Binary Hypothesis Testing With Byzantine Sensors: Fundamental Tradeoff Between Security and Efficiency

Xiaoqiang Ren, Jiaqi Yan, and Yilin Mo

***Abstract*—This paper studies binary hypothesis testing based on measurements from a set of sensors, a subset of which can be compromised by an attacker. The measurements from a compromised sensor can be manipulated arbitrarily by the adversary. The asymptotic exponential rate, with which the probability of error goes to zero, is adopted to indicate the detection performance of a detector. In practice, we expect the attack on sensors to be sporadic, and therefore the system may operate with all the sensors being benign for an extended period of time. This motivates us to consider the tradeoff between the detection performance of a detector, i.e., the probability of error, when the attacker is absent (defined as efficiency) and the worst case detection performance when the attacker is present (defined as security). We first provide the fundamental limits of this tradeoff, and then propose a detection strategy that achieves these limits. We then consider a special case, where there is no tradeoff between security and efficiency. In other words, our detection strategy can achieve the maximal efficiency and the maximal security simultaneously. Two extensions of the secure hypothesis testing problem are also studied and fundamental limits and achievability results are provided: first, a subset of sensors, namely "secure" sensors, are assumed to be equipped with better security countermeasures and hence are guaranteed to be benign; and second, detection performance with unknown number of compromised sensors. Numerical examples are given to illustrate the main results.**

***Index Terms*—Hypothesis testing, security, secure detection, efficiency, trade-off, Byzantine attacks, fundamental limits.**

I. Introduction

BACKGROUND AND MOTIVATIONS: Network embedded sensors, which are pervasively used to monitor the system, are vulnerable to malicious attacks due to their limited capacity and sparsely spatial deployment. An attacker may get access to the sensors and send arbitrary messages, or break the communication channels between the sensors and the system operator to tamper with the transmitted data. Such integrity attacks have motivated many researches on how to infer useful information from corrupted sensory data in a secure manner [1]–[3]. In this paper, we follow this direction but with the focus on the trade-off between the performance of the inference algorithm when the attacker is absent and the "worst-case" performance when the attacker, which has the knowledge of the inference algorithm, is present. We define two metrics, *efficiency* and *security*, to characterize the performance of the hypothesis testing algorithm (or detector) under the two scenarios respectively and analyze the trade-off between security and efficiency.

Our Work and its Contributions: We consider the sequential binary hypothesis testing based on the measurements from m sensors. It is assumed that n out of these m sensors may be compromised by an attacker, the set of which is chosen by the attacker and fixed over time. The adversary can manipulate the measurements sent by the compromised sensors arbitrarily. According to Kerckhoffs's principle [4], i.e., the security of a system should not rely on its obscurity, we assume that the adversary knows exactly the hypothesis testing algorithm used by the fusion center. On the other hand, the fusion center (i.e., the system operator) only knows the number of malicious sensors n, but does not know the exact set of the compromised sensors.

At each time k, the fusion center needs to make a decision about the underlying hypothesis based on the possibly corrupted measurements collected from all sensors until time k. Given a hypothesis testing algorithm at the fusion center (i.e., a measurements fusion rule), the worst-case probability of error is investigated, and the asymptotic exponential decay rate of the error, which we denote as the "security" of the system, is adopted to indicate the detection performance. On the other hand, when the attacker is absent, the detection performance of a hypothesis testing algorithm, i.e., the asymptotic exponential decay rate of the error probability, is denoted by the "efficiency".

We focus on the trade-off between efficiency and security. In particular, we are interested in characterizing the fundamental limits of the trade-offs between efficiency and security and the detectors that achieve these limits.

The main contributions of this work are summarized as follows:

1) To the best of our knowledge, this is the first work that studies the trade-off between the efficiency and security of any inference algorithm.
2) With mild assumptions on the probability distributions of the measurements, we provide the fundamental limits of

Manuscript received July 27, 2017; revised November 25, 2017; accepted December 15, 2017. Date of publication January 1, 2018; date of current version February 1, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chandra Murthy. *(Corresponding author: Yilin Mo.)*

The authors are with the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798 (e-mail: xren@ntu.edu.sg; jyan004@e.ntu.edu.sg; ylmo@ntu.edu.sg).

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/TSP.2017.2788420

the trade-off between the efficiency and security (Corollaries 1 and 2). Furthermore, we present detectors, with low computational complexity, that achieve these limits (Theorem 4). Therefore, the system operator can easily adopt the detectors we proposed to obtain the best trade-off between efficiency and security. Interestingly, in some cases, e.g., Gaussian random variables with same variance and different mean, the maximal efficiency and the maximal security can be achieved simultaneously (Theorem 5).

3) Similar results, i.e., the fundamental limits of the trade-off and the detectors that possess these limits, are established with several different problem settings (Section V). This shows that our analysis techniques are insightful and may be helpful for the future related studies.

Related Literature: A sensor is referred to as a Byzantine sensor if its messages to the fusion center are fully controlled by an adversary.[1] Recently, detection with Byzantine sensors has been studied extensively in [5]–[14], among which [5]–[7] took the perspective of an attacker and aimed to find the most effective attack strategy, [8]–[11] focused on designs of resilient detectors, and [12]–[14] formulated the problem in a game-theoretic way. The main results of [5] are the critical fraction of Byzantine sensors that blinds the fusion center, which is just the counterpart of the breakdown point in robust statistics [15], and the most effective attack strategy that minimizes the asymptotic error exponent in the Neyman-Pearson setting, i.e., the Kullback–Leibler (K–L) divergence. Since the Byzantine sensors are assumed to generate independent and identical distributed (i.i.d.) data, the resulting measurements with minimum K–L divergence and the corresponding robust detector coincide with those in [16]. Similar results were obtained in [6], [7] by considering non-asymptotic probability of error in Bayesian setting and asymptotic Bayesian performance metric, i.e., Chernoff information, respectively. The authors in [8] focused on computation efficient algorithms to determine optimal parameters of the q-out-of-m procedure [17] in large scale networks for different fractions of Byzantine sensors. More than two types of sensors were assumed in [9], [10]. The authors thereof proposed a maximum likelihood procedure, which is based on the iterative expectation maximization (EM) algorithm [18], simultaneously classifying the sensor nodes and performing the hypothesis testing. The authors in [11] showed that the optimal detector is of a threshold structure when the fraction of Byzantine sensors is less than 0.5. A zero-sum game was formulated in each of [12]–[14], among which a closed-form equilibrium point of attack strategy and detector was obtained in [14], computation efficient and nearly optimal equilibrium point (exact equilibrium point only in certain cases) was obtained in [12], and numerical simulations were used to study the equilibrium point in [13].

While in [5]–[10], [13] the Byzantine sensors are assumed to generate malicious data independently, this work, as in [11], [12], [14], assumes that the Byzantine sensors may collude with each other. The collusion model is more reasonable since the attacker is malicious and will arbitrarily change the messages of the sensors it controlls. Notice also that compared to the independence model, the collusion model complicates the analysis significantly. Unlike [6]–[10], [12], [13], where the sensors only send binary messages, this work, as in [5], [11], [14], assumes that the measurements of a benign sensor can take any value. Since the binary message model simplifies the structure of corrupted measurements, and, hence, implicitly limits the capability of an attacker, it is easier to be dealt with. This work differs from [11], [14] as follows. The authors in [11] focused on one time step scenario. The analysis is thus fundamentally different and more challenging. On the contrary, in this work the hypothesis testing is performed sequentially and an asymptotic regime performance metric, i.e., the Chernoff information, is concerned. A similar setting as in this work was considered in our recent work [14]. However, [14] focused on the equilibrium point. The performance (i.e., the security and efficiency) of the obtained equilibrium detection rule is merely one point of the admissible set that will be characterized in this paper.

Finally, we should remark that the aforementioned literature mainly focuses on designing algorithms in adversarial environment. However, those algorithms may perform poorly in the absence of the adversary comparing to the classic Neyman-Pearson detector or Naive Bayes detector. A fundamental question, which we seek to answer in this paper, is that how to design a detection strategy which performs "optimally" regardless of whether the attacker is present.

Organization: In Section II, we formulate the problem of binary hypothesis testing in adversarial environments, in which the attack model, the performance indices and the notion of the efficiency and security are defined. For the sake of completeness, we give a brief introduction the large deviation theory in Section III, which is a key supporting technique for the later analysis. The main results are presented in Section IV. We first provide the fundamental limits of the trade-off between the efficiency and security. We then propose detectors that achieve these limits. At last, we show that the maximal efficiency and the maximal security can be achieved simultaneously in some cases. Two extensions are investigated in Section V. After providing numerical examples in Section VI, we conclude the paper in Section VII.

Notations: $\mathbb{R}$ ($\mathbb{R}_+$) is the set of (nonnegative) real numbers. $\mathbb{Z}_+$ is the set of positive integers. The cardinality of a finite set $\mathcal{I}$ is denoted as $|\mathcal{I}|$. For a set $\mathcal{A} \in \mathbb{R}^n$, $\mathrm{int}(\mathcal{A})$ denotes its interior. For any sequence $\{x(k)\}_{k=1}^{\infty}$, we denote its average at time k as $\bar{x}(k) \triangleq \sum_{t=1}^{k} x(t)/k$. For a vector $\mathbf{x} \in \mathbb{R}^n$, the support of $\mathbf{x}$, denoted by $\mathrm{supp}(\mathbf{x})$, is the set of indices of nonzero elements:

$$\mathrm{supp}(\mathbf{x}) \triangleq \{i \in \{1, 2, \ldots, n\} : \mathbf{x}_i \neq 0\}.$$

II. Problem Formulation

Consider the problem of detecting a binary state $\theta \in \{0,\ 1\}$ using m sensors' measurements. Define the measurement $\mathbf{y}(k)$ at time k to be a row vector:

$$\mathbf{y}(k) \triangleq \begin{bmatrix} y_1(k) & y_2(k) & \cdots & y_m(k) \end{bmatrix} \in \mathbb{R}^m, \qquad (1)$$

[1] In practice, to manipulate the data of a sensor, an adversary may attack the sensor node itself or break the communication channel between the sensor and the fusion center. In this paper, we do not distinguish these two approaches.

where $y_i(k)$ is the scalar measurement from sensor i at time k. For simplicity, we define $\mathbf{Y}(k)$ as a vector of all measurements from time 1 to time k:

$$\mathbf{Y}(k) \triangleq \begin{bmatrix} \mathbf{y}(1) & \mathbf{y}(2) & \cdots & \mathbf{y}(k) \end{bmatrix} \in \mathbb{R}^{mk}. \tag{2}$$

Given θ, we assume that all measurements $\{y_i(k)\}_{i=1,\ldots,m,\ k=1,2,\ldots}$ are independent and identically distributed (i.i.d.). The probability measure generated by $y_i(k)$ is denoted as ν when $\theta = 0$ and it is denoted as μ when $\theta = 1$. In other words, for any Borel-measurable set $\mathcal{A} \subseteq \mathbb{R}$, the probability that $y_i(k) \in \mathcal{A}$ equals $\nu(\mathcal{A})$ when $\theta = 0$ and equals $\mu(\mathcal{A})$ when $\theta = 1$. We denote the probability space generated by all measurements $\mathbf{y}(1),\ \mathbf{y}(2),\ \ldots$ as $(\Omega_y,\ \mathcal{F}_y,\ \mathbb{P}_\theta^o)$,[2] where for any $l \geq 1$

$$\mathbb{P}_\theta^o(y_{i_1}(k_1) \in \mathcal{A}_1, \ldots, y_{i_l}(k_l) \in \mathcal{A}_l) = \begin{cases} \nu(\mathcal{A}_1)\nu(\mathcal{A}_2)\ldots\nu(\mathcal{A}_l) & \text{if } \theta = 0 \\ \mu(\mathcal{A}_1)\mu(\mathcal{A}_2)\ldots\mu(\mathcal{A}_l) & \text{if } \theta = 1 \end{cases},$$

when $(i_j, k_j) \neq (i_{j'}, k_{j'})$ for all $j \neq j'$. The expectation taken with respect to $\mathbb{P}_\theta^o$ is denoted by $\mathbb{E}_\theta^o$. We further assume that ν and μ are absolutely continuous with respect to each other. Hence, the log-likelihood ratio $\lambda : \mathbb{R} \to \mathbb{R}$ of $y_i(k)$ is well defined as

$$\lambda(y_i) \triangleq \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y_i)\right), \tag{3}$$

where $d\mu/d\nu$ is the Radon-Nikodym derivative.

