On Minimal Tests of Sensor Veracity for Dynamic Watermarking-Based Defense of Cyber-Physical Systems

Bharadwaj Satchidanandan, and P. R. Kumar, *Fellow, IEEE*

***Abstract*—We address the problem of security of cyber-physical systems where some sensors may be malicious. We consider a multiple-input, multiple-output stochastic linear dynamical system controlled over a network of communication and computational nodes which contains (i) a controller that computes the inputs to be applied to the physical plant, (ii) actuators that apply these inputs to the plant, and (iii) sensors which measure the outputs of the plant. Some of these sensors, however, may be malicious. The malicious sensors do not report the true measurements to the controller. Rather, they report false measurements that they fabricate, possibly strategically, so as to achieve any objective that they may have, such as destabilizing the closed-loop system or increasing its running cost. Recently, it was shown that under certain conditions, an approach of "dynamic watermarking" can secure such a stochastic linear dynamical system in the sense that either the presence of malicious sensors in the system is detected, or the malicious sensors are constrained to adding a distortion that can only be of zero power to the noise already entering the system. The first contribution of this paper is to generalize this result to partially observed MIMO systems with both process and observation noises, a model which encompasses some of the previous models for which dynamic watermarking was established to guarantee security. This result, similar to the prior ones, is shown to hold when the controller subjects the reported sequence of measurements to two particular tests of veracity. The second contribution of this paper is in showing, via counterexamples, that both of these tests are needed in order to secure the control system in the sense that if any one of these two tests of sensor veracity is dropped, then the above guarantee does not hold. The proposed approach has several potential applications, including in smart grids, automated transportation, and process control.**

***Index Terms*—Dynamic Watermarking, Minimal Veracity Tests, Secure control, Cyber-physical systems**

I. Introduction

A major concern that has risen to the fore with the advent of societal scale cyber-physical systems (CPS) capable of meeting global challenges in areas such as energy, water, healthcare, and transportation, is their increased vulnerability to security breaches. Many recent attacks on industrial-grade control systems reinforce this concern. In the year 2010, a computer worm known as Stuxnet subverted the computers controlling the centrifuges in Iran's uranium enrichment facility and issued control commands that caused them to spin at abnormal speeds and tear themselves apart [1]. In order to ensure that the human operators in the facility did not come to know of the attack, Stuxnet recorded sensor measurements under normal operating conditions prior to each attack, and replayed those measurements in the control room in a loop during the attack. This attack is referred to as the replay attack in the literature [2], [3]. Another example is the attack on Davis-Besse nuclear power plant in Ohio, where the computers controlling the safety display systems were infected by the Slammer worm, causing them to shut down [4]. While the Slammer worm was not designed to target the power plant, the use of commodity IT solutions in computers controlling the power plant resulted in their vulnerability to generic cyber attacks. Owing to the many advantages that commodity IT solutions bring to Industrial Control Systems (ICS), such as rapid deployability and scalability, their use in ICS, and consequently the latter's vulnerability to cyber atacks, is only expected to increase in the coming years. While the aforementioned attacks originated from security breaches in the cyber layer, and could in principle be addressed by advanced network and information security mechanisms, the following incident illustrates the inadequacy of such an approach in which only the cyber layer is secured in order to secure the cyber-physical system. In the year 2000, a disgruntled employee of a sewage treatment facility in Maroochy-Shire, Australia, issued malicious control commands resulting in $800,000$ litres of raw sewage spilling out [5]. Since this attack was carried out by an insider who had valid authentication credentials and access control, network or information security mechanisms could not have prevented this security breach. This shows the need to fundamentally secure a cyber-physical system from attacks that fall entirely within the domain of the physical layer, i.e., attacks on the plant's physical signals that can be carried out via attacks on its sensors, controllers, etc. It is this topic that is addressed in this paper. Specifically, we extend to partially observed MIMO systems the approach of Dynamic Watermarking [2], [25] which secures the physical layer of a cyber-physical system.

Consider a multiple-input, multiple-output partially observed stochastic linear dynamical system controlled over a network of communication and computational nodes. Fig. 1 illustrates the architecture of such a system. At the heart of the system is a physical plant actuated by m actuators and whose outputs are measured by n sensors. Some of these sensors may be malicious. The malicious sensors may not report their measurements truthfully to the controller. Rather, they may report false measurements that are fabricated so as to achieve some

This material is based upon work partially supported by NSF under Contract Nos. CNS-1646449, CCF-1619085 and Science & Technology Center Grant CCF-0939370, the U.S. Army Research Office under Contract No. W911NF-15-1-0279, and NPRP grant NPRP 8-1531-2-651 from the Qatar National Research Fund, a member of Qatar Foundation.

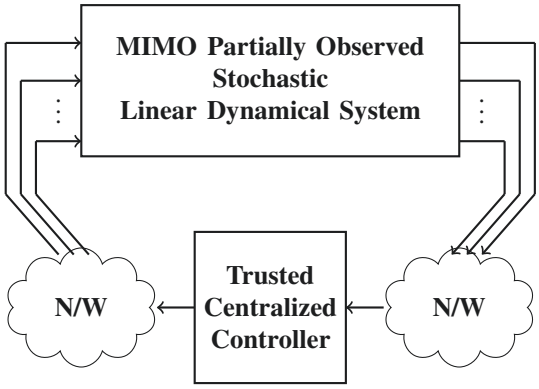

Fig. 1. A Networked Cyber-Physical System

malicious objective that they may have, such as destabilizing the closed-loop system or increasing its running cost. A trusted centralized controller receives measurements from the sensors, and based on these measurements and a specified control policy, computes the inputs to be applied by the actuators. The controller then communicates this information to the actuators which ultimately apply the inputs. The actuators are assumed to have minimal or no computational capabilities, so that they can be assumed as being honest. All the communication may take place over an underlying communication network such as the Internet. If the communication network is assumed to be secured using a combination of traditional approaches such as cryptography, and more recent ideas such as those reported in [6], [7], then one may abstract the communication network as a set of secure, reliable, delay-guaranteed bit pipes between all pairs of nodes in the system. However, it is worth noting that the results reported in this paper also apply to the scenario where only the communication links between the controller and the actuators are secure if they are not collocated, while the other links in the communication network may not be so. We show that for such a system, under certain conditions, dynamic watermarking ensures that the malicious sensors are restricted to adding a distortion to the system's innovations process, the only information required by the controller for controlling the system, that can only be of zero power. This is the fundamental security guarantee provided by dynamic watermarking. It follows that for the class of systems that are open-loop stable, the above result is sufficient to ensure that the malicious sensors cannot destabilize the system, or under a nominal linear control law, cause a quadratic cost of the system to deviate from its optimal value.

