

$$f_{U_2|X}(u_2|\mathsf{x}) = f_{U_3|X}(u_3|\mathsf{x}). \tag{49}$$

■

We now return to the proof of Proposition 2. Since the detectable class $T = T_1 \cup T_2$, we have $T^c = T_1^c \bigcap T_2^c$. For the set $T_1^c$, $f_{Y|X}(y|\mathsf{x}; f_{V|U}, \mathbf{h})$ is identical to $f^0_{Y|X}(y|\mathsf{x}; \mathbf{g})$. Following (1), (2) and (5), we have

$$f_{h_{rs}V+N_{rs}|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{g_{rs}(g_{sr}X+N_{sr})+N_{rs}|X}(t|\mathsf{x}; \mathbf{g}). \tag{50}$$

According to Lemma 1, it is easy to obtain that

$$f_{h_{rs}V|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = f^0_{g_{rs}(g_{sr}X+N_{sr})|X}(t|\mathsf{x}; \mathbf{g}), \tag{51}$$

and if we note that $f^0_{g_{rs}(g_{sr}X+N_{sr})|X}(t|\mathsf{x}; \mathbf{g}) = \frac{1}{\pi\sigma_{sr}^2 g_{rs}^2} \exp(-\frac{\|t-g_{sr}g_{rs}\mathsf{x}\|^2}{\sigma_{sr}^2 g_{rs}^2})$, then we have

$$f_{V|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = \frac{h_{rs}^2}{\pi\sigma_{sr}^2 g_{rs}^2} \exp(-\frac{\|h_{rs}t - g_{sr}g_{rs}\mathsf{x}\|^2}{\sigma_{sr}^2 g_{rs}^2}). \tag{52}$$

Following (2b), (52) can be re-expressed as

$$f_{Y_2|X}(t|\mathsf{x}; f_{V|U}, \mathbf{h}) = \frac{1}{\pi(K^2\sigma_{sr}^2 + \sigma_{rd}^2)} \exp(-\frac{\|t - g_{sr}K\mathsf{x}\|^2}{(K^2\sigma_{sr}^2 + \sigma_{rd}^2)}), \tag{53}$$

where $K = g_{rs}h_{rd}/h_{rs}$ is unknown. According to (12), we have

$$f^0_{Y_2|X}(t|\mathsf{x}; \mathbf{g}) = \frac{1}{\pi\sigma_2^2} \exp(-\frac{\|t - g_{sr}g_{rd}\mathsf{x}\|^2}{\sigma_2^2}), \tag{54}$$

where $\sigma_2^2 = g_{rd}^2\sigma_{sr}^2 + \sigma_{rd}^2$.

Let us now consider $T_2^c$. For any $y_1$ and $y_2$, we obtain that

$$f_{Y_2|Y_1}(y_2|y_1; f_{V|U}, \mathbf{h}) = f^0_{Y_2|Y_1}(y_2|y_1; \mathbf{g}). \tag{55}$$

Furthermore, since $(Y_1, X, Y_2)$ forms a Markov chain as $Y_1 \to X \to Y_2$, we have

$$\sum_{\mathsf{x}\in\mathcal{X}} \Pr(X = \mathsf{x}|Y_1 = y_1) f_{Y_2|X}(Y_2 = y_2|X = \mathsf{x}; f_{V|U}, \mathbf{h}) = \sum_{\mathsf{x}\in\mathcal{X}} \Pr(X = \mathsf{x}|Y_1 = y_1) f^0_{Y_2|X}(Y_2 = y_2|X = \mathsf{x}; \mathbf{g}). \tag{56}$$

Note that $\Pr(X=\mathsf{x}|Y_1=y_1)$ in (56) can be written as

$$\begin{aligned}&\Pr(X=\mathsf{x}|Y_1=y_1)\\&=\frac{\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}{\Pr(Y_1=y_1)}\\&=\frac{\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}{\sum_{\mathsf{x}\in\mathcal{X}}\Pr(Y_1=y_1|X=\mathsf{x})\Pr(X=\mathsf{x})}\\&=\frac{1}{1+\sum_{X\neq\mathsf{x}}\exp(y_1h_{sd}(x-\mathsf{x})/\sigma_{sd}^2)}.\end{aligned}\tag{57}$$

Without loss of generality, we consider $X\in(-1,+1)$. If $\mathsf{x}=+1$, it is easy to show that $\Pr(X=\mathsf{x}|Y_1=y_1)$ becomes very small when $y_1$ is far less than 0. When $y_1\to-\infty$, we have $\lim_{y_1\to-\infty}\Pr(X=\mathsf{x}|Y_1=y_1)=0$ and $\lim_{y_1\to-\infty}\Pr(X\neq\mathsf{x}|Y_1=y_1)=1$. Therefore, (56) can be reduced to

$$\begin{aligned}&f_{Y_2|X}(Y_2=y_2|X=\mathsf{x};f_{V|U},\mathbf{h})\\&\qquad=f^0_{Y_2|X}(Y_2=y_2|X=\mathsf{x};\mathbf{g}).\end{aligned}\tag{58}$$

Substituting (53) and (54) into (58), we can obtain $K=g_{rd}$, which means that $f_{Y_2|X}(y_2|\mathsf{x};f_{V|U},\mathbf{h})$ can be expressed only by the known unreliable CSI.