We define $f_k : \mathbb{R}^{mk} \to [0,\ 1]$, the detector at time k, as a mapping from the measurement space $\mathbf{Y}(k)$ to the interval $[0,\ 1]$. When $f_k(\mathbf{Y}(k)) = 0$, the system makes a decision $\hat{\theta} = 0$, and when $f_k(\mathbf{Y}(k)) = 1, \hat{\theta} = 1$. When $f_k(\mathbf{Y}(k)) = \gamma \in (0,\ 1)$, the system then "flips a biased coin" to choose $\hat{\theta} = 1$ with probability γ and $\hat{\theta} = 0$ with probability $1 - \gamma$. The system's strategy $f \triangleq (f_1,\ f_2,\ \cdots)$ is defined as an infinite sequence of detectors from time 1 to ∞.

A. Attack Model

Let the (*manipulated*) measurements received by the fusion center at time k be

$$\mathbf{y}'(k) = \mathbf{y}(k) + \mathbf{y}^a(k), \tag{4}$$

where $\mathbf{y}^a(k) \in \mathbb{R}^m$ is the bias vector injected by the attacker at time k. In the following, Assumptions 1–3 are made on the attacker, among which Assumption 1 is in essence the only limitation we pose.

Assumption 1 (Spare Attack): There exists an index set $\mathcal{I} \subset \mathcal{M} \triangleq \{1, 2, \ldots, m\}$ with $|\mathcal{I}| = n$ such that $\bigcup_{k=1}^{\infty} \mathrm{supp}\,(\mathbf{y}^a(k)) = \mathcal{I}$. Furthermore, the system knows the number n, but it does not know the set $\mathcal{I}$.

We should remark that the above assumption does not pose any restrictions on the value of $y_i^a(k)$ if sensor i is compromised at time k, i.e., the bias injected into the data of a compromised sensor can be arbitrary.

[2]The superscript "o" stands for original, which is contrasted with corrupted measurements.

Assumption 1 says that the attacker can compromise up to n out of m sensors at each time. It is practical to assume that the attacker possesses limited resources, i.e., the number of compromised sensors is (non-trivially) upper bounded, since otherwise it would be too pessimistic and the problem becomes trivial. The quantity n might be determined by the *a priori* knowledge about the quality of each sensor. Alternatively, the quantity n may be viewed as a design parameter, which indicates the resilience level that the system is willing to introduce; the details of which are in Remark 1. Notice also that since the worst-case attacks (over the set of compromised sensors and the attack strategy) are concerned (the performance metric will be introduced shortly), it is equivalent to replace the cardinality requirement $|\mathcal{I}| = n$ by $|\mathcal{I}| \leq n$. We should note that in [6], [8], [12], it was also assumed that the number/fraction of malicious sensor nodes is known to the system operator.

Moreover, the set of compromised sensors is assumed to be fixed over time. Notice that if we assume that the set of compromised sensors has a fixed cardinality but is time-varying, i.e., there exists no a set like $\mathcal{I}$ to bound the compromised sensors, the attacker would be required to abandon the sensor nodes it has compromised, which is not sensible. Notice that in [8]–[10], it was assumed the set of malicious/misbehaving sensors is fixed as well. We should also note that though this work is concerned with asymptotic performances (i.e., the security and efficiency introduced later), the numerical simulations in Section VI show that our algorithm indeed perform quite well in an non-asymptotic setup. Actually, if a finite-time horizon problem is considered and the time required for an attacker to control a benign sensor is large enough, then it is reasonable to assume that the set of compromised sensors is fixed.

In fact, the exactly same sparse attack model as in Assumption 1 has been widely adopted by literature dealing with Byzantine sensors, e.g., state estimation [19], [20], and quickest change detection [21].

Finally, we should note that we do not assume any pattern of the bias $y_i^a(k)$ for $i \in \mathcal{I}$, i.e., the malicious bias injected may be correlated across the compromised sensors and correlated over time. Compared to the independence assumption in [5]–[10], [13], our assumption improves the effectiveness of the attacker and is more realistic in the sense that the attacker is malicious and will do whatever it wants.

Remark 1: The parameter n can also be interpreted as how many bad sensors the system can and is willing to tolerate, which is a design parameter for the system operator. In general, increasing n will increase the resilience of the detector under attack. However, as is shown in the rest of the paper, a larger n may result in more conservative design and is likely to cause a performance degradation during normal operation when no sensor is compromised.

Assumption 2 (Model Knowledge): The attacker knows the probability measure μ and ν and the true state θ.

By the knowledge about the sensor, the attacker can develop the probability measure μ and ν. To obtain the true state, the attacker may deploy its own sensor network. Though it might be difficult to satisfy in practice, this assumption is in fact conventional in literature concerning the worst-case attacks, e.g., [5],

[21]. Nevertheless, this assumption is in accordance with the Kerckhoffs's principle.

Assumption 3 (Measurement Knowledge): At time k, the attacker knows the current and all the historical measurements available at the compromised sensors.

Since the attacker knows the true measurement of a compromised sensor i, $y_i(k)$, it may set the fake measurement arrived at the fusion center $y_i'(k)$ to any value it wants by injecting $y_i^a(k)$. One may also verify that all the results in this paper remain even if the attacker is "strong" enough where at time k, it knows measurements from all the sensors $\mathbf{Y}(k)$.

An admissible attack strategy is any causal mapping from the attacker's available information to a bias vector that satisfies Assumption 1. This is formalized as follows. Let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$. Define the true measurements of the compromised sensors from time 1 to k as

$$\mathbf{Y}_{\mathcal{I}}(k) \triangleq \begin{bmatrix} \mathbf{y}_{\mathcal{I}}(1) & \mathbf{y}_{\mathcal{I}}(2) & \cdots & \mathbf{y}_{\mathcal{I}}(k) \end{bmatrix} \in \mathbb{R}^{|\mathcal{I}|k}$$

with

$$\mathbf{y}_{\mathcal{I}}(k) \triangleq \begin{bmatrix} y_{i_1}(k) & y_{i_2}(k) & \cdots & y_{i_n}(k) \end{bmatrix} \in \mathbb{R}^{|\mathcal{I}|}.$$

Similar to $\mathbf{Y}(k)$, $\mathbf{Y}'(k)$ ($\mathbf{Y}^a(k)$) is defined as all the manipulated (bias vector) from time 1 to k. The bias $\mathbf{y}^a(k)$ is chosen as a function of the attacker's available information at time k:

$$\mathbf{y}^a(k) \triangleq g(\mathbf{Y}_{\mathcal{I}}(k), \mathbf{Y}^a(k-1), \mathcal{I}, \theta, k), \tag{5}$$

where g is a function[3] of $\mathbf{Y}_{\mathcal{I}}(k), \mathbf{Y}^a(k-1), \mathcal{I}, \theta, k$ such that $\mathbf{y}^a(k)$ satisfies Assumption 1. We denote g as an admissible attacker's strategy. Notice that since time k is an input variable and the available measurements $\mathbf{Y}_{\mathcal{I}}(k), \mathbf{Y}^a(k-1)$ are "increasing" with respect to time k, the definition in (5) does not exclude the time-varying attack strategy. Denote the probability space generated by all manipulated measurements $\mathbf{y}'(1)$, $\mathbf{y}'(2)$, ... as $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$. The expectation taken with respect to the probability measure $\mathbb{P}_\theta$ is denoted by $\mathbb{E}_\theta$.

B. Asymptotic Detection Performance

Given the strategy of the system and the attacker, the probability of error at time k can be defined as

$$e(\theta, \mathcal{I}, k) \triangleq \begin{cases} \mathbb{E}_0 f_k(\mathbf{Y}'(k)) & \text{when } \theta = 0, \\ 1 - \mathbb{E}_1 f_k(\mathbf{Y}'(k)) & \text{when } \theta = 1. \end{cases} \tag{6}$$

Notice that f_k could take any value from [0,1]. Hence, the expected value of f_k is used to compute the probability of error. In this paper, we are concerned with the worst-case scenario. As a result, let us define

$$\epsilon(k) \triangleq \max_{\theta=0,1,|\mathcal{I}|=n} e(\theta, \mathcal{I}, k). \tag{7}$$

In other words, $\epsilon(k)$ indicates the worst-case probability of error considering all possible sets of compromised sensors and the state θ given the detection rule f and attack strategy g. Notice also that in accordance with Assumption 1, the set $\mathcal{I}$ in the above equation is fixed over time.

Ideally, for each time k, the system wants to design a detector f_k to minimize $\epsilon(k)$. However, such a task can hardly be accomplished analytically since the computation of the probability of error usually involves numerical integration. Thus, in this article, we consider the asymptotic detection performance in hope to provide more insight on the detector design. Define the rate function as

$$\rho \triangleq \liminf_{k\to\infty} -\frac{\log \epsilon(k)}{k}. \tag{8}$$

Clearly, ρ is a function of both the system strategy f and the attacker's strategy g. As such, we will write ρ as $\rho(f, g)$ to indicate such relations. Since ρ indicates the rate with which the probability of error goes to zero, the system would like to maximize ρ in order to minimize the detection error. On the contrary, the attacker wants to decrease ρ to increase the detection error.

C. Interested Problems

In practice, the attacker may not be present consistently. As a result, the system will operate for an extended period of time with all sensors being benign. Thus, a natural question arises: is there any detection rule that has "decent" performance regardless of the presence of the attacker? Or is there a fundamental trade-off between security and efficiency? In other words, a detector that is "good" in the presence of an adversary will be "bad" in a benign environment. This paper is devoted to answering this question.

Informally, the performance of a detection rule when there is no attacker at all is referred to by "efficiency", while the performance when the worst-case attacker (provided that the attacker knows the detection rule used by the system) is present is referred to by "security". Mathematically speaking, given a system strategy f, denote by $\mathcal{E}(f)$ and $\mathcal{S}(f)$ its efficiency and security respectively, which are formalized as follows:

$$\mathcal{E}(f) \triangleq \rho(f, g = \mathbf{0}), \tag{9}$$

$$\mathcal{S}(f) \triangleq \inf_g \rho(f, g) \tag{10}$$

where $\mathbf{0} \in \mathbb{R}^m$ is the zero vector.

III. Preliminary: Large Deviation Theory

In this section, we introduce the large deviation theory, which is a key supporting technique of this paper.

To proceed, we first introduce some definitions. Let $M_\omega(w) \triangleq \int_{\mathbb{R}^d} e^{w\cdot X} d\omega(X), w \in \mathbb{R}^d$ be the moment generating function for the random vector $X \in \mathbb{R}^d$ that has the probability measure ω, where $w \cdot X$ is the dot product. Let $\mathrm{dom}_\omega \triangleq \{w \in \mathbb{R}^d | M_\omega(w) < \infty\}$ be the support such that $M_\omega(w)$ is finite. Define the Fenchel–Legendre transform of the function $\log M_\omega(w)$ as

$$I_\omega(x) = \sup_{w\in\mathbb{R}^d} \{x \cdot w - \log M_\omega(w)\},\ x \in \mathbb{R}^d. \tag{11}$$

[3]The function g is possibly random. For example, given the available information, the adversary can flip a coin to decide whether to change the measurement or not.