The rest of the paper is organized as follows. Section-II provides an account of related work in this area. Section-III describes dynamic watermarking and establishes (i) the fundamental security guarantee provided by dynamic watermarking for partially observed MIMO systems, and (ii) the minimality of the two particular tests of sensor veracity that the aforementioned results are based upon. Section-IV presents simulation results.

II. Related Work

Initial work on secure CPS involved defining the objective of secure control and identifying distinctive features that make it different from fields such as network and information security [8], [9]. Certain key operational goals are identified such as closed-loop stability, and it is proposed that secure control constitutes the maintenance of these key operational goals even under attack, or in the case of cost functions, their graceful degradation. A model for two well-known attacks, viz., the Denial-of-Service (DoS) attack and the deception attack, are presented in [9]. Reference [10] addresses the problem of optimal control under DoS attack.

A standard detection algorithm employed in SCADA systems is the bad data detection (BDD) algorithm [11]. An unobservable attack is defined in [12] as an attack that cannot be detected by a BDD. In order to assess the vulnerability of a given system to unobservable attacks, [13] defines an index termed as "security index". Computation of this index is in general NP-hard, and a method for its efficient, approximate computation is presented in [14].

A particular attack strategy called the packet reordering integrity attack is studied in [15]. In this attack model, the adversary is assumed to have subverted the underlying communication network. The adversary then reorders the sequence of innovations process that the sensors send to the controller, so that the statistical properties of the reported innovations sequence are no different from that of the actual innovations sequence. This ensures that the residue-based anomaly detection algorithms that may be employed in the system do not detect the attack. The effect of packet reordering attack on the state estimation error is then analyzed, and the optimal packet reordering strategy which maximizes the state estimation error is derived.

Reference [16] considers the setup where the adversary has the capability to read and modify all sensor measurements. It restricts the adversary to be linear, i.e., the reported measurement is some linear transformation of the actual data observed by the sensors, and determines linear attack strategy which maximizes the state estimation error subject to the constraint that the attack is not detected by a residue-based detection algorithm.

The disclosure resources of an attacker denote those resources that enable the attacker to gather certain real-time system data. For example, a subverted communication link through which a sensor transmits its measurements constitutes a disclosure resource since it enables the adversary to gather real-time sensor measurements. Similarly, the disruption resources of an attacker are defined as those resources that enable the attacker to inject malicious signals into the system. Based on the attacker's (i) system knowledge, (ii) disclosure resouces, and (iii) disruption resources, an attack space is defined in [17], and commonly known attacks such as DoS attack, replay attack, zero-dynamics attack, etc., are mapped into the attack space and analyzed.

At a high level, the aforementioned papers analyze the security vulnerability of CPS, mathematically model known attacks for CPS, present novel attack strategies on CPS,

and analyze their consequences. A parallel body of research focuses on defending a CPS from such attacks. One of the fundamental problems encountered is that of attack detection. Fundamental limits of attack detection and identification for three classes of detectors, viz., static, dynamic, and active, are presented in [18]. Here, attack detection refers to detection of the presence of adversarial nodes in the system, whereas attack identification refers to determining the identity of malicious nodes in the system. A static monitor refers to a detection algorithm that does not exploit system dynamics. A typical example of a static monitor is the Bad Data Detector [11]. A dynamic monitor, on the other hand, processes a time series of measurements and uses its knowledge of the system dynamics to determine whether or not the system is under attack. An active monitor is a dynamic monitor which also excites the system using inputs unknown to other entities in the system. The approach of dynamic watermarking falls in this category.

The resilience of state estimation in CPS to malicious sensors is characterized in [19], [20], and for the case when the number of malicious sensors is lesser than or equal to an appropriate measure of resiliency, an algorithm for optimal state estimation is developed. Attacks that cannot be detected using the system's inputs and outputs are termed as zero-dynamics attacks. An approach of perturbing system parameters to detect zero-dynamics attacks is presented in [21]. A data verification framework for detection and removal of malicious measurements from a wireless sensor network is presented in [22]. The basic idea is to exploit correlations between the measurements of different sensors to identify malicious reports.

Qualitatively, the aforementioned defense mechanisms can be classified as passive defense mechanisms. By and large, this is the approach that has dominated the literature thus far. An alternate paradigm to securing cyber-physical systems, called "Physical Watermarking" in [23], is that of active defense, where the controller injects a certain random signal into the system in addition to the control policy-specified input. We term this random signal the controller's "private excitation." The actual realization of the private excitation is unknown to other nodes in the system. This private excitation which is applied by the actuators evokes a particular response from the sensors in the system in accordance with the system dynamics. Therefore, by subjecting the reported sequence of measurements to some carefully designed tests, one can check if the reported measurements are appropriately correlated with the private excitation. This in turn can reveal the presence of malicious sensors in the system. The papers [2], [3], [24] were the first, to our knowledge, to investigate this idea of active defense, and used it to detect replay attacks. The idea was then extended in [23] to detect more intelligent attack strategies. However, a common aspect of these papers is that the reported sequence of measurements are subjected to only one test of sensor veracity, essentially Test 1 described in this paper. While this may ensure that certain *specific* attack policies don't pass the test and are hence detected, it need not be effective in the face of *arbitrary* attack policies. In Section-III, we construct an explicit attack policy that causes significant performance degradation to the control system, but nevertheless passes the above test. In [25], [26], it has been shown that by subjecting the reported sequence of sensor measurements to an additional test of sensor veracity, one can in fact ensure that even malicious sensors employing *arbitrary* attack strategies are constrained to adding a distortion that can only be of zero power to the process noise already entering the system. This is the fundamental security guarantee provided by this method that is called "dynamic watermarking" in [25]. In this paper, we generalize this result to a more general case of partially observed MIMO systems. We also show that both of the tests are needed in our approach in the sense that neither of them can be dropped from the set if the aforementioned result is to hold.