In addition, since the direct link S-D and the relay link S-R-D are independent of each other, we have

$$\begin{aligned}&f_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};f_{V|U},\mathbf{h}))\\&=f_{Y_1|X}(y_1|\mathsf{x})f_{Y_2|X}(y_2|\mathsf{x};f_{V|U},\mathbf{h})\\&=\frac{1}{\pi(K^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}K\mathsf{x}\|^2}{K^2\sigma_{sr}^2+\sigma_{rd}^2}\right)\\&=\frac{1}{\pi(g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}g_{rd}\mathsf{x}\|^2}{g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2}\right).\end{aligned}\tag{59}$$

On the other hand, according to (54), we have

$$\begin{aligned}&f^0_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};\mathbf{g})\\&=f_{Y_1|X}(y_1|\mathsf{x})f^0_{Y_2|X}(y_2|\mathsf{x};\mathbf{g})\\&=\frac{1}{\pi(g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2)\sigma_{sd}^2}\\&\quad\times\exp\left(-\frac{\|y_1-h_{sd}\mathsf{x}\|^2}{\sigma_{sd}^2}-\frac{\|y_2-g_{sr}g_{rd}\mathsf{x}\|^2}{g_{rd}^2\sigma_{sr}^2+\sigma_{rd}^2}\right).\end{aligned}\tag{60}$$

From (59) and (60), we see that $f_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};f_{V|U},\mathbf{h})=f^0_{Y_1,Y_2|X}(y_1,y_2|\mathsf{x};\mathbf{g})$, which completes the proof of Proposition 2. ■

## REFERENCES


[1] N. Yang, L. Wang, G. Geraci, M. Elkashlan, J. Yuan, and M. Di Renzo, "Safeguarding 5G wireless communication networks using physical layer security," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 20–27, Apr. 2015.

[2] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity. Part I. System description," *IEEE Trans. Commn.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[3] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[4] R. U. Nabar, H. Bolcskei, and F. W. Kneubuhler, "Fading relay channels: Performance limits and space-time signal design," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 1099–1109, Aug. 2004.

[5] S. Dehnie, H. T. Sencar, and N. Memon, "Cooperative diversity in the presence of a misbehaving relay: Performance analysis," in *Proc. IEEE Sarnoff Symp.*, Princeton, NJ, USA, May 2007, pp. 1–7.

[6] P. Papadimitratos and Z. J. Haas, "Secure data communication in mobile ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 2, pp. 343–356, Feb. 2006.

[7] S. Capkun, L. Buttyan, and J. P. Hubaux, "Self-organized public-key management for mobile ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 2, no. 1, pp. 52–64, Jan. 2003.

[8] Y.-C. Hu, A. Perrig, and D. B. Johnson, "Ariadne: A secure on-demand routing protocol for ad hoc networks," *Wireless Netw.*, vol. 11, nos. 1–2, pp. 21–38, Jan. 2005.

[9] Y. Mao and M. Wu, "Tracing malicious relays in cooperative wireless communications," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 2, pp. 198–212, Jun. 2007.

[10] L.-C. Lo and W.-J. Huang, "Misbehavior detection without channel information in cooperative networks," in *Proc. IEEE 74th Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.

[11] L.-C. Lo, Z.-J. Wang, and W.-J. Huang, "Noncoherent misbehavior detection in space-time coded cooperative networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 3061–3064.

[12] L.-C. Lo, W.-J. Huang, R. Y. Chang, and W.-H. Chung, "Noncoherent detection of misbehaving relays in decode-and-forward cooperative networks," *IEEE Commun. Lett.*, vol. 19, no. 9, pp. 1536–1539, Sep. 2015.

[13] W. Hou, X. Wang, and A. Refaey, "Misbehavior detection in amplify-and-forward cooperative OFDM systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Budapest, Hungary, Jun. 2013, pp. 5345–5349.

[14] S. W. Kim, "Physical integrity check in cooperative relay communications," *IEEE Trans. Wireless commun.*, vol. 14, no. 11, pp. 6401–6413, Nov. 2015.

[15] S. Dehnie, H. T. Sencar, and N. Memon, "Detecting malicious behavior in cooperative diversity," in *Proc. 41st IEEE Annu. Conf. Inf. Sci. Syst.*, Baltimore, MD, USA, Mar. 2007, pp. 895–899.

[16] E. Graves and T. F. Wong, "Detection of channel degradation attack by intermediary node in linear networks," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 747–755.