Theorem 1 (Multidimensional Cramér's Theorem [22]): Suppose $X(1),\dots,X(k),\dots$ be a sequence of i.i.d. random vectors and $X(k)\in\mathbb{R}^d$ has the probability measure ω. Let $\overline{X}(k)\triangleq\sum_{t=1}^{k}X(t)/k, k\in\mathbb{Z}_+$ be the empirical mean. Then if $0\in\mathrm{int}(\mathrm{dom}_\omega)$, the probability $\mathbb{P}(\overline{X}(k)\in\mathcal{A})$ with $\mathcal{A}\subseteq\mathbb{R}^d$ satisfies the large deviation principle, i.e.,

1) if $\mathcal{A}$ is closed,

$$\limsup_{k\to\infty}\frac{1}{k}\log\mathbb{P}(\overline{X}(k)\in\mathcal{A})\le-\inf_{x\in\mathcal{A}}I_\omega(x).$$

2) if $\mathcal{A}$ is open,

$$\liminf_{k\to\infty}\frac{1}{k}\log\mathbb{P}(\overline{X}(k)\in\mathcal{A})\ge-\inf_{x\in\mathcal{A}}I_\omega(x).$$

IV. Main Results

A. Technical Preliminaries

Denote the moment generating function of the log-likelihood ratio λ under each hypothesis as:

$$M_0(w)\triangleq\int_{y=-\infty}^{\infty}\exp(w\lambda(y))\mathrm{d}\nu(y),\tag{12}$$

$$M_1(w)\triangleq\int_{y=-\infty}^{\infty}\exp(w\lambda(y))\mathrm{d}\mu(y).\tag{13}$$

Furthermore, define dom_0 as the region where $M_0(w)$ is finite and $I_0(x)$ as the Fenchel–Legendre transform of $\log M_0(w)$. The quantities $M_1(w)$, dom_1 and $I_1(x)$ are defined similarly.

Denote the the Kullback-Leibler (K–L) divergences by

$$D(1\|0)\triangleq\int_{y=-\infty}^{\infty}\lambda(y)\mathrm{d}\mu,\ D(0\|1)\triangleq-\int_{y=-\infty}^{\infty}\lambda(y)\mathrm{d}\nu.$$

To apply the multidimensional Cramér's Theorem and avoid degenerate problems, we adopt the following assumptions:

Assumption 4: $0\in\mathrm{int}(\mathrm{dom}_0)$ and $0\in\mathrm{int}(\mathrm{dom}_1)$.

Assumption 5: The K–L divergences are well-defined, i.e., $0<D(1\|0),D(0\|1)<\infty$.

With the above assumptions, we have the following properties of $I_0(x)$ and $I_1(x)$. the proof of which is provided in Appendix A.

Theorem 2: Under Assumptions 4 and 5, the followings hold:

1) $I_0(x)$ $(I_1(x))$ is twice differentiable, strictly convex and strictly increasing (strictly decreasing) on $[-D(0\|1),D(1\|0)]$.
2) The following equalities hold:

$$I_1(D(1\|0))=0,\tag{14}$$

$$I_0(D(1\|0))=D(1\|0),\tag{15}$$

$$I_0(-D(0\|1))=0,\tag{16}$$

$$I_1(-D(0\|1))=D(0\|1).\tag{17}$$

$$I_0(0)=I_1(0).\tag{18}$$

Since $I_0(0)=I_1(0)$, let us define

$$C\triangleq I_0(0).\tag{19}$$

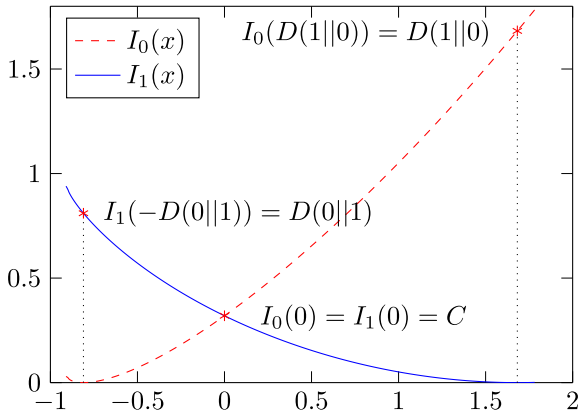


Fig. 1. Illustration of $I_0(x)$ and $I_1(x)$. The figure is plotted by assuming $y_1(1)$ to be Bernoulli distributed under both hypotheses with $\mathbb{P}_0^o(y_1(1)=1)=0.02$ and $\mathbb{P}_1^o(y_1(1)=1)=0.6$.

To make the presentation clear, we illustrate $I_0(x)$ and $I_1(x)$ in Fig. 1.

The "inverse functions" of $I_0(x)$ and $I_1(x)$ are defined as follows: for $z\ge0$,

$$I_0^{-1}(z)=\max\{x\in\mathbb{R}:I_0(x)=z\},$$

$$I_1^{-1}(z)=\min\{x\in\mathbb{R}:I_1(x)=z\}.$$

Let $D_{\min}\triangleq\min\{D(0\|1),D(1\|0)\}$. We further define $h(z):(0,(m-n)D_{\min})\mapsto(0,(m-n)D_{\min})$ as

$$h(z)\triangleq(m-n)\min\{I_0(I_1^{-1}(z/(m-n))),I_1(I_0^{-1}(z/(m-n)))\}.\tag{20}$$

B. Fundamental Limits

We are ready to provide the fundamental limitations between efficiency and security. The proof is provided in Appendix B.

Theorem 3: For any detection rule f, the following statements on $\mathcal{E}(f)$ and $\mathcal{S}(f)$ are true:

1) $\mathcal{E}(f)\le mC$,
2) $\mathcal{S}(f)\le(m-2n)^+C$, where $(m-2n)^+=\max\{0,m-2n\}$.
3) $\mathcal{S}(f)\le\mathcal{E}(f)$,
4) Let $\mathcal{E}(f)=z$, we have

$$\mathcal{S}(f)\le\begin{cases}h(z) & \text{if } 0<z<(m-n)D_{\min}\quad(21a)\\ 0 & \text{if } z\ge(m-n)D_{\min}.\quad(21b)\end{cases}$$

Remark 2: Theorem 3 indicates that mC is the maximum efficiency that can be achieved by any detector, while $(m-2n)^+C$ is the maximum security that can be achieved. Therefore, if $m\le2n$, i.e., no less that half of the sensors are compromised, then $\mathcal{S}(f)=0$ for any f, which implies that all detectors will have zero security. In that case, the naive Bayes detector will be the optimal choice since it has the optimal efficiency and the analysis becomes trivial. Therefore, without any further notice, we assume $m>2n$ for the rest of the paper.

Notice that fourth constraint in Theorem 3 indicates a trade-off between security and efficiency. For general cases, the maximum security and efficiency may not be achieved simultaneously. However, in Section IV-D, we will prove that for a special case, there exist detectors that can achieve the maximum security and efficiency at the same time.

Notice that $I_0(x)$ ($I_1(x)$) is strictly increasing (decreasing) on $[-D(0\|1), D(1\|0)]$. Therefore, combining (15) and (17), one obtains that $h(z)$ is strictly decreasing. Then the dual version of (21a) is obtained as follows. Let $\mathcal{S}(f) = z$ we have that if $0 < z \leq (m-2n)C$,

$$\mathcal{E}(f) \leq h^{-1}(z) = h(z), \tag{22}$$

where $h^{-1}(z)$ is the inverse function of $h(z)$, and the equality holds because $h(z)$ is an involutory function, i.e., $h(h(z)) = z$ for every $z \in (0, (m-n)D_{\min})$.

We then have the following two corollaries. The results follow straightforwardly from Theorem 3 and (22), we thus omit the proofs.

Corollary 1: Suppose the security of a detector f satisfies

$$\mathcal{S}(f) = z \in [0, (m-2n)C],$$

then the maximum efficiency of f satisfies the following inequality:

$$\max_{f \in \{f: \mathcal{S}(f) = z\}} \mathcal{E}(f) \leq \begin{cases} mC & \text{if } z = 0 \\ h_e(z) & \text{if } z > 0 \end{cases},$$

where $h_e(z) \triangleq \min\{mC, h(z)\}$.

Corollary 2: Suppose the efficiency of a detector f satisfies

$$\mathcal{E}(f) = z \in [0, mC],$$

then the maximum security of f satisfies the following inequality:

$$\max_{f \in \{f: \mathcal{E}(f) = z\}} \mathcal{S}(f) \leq \begin{cases} h_s(z) & \text{if } 0 < z < z', \\ 0 & \text{if } z \geq z' \text{ or } z = 0, \end{cases}$$

where $h_s(z) = \min\{z, (m-2n)C, h(z)\}$, and $z' = (m-n)D_{\min}$.

C. Achievability

In this section, we propose a detector that achieves the upper bounds in Corollaries 1 and 2.

Let $z_s \leq (m-2n)C$. At time $k \geq 1$, the algorithm, denoted by $f^*_{z_s}$, is implemented as follows.

Remark 3: We here discuss about the computational complexity of the detection rule $f^*_{z_s}$. The computational complexity for the step 1 is $O(m)$. Notice that the quantity $\bar{\lambda}_i(k)$ is computed in a recursive fashion. The complexity for the step 2 is $O(m \log m)$. To compute $\delta(0,k)$ and $\delta(1,k)$, we can first sort $I_0(\bar{\lambda}_i(k))$ and $I_1(\bar{\lambda}_i(k))$ in ascending order, respectively, and then sum the first $m-2n$ elements of each. The computational complexity for the step 3 and step 4 is fixed, and the step 5 has computational complexity $O(m)$. Therefore, the total computational complexity for each time step is $O(m \log m)$.

We now show the performance of $f^*_{z_s}$ and the proof is provided in Appendix C.

Algorithm 1: Hypothesis Testing Algorithm $f^*_{z_s}$.

1: Compute the empirical mean of the likelihood ratio from time 1 to time k for each sensor i:

$$\begin{aligned} \bar{\lambda}_i(k) &\triangleq \sum_{t=1}^{k} \lambda(y_i'(t))/k \\ &= \frac{k-1}{k}\bar{\lambda}_i(k-1) + \frac{1}{k}\lambda(y_i'(k)) \end{aligned} \tag{23}$$

with $\bar{\lambda}_i(0) = 0$.

2: Compute $I_0(\bar{\lambda}_i(k))$ and $I_1(\bar{\lambda}_i(k))$ for each i. Compute the following sum:

$$\delta(0,k) \triangleq \min_{|\mathcal{O}|=m-n, \mathcal{O}\subset\mathcal{M}} \sum_{i\in\mathcal{O}} I_0(\bar{\lambda}_i(k)),$$

$$\delta(1,k) \triangleq \min_{|\mathcal{O}|=m-n, \mathcal{O}\subset\mathcal{M}} \sum_{i\in\mathcal{O}} I_1(\bar{\lambda}_i(k)).$$

3: If $\delta(0,k) < z_s$, make a decision $\hat{\theta} = 0$; go to the next step otherwise.

4: If $\delta(1,k) < z_s$, make a decision $\hat{\theta} = 1$; go to the next step otherwise.

5: If $\sum_{i=1}^{m} \bar{\lambda}_i(k) < 0$, make a decision $\hat{\theta} = 0$; make a decision $\hat{\theta} = 1$ otherwise.

Definition 1: (z_e, z_s) are called an admissible pair if the following inequalities holds:

$$0 \leq z_s \leq (m-2n)C,$$

$$z_e \leq \begin{cases} mC & \text{if } z_s = 0 \\ h_e(z_s) & \text{if } z_s > 0 \end{cases},$$

where $h_e(z_s)$ is defined in Corollary 1.

Theorem 4: Let (z_e, z_s) be any admissible pair of efficiency and security. Then there holds

$$\mathcal{E}(f^*_{z_s}) \geq z_e, \quad \mathcal{S}(f^*_{z_s}) \geq z_s.$$

The above theorem means that the upper bounds in Corollaries 1 and 2 are achieved by $f^*_{z_s}$. Hence, we provide a tight characterization on admissible efficiency and security pair. We illustrate the shape of admissible region in Fig. 2.