III. Dynamic Watermarking: The case of Partially Observed MIMO Systems

Consider a p^{th} order $m \times n$ partially observed MIMO stochastic linear dynamical system described by

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + B\mathbf{u}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1] + \mathbf{n}[t+1], \end{aligned} \tag{1}$$

where $\mathbf{x}[t] \in \mathbb{R}^p$ is the system's state at time t, $\mathbf{u}[t] \in \mathbb{R}^m$ and $\mathbf{y}[t] \in \mathbb{R}^n$ are respectively the system's input and output at time t, $\mathbf{w}[t] \sim \mathcal{N}(0, Q)$ and $\mathbf{n}[t] \sim \mathcal{N}(0, R)$ with $R > 0$ are respectively the process and observation noises at time t, and A, B, C are known matrices of appropriate dimensions which specify the system dynamics. We assume that the random processes $\{\mathbf{w}\}$ and $\{\mathbf{n}\}$ are independent, and also that each of them is i.i.d. across time.

We denote by $\mathbf{z}[t]$ the measurements reported by the sensors at time t to the controller. A truthful sensor is supposed to report $\mathbf{z} \equiv \mathbf{y}$, but a malicious sensor may report any values for $\{\mathbf{z}\}$. We assume the existence of a general history-dependent control policy $\{g_t\}$ according to which the controller computes the input that the actuators should apply at each time t. This control policy is made public, meaning that it is known to all the nodes in the system. While the control policy is supposed to be applied on the actual output sequence $\{\mathbf{y}\}$, since the controller does not directly measure the plant's outputs, it is implemented on $\{\mathbf{z}\}$ as reported by the sensors. Additionally, as outlined in the previous section, in order to secure the system from malicious sensors, the controller commands the actuators to superimpose a private excitation sequence $\{\mathbf{e}\}$ on the sequence of control policy-specified inputs. Hence, the net input applied to the system at time t is given by $u[t] = g_t(\mathbf{z}^t) + \mathbf{e}[t]$, where $\mathbf{e}[t] \sim \mathcal{N}(0, \sigma_e^2 I)$, i.i.d across time, is the controller's private excitation, and $\mathbf{z}^t := (\mathbf{z}[0], \mathbf{z}[1], ..., \mathbf{z}[t])$ denotes the past values of $\{\mathbf{z}\}$. Consequently, the system evolves as

$$\begin{aligned} \mathbf{x}[t+1] &= A\mathbf{x}[t] + Bg_t(\mathbf{z}^t) + B\mathbf{e}[t] + \mathbf{w}[t+1], \\ \mathbf{y}[t+1] &= C\mathbf{x}[t+1] + \mathbf{n}[t+1], \end{aligned} \tag{2}$$

where $g_t(\mathbf{z}^t) \coloneqq [g_t^1(\mathbf{z}^t)\ , g_t^2(\mathbf{z}^t), ..., g_t^m(\mathbf{z}^t)]^T$.

Assume that (A, C) is observable, and $(A, Q^{\frac{1}{2}})$ is reachable. The controller performs Kalman filtering on the reported sequence of measurements as follows. Let $\mathbf{x}_F(k|k)$ denote the estimate of the state $\mathbf{x}[k]$ given the information upto time k,

i.e., $(\mathbf{z}^k, \mathbf{e}^{k-1})$, and $\mathbf{x}_F(k|k-1)$ denote the estimate given the information upto time $k-1$. They are given by the Kalman filtering equations:

$$\mathbf{x}_F(k+1|k+1) = A\mathbf{x}_F(k|k) + Bg_k(\mathbf{z}^k) + B\mathbf{e}[k] + K_{k+1}\boldsymbol{\nu}_F[k+1], \quad (3)$$

where $\nu_F[k+1] \coloneqq \mathbf{z}[k+1] - C\mathbf{x}_F(k+1|k)$. We note that if the sensors were truthful, the estimates above would be the conditional mean estimates, and $\boldsymbol{\nu}[t+1]$ would be the innovations process [27] at time t. However, the sensor may be malicious, and so we refer to $\boldsymbol{\nu}_F[t+1]$ as the "false innovations" at time $t+1$. For the purpose of analysis, we also define the "true" Kalman filter which operates on $\{\mathbf{y}\}$:

$$\mathbf{x}_T(k+1|k+1) = A\mathbf{x}_T(k|k) + Bg_k(\mathbf{z}^k) + B\mathbf{e}[k] + K_{k+1}\boldsymbol{\nu}_T[k+1], \quad (4)$$

where $\boldsymbol{\nu}_T[k+1] \coloneqq \mathbf{y}[k+1] - C\mathbf{x}_T(k+1|k)$ is the "true innovations" at time $k+1$. We suppose that the Kalman filters are initialized with the Kalman gain K_0 set to its steady-state value K so that they behave as time-invariant filters [28].

The dynamic watermarking tests we employ are based on the following two observations that hold for the true Kalman filter:

1) $\mathbf{e}[k]$ is independent of $K\boldsymbol{\nu}_T[k+1]$, and
2) the sequence of conditional estimates $\{\mathbf{x}_T(k|k)\}$ satisfies

$$\{\mathbf{x}_T(k+1|k+1) - A\mathbf{x}_T(k|k) - Bg_k(\mathbf{y}^k) - B\mathbf{e}[k]\} \sim \mathcal{N}(0, K\Sigma K^T),$$

where

$$K = PC^T(CPC^T + R)^{-1} \quad (5)$$

is the steady-state Kalman gain of the Kalman filter,

$$\Sigma = CPC^T + R \quad (6)$$

is the steady-state covariance matrix of the true innovations process, and P is the unique nonnegative definite solution of the discrete algebraic Riccati equation $P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T$. The unique nonnegative definite solution P is guaranteed to exist for the above Riccati equation since (A, C) is observable, $(A, Q^{\frac{1}{2}})$ is reachable, and $R > 0$ [28]. The matrix P has the interpretation of the covariance matrix of the one-step ahead state prediction error of the Kalman filter [28].

