2016 IEEE 55th Conference on Decision and Control (CDC)
ARIA Resort & Casino
December 12-14, 2016, Las Vegas, USA

Secure Control of Networked Cyber-Physical Systems

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

***Abstract*—For networked cyber-physical systems to proliferate, it is important to ensure that the resulting control system is secure. We consider a physical plant, abstracted as a single-input-single-output stochastic linear dynamical system, in which a sensor node can exhibit malicious behavior. A malicious sensor may report false or distorted sensor measurements. For such compromised systems, we propose a technique which ensures that malicious sensor nodes cannot introduce any significant distortion without being detected. The crux of our technique consists of the actuator node superimposing a random signal, whose realization is unknown to the sensor, on the control law-specified input. We show that in spite of a background of process noise, the above method can detect the presence of malicious nodes. Specifically, we establish that by injecting an arbitrarily small amount of such random excitation into the system, one can ensure that either the malicious sensor is detected, or it is restricted to add distortion that is only of zero-power to the noise entering the system. The proposed technique is potentially usable in applications such as smart grids, intelligent transportation, and process control.**

***Index Terms*—Networked Cyber-Physical Systems, Secure Control, Stochastic Systems.**

I. Introduction

Cyber-Physical Systems (CPS) are comprised of a tight interplay between communication, computation and control. Potential application areas include the smart grid, automated transportation, and advanced manufacturing systems [1]. While they are poised to address some of the system-building challenges of the coming century, a potential stumbling block is their vulnerability to security breaches. Many recent instances of such attacks reinforce this concern. A well-known example is the Stuxnet worm [2], which provided malicious control commands while simultaneously spoofing the sensor measurements so as to appear normal in the control room. Specifically, Stuxnet recorded the sensor measurements for a few seconds before initiating an attack, and replayed them in the control room during the attack [2]. Another example is the Maroochy-Shire incident, in which an employee of a sewage treatment plant is alleged to have issued malicious commands to the control systems [3]. A third example is the Davis-Besse nuclear power plant in Ohio, where the Slammer worm affected some of the safety display systems in the facility. Though the Slammer worm was not designed to target the power plant, use of commodity IT software made the computers in the power plant vulnerable to generic cyber attacks [4]. The reader is referred to [5] for further instances of attacks on Industrial Control Systems (ICS). The increasing deployment of commodity IT software in ICS to achieve rapid scaling and easy implementation has increased their vulnerability to cyber attacks.

This paper is partially based on work supported by NSF Science and Technology Center Grant CCF-0939370, NSF Contract Nos. CNS-1646449, CCF-1619085, and CNS-1302182, and the US Army Research Office under Contract No. W911NF-15-1-0279.

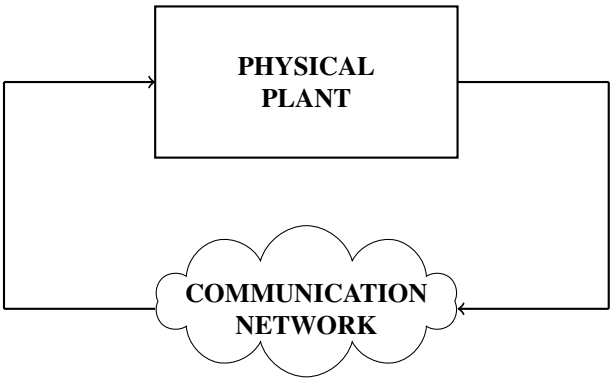

Fig. 1. A Networked Cyber-Physical System

Incidents such as the Maroochy-Shire attack, where the system is compromised not as a result of an attack on the cyber layer, but due to alleged malicious actions of authorized personnel, bring out the fact that there are two aspects to securing a cyber-physical system:

1) Securing the cyber layer, the underlying communication network.
2) Securing the physical layer, composed of sensors and actuators interfacing with the physical world.

Fig. 1 shows such an abstraction of a networked CPS consisting of interacting cyber and physical layers. At its heart is a physical plant consisting of an actuator and a sensor (the physical layer), exchanging information over an underlying communication network composed of relays, routers, switches, etc. (the cyber layer). Both layers need to be secured if the overall system is to be secured.

Security of the cyber layer includes the traditional notions of confidentiality, integrity, and availability of data. Confidentiality requires that the data transmitted by one node to another in the network is not known to any other node, even though the packets may be routed through multiple nodes in the system. Integrity requires that a packet received by a node is not tampered with by any node in its route, so that the data it receives is the data that was transmitted by the node where the packet originated. Availability refers to data being available when required. For example, availability requires that a node not abstain from forwarding a packet that it receives. Network and information security, which amounts to security of the cyber layer, is a relatively mature field and

978-1-5090-1837-6/16/$31.00 ©2016 IEEE

can, in principle, be accomplished using techniques such as cryptography, and those reported in [6], [7].

Relatively less is known when it comes to security of the physical layer - the problem of secure control with malicious Byzantine behavior. This paper addresses this particular problem. Specifically, we assume that the cyber layer is secured using techniques such as [6] and [7], so that the complexities of the underlying communication network in Fig. 1 can be abstracted as bidirectional, secure, reliable, delay-guaranteed bit pipes between the sensor and the actuator nodes. The problem of securing the CPS then translates to securing the physical layer.

II. The Model

We model the physical plant as a single-input-single-output stochastic linear dynamical system. While most real-world systems do not conform to properties of linearity or time-invariance, the class of LTI systems serves as a natural starting point for the development of any new theory, which in this case is a theory of security. This is because they are simple enough for analytical tractability, and yet retain appropriate features so that insights gained by analyzing them carry over to more general classes of systems.