[17] R. Cao, E. Graves, T. F. Wong, and T. Lv, "Detecting substitution attacks against non-colluding relays," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 1856–1861.

[18] R. Cao, S. Huang, and Y. Lu, "Detecting and tracing i.i.d. Attacks in networks with any number of relays," *IEEE Access*, vol. 4, pp. 6757–6765, Oct. 2016.

[19] J. K. Tugnait, "Self-contamination for detection of pilot contamination attack in multiple antenna systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 5, pp. 525–528, Oct. 2015.

[20] K. P. Peppas, G. C. Alexandropoulos, and P. T. Mathiopoulos, "Performance analysis of dual-hop AF relaying systems over mixed $\eta-\mu$ and $\kappa-\mu$ fading channels," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3149–3163, Sep. 2013.

[21] J. N. Laneman and G. W. Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Chicago, IL, USA, Sep. 2000, pp. 7–12.

[22] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.

[23] R. Pabst *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Wireless Commun. Mag.*, vol. 42, no. 9, pp. 80–89, Sep. 2004.

[24] B. Zafar, S. Gherekhloo, and M. Haardt, "Analysis of multihop relaying networks: Communication between range-limited and cooperative nodes," *IEEE Veh. Technol. Mag.*, vol. 7, no. 3, pp. 40–47, Sep. 2012.

[25] N. Zlatanov, A. Ikhlef, T. Islam, and R. Schober, "Buffer-aided cooperative communications: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 146–153, Apr. 2014.

**Tiejun Lv** (M'08–SM'12) received the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 1997 and 2000, respectively. From 2001 to 2002, he was a Post-Doctoral Fellow with Tsinghua University, Beijing, China. In 2005, he was promoted to a Full Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. From 2008 to 2009, he was a Visiting Professor with the Department of Electrical Engineering, Stanford University, Stanford, CA, USA. He has authored over 50 published IEEE journal papers and 170 conference papers on the physical layer of wireless mobile communications. His current research interests include signal processing, communications theory, and networking. He was a recipient of the Program for New Century Excellent Talents in University Award from the Ministry of Education, China, in 2006. He received the Nature Science Award from the Ministry of Education of China for the hierarchical cooperative communication theory and technologies in 2015.

**Yajun Yin** received the B.Eng. degree in electronic and information engineering from the Harbin Institute of Technology at Weihai, China, in 2015. He is currently pursuing the M.S. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include physical-layer security and Web anti-spam.

**Yueming Lu** received the B.S. and M.S. degrees in computer science from the Xi'an University of Architecture and Technology in 1994 and 1997, respectively, and the Ph.D. degree in computer architecture from Xi'an Jiaotong University in 2000. He was a Researcher of Lucent from 2000 to 2003. He is currently a Professor with the Beijing University of Posts and Telecommunications. His research interests include network design, network security, and distributed computing.

**Shaoshi Yang** (S'09–M'13) received the B.Eng. degree in information engineering from the Beijing University of Posts and Telecommunications, China, in 2006, and the Ph.D. degree in electronics and electrical engineering from the University of Southampton, U.K., in 2013. From 2008 to 2009, he was an Intern Research Fellow of Intel Labs China, where he was involved in the mobile WiMAX standardization. From 2013 to 2016, he was a Research Fellow with the School of Electronics and Computer Science, University of Southampton. He is currently a Principal Engineer with Huawei Technologies Co., Ltd., China. He is also a member of the Isaac Newton Institute for Mathematical Sciences, Cambridge University. His research interests include high-dimensional signal processing for communications, green radio, wireless video transmission, cross-layer system design, mathematical optimization and its applications. He was recognized by the prestigious National 1000-Young-Talent Fellowship of China and the Dean's Award for Early Career Research Excellence at the University of Southampton. He was a Guest Associate Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (http://shaoshiyang.weebly.com/).

**Enjie Liu** received the Ph.D. degree in telecommunications from the Queen Mary University of London, London, U.K., in 2002. She is currently a Reader in network applications with the School of Computer Science and Technology, University of Bedfordshire, U.K. Her current research interests include power management at protocol stack for mobile terminals, MAC and network layer protocol design, vehicular ad hoc networks, small-cell propagation modeling, and Internet of Things protocols and applications.

**Gordon Clapworthy** received the B.Sc. degree (Hons.) in mathematics from the University of London, the M.Sc. degree (Hons.) in computer science from the City, University of London, and the Ph.D. degree in aeronautical engineering from the University of London. He is currently a Professor of computer graphics with the School of Computer Science and Technology, University of Bedfordshire, U.K. He has published over 300 refereed papers and has been a Principal Investigator in 28 European projects, being a Project Coordinator in eight of them. His previous interests have included computer animation, biomechanics, space robots, transonic aerodynamics, virtual reality, surface modeling, and fundamental computer graphics algorithms, but his most recent work has focused on the technological aspects of next-generation medical systems.