Remark 4: The optimal detector may not be necessarily unique, in the sense that there may exist other detectors, other than the one defined by Algorithm 1, that can achieve the same efficiency and security limits. By definition, the detectors achieving the limits have the same asymptotic performance. However, the finite-time performance (in terms of detection error) may be different and we are planning to investigate this in the future work.

D. Special Case: Symmetric Distribution

In this subsection, we discuss a case where the maximum security and efficiency can be achieved simultaneously by a detector.

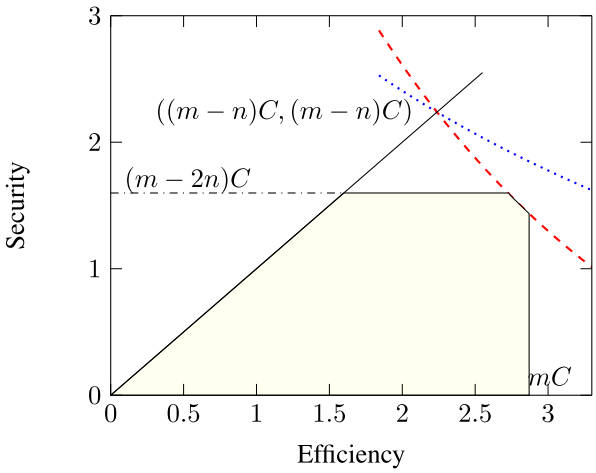


Fig. 2. Achievable efficiency and security region for any detector. The figure is plotted by assuming $y_1(1)$ to be Bernoulli distributed under both hypotheses with $\mathbb{P}_0^o(y_1(1)=1)=0.02$ and $\mathbb{P}_1^o(y_1(1)=1)=0.6$. The shaded area is the admissible pair $(\mathcal{E}(f),\mathcal{S}(f))$ for any detector f. The red dashed line is the function $(m-n)I_0(I_1^{-1}(z/(m-n)))$, while the blue dotted line the function $(m-n)I_1(I_0^{-1}(z/(m-n)))$.

Notice that by the definition of admissible pair, if we have

$$h((m-2n)C)\geq mC, \tag{24}$$

then we know that $(z_e=mC, z_s=(m-2n)C)$ is an admissible pair and hence the detector $f^*_{(m-2n)C}$ defined in Section IV-C can achieve maximum security $(m-2n)C$ and efficiency mC simultaneously. In other words, adding security will not deteriorate the performance of the system in the absence of the adversary.

The following theorem provides a sufficient condition for (24), which is based on the first order derivative of $I_0(\cdot)$ and $I_1(\cdot)$. The proof is presented in Appendix D for the sake of legibility.

Theorem 5: If $I_0^{(1)}(x)|_{x=0}=-I_1^{(1)}(x)|_{x=0}$, then $h((m-2n)C)\geq mC$ holds. Therefore, $f^*_{(m-2n)C}$ possesses not only the maximal security but also the maximal efficiency.

Notice that whether or not the above sufficient condition is satisfied merely depends on the probability distribution of the original observations, which is independent of the number of the compromised sensors.

If there exists "symmetry" between distribution μ and ν, then the sufficient condition can be satisfied. To be specific, if there exists a constant a such that for any Borel measurable set $\mathcal{A}$, we have

$$\mu(a+\mathcal{A})=\nu(a-\mathcal{A}),$$

then one can prove that

$$M_0(w)=M_0(-w),$$

which further implies that

$$I_0(x)=I_1(-x)\Rightarrow I_0^{(1)}(x)|_{x=0}=-I_1^{(1)}(x)|_{x=0}.$$

We provide two examples of pairs of "symmetric" distributions as follows:

1) Each $y_i(k)$ is i.i.d. Bernoulli distributed and

$$y_i(k)=\begin{cases}\theta & \text{with probability } p_0\\ 1-\theta & \text{with probability } 1-p_0\end{cases}$$

2) Each $y_i(k)$ satisfies the following equation:

$$y_i(k)=a\theta+v_i(k),$$

where $a\neq 0$ and $v_i(k)\sim\mathcal{N}(\bar{v},\sigma^2)$ is i.i.d. Gaussian distributed.

V. Extension

In this section, we consider two extensions to the problem settings discussed in Section IV.

A. Secure Sensors

Consider that there is a subset of "secure" sensors that are well protected and cannot be compromised by the attacker. We would like to study the trade-off between security and efficiency when those "secure" sensors are deployed.

Let m_s out of the total m sensors be "secure" and the remaining $m-m_s$ sensors are "normal" ones that can be compromised by an adversary. In this subsection, n can take any value in $\{0,1,\ldots,m-m_s\}$ and does not necessarily satisfy $2n<m$. The other settings are the same as in Section II. Denote by $\mathcal{E}_s(f)$ and $\mathcal{S}_s(f)$ the efficiency and security of a detection rule f under such case.

Then one obtains the following results as in Theorem 3.

Theorem 6: For any detection rule f, the following statements on $\mathcal{E}_s(f)$ and $\mathcal{S}_s(f)$ are true:

1) $\mathcal{E}_s(f)\leq mC$,
2) $\mathcal{S}_s(f)\leq\max((m-2n)C, m_sC)$,
3) $\mathcal{S}_s(f)\leq\mathcal{E}_s(f)$,
4) Let $\mathcal{E}_s(f)=z$, we have

$$\mathcal{S}_s(f)\leq\begin{cases}h(z) & \text{if } 0<z<(m-n)D_{\min}\\ 0 & \text{if } z\geq(m-n)D_{\min}\end{cases}.$$

The above theorem is proved in the same manner as in Appendix B. Notice that the essential difference is the range of $\mathcal{S}_s(f)$, i.e., the statement in the second bullet. This is due to the fact that the m_s secure sensors cannot be compromised.

Remark 5: From the above theorem, one sees that replacing m_s normal sensors with secure sensors does not change the fundamental trade-off between the security and efficiency. However, the benefit of these m_s secure sensors are that the security itself is improved when $2n>m-m_s$. Also, one notice that when $m-m_s\geq 2n$, there are no gains of deploying secure sensors. Intuitively, in such case the redundancy of the $m-m_s$ normal sensors is enough.

Furthermore, the detector $f^s_{z_s}$ in Algorithm 2, which is a slight variation of $f^*_{z_s}$ and treats the m_s secure sensors separately, achieves the limits. This is stated in the following theorem, which is proved in the same manner as in Appendix C.

Algorithm 2: Hypothesis Testing Algorithm $f^s_{z_s}$ when there are Secure Sensors.

1: Compute $\bar{\lambda}_i(k)$ for each sensor.
2: Compute $I_0(\bar{\lambda}_i(k))$ and $I_1(\bar{\lambda}_i(k))$ for each sensor.
Compute the minimum sum from the "normal" sensors:

$$\delta(0,k) = \min_{|\mathcal{O}|=m-m_s-n, \mathcal{O}\subset\{1,\ldots,m-m_s\}} \sum_{i\in\mathcal{O}} I_0(\bar{\lambda}_i(k)),$$

$$\delta(1,k) = \min_{|\mathcal{O}|=m-m_s-n, \mathcal{O}\subset\{1,\ldots,m-m_s\}} \sum_{i\in\mathcal{O}} I_1(\bar{\lambda}_i(k)).$$

3: If $\delta(0,k) + \sum_{m-m_s+1}^{i=m} I_0(\bar{\lambda}_i(k)) < z_s$, make a decision $\hat{\theta}=0$; go to the next step otherwise.
4: If $\delta(1,k) + \sum_{m-m_s+1}^{i=m} I_1(\bar{\lambda}_i(k)) < z_s$, make a decision $\hat{\theta}=1$; go to the next step otherwise.
5: If $\sum_{i=1}^m \bar{\lambda}_i(k) < 0$, make a decision $\hat{\theta}=0$; make a decision $\hat{\theta}=1$ otherwise.

Theorem 7: If the pair (z_e, z_s) satisfies

$$0 \le z_s \le \max((m-2n)C, m_s C),$$

$$z_e \le \begin{cases} mC & \text{if } z_s = 0 \\ h_e(z_s) & \text{if } z_s > 0 \end{cases},$$

then there holds

$$\mathcal{E}_s(f^s_{z_s}) \ge z_e, \quad \mathcal{S}_s(f^s_{z_s}) \ge z_s.$$

B. Unknown Number of Compromised Sensors

In the previous section, we assume that if the system is being attacked, then n sensors are compromised. However, in practice, the exact number of compromised sensors is likely to be unknown. In this subsection, we assume that we know an estimated upper bound on the compromised sensors, denoted by n. Let n_a denote the number of the sensors that are actually compromised. Therefore, n_a may take value in $\mathcal{N}_a \triangleq \{0,1,2,\ldots n\}$.[4]

Given a detector f, denote by $\mathcal{D}_{n_a}(f)$ the detection performance when the number of compromised sensor is n_a. Then, one has $\mathcal{D}_0(f) = \mathcal{E}(f)$ and $\mathcal{D}_n(f) = \mathcal{S}(f)$. In the following, we present the pairwise trade-off between $\mathcal{D}_{n_a}(f)$ and $\mathcal{D}_{n'_a}(f)$ for any $0 \le n_a, n'_a \le n$, and propose an algorithm to achieve these performance limits. A similar argument as in Section IV can be adopted to obtain these results, the details of which are omitted.

We define $\tilde{h} : \mathcal{N}_a \times \mathcal{N}_a \times (0,\infty) \mapsto (0,\infty)$ as

$$\tilde{h}(n_a, n'_a, z) \triangleq (m-\tilde{n}_a)\min\left\{\tilde{h}_0(z/(m-\tilde{n}_a)), \tilde{h}_1(z/(m-\tilde{n}_a))\right\},$$

[4]In Section II-A, we remark that the requirement $|\mathcal{I}| = n$ can be equivalently replaced by $|\mathcal{I}| \le n$. The implicit assumption is that the estimated upper bound n is tight and the worst-case number of compromised sensors is in indeed n. Therefore, n_a in this section may also be interpreted as the tight upper bound of the number of actually compromised senors.

Algorithm 3: Hypothesis Testing Algorithm $f^*_{\mathbf{z}}$.

initialization: $n_a = n$.
1: Compute $\bar{\lambda}_i(k), I_0(\bar{\lambda}_i(k)), I_1(\bar{\lambda}_i(k))$ for each sensor i.
2: **While** $n_a \ge 1$
1) Compute these two minima:

$$\tilde{\delta}(0,k,n_a) \triangleq \min_{|\mathcal{O}|=m-n_a, \mathcal{O}\subset\mathcal{M}} \sum_{i\in\mathcal{O}} I_0(\bar{\lambda}_i(k)),$$

$$\tilde{\delta}(1,k,n_a) \triangleq \min_{|\mathcal{O}|=m-n_a, \mathcal{O}\subset\mathcal{M}} \sum_{i\in\mathcal{O}} I_1(\bar{\lambda}_i(k)).$$

2) If $\tilde{\delta}(0,k,n_a) < z_{n_a}$, make a decision $\hat{\theta}=0$ and stop.
3) If $\tilde{\delta}(1,k,n_a) < z_{n_a}$, make a decision $\hat{\theta}=1$ and stop.
4) Replace n_a with $n_a - 1$.
3: If $\sum_{i=1}^m \bar{\lambda}_i(k) < 0$, make a decision $\hat{\theta}=0$; make a decision $\hat{\theta}=1$ otherwise.

where $\tilde{n}_a = n_a + n'_a$, and

$$\tilde{h}_0(z) = \begin{cases} I_0(I_1^{-1}(z)) & \text{if } 0 < z < D(0||1) \\ 0 & \text{if } z \ge D(0||1) \end{cases},$$

$$\tilde{h}_1(z) = \begin{cases} I_1(I_0^{-1}(z)) & \text{if } 0 < z < D(1||0) \\ 0 & \text{if } z \ge D(1||0) \end{cases}.$$

Then one obtains that for any detector f and $n_a, n'_a \in \mathbb{N}$, there hold

$$\mathcal{D}_{n_a}(f) \le (m-2n_a)C, \tag{25}$$

$$\mathcal{D}_{n_a}(f) \le \tilde{h}\left(n_a, n'_a, \mathcal{D}_{n'_a}(f)\right). \tag{26}$$

Let $\mathbf{z} \triangleq (z_0, z_1, \ldots, z_n)$ be a n-tuplet of admissible detection performance, i.e.,

$$z_{n_a} \le (m-2n_a)C,$$

$$z_{n_a} \le \tilde{h}(n_a, n'_a, z_{n'_a}).$$

Then the detector in Algorithm 3, which is a variation of $f^*_{z_s}$ in Section IV-C and is denoted by $f^*_{\mathbf{z}}$, can achieve these performance, i.e., $\mathcal{D}_{n_a}(f^*_{\mathbf{z}}) \ge z_{n_a}$ for any $n_a \in \mathbb{N}$.