The controller therefore performs the following two tests on the reported sequence of observations $\{\mathbf{z}\}$:

1) **Controller Test 1:** Check if the sequence of reported measurements satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} (\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k])(\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k])^T = K\Sigma K^T. \quad (7)$$

2) **Controller Test 2:** Check if the sequence of reported observations satisfies

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k](\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k])^T = 0. \quad (8)$$

The above tests are equivalent to the tests proposed in [25]. In particular, the second test above can be shown, via straightforward algebraic manipulations, to be equivalent to the corresponding test for first-order SISO systems considered in [25]. We define

$$\mathbf{v}[k+1] \coloneqq \mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k] - K\boldsymbol{\nu}_T[k+1], \quad (9)$$

and note that if there are no malicious sensors in the system, $\mathbf{v} \equiv 0$. We will call the quantity

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathbf{v}[k]\|^2$$

as the *additive distortion power* of the malicious sensors, for reasons explained later in the paper.

The following theorem, which is a generalization of the results in [25], establishes the fundamental security guarantee provided by dynamic watermarking.

Theorem 1. *Suppose that (A, C) is observable, $(A, Q^{\frac{1}{2}})$ is reachable, and $R > 0$. Further suppose that the matrix CB is of rank n. Then, if the reported measurements $\{\mathbf{z}\}$ pass both (8) and (7), it can be guaranteed that the additive distortion is of zero power, i.e.,*

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathbf{v}[k]\|^2 = 0. \quad (10)$$

Proof. We appeal to the following lemma.

Lemma 1: Define $M \coloneqq \Sigma(\Sigma + \sigma_e^2 CBB^T C^T)^{-1}$. If CB is of rank n, then, $(I - M)^{-1}$ exists.

Proof.

$$I - M = I - \Sigma[\Sigma + \sigma_e^2 CBB^T C^T]^{-1}. \quad (11)$$

In the above, $[\Sigma + \sigma_e^2 CBB^T C^T]^{-1}$ is guaranteed to exist since $\Sigma > 0$ (because $R > 0$). Its inverse is

$$(I - M)^{-1} = I + \Sigma(\sigma_e^2 CBB^T C^T)^{-1}, \quad (12)$$

since $(\sigma_e^2 CBB^T C^T)^{-1}$ exists because CB has rank n. □

Since the reported measurements pass (8), we have using (9),

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k](K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1])^T = 0.$$

Since $\mathbf{e}[k]$ is independent of the innovations $\boldsymbol{\nu}_T[k+1]$, the above simplifies to

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathbf{e}[k]\mathbf{v}^T[k+1] = 0. \quad (13)$$

Since the reported measurements also pass (7), we have using (9),

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}(K\boldsymbol{\nu}_T[k+1]+\mathbf{v}[k+1])$$
$$(K\boldsymbol{\nu}_T[k+1]+\mathbf{v}[k+1])^T = K\Sigma K^T.$$

Simplifying the above gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=0}^{T-1}K\boldsymbol{\nu}_T[k+1]\mathbf{v}^T[k+1]$$
$$+(K\boldsymbol{\nu}_T[k+1]\mathbf{v}^T[k+1])^T+\mathbf{v}[k+1]\mathbf{v}^T[k+1]=0. \quad (14)$$

Define the σ-algebra $\mathcal{S}_{k+1} \coloneqq \sigma(\mathbf{y}^{k+1},\mathbf{z}^{k+1},\mathbf{e}^{k-1},\mathbf{x}_T^{k|k})$, where $\mathbf{x}_T^{k|k} \coloneqq (\mathbf{x}_T(0|0),\mathbf{x}_T(1|1),...,\mathbf{x}_T(k|k))$. We also define $\widehat{\boldsymbol{\nu}}_T[k] \coloneqq \mathbb{E}[\boldsymbol{\nu}_T[k]\big|\mathcal{S}_k]$, and $\widetilde{\boldsymbol{\nu}}_T[k] \coloneqq \boldsymbol{\nu}_T[k]-\widehat{\boldsymbol{\nu}}_T[k]$. Then, from the definition of the innovations at time $k+1$, we have

$$\boldsymbol{\nu}_T[k+1] \coloneqq \mathbf{y}[k+1]-C\mathbf{x}_T(k+1|k)$$
$$=\mathbf{y}[k+1]-CA\mathbf{x}_T(k|k)-CBg_k(\mathbf{z}^k)-CB\mathbf{e}[k]. \quad (15)$$

From the above, it follows that

$$(\mathbf{y}^k,\mathbf{e}^{k-1},\mathbf{x}_T^{k-1|k-1}) \to$$
$$(\mathbf{x}_T(k|k),\mathbf{y}[k+1],\mathbf{z}^{k+1}) \to \boldsymbol{\nu}_T[k+1]$$

forms a Markov chain. Therefore,

$$\widehat{\boldsymbol{\nu}}_T[k+1] \coloneqq \mathbb{E}[\boldsymbol{\nu}_T[k+1]\big|\sigma(\mathbf{y}^{k+1},\mathbf{z}^{k+1},\mathbf{e}^{k-1},\mathbf{x}_T^{k|k})]$$
$$=\mathbb{E}[\boldsymbol{\nu}_T[k+1]\big|\sigma(\mathbf{x}_T(k|k),\mathbf{y}[k+1],\mathbf{z}^{k+1})].$$

Combining the above with (15), we have the MMSE estimate as [28, Chapter 7, Lemma 2.5]

$$\widehat{\boldsymbol{\nu}}_T[k+1]=M(CB\mathbf{e}[k]+\boldsymbol{\nu}_T[k+1]). \quad (16)$$

Hence,

$$\boldsymbol{\nu}_T[k+1]=\widehat{\boldsymbol{\nu}}_T[k+1]+\widetilde{\boldsymbol{\nu}}_T[k+1]$$
$$=MCB\mathbf{e}[k]+M\boldsymbol{\nu}_T[k+1]+\widetilde{\boldsymbol{\nu}}_T[k+1]$$

Rearranging and using Lemma 1, we have

$$\boldsymbol{\nu}_T[k+1]=(I-M)^{-1}MCB\mathbf{e}[k]+(I-M)^{-1}\widetilde{\boldsymbol{\nu}}_T[k+1]. \quad (17)$$

Now, the RHS of (15), and hence $\boldsymbol{\nu}_T[k+1]$, is measurable with respect to $\mathcal{S}_{k+2}$. Also, since $\widehat{\boldsymbol{\nu}}_T[k+1]\in\mathcal{S}_{k+1}\subset\mathcal{S}_{k+2}$, it follows that $\widetilde{\boldsymbol{\nu}}_T[k+1]\in\mathcal{S}_{k+2}$. Clearly, $\mathbb{E}[\widetilde{\boldsymbol{\nu}}_T[k+2]\big|\mathcal{S}_{k+2}]=0$. Hence, we have that $(\widetilde{\boldsymbol{\nu}}_T[k+1],\mathcal{S}_{k+2})$ is a martingale difference sequence. Moreover, since $\mathbf{v}[k+1]$, after some algebra, can be expressed as $K(\mathbf{z}[k+1]-\mathbf{y}[k+1])-KCA(\mathbf{x}_F(k|k)-\mathbf{x}_T(k|k))$, we have $\mathbf{v}[k+1]\in\mathcal{S}_{k+1}$. Hence, Martingale Stability Theorem (MST) [29, Lemma 2(iii)] holds, and we have

$$\sum_{k=0}^{T-1}\widetilde{\boldsymbol{\nu}}_T[k+1]\mathbf{v}^T[k+1]=$$
$$\begin{bmatrix}o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k])\\ \vdots & \vdots & \vdots\\ o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k])\end{bmatrix}+[O(1)]_{p\times p}, \quad (18)$$

where $[O(1)]_{p\times p}$ denotes a $p\times p$ matrix all of whose entries are $O(1)$. Substituting (17) in (14) and using (13) and (18) yields

$$\sum_{k=1}^T\mathbf{v}[k]\mathbf{v}^T[k]+\begin{bmatrix}o(T) & \dots & o(T)\\ \vdots & \vdots & \vdots\\ o(T) & \dots & o(T)\end{bmatrix}$$
$$+\begin{bmatrix}o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k])\\ \vdots & \vdots & \vdots\\ o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k])\end{bmatrix}$$
$$+\begin{bmatrix}o(\sum_{k=1}^T v_1^2[k]) & \dots & o(\sum_{k=1}^T v_1^2[k])\\ \vdots & \vdots & \vdots\\ o(\sum_{k=1}^T v_p^2[k]) & \dots & o(\sum_{k=1}^T v_p^2[k])\end{bmatrix}=o(T). \quad (19)$$

Dividing the above by T, equating the trace, and letting $T\to\infty$ completes the proof. □

We now show that if one drops either of the two controller tests (8) or (7), then the guarantee does not hold. We do so by explicitly constructing two attack strategies, each of which passes exactly each one of the tests, and yet, $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^T\|\mathbf{v}[k]\|^2\neq 0$ for both the attacks.

Consider a special case of system (1), viz., a SISO first-order perfectly observed system ($p=m=n=C=1, R=0, Q=\sigma_w^2\in\mathbb{R}_+$). In that case, $\mathbf{x}_F(k|k)$ reduces to $\mathbf{z}[k]$, $\boldsymbol{\nu}_T[k]$ to $w[k]$, K to 1, and Σ to σ_w^2. Consequently, tests (7) and (8) reduce respectively to

1) **Test 1:** Check if the reported sequence of measurements $\{z\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^T(z[k+1]-Az[k]-Bg_k(z^k)-Be[k])^2$$
$$=\sigma_w^2, \quad (20)$$

2) **Test 2:** Check if the reported sequence of measurements $\{z\}$ satisfies

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^T e[k](z[k+1]-Az[k]-Bg_k(z^k)$$
$$-Be[k])=0, \quad (21)$$

and $v[t+1]=z[t+1]-Az[t]-Bg_t(z^t)-Be[t]-w[t+1]$.

Counterexample showing Controller Test 1 alone is not sufficient: Suppose that the reported measurements are subjected to (20) alone, the first test. To show that this is not sufficient to guarantee zero additive distortion power by the malicious sensor, consider the following attack.

Suppose that the malicious sensor reports measurements $\{z\}$ generated as

$$z[k+1] = Az[k] + Bg_k(z^k) + (\frac{B^2\sigma_e^2 - \sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2})(y[k+1] - Ay[k] - Bg_k(z^k)). \quad (22)$$

We now show that the sequence $\{z\}$ so generated passes (20), the first test.

Define $\gamma[k] \coloneqq z[k] - Az[k-1] - Bu^g[k-1] - Be[k-1]$, the quantity whose second moment is being empirically tested in (20). Then, we have

$$\begin{aligned}\gamma[k] &= z[k] - Az[k-1] - Bg_{k-1}(z^{k-1}) - Be[k-1] \\ &= (\frac{B^2\sigma_e^2 - \sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2})(y[k] - Ay[k-1] - Bg_{k-1}(z^{k-1}) \\ &\quad - Be[k-1]) \\ &= (\frac{B^2\sigma_e^2 - \sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2})(Be[k-1] + w[k]) - Be[k-1] \\ &= -\frac{2\sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2}Be[k-1] + (\frac{B^2\sigma_e^2 - \sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2})w[k]. \quad (23)\end{aligned}$$

From the above, it is clear that $\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} \gamma^2[k]$ of (20) is simply the variance of the RHS of the above, given by

$$(\frac{-2\sigma_w^2 B}{B^2\sigma_e^2 + \sigma_w^2})^2 + (\frac{B^2\sigma_e^2 - \sigma_w^2}{B^2\sigma_e^2 + \sigma_w^2})^2.$$

This simplifies to σ_w^2, and hence, this attack passes Test 1.