VI. Numerical Examples

A. Asymptotic Performance

We simulate the performance of the detector $f^*_{z_s}$ proposed in Section IV-C (i.e., its efficiency and security) and compare the empirical results to the theoretical ones shown in Fig. 2.

The same parameters as in Fig. 2 are used, i.e., $\mathbb{P}^o_0(y_1(1)=1) = 0.02$, $\mathbb{P}^o_1(y_1(1)=1) = 0.6$, $m=9$ and $n=2$. To simulate the security, it is assumed that the following attack strategy is adopted. If $\theta=0$, the attacker modifies the observations of the compromised sensors such that for every $i \in \mathcal{I}$ and $k \ge 1$

$$I_0(\bar{\lambda}_i(k)) \ge z_s.$$

On the other hand, if $\theta=1$, the attack strategy is such that $I_1(\bar{\lambda}_i(k)) \ge z_s$ holds for every $i \in \mathcal{I}$ and $k \ge 1$.

To simulate the performance with high accuracy, we adopt the importance sampling approach [23]. To plot Fig. 3, we let z_s

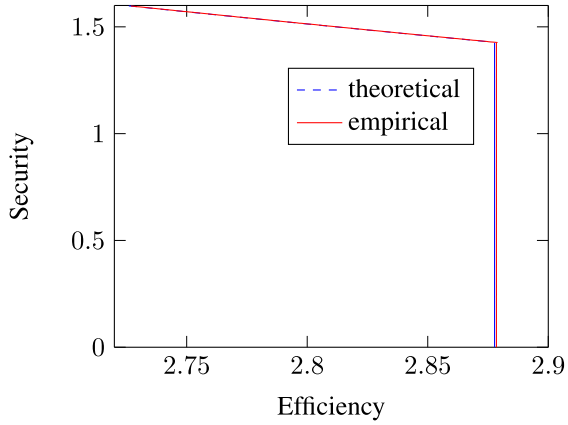


Fig. 3. Comparison between the empirical and theoretical performance of the detector $f^*_{z_s}$ when $z_s \in [0, (m-2n)C]$.

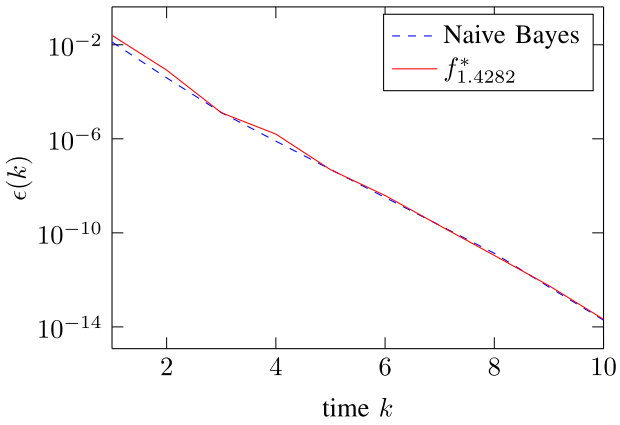


Fig. 4. Finite-time performance of $f^*_{z_s}$ in the absence of the adversary.

be in $[0, (m-2n)C = 1.5987]$. Notice that the theoretical performance of $f^*_{z_s}$ coincides exactly with the fundamental limits in Fig. 2. Therefore, Fig. 3 verifies that our algorithm $f^*_{z_s}$ indeed achieves the fundamental limits.

B. Non-Asymptotic Performance

We have proved that our algorithm is optimal in the sense that it achieves the fundamental trade-off between the security and efficiency. However, notice that both the security and efficiency are asymptotic performance metrics. In this example, we show that our algorithm possesses quite "nice" finite-time performance as well by comparing it to the naive Bayes detector. We should remark that while the Bayes detector is strictly optimal (i.e., optimal for any time horizon) in the absence of attackers, its security is zero. The results are in Fig. 4, where z_s is chosen to be 1.4282. Fig. 4 illustrates that the algorithm $f^*_{z_s}$ with $z_s = 1.4282$ has a finite-time detection performance comparable to that of naive Bayes detector when the attacker is absent. The finite-time performance metric $\epsilon(k)$ is defined in (7), where the attacker is absent, i.e., $g = 0$, and the detector is $f^*_{z_s=1.4282}$ or the naive Bayes. One should note that the security of $f^*_{z_s}$ is 1.4282. As a result, adopting the secure detector $f^*_{z_s}$ increases the security of the system while introducing minimum performance loss in the absence of the adversary.

TABLE I
THE ASYMPTOTIC PERFORMANCES OF OUR ALGORITHM $f^*_{z_s}$ WITH $z_s = 1.4282$, THE TRIMMED MEAN DETECTOR IN [14] $f_{\rm trim}$, AND THE OPTIMAL Q-OUT-M PROCEDURE $f_{\rm qom}(\mathbf{q}^*)$

	$f^*_{z_s=1.4282}$	$f_{\rm trim}$	$f_{\rm qom}(\mathbf{q}^*)$
security	1.43	1.43	0.69
efficiency	2.88	2.00	1.68

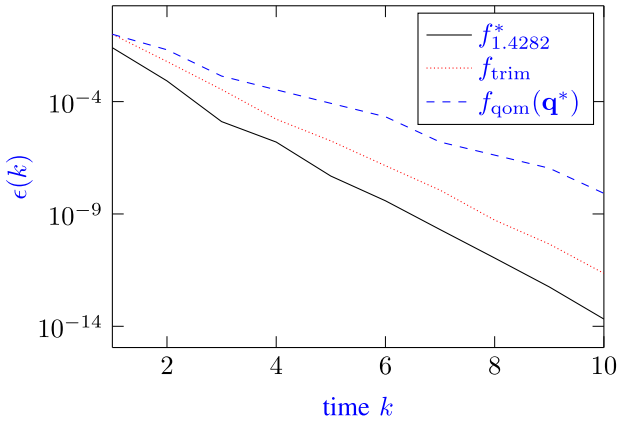


Fig. 5. Finite-time performance of $f^*_{z_s}$, $f_{\rm trim}$, and $f_{\rm qom}(\mathbf{q}^*)$ in the absence of the adversary.

C. Comparison With Other Detectors

To simulate the detectors introduced later, we use the same sensor network model as in Fig. 3. The asymptotic performances, i.e., the efficiency and security, are summarized in Table I, while the non-asymptotic performances when the attacker is absent are in Fig. 5. Table I is consistent with the statement that our algorithm achieves the best trade-off between the security and efficiency, while Fig. 5 shows that it is preferable to adopt our algorithm as well with respect to the finite-time performance when the attacker is absent. In the following, we present the two detectors to be compared to in detail.

The first detector is the equilibrium detection rule that is proposed in [14] for cases where $m > 2n$. This detection rule, which shares the same spirit with the α-trimmed mean in robust statistics [15], first removes the largest n and smallest n log-likelihood ratios, and then compares the mean of the remaining $m-2n$ log-likelihood ratios to 0, just as in the classic probability ratio test. The details of the detection rule, denoted by $f_{\rm trim}$, are formalized as follows.

$$f_{\rm trim}(\mathbf{Y}'(k)) = \begin{cases} 0 & \text{if } \sum_{i=n+1}^{m-n} \bar{\lambda}_{[i]}(k) < 0, \\ 1 & \text{if } \sum_{i=n+1}^{m-n} \bar{\lambda}_{[i]}(k) \geq 0, \end{cases}$$

where $\bar{\lambda}_{[i]}(k)$ is the i-th smallest element of $\{\bar{\lambda}_1(k), \bar{\lambda}_2(k), \ldots, \bar{\lambda}_m(k)\}$ with $\bar{\lambda}_i(k)$ being the empirical mean of log-likelihood ratio from time 1 to k for senor i, which is defined in (23).

It was shown in [14] that the security and efficiency of $f_{\rm trim}$ are

$$\mathcal{S}(f_{\rm trim}) = (m-2n)C, \quad \mathcal{E}(f_{\rm trim}) = (m-n)C.$$

Since for any $z < C$, there hold

$$I_0(I_1^{-1}(z)) > C, \quad I_1(I_0^{-1}(z)) > C.$$

Then by the definition of $h(z)$ and Theorem 4, one obtains that if the security of our algorithm is $(m-2n)C$, its efficiency is larger than $(m-n)C$, i.e.,

$$\mathcal{E}(f^*_{z_s=(m-2n)C}) > (m-n)C.$$

Therefore, our algorithm is preferable since, with the same security, it achieves the larger efficiency than the algorithm $f_{\rm trim}$. In particular, by Theorem 5, the efficiency gain of our algorithm in certain cases is nC.

The next detector is the q-out-of-m procedure [17], which has been studied in [8] by assuming that the malicious sensor nodes generate fictitious data randomly and independently, and the probability that the compromised sensor flips the binary message is known. The q-out-of-m procedure is simple and works as follows. At time k, after receiving the mk (binary) messages, the fusion center makes a decision

$$\hat{\theta} = \begin{cases} 1 & \text{if } \sum_{t=1}^{k}\sum_{i=1}^{m} y_i'(k) \geq q_k, \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

Let $\mathbf{q} = [q_1, \ldots, q_k, \ldots]$ be a sequence of thresholds used in the above detector from time 1 to infinity. In the sequel, we denote the above detector as $f_{\rm qom}(\mathbf{q})$. Notice that $f_{\rm qom}(\mathbf{q})$ is just the naive Bayesian detector, which minimizes the weighted sum of miss detection and false alarm at each time k (the weight is determined by q_k). It is clear that if $f_{\rm qom}(\mathbf{q})$ is used at the fusion center, the worst-case attack is always sending 0 (1) if the true state θ is 1 (0). Therefore, at time k, the performance (i.e., the probability of detection error) of the detector $f_{\rm qom}(\mathbf{q})$ under the worst-case attacks is as follows.

$$\mathbb{P}_1(f_{\rm qom}(\mathbf{q}) = 0) = \sum_{j=0}^{q_k} \binom{mk-nk}{j} p_1^j (1-p_1)^{mk-nk-j},$$

$$\mathbb{P}_0(f_{\rm qom}(\mathbf{q}) = 1)$$

$$= \sum_{j=\max(0,q_k-nk)}^{mk-nk} \binom{mk-nk}{j} p_0^j (1-p_0)^{mk-nk-j},$$

where $p_0 \triangleq \mathbb{P}_0^o(y_1(1)=1) = 0.02, p_1 \triangleq \mathbb{P}_1^o(y_1(1)=1) = 0.6$. Then it is reasonable to set $nk < q_k < mk - nk$, since otherwise the worst-case (over θ) detection error will be 1. However it is challenging to obtain the optimal q_k analytically to minimize the worst-case detection error; we do this by brute-force numerical simulations. By varying the time k from 1 to 40, we obtain the (approximate) security and the optimal parameters $q_1^*, \ldots, q_{40}^*$. Then we further simulate the performance of the q-out-m algorithm when the optimal parameters obtained above are used and the attacker is absent.