Finally, for the above attack, it is easy to see that $v[k+1] = \gamma[k+1] - w[k+1] = -\frac{2\sigma_w^2}{B^2\sigma_e^2+\sigma_w^2}(Be[k]+w[k+1])$, and hence, $\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} v^2[k] = \frac{4\sigma_w^4}{B^2\sigma_e^2+\sigma_w^2} \neq 0$. □

Counterexample showing Controller Test 2 alone is not sufficient: Now suppose that the reported measurements are subjected to (21) alone. To show that this is not sufficient to guarantee zero additive distortion power by the malicious sensor, consider the following attack. The sensor reports measurements $\{z\}$ generated as

$$z[k+1] = Az[k] + Bg_k(z^k) + (y[k+1] - Ay[k] - Bg_k(z^k) + \lambda[k+1]), \quad (24)$$

where $\lambda[k+1] \sim \mathcal{N}(0, \sigma_\lambda^2)$ is chosen by the sensor in an i.i.d. fashion across time, and also independently of all random variables that it has observed till then.

To show that the sequence $\{z\}$ so generated passes (21), the second test, note that

$$\begin{aligned}&\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} e[k](z[k+1] - Az[k] - Bg_k(z^k) - Be[k]) \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} e[k](y[k+1] - Ay[k] - Bg_k(z^k) \\ &\qquad + \lambda[k+1] - Be[k]) \\ &= \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} e[k](w[k+1] + \lambda[k+1]).\end{aligned}$$

Since $w[k+1]$ and $\lambda[k+1]$ are independent of $e[k]$, the above reduces to 0, thereby passing (21), and hence, (8).

However, for the above attack, it is easy to see that $v[k+1] = \lambda[k+1]$, and hence, $\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} v^2[k] = E[\lambda^2[k]] = \sigma_\lambda^2 \neq 0$. □

Remark: It is well known that the innovations process of a stochastic process is a causal and causally invertible transformation of the stochastic process with the property that it is uncorrelated across time [27]. Hence, the innovations at time t can be thought of as summarizing all the "new" information provided by the sensors at time t, information that could not have been predicted from the past. Therefore, one can think of the honest sensors' purpose as being to report the innovations at each time t. Now, from (9), we have $\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k] = K\boldsymbol{\nu}_T[k+1] + \mathbf{v}[k+1]$. The LHS of the above can be computed by the controller. Hence, $\mathbf{v}[k+1]$ has a physical interpretation as the distortion added by the malicious sensors to the true innovations at time $k+1$ (hence the nomenclature for additive distortion power). What the above theorem says is that the malicious sensors cannot distort the true innovations process beyond adding a zero-power sequence to it if they wish to remain undetected.

We now consider linear control designs that provide some guarantee on the quadratic state tracking error. The design need not be an optimal LQG design, but one that merely aims at providing some upper bound on the aforementioned quantity. The following theorem shows that for stable systems, guaranteeing that any additive distortion is of power zero is sufficient to ensure that the malicious sensors do not increase the quadratic cost of the state from its design value in the case of linear designs.

Theorem 2. *Suppose that the system (1) is open-loop stable, i.e., A has all its eigenvalues in the open left half-plane, and define*

$$\mathbf{d}[k] \coloneqq \mathbf{x}_F(k|k) - \mathbf{x}_T(k|k). \quad (25)$$

If the reported measurements pass (8) and (7), then,

1) $$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{d}[k]\|^2 = 0. \quad (26)$$

2) *Suppose that the control policy is a linear feedback policy $g_t(\mathbf{z}^t) = F\mathbf{x}_F(t|t)$, and a control objective is quadratic regulation. Then, the true quadratic regulation performance $\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2$ of the system is no different from what the controller thinks it is, in the sense that*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 = \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2 \quad (27)$$

3) *Under the same conditions as (2) above, the malicious sensors cannot increase the quadratic regulation cost of the system $\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2$ from the value that would be obtained if all the sensors were honest.*

Proof. Subtracting (4) from (3), we have $\mathbf{d}[k+1] = A\mathbf{d}[k] + K(\boldsymbol{\nu}_F[k+1] - \boldsymbol{\nu}_T[k+1])$. From (3), we have $K\boldsymbol{\nu}_F[k+1] =$

$\mathbf{x}_F(k+1|k+1) - A\mathbf{x}_F(k|k) - Bg_k(\mathbf{z}^k) - B\mathbf{e}[k]$. Combining this with (9) gives $K(\boldsymbol{\nu}_F[k+1] - \boldsymbol{\nu}_T[k+1]) = \mathbf{v}[k+1]$. Substituting this in the above equation gives $\mathbf{d}[k+1] = A\mathbf{d}[k] + \mathbf{v}[k+1]$. Since $\{\mathbf{v}\}$ is of zero power and A is stable, result (1) follows.

Now, from (25), we have $\mathbf{x}_F(k|k) = \mathbf{x}_T(k|k) + \mathbf{d}[k]$. By triangular inequality, we have $\|\mathbf{x}_F(k|k)\| \leq \|\mathbf{x}_T(k|k)\| + \|\mathbf{d}[k]\|$. Hence, $\|\mathbf{x}_F(k|k)\|^2 \leq \|\mathbf{x}_T(k|k)\|^2 + \|\mathbf{d}[k]\|^2 + 2\|\gamma\mathbf{x}_T(k|k)\|\|\gamma^{-1}\mathbf{d}[k]\|$ for all $\gamma > 0$. Since $2\|\gamma\mathbf{x}_T(k|k)\|\|\gamma^{-1}\mathbf{d}[k]\| \leq \|\gamma\mathbf{x}_T(k|k)\|^2 + \|\gamma^{-1}\mathbf{d}[k]\|^2$, substituting this in the above yields $\|\mathbf{x}_F(k|k)\|^2 \leq (1+\gamma^2)\|\mathbf{x}_T(k|k)\|^2 + (1+\gamma^{-2})\|\mathbf{d}[k]\|^2$. Hence,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 \leq \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} (1+\gamma^2)\|\mathbf{x}_T(k|k)\|^2 + \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} (1+\gamma^{-2})\|\mathbf{d}[k]\|^2$$

The second term reduces to zero from the previous result. Since the above is true for all $\gamma > 0$, taking $\gamma \to 0$ gives

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2 \leq \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2. \quad (28)$$

Similarly, from (25), we have $\mathbf{x}_T(k|k) = \mathbf{x}_F(k|k) - \mathbf{d}[k]$. Hence, $\|\mathbf{x}_T(k|k)\| = \|\mathbf{x}_F(k|k) - \mathbf{d}[k]\| \leq \|\mathbf{x}_F(k|k)\| + \|\mathbf{d}[k]\|$. Continuing as above, we arrive at

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_T(k|k)\|^2 \leq \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} \|\mathbf{x}_F(k|k)\|^2. \quad (29)$$

Combining the above with (28) gives the second result.

It follows from the above result that even though the controller does not have access to the true measurements $\{\mathbf{y}\}$, it can empirically compute the true quadratic regulation cost $\mathbb{E}[\|\mathbf{x}_T(k|k)\|^2]$ of the system. It follows that the malicious sensors cannot increase the true quadratic regulation cost of the system from its design value without exposing their presence. □

IV. Simulation Results

This section presents the simulation results of the attacks presented in the previous section. The attacks are simulated for system parameters $A = 0.5, B = 1, \sigma_w^2 = 2$, and with $\sigma_e^2 = 1$, $g_k(z^k) = -0.1z[k]$.

We first present the results for the attack that passes Test 1 alone. We call this "Attack 1." Fig. 2 plots $\frac{1}{t}\sum_{k=1}^{t} \gamma^2[k]$ as a function of t, where $\gamma[k]$ is computed using the measurements generated using (22). It can be seen that it approaches the value of σ_w^2, thereby passing Test 1.

Also, for Attack 1, we have the additive distortion power $\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} v^2[k] = \frac{4\sigma_w^4}{B^2\sigma_e^2+\sigma_w^2} = \frac{16}{3}$. Fig. 3 plots $\frac{1}{t}\sum_{k=1}^{t} v^2[k]$ as a function of t, and it can be seen that it indeed approaches $\frac{16}{3} = 5.3$, showing that the additive distortion power is not equal to zero. This shows that Test 1 alone is not sufficient to guard against Attack 1.

Next, we present analogous results for the attack passing Test 2 alone. We call this "Attack 2". The adversary is simulated with $\sigma_\lambda^2 = 4$. Fig. 4 plots $\frac{1}{t}\sum_{k=1}^{t} e[k](z[k+1] - Az[k] - Bg_k(z^k) - Be[k])$ as a function of t, where $\{z\}$ is computed using (24). It can be seen that it approaches the value of 0, thereby passing Test 2.

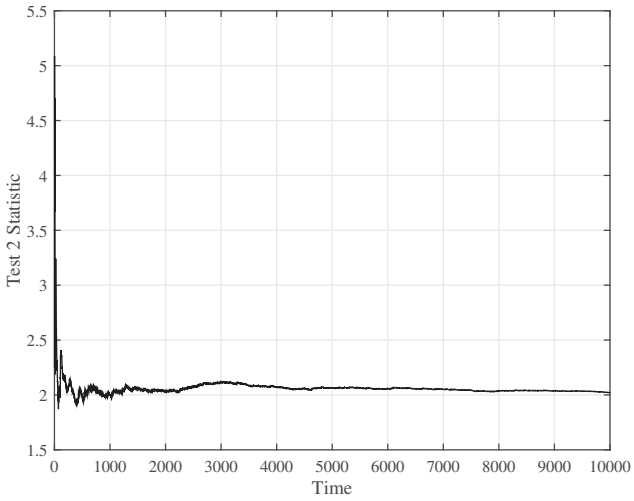

Fig. 2. Test statistic of Test 1 vs. time

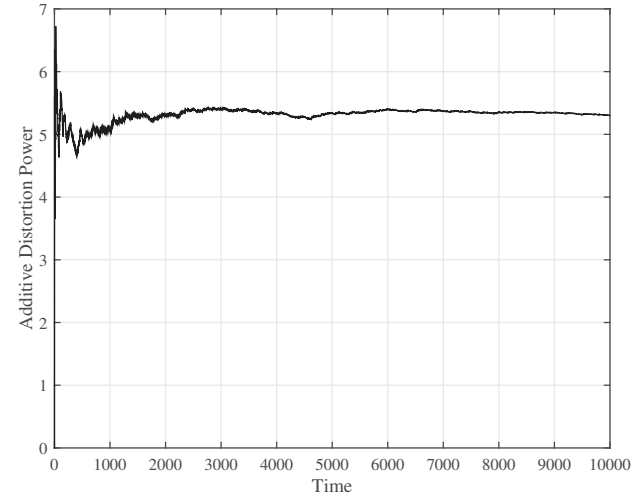

Fig. 3. Additive Distortion Power of Attack 1 vs. time

Finally, for Attack 2, the additive distortion power $\lim_{T\to\infty} \frac{1}{T}\sum_{k=1}^{T} v^2[k] = \sigma_\lambda^2 = 4$. Fig. 5 plots $\frac{1}{t}\sum_{k=1}^{t} v^2[k]$ as a function of t, and it can be seen that it indeed approaches the value of 4, showing that the additive distortion power is not equal to zero. This shows that Test 2 alone is not sufficient to guard against Attack 2.