VII. Conclusion and Future Work

In this paper, we studied the trade-off between the detection performance of a detector when the attacker is absent (termed efficiency) and the "worst-case" detection performance when the attacker, knowing the detector, is present (termed security). The setting is that a binary hypothesis testing is conducted based on measurements from a set of sensors, some of which can be compromised by an attacker and their measurements can be manipulated arbitrarily. We first provided the fundamental limits of the trade-off between the efficiency and security of any detector. We then presented detectors that possesses the limits of the efficiency and security. Therefore, a clear guideline on how to balance the efficiency and security has been established for the system operator. An interesting point of the fundamental trade-off is that in some cases, the maximal efficiency and the maximal security can be achieved simultaneously, i.e., the maximal efficiency (security) can be achieved without compromising any security (efficiency). In addition, two extensions were investigated: secure sensors are assumed for the first one, and the detection performance beyond the efficiency and security is concerned for the second one. The main results were verified by numerical examples. Investigating the problem when the measurements from the benign sensors are not i.i.d. is a future direction.

Appendix A
The Proof of Theorem 2

The following lemma is needed to prove Theorem 2:

Lemma 1: If Assumption 4 and 5 hold, then the following statement is true:

1) For any w,
$$M_0(w+1) = M_1(w). \tag{28}$$
2) There exists a small enough $\epsilon > 0$, such that $\log M_0(w)$ is well-defined on $[-\epsilon, 1+\epsilon]$, and $\log M_1(w)$ is well-defined on $[-1-\epsilon, \epsilon]$.
3) $\log M_0(w)$, $\log M_1(w)$ are strictly convex.
4) The derivative of $\log M_0(w)$ and $\log M_1(w)$ satisfy
$$(\log M_0(w))^{(1)}|_{w=1} = D(1\|0). \tag{29}$$
$$(\log M_0(w))^{(1)}|_{w=0} = -D(0\|1). \tag{30}$$
$$(\log M_1(w))^{(1)}|_{w=0} = D(1\|0). \tag{31}$$
$$\log M_1(w))^{(1)}|_{w=-1} = -D(0\|1). \tag{32}$$

Proof: By definition,

$$M_0(w+1) = \int_{-\infty}^{\infty} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y)\right)^w \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y)\mathrm{d}\nu(y)$$
$$= \int_{-\infty}^{\infty} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y)\right)^w \mathrm{d}\mu(y) = M_1(w),$$

which proves (28).

Assuming $M_0(w_1)$, $M_0(w_2) < \infty$ and $w_1 < w_2$, by the convexity of the exponential function, we know that for any λ and

$0 < \alpha, \beta < 1$ and $\alpha + \beta = 1$,

$$0 < \exp\left[(\alpha w_1 + \beta w_2)\lambda\right] \leq \alpha e^{w_1 \lambda} + \beta e^{w_2 \lambda}.$$

Therefore $0 < M_0(\alpha w_1 + \beta w_2) \leq \alpha M_0(w_1) + \beta M_0(w_2)$ is well-defined, which proves that the domain of $\log M_0(w)$ is convex.

Furthermore, by Assumption 4, $0 \in \mathrm{int}(\mathrm{dom}_1)$ gives

$$1 \in \mathrm{int}(\mathrm{dom}_0), \tag{33}$$

Hence, $[0, 1] \subset \mathrm{int}(\mathrm{dom}_0)$, which proves that $\log M_0(w)$ is well-defined on $[-\epsilon, 1 + \epsilon]$ if ϵ is small enough.

It is well known that $\log M_0(w)$ is *infinitely differentiable* on $\mathrm{int}(\mathrm{dom}_0)$ (see [22, Exercise 2.2.24]). Basic calculations give that

$$(\log M_0(w))^{(2)} \tag{34}$$

$$= \int_{\mathbb{R}} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y)\right)^w \left(\log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y)\right)\right)^2 \mathrm{d}\nu(\mathrm{y}) > 0 \tag{35}$$

always holds, where $(\log M_0(w))^{(2)}$ is the second derivative. The above quantity is strictly positive since the KL divergence between probability measure μ and ν are strictly positive by Assumption 5. Therefore, $\log M_0(w)$ is strictly convex.

The domain and the strict convexity of $\log M_1(w)$ can be proved similarly.

Take the derivative of $\log M_0(w)$ at $w = 1$ yields

$$\begin{aligned}(\log M_0(w))^{(1)}|_{w=1} &= \int_{\mathbb{R}} \lambda(y) \frac{\mathrm{d}\mu}{\mathrm{d}\nu}(y) \mathrm{d}\nu(y) \\ &= \int_{\mathbb{R}} \lambda(y) \mathrm{d}\mu(y) = D(1\|0).\end{aligned}$$

Equations (30), (31) and (32) can be proved similarly. ■

We are now ready to prove Theorem 2:

Proof of Theorem 2: Define the derivative of $\log M_0(w)$ to be $\psi(w)$. Since $\log M_0(w)$ is strictly convex, we know that $\psi(w)$ is strictly increasing and therefore, its inverse function is well defined on $[-D(0\|1), D(1\|0)]$. Denote the inverse function as $\varphi(x)$. By the convexity of $\log M_0(w)$, we have that

$$\log M_0(w) \geq \log M_0(w_*) + \psi(w_*)(w - w_*). \tag{36}$$

Hence, for any $x \in [-D(0\|1), D(1\|0)]$, suppose that $\psi(w_*) = x$, we have

$$\begin{aligned}wx - \log M_0(w) &= [w_* \psi(w_*) - \log M_0(w_*)] \\ &\quad + [(w - w_*)\psi(w_*) + \log M_0(w_*) - \log M_0(w)].\end{aligned}$$

Notice the last term on the RHS of the equation is non-positive. Hence, we can prove that

$$I_0(x) = w_* \psi(w_*) - \log M_0(w_*) = \varphi(x)x - \log M_0(\varphi(x)). \tag{37}$$

Take the derivative and second order derivative of $I_0(x)$ we have

$$\frac{\mathrm{d}I_0(x)}{\mathrm{d}x} = \varphi(x), \ \frac{\mathrm{d}^2 I_0(x)}{\mathrm{d}x^2} = \frac{1}{\psi^{(1)}(\varphi(x))} > 0,$$

where the last inequality is due to the fact that $\log M_0(w)$ is strictly convex, and thus its second derivative $\psi^{(1)}$ is strictly positive. Hence we prove that $I_0(x)$ is twice differentiable and strictly convex on $[-D(0\|1), D(1\|0)]$. Notice that

$$\left.\frac{\mathrm{d}I_0(x)}{\mathrm{d}x}\right|_{x=-D(0\|1)} = \varphi(-D(0\|1)) = 0,$$

we can prove that $I_0(x)$ is also strictly increasing. Similarly we can prove the properties for $I_1(x)$.

Combining (37), (29) and (30), we can prove (14) and (15). Equations (16) and (17) can be proved similarly.

Since

$$I_0(0) = \sup_w 0 \cdot w - \log M_0(w) = \sup_w -\log M_0(w),$$

$$I_1(0) = \sup_w 0 \cdot w - \log M_1(w) = \sup_w -\log M_1(w),$$

and

$$M_0(w + 1) = M_1(w),$$

We can conclude $I_0(0) = I_1(0)$. ■

Appendix B
The Proof of Theorem 3

The proof is divided into four parts, each of which is devoted to one of the statements in Theorem 3.

Part I: For any index set $\mathcal{O} \subset \mathcal{M}$ and $\chi \in \mathbb{R}$, define the following Bayesian like detector:

$$f_{k,\chi,\mathcal{O}}(\mathbf{Y}'(k)) = \begin{cases} 0 & \text{if } \sum_{i\in\mathcal{O}} \bar{\lambda}_i(k) < \chi, \\ 1 & \text{if } \sum_{i\in\mathcal{O}} \bar{\lambda}_i(k) \geq \chi, \end{cases} \tag{38}$$

where $\bar{\lambda}_i(k)$ is the empirical mean of the log-likelihood ratio from time 1 to k for sensor i, which is defined in (23). Denote

$$f_{\chi,\mathcal{O}} = (f_{1,\chi,\mathcal{O}}(\mathbf{Y}'(1)), \cdots, f_{k,\chi,\mathcal{O}}(\mathbf{Y}'(k)), \cdots)$$

and

$$f^*_{\mathcal{O}} \triangleq f_{0,\mathcal{O}}. \tag{39}$$

It is well known that $f^*_{\mathcal{M}}$ minimize the average error probability [24]: $e(\theta = 0, \mathcal{O} = \emptyset, k) + e(\theta = 1, \mathcal{O} = \emptyset, k)$, where, recall, $e(\theta, \mathcal{O}, k)$ is defined in (6). Notice that

$$\begin{aligned}&\liminf_{k\to\infty} -\frac{\log(e(\theta = 0, \mathcal{O} = \emptyset, k) + e(\theta = 1, \mathcal{O} = \emptyset, k))}{k} \\ &= \liminf_{k\to\infty} -\frac{\log \max_\theta e(\theta, \mathcal{O} = \emptyset, k)}{k}.\end{aligned}$$

Hence, when the attacker is absent, $f^*_{\mathcal{M}}$ is optimal in the sense that the rate ρ defined in (8) is maximized. Furthermore, Cramér's Theorem gives that $\mathcal{E}(f^*_{\mathcal{M}}) = mI_0(0) = mC$. Therefore, $\mathcal{E}(f) \leq mC$ holds for any detector f.

Part II: In this part, we show $\mathcal{S}(f) \leq (m - 2n)^+ C$. The proof is by construction: we construct a attack strategy g^* such that, for any detection rule f, the following inequality holds:

$$\rho(f, g^*) \leq (m - 2n)^+ C. \tag{40}$$

Let $\mathcal{O}' = \{1, \ldots, n\}$ and $\mathcal{O}'' = \{m - n + 1, \ldots, m\}$. The attack strategy g^* is as follows.

i) When $\theta = 0$, sensors in $\mathcal{O}'$ are compromised and the distributions are flipped, i.e., the measurements of sensors in $\mathcal{O}'$ are i.i.d. as μ.
ii) When $\theta = 1$, sensors in $\mathcal{O}'' \setminus \mathcal{O}'$ are compromised and the distributions are flipped.

Thus, under attack g^*, for either $\theta = 1$ or $\theta = 0$, sensors in $\mathcal{O}'$ will follow distribution μ and sensors in $\mathcal{O}'' \setminus \mathcal{O}'$ will follow distribution ν. In other words, only sensors in $\mathcal{M} \setminus (\mathcal{O}' \cup \mathcal{O}'')$ have different distributions under different θ. Notice that when $m \leq 2n$, $\mathcal{M} \setminus (\mathcal{O}' \cup \mathcal{O}'') = \emptyset$, which means that $\rho(f, g^*) = 0$. If $m > 2n$, by the optimality of the detection rule $f^*_{\mathcal{M}\setminus(\mathcal{O}'\cup\mathcal{O}'')}$ defined in (39), one obtains $\rho(f, g^*) \leq (m - 2n)C$. Equation (40) is thus obtained.

Part III: It is clear from the definitions of $\mathcal{E}(f)$ and $\mathcal{S}(f)$ that $\mathcal{S}(f) \leq \mathcal{E}(f)$ holds.

Part IV: Consider the following product measures:

$$\mu_a = \underbrace{\mu \times \mu \ldots \times \mu}_{m-n} \times \underbrace{\nu \times \nu \ldots \times \nu}_{n},$$
$$\mu_a^k = \underbrace{\mu_a \times \mu_a \ldots \times \mu_a}_{k}$$

and

$$\nu_* = \underbrace{\nu \times \nu \ldots \times \nu}_{m},$$
$$\nu_*^k = \underbrace{\nu_* \times \nu_* \ldots \times \nu_*}_{k}.$$

The measure μ_a is generated by an attack that flips the distribution on the last n sensors, when the true hypothesis is $\theta = 1$. The measure ν_* is generated by benign sensors when the true hypothesis is $\theta = 0$.

Now let us consider the following problem: given $\phi > 0$, find the detection rule f such that

$$\mathbb{E}_{\nu_*^k} f_k + \phi^k \mathbb{E}_{\mu_a^k}(1 - f_k) \tag{41}$$

is minimized for every $k \geq 1$. Let $f_\phi = (f_{\phi,1}, \ldots, f_{\phi,k}, \ldots)$ with $f_{k,\phi}$ given by

$$f_{k,\phi}(\mathbf{Y}'(k)) = f_{k,-\log\phi,\{1,2,\ldots,m-n\}}(\mathbf{Y}'(k)), \tag{42}$$

where, recall, the function $f_{k,\chi,\mathcal{O}}(\mathbf{Y}'(k))$ is defined in (38). Then by the Bayesian decision-theoretic detection theory, f_ϕ is a solution to the above problem. Let

$$E_\phi \triangleq \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\nu_*^k} f_{\phi,k}}{k}$$

and

$$S_\phi \triangleq \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\mu_a^k}(1 - f_{\phi,k})}{k}.$$

Then from the optimality of f_ϕ, for any $\phi > 0$ and any detector $f = (f_1, \ldots, f_k, \ldots)$, the following hold for any k:

$$\text{If } \mathbb{E}_{\nu_*^k} f_k \leq \mathbb{E}_{\nu_*^k} f_{\phi,k}, \text{ then } \mathbb{E}_{\mu_a^k}(1 - f_k) \geq \mathbb{E}_{\mu_a^k}(1 - f_{\phi,k}).$$

This implies that

$$\text{If } \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\nu_*^k} f_k}{k} \geq E_\phi,$$
$$\text{then } \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\mu_a^k}(1 - f_k)}{k} \leq S_\phi.$$

Furthermore, the definitions of $\mathcal{E}(f)$ and $\mathcal{S}(f)$ yield

$$\mathcal{E}(f) \leq \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\nu_*^k} f_k}{k},$$
$$\mathcal{S}(f) \leq \liminf_{k\to\infty} -\frac{\log \mathbb{E}_{\mu_a^k}(1 - f_k)}{k}.$$

Therefore, for any $\phi > 0$ and any detector f, the following hold:

$$\text{If } \mathcal{E}(f) \geq E_\phi, \text{ then } \mathcal{S}(f) \leq S_\phi.$$

Now let us evaluate E_ϕ and S_ϕ. Let $\tilde{\phi} = -\log\phi/(m-n)$, then Cramér's theorem yields

$$E_\phi = \begin{cases} 0 & \text{if } \tilde{\phi} \leq -D(0\|1), \\ (m-n)I_0(\tilde{\phi}) & \text{if } \tilde{\phi} > -D(0\|1), \end{cases}$$

and

$$S_\phi = \begin{cases} 0 & \text{if } \tilde{\phi} \geq D(1\|0), \\ (m-n)I_1(\tilde{\phi}) & \text{if } \tilde{\phi} < D(1\|0). \end{cases}$$

Notice that the monotonicity of $I_0(\cdot)$ on $[-D(0\|1), \infty)$ implies that if $0 < E_\phi < (m-n)I_0(D(1\|0)) = (m-n)D(1\|0)$, $\tilde{\phi} \in (-D(0\|1), D(1\|0))$ holds. Therefore, if $0 < E_\phi < (m-n)D(1\|0)$, there holds

$$S_\phi = (m-n)I_1(I_0^{-1}(E_\phi/(m-n))).$$

One thus obtains that for any detector f, if $0 < \mathcal{E}(f) < (m-n)D(1\|0)$

$$\mathcal{S}(f) \leq (m-n)I_1(I_0^{-1}(\mathcal{E}(f)/(m-n))). \tag{43}$$

Also, it is easy to see that if $E_\phi \geq (m-n)D(1\|0)$, $S_\phi = 0$ holds. Thus,

$$\mathcal{S}(f) = 0 \text{ if } \mathcal{E}(f) \geq (m-n)D(1\|0). \tag{44}$$

Similarly, one considers the detection problem for the measures μ_*^k and ν_a^k and obtains that for any detector f, if $0 < \mathcal{E}(f) < (m-n)D(0\|1)$

$$\mathcal{S}(f) \leq (m-n)I_0(I_1^{-1}(\mathcal{E}(f)/(m-n))) \tag{45}$$

and

$$\mathcal{S}(f) = 0 \text{ if } \mathcal{E}(f) \geq (m-n)D(0\|1). \tag{46}$$

Then equation (21a) follows from (43) and (45), and equation (21b) from (44) and (46).

APPENDIX C
THE PROOF OF THEOREM 4

This theorem is proved by showing that $f^*_{z_s} = 0$ (or 1) if certain conditions are satisfied (i.e., Lemma 4). Furthermore, the special structure of these conditions can ensure that, under *any* attacks, $f^*_{z_s} = 0$ (or 1) if the measurements of sensors in

an attack free environment belong to a certain set, to which the Cramér's Theorem is applied.

Before proceeding, we need to define the following subsets of $\mathbb{R}^m$:

Definition 2: Define $\mathcal{B}^-,\ \mathcal{B}^+ \subset \mathbb{R}^m$ as

$$\mathcal{B}^- \triangleq \left\{\lambda \in \mathbb{R}^m : \sum_{i=1}^m \lambda_i < 0\right\}, \mathcal{B}^+ \triangleq \left\{\lambda \in \mathbb{R}^m : \sum_{i=1}^m \lambda_i \geq 0\right\}.$$

Definition 3: Let $\mathcal{O} \subset \mathcal{M}$, $j \in \{0,1\}$ and $z \in \mathbb{R}_+$, define a ball as

$$\mathrm{Bal}(\mathcal{O}, j, z) = \left\{\lambda \in \mathbb{R}^m : \sum_{i\in\mathcal{O}} I_j(\lambda_i) < z\right\}.$$

Definition 4: Let $j \in \{0,1\}$ and $z \in \mathbb{R}_+$, define an extended ball as

$$\mathrm{EBal}(j, z, n) \triangleq \bigcup_{|\mathcal{O}|=m-n} \mathrm{Bal}(\mathcal{O}, j, z).$$

From the definition of extended balls, it is clear that

$$\left[\lambda_1 \ \ldots \ \lambda_m\right] \in \mathrm{EBal}(j, z, n)$$

if and only if the following inequality holds:

$$\min_{|\mathcal{O}|=m-n} \sum_{i\in\mathcal{O}} I_j(\lambda_i) < z.$$

Combining with the definition of $f^*_{z_s}$, we know that at time k, the output of $f^*_{z_s}$ is 0 if and only if

$$\bar{\lambda}(k) \triangleq \left[\bar{\lambda}_1(k) \cdots \bar{\lambda}_m(k)\right] \in \lambda^-(z_s),$$

where $\lambda^-(z_s)$ is defined as

$$\lambda^-(z_s) \triangleq \mathrm{EBal}(0, z_s, n) \bigcup \left(\mathcal{B}^- \setminus \mathrm{EBal}(1, z_s, n)\right).$$

The output is 1 if $\bar{\lambda}(k) \in \lambda^+(z_s)$, where

$$\begin{aligned}\lambda^+(z_s) &\triangleq \mathbb{R}^m \setminus \lambda^-(z_s)\\ &= \left(\mathcal{B}^+ \bigcup \mathrm{EBal}(1, z_s, n)\right) \setminus \mathrm{EBal}(0, z_s, n)\end{aligned}$$

We first need the following supporting lemma.

Lemma 2: Given $\mathcal{O}_1,\ \mathcal{O}_2 \subset \mathcal{M} \triangleq \{1, 2, \ldots, m\}$ with $|\mathcal{O}_1 \bigcap \mathcal{O}_2| = p > 0$, $z \leq pD(1\|0)$, the optimal value of the following optimization problem is given by $pI_1(I_0^{-1}(z/p))$:

$$\begin{aligned}\inf_{x\in\mathbb{R}^m} \quad & \sum_{i\in\mathcal{O}_1} I_1(x_i)\\ \text{s.t.} \quad & \sum_{i\in\mathcal{O}_2} I_0(x_i) < z.\end{aligned} \tag{47}$$

Proof: Since $I_1(\cdot)$ is nonnegative, $I_1(D(1\|0)) = 0$ and x_i can take any value when $i \notin \mathcal{O}_2$, one can equivalently rewrite (47) as

$$\begin{aligned}\inf_{x\in\mathbb{R}^m} \quad & \sum_{i\in\mathcal{O}_1\cap\mathcal{O}_2} I_1(x_i)\\ \text{s.t.} \quad & \sum_{i\in\mathcal{O}_2} I_0(x_i) < z,\\ & x_i = D(1\|0), i \in \mathcal{O}_1 \setminus \mathcal{O}_2.\end{aligned}$$

By the nonnegativity of $I_0(\cdot)$, the above equation is equivalent to

$$\begin{aligned}\inf_{x\in\mathbb{R}^m, 0<z'\leq z} \quad & \sum_{i\in\mathcal{O}_1\cap\mathcal{O}_2} I_1(x_i)\\ \text{s.t.} \quad & \sum_{i\in\mathcal{O}_1\cap\mathcal{O}_2} I_0(x_i) < z',\\ & \sum_{i\in\mathcal{O}_2\setminus\mathcal{O}_1} I_0(x_i) \leq z - z',\\ & x_i = D(1\|0), i \in \mathcal{O}_1 \setminus \mathcal{O}_2.\end{aligned} \tag{48}$$

To obtain the solution to the above equation, let us fist focus on the following optimization problem:

$$\begin{aligned}\min_{x\in\mathbb{R}^m} \quad & \sum_{i\in\mathcal{O}} I_1(x_i)\\ \text{s.t.} \quad & \sum_{i\in\mathcal{O}} I_0(x_i) = z',\end{aligned} \tag{49}$$

where $\mathcal{O} = \mathcal{O}_1 \cap \mathcal{O}_2$. Denotes its optimal value by $\psi(z')$. In the following, we show that

$$\psi(z') = pI_1(I_0^{-1}(z'/p)). \tag{50}$$

We claim that a solution to (49) is

$$x_i = \begin{cases} I_0^{-1}(z'/p) & \text{if } i \in \mathcal{O}, \quad (51a)\\ \text{whatever} & \text{if } i \notin \mathcal{O}. \quad (51b)\end{cases}$$

With this claim, (50) clearly holds. In the following we show that this claim is correct. Equation (51b) is trivial. We then focus on (51a). Due to the convexity of the functions $I_0(\cdot)$ and $I_1(\cdot)$, one obtains that for any $x \in \mathbb{R}^m$,

$$\begin{aligned} pI_0\left(\sum_{i\in\mathcal{O}} x_i/p\right) &\leq \sum_{i\in\mathcal{O}} I_0(x_i),\\ pI_1\left(\sum_{i\in\mathcal{O}} x_i/p\right) &\leq \sum_{i\in\mathcal{O}} I_1(x_i)\end{aligned}$$

Therefore, without any performance loss, one may restrict the solution to the set $\mathbb{X}^*$ as follows:

$$\mathbb{X}^* \triangleq \{x \in \mathbb{R}^m : x_1 = x_2 = \cdots = x_p\}.$$

Then it is clear from the monotonicity of I_0 and I_1 that (51a) holds. This thus proves (50).