V. Conclusion

This paper addresses the problem of securing the physical layer of a cyber-physical system from malicious sensors. The approach of dynamic watermarking was developed for partially observed stochastic MIMO linear dynamical system, a model which encompasses some of the previous models for which dynamic watermarking was established to guarantee

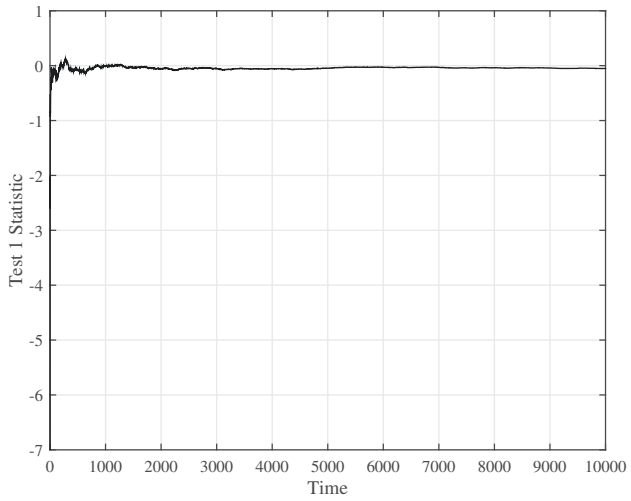

Fig. 4. Test statistic of Test 2 vs. time

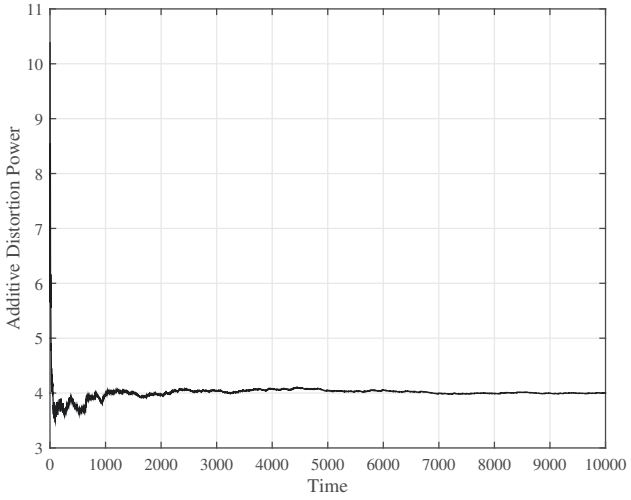

Fig. 5. Additive Distortion Power of Attack 2 vs. time

security. These guarantees are contingent on the controller conducting two particular tests of sensor veracity, and it was shown via explicit construction of two attack strategies that both of these tests are required in that neither can be dropped if the security guarantees are to hold.

REFERENCES

[1] R. Langner, "Stuxnet: Dissecting a Cyberwarfare Weapon," *Security & Privacy, IEEE*, vol. 9, no. 3, pp. 49–51, 2011.

[2] Y. Mo and B. Sinopoli, "Secure Control Against Replay Attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Sept 2009.

[3] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting Integrity Attacks on SCADA Systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[4] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for Securing Cyber Physical Systems," in *Workshop on future directions in cyber-physical systems security*, 2009.

[5] M. Abrams, "Malicious Control System Cyber Security Attack Case Study-Maroochy Water Services, Australia," 2008.

[6] J. Ponniah, Y.-C. Hu, and P. R. Kumar, "A Clean Slate Approach to Secure Wireless Networking," *Foundations and Trends in Networking*, vol. 9, no. 1, pp. 1–105, 2014. [Online]. Available: http://dx.doi.org/10.1561/1300000037

[7] I.-H. Hou, V. Borkar, and P. R. Kumar, "A Theory of QoS for Wireless," in *IEEE INFOCOM*. IEEE, 2009.

[8] A. A. Cardenas, S. Amin, and S. Sastry, "Secure Control: Towards Survivable Cyber-Physical Systems," in *The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008.

[9] ——, "Research Challenges for the Security of Control Systems."

[10] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and Secure Networked Control Systems under Denial-of-Service Attacks," in *Hybrid Systems: Computation and Control*. Springer, 2009.

[11] A. Abur and A. G. Exposito, *Power system state estimation: theory and implementation*. CRC press, 2004.

[12] K. C. Sou, H. Sandberg, and K. H. Johansson, "Data attack isolation in power networks using secure voltage magnitude measurements," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 14–28, 2014.

[13] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First Workshop on Secure Control Systems (SCS), Stockholm, 2010*, 2010.

[14] J. M. Hendrickx, K. H. Johansson, R. M. Jungers, H. Sandberg, and K. C. Sou, "Efficient computations of a security index for false data attacks in power networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3194–3208, 2014.

[15] Z. Guo, K. H. Johansson, and L. Shi, "A study of packet-reordering integrity attack on remote state estimation," in *2016 35th Chinese Control Conference (CCC)*, July 2016, pp. 7250–7255.

[16] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation."

[17] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[18] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack Detection and Identification in Cyber-Physical Systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[19] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure State-Estimation for Dynamical Systems under Active Adversaries," in *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011.

[20] ——, "Secure Estimation and Control for Cyber-Physical Systems under Adversarial Attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[21] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, Oct 2012, pp. 1806–1813.

[22] S. Gisdakis, T. Giannetsos, and P. Papadimitratos, "SHIELD: A Data Verification Framework for Participatory Sensing Systems," in *Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, ser. WiSec '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2766498.2766503

[23] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting Integrity Attacks on Control Systems using Robust Physical Watermarking," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 3757–3764.

[24] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical Authentication of Control Systems: Designing Watermarked Control Inputs to Detect Counterfeit Sensor Outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, Feb 2015.

[25] B. Satchidanandan and P. R. Kumar, "Dynamic Watermarking: Active Defense of Networked Cyber-Physical Systems," *Proceedings of the IEEE, to appear*.

[26] ——, "Secure Control of Networked Cyber-Physical Systems," in *55th Conference on Decision and Control (CDC), to appear*. IEEE, 2016.

[27] T. Kailath, "The innovations approach to detection and estimation theory," *Proceedings of the IEEE*, vol. 58, no. 5, pp. 680–695, 1970.

[28] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. SIAM Classics in Applied Mathematics, SIAM, Philadelphia, PA, USA, 2015.

[29] T. L. Lai and C. Z. Wei, "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems," *The Annals of Statistics*, pp. 154–166, 1982.