Notice that $\psi(z')$ in (50) is decreasing with respect to z'. Then the fact that $I_0(-D(0\|1)) = 0$ yields that (48) is equivalent to

$$\begin{aligned}\min_{x\in\mathbb{R}^m} \quad & \sum_{i\in\mathcal{O}_1\cap\mathcal{O}_2} I_1(x_i)\\ \text{s.t.} \quad & \sum_{i\in\mathcal{O}_1\cap\mathcal{O}_2} I_0(x_i) = z,\\ & x_i = -D(0\|1), i \in \mathcal{O}_2 \setminus \mathcal{O}_1,\\ & x_i = D(1\|0), i \in \mathcal{O}_1 \setminus \mathcal{O}_2,\end{aligned}$$

which concludes Lemma 2 by (50). ■

Lemma 3: Assume that (z_e, z_s) are an admissible pair, then the following statements are true:

1) $\mathrm{Bal}(\mathcal{M}, 0, z_e) \subseteq \mathcal{B}^-$.
2) $\mathrm{Bal}(\mathcal{M}, 1, z_e) \subseteq \mathcal{B}^+$.
3) $\mathrm{EBal}(0, z_s, n) \bigcap \mathrm{EBal}(1, z_s, n) = \emptyset$.
4) $\mathrm{EBal}(1, z_s, n) \bigcap \mathrm{Bal}(\mathcal{M}, 0, z_e) = \emptyset$.
5) $\mathrm{EBal}(0, z_s, n) \bigcap \mathrm{Bal}(\mathcal{M}, 1, z_e) = \emptyset$.

Proof: 1): It suffices to prove that given any $x \in \mathbb{R}^m$, if $x \in \mathcal{B}^+$, then $x \notin \mathrm{Bal}(\mathcal{M}, 0, z_e)$. By the convexity of $I_0(x)$, one obtains that

$$\sum_{i\in\mathcal{M}} I_0(x_i) \geq mI_0(1/m \sum_{i\in\mathcal{M}} x_i) \geq mI_0(0) = mC,$$

where the second inequality follows from $x \in \mathcal{B}^+$ and the fact that $I_0(x)$ is increasing when $x \geq 0$. Notice that by its definition, $z_e \leq mC$ holds. The proof is done.

2): This can be proved similarly to 1).

3): By the definition of EBal, we need to prove that for any $\mathcal{O}_1, \mathcal{O}_2$ with $|\mathcal{O}_1| = |\mathcal{O}_2| = m - n$, $\mathrm{Bal}(\mathcal{O}_1, 0, z_s) \cap \mathrm{Bal}(\mathcal{O}_2, 1, z_s) = \emptyset$ holds. Notice that when $z \leq pD(1\|0)$, $pI_1(I_0^{-1}(z/p))$ is increasing with respect to p. Thus by Lemma 2, it suffices to prove that $(m-2n)I_1(I_0^{-1}(z_s/(m-2n))) \geq z_s$, which is true because $0 \leq z_s \leq (m-2n)C$, $pI_1(I_0^{-1}(z/p))$ is decreasing with respect to z when $z \leq pD(1\|0)$, and $(m-2n)I_1(I_0^{-1}(0)) = (m-2n)C$.

4): Similar to 3), it suffices to prove that for any $\mathcal{O}_1$ with $|\mathcal{O}_1| = m - n$, $\mathrm{Bal}(\mathcal{O}_1, 0, z_s) \cap \mathrm{Bal}(\mathcal{M}, 0, z_e) = \emptyset$ holds. By Lemma 2, it suffices to prove that $(m-n)I_1(I_0^{-1}(z_e/(m-n))) \geq z_s$. Then it is equivalent to prove that $(m-n)I_1(I_0^{-1}(h_e(z_s)/(m-n))) \geq z_s$, which follows from the definition of $h_e(z)$ and the fact that $pI_1(I_0^{-1}(z/p))$ is decreasing with respect to z when $z \leq pD(1\|0)$.

5): This can be proved similarly to 4). ■

From Lemma 3, one obtains straightforwardly the following lemma.

Lemma 4: Assume that (z_e, z_s) are an admissible pair, then the following set inclusions are true:

1) $\mathrm{EBal}(0, z_s, n) \subseteq \lambda^-(z_s)$.
2) $\mathrm{EBal}(1, z_s, n) \subseteq \lambda^+(z_s)$.
3) $\mathrm{Bal}(\mathcal{M}, 0, z_e) \subseteq \lambda^-(z_s)$.
4) $\mathrm{Bal}(\mathcal{M}, 1, z_e) \subseteq \lambda^+(z_s)$.

We are now ready to prove Theorem 4.

Proof of Theorem 4: We focus on the proof of $\mathcal{S}(f^*_{z_s}) \geq z_s$, and a similar (and simpler) approach can be used to prove $\mathcal{E}(f^*_{z_s}) \geq z_e$. Notice that $\mathrm{EBal}(0, z_s, n) \subseteq \lambda^-(z_s)$ in Lemma 4 gives that, under any attacks, there holds $\mathrm{Bal}(\mathcal{M}, 0, z_s) \subseteq \lambda^-(z_s)$. Therefore,

$$\begin{aligned}
&\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}_0(f^*_{z_s,k} = 1) \\
&\leq \limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}_0^o(\bar{\lambda}(k) \in \mathbb{R}^m \setminus \mathrm{Bal}(\mathcal{M}, 0, z_s)) \\
&\leq -\inf_{x\in\mathbb{R}^m \setminus \mathrm{Bal}(\mathcal{M},0,z_s)} \sum_{i=1}^m I_0(x_i) \\
&= -z_s, \qquad (52)
\end{aligned}$$

where the second inequality holds because of the Cramér's Theorem and the fact that $\mathbb{R}^m \setminus \mathrm{Bal}(\mathcal{M}, 0, z_s)$ is closed.

Similarly, by $\mathrm{EBal}(1, z_s, n) \subseteq \lambda^+(z_s)$ in Lemma 4, one obtains

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}_1(f^*_{z_s,k} = 0) \leq -z_s. \qquad (53)$$

It follows from (52) and (53) that $\mathcal{S}(f^*_{z_s}) \geq z_s$. The proof is thus complete. ■

Appendix D
The Proof of Theorem 5

Define the following two functions $h_0(z), h_1(z) : (0, D_{\min}) \mapsto (0, D_{\min})$:

$$h_0(z) = I_0(I_1^{-1}(z)),$$
$$h_1(z) = I_1(I_0^{-1}(z)).$$

Then we have the following two lemmas on $h_0(z)$ and $h_1(z)$.

Lemma 5: Both $h_0(z)$ and $h_1(z)$ are convex. Furthermore, the following equality holds:

$$h_0(C) = h_1(C) = C. \qquad (54)$$

Proof: The equation (54) follows directly from (18) and (19). To prove the convexity of $h_0(z)$ and $h_1(z)$, we first need to prove that $I_0^{-1}(x)$ is convex and $I_1^{-1}(x)$ is concave on $[0,\ D_{\min}]$. Notice that if ψ is the inverse function of ϕ and ϕ are twice differentiable, then by chain rule

$$\psi^{(2)}(x) = -\frac{\phi^{(2)}(\psi(x))}{\left[\phi^{(1)}(\psi(x))\right]^3}.$$

Therefore, since $I_0(x)$ ($I_1(x)$) is strictly convex and strictly decreasing (increasing) on $[-D(0\|1),\ D(1\|0)]$, $I_0^{-1}(x)$ ($I_1^{-1}(x)$) is convex (concave) on $[0,\ D_{min}]$.

The convexity of $h_0(z)$ and $h_1(z)$ then follows the fact that the composition of a convex and increasing (decreasing) function with a convex (concave) function is convex [25]. ■

We are now ready to prove Theorem 5

Proof: By chain rule, we know that

$$\frac{\mathrm{d}h_0(z)}{\mathrm{d}z}|_{z=C} = I_0^{(1)}(x)|_{x=0} \times \frac{1}{I_1^{(1)}(x)|_{x=0}} = -1.$$

Therefore, by the convexity of $h_0(z)$, we know that

$$h_0(z) \geq h_0(C) - (z - C) \times \frac{\mathrm{d}h_0(z)}{\mathrm{d}z}|_{z=C} = 2C - z.$$

Similarly, one can prove that

$$h_1(z) \geq 2C - z.$$

Hence, by the definition of $h(z)$,

$$h(z) \geq 2(m-n)C - z,$$

which implies that $h((m-2n)C) \geq mC$ holds and $f^*_{(m-2n)C}$ achieves maximum security and efficiency simultaneously. ■

REFERENCES

[1] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 774–786, Feb. 2011.

[2] Y. Mo and B. Sinopoli, "Secure estimation in the presence of integrity attacks," *IEEE Trans. Automat. Control*, vol. 60, no. 4, pp. 1145–1151, Apr. 2015.

[3] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Secure control systems: A quantitative risk management approach," *IEEE Control Syst.*, vol. 35, no. 1, pp. 24–45, Feb. 2015.

[4] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Tech. J.*, vol. 28, no. 4, pp. 656–715, 1949.

[5] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of byzantine attacks," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 16–29, Jan. 2009.

[6] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed Bayesian detection in the presence of Byzantine data," *IEEE Trans. Signal Process.*, vol. 63, no. 19, pp. 5250–5263, Oct. 2015.

[7] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Asymptotic analysis of distributed Bayesian detection with byzantine data," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 608–612, May 2015.

[8] M. Abdelhakim, L. E. Lightfoot, J. Ren, and T. Li, "Distributed detection in mobile access wireless sensor networks under byzantine attacks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 4, pp. 950–959, Apr. 2014.

[9] E. Soltanmohammadi, M. Orooji, and M. Naraghi-Pour, "Decentralized hypothesis testing in wireless sensor networks in the presence of misbehaving nodes," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 1, pp. 205–215, Jan. 2013.

[10] E. Soltanmohammadi and M. Naraghi-Pour, "Fast detection of malicious behavior in cooperative spectrum sensing," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 3, pp. 377–386, Mar. 2014.

[11] Y. Mo, J. P. Hespanha, and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 31–43, Jan. 2014.

[12] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Detection in adversarial environments," *IEEE Trans. Automat. Control*, vol. 59, no. 12, pp. 3209–3223, Dec. 2014.

[13] A. Abrardo, M. Barni, K. Kallas, and B. Tondi, "A game-theoretic framework for optimum decision fusion in the presence of Byzantines," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 6, pp. 1333–1345, Jun. 2016.

[14] J. Yan, X. Ren, and Y. Mo, "Sequential detection in adversarial environments," in *Proc. 56th IEEE Conf. Decision Control*, Dec. 2017, pp. 170–175.

[15] P. J. Huber, *Robust Statistics*. New York, NY, USA: Springer, 2011.

[16] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, 1965.

[17] R. Viswanathan and V. Aalo, "On counting rules in distributed detection," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 5, pp. 772–775, May 1989.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 39, pp. 1–38, 1977.

[19] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Automat. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.

[20] S. Mishra, Y. Shoukry, N. Karamchandani, S. N. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 49–59, Mar. 2017.

[21] G. Fellouris, E. Bayraktar, and L. Lai, "Efficient Byzantine sequential change detection," in *IEEE Trans. Inf. Theory*, 2017, preprint, doi: 10.1109/TIT.2017.2755025.

[22] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38. New York, NY, USA: Springer, 2009.

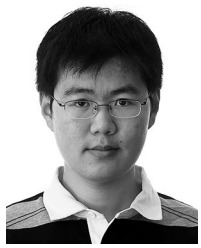
[23] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*. Hoboken, NJ, USA: Wiley, 2016.

[24] S. Key, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Xiaoqiang Ren received the B.E. degree from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, and the Ph.D. degree from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, in 2012 and 2016, respectively. From September to November 2016, he was a Research Associate with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology. He is currently a Research Fellow in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. From January 2015 to June 2015, he was a visiting student in the ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden. From August 2015 to January 2016, he was a visiting student in NeSC group, Zhejiang University. His research interests include sequential detection, security of cyber-physical systems, and networked estimation and control.

Jiaqi Yan received the B.S. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2016. She is currently working toward the Ph.D. degree at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Her research focuses on security of cyber-physical system and networked control system.



Yilin Mo received the Bachelor of Engineering degree from the Department of Automation, Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2007 and 2012, respectively. He is currently an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Prior to his current position, he was a Postdoctoral Scholar at Carnegie Mellon University in 2013 and California Institute of Technology from 2013 to 2015. His research interests include secure control systems and networked control systems, with applications in sensor networks and power grids.