Consider a single-input-single-output stochastic linear dynamical system described by an ARMAX model of the form

$$y[t] = -\sum_{i=1}^{m} a_i y[t-i] + \sum_{i=1}^{n} b_i u[t-i] + \sum_{i=0}^{p} c_i w[t-i], \quad (1)$$

where $y[t] \in \mathbb{R}$ denotes the output of the system at time t, $u[t] \in \mathbb{R}$ denotes the input applied to the system at time t, and $\{w\}$ is a sequence of independent and identically distributed Gaussian random variables with zero mean and variance σ_w^2. Without loss of generality, one can assume $c_0 = 1$. We also assume that $b_1 \neq 0$.

A sensor measures the outputs $y[t]$ and is supposed to convey them truthfully over the network to an actuator that is then supposed to implement a general history-dependent control law $g = (g_1, g_2, \ldots, g_t, \ldots)$, so that it applies the input $u^g[t] = g_t(y[0], y[1], \ldots, y[t])$ to the plant.

However, the sensor may be malicious and not report the true measurements, which can cause problems as the following example illustrates.

Example: Consider the system $y[t] = 0.9y[t-1] + u[t-1] + w[t]$. A minimum variance control law [8] $u[t] = -0.9y[t]$ is designed to be applied. If it were faithfully implemented, it would result in $y[t] = w[t]$ which has the minimum variance $var(y[t]) = \sigma_w^2$ for all t, and the input sequence $u[t] = -0.9w[t]$ for all t.

Suppose now that the sensor is malicious. It generates a random sequence $z[t]$ that is i.i.d., $N(0, \sigma_w^2)$, and independent of $\{w[t]\}$ and $y[0]$. It then falsely reports $z[t]$ as the measurement that it made at time t. The actuator, which is not malicious, relies on this reported measurement of the output, and applies the input $u[t] = -0.9z[t]$. From the actuator's point of view, the joint distribution (z, u) of its reported output and its own input , $(z, -0.9z)$, is exactly the same as the joint distribution $(w, -0.9w)$ that would have resulted if the sensor were not malicious, and so it has no reason to suspect that the sensor has lied since it has absolutely no other information to rely on. However, due to the malicious behavior of the sensor, the actual output of the system that results is $y[t] = 0.9y[t-1] - 0.9z[t-1] + w[t]$. This results in an asymptotic variance of $\frac{181}{19}\sigma_w^2$, which is a significant deterioration of performance. □

In this paper we will show how the actuator can defend against such a scenario where the sensor may be malicious. We denote by $z[t]$ the measurement reported by the sensor at time t. Since the sensor is malicious, $z[t]$ may not be equal to $y[t]$. While the history-dependent control law $g = (g_0, g_1, ...)$ is meant to be applied on the actual measurements, the fact that the measurements received by the actuator may be falsely reported implies that the control law is applied on the reported measurements which may or may not be the actual measurements. Therefore, $u^g[t] = g_t(z^t)$, where $z^t := (z[0], z[1], ..., z[t])$.

Our goal in this paper is to restrict the set of actions that the malicious sensor can engage in if it aims to be undetected. Towards this end, we pursue the technique of *dynamic watermarking* of signals, which ensures that if the malicious nodes wish to remain undetected, they are confined to adding a distortion to the noise entering the system that has only zero power. The results are equally applicable to fault detection where sensors may have failed or degraded rather than have been subverted.

III. Prior Work

Early efforts in security of cyber-physical systems identified key features that render it different from the problem of network security or information security. One is that a cyber-physical system may not lend itself to periodic patching and security updates, routinely done in IT security, since the presence of a physical plant in the loop may restrict controllers and sensors from going offline [5]. Also, the traditional notion of information security, of which data availability is a defining feature, does not put constraints on how soon after a data request the data should be available to an authorized party. However, the presence of a dynamical system in the loop imposes deadlines for each data packet [5]. Traditional solutions for IT security are not designed to handle such constraints. A mathematical model for the evolution of a system under attack could aid in a theoretical study of secure control. Towards this, [5] proposes models for two specific types of attacks, viz., deception and Denial-of-Service (DoS) attacks. Some definitions of security for cyber-physical systems are presented in [9]. Specifically, it is proposed that a cyber-physical system is secure if it its operational goals are maintained or at least, undergo a graceful degradation when attacked [9]. Related works in IT security and classical control are also identified that can potentially aid in addressing security of CPS. The problem of optimal control against DoS attacks is examined in [10] where the

adversary may jam a packet randomly and independently of other packets with some probability.

Conditions for the detection and identification of attacks by different classes of algorithms are presented in [11]. Here, attack detection refers to the knowledge of just the presence of an attack, whereas attack identification refers to the knowledge of the identity of malicious actuators/sensors in the system. Using techniques from compressed sensing, an approach for state estimation for cyber-physical systems when a subset of actuators and sensors are adversarial is presented in [12], [13]. Several studies have also been carried out on stealthy attacks [14]–[16]. A technique for filtering measurements from faulty sensors is presented in [17], which exploits spatial correlation of the sensed signal to identify outliers. Though not presented in the context of a feedback control system, the ideas presented in [17] could be employed in cyber-physical systems as well.

The aforementioned techniques can be classified as "passive defense mechanisms," allowing for the adversarial nodes to carry out certain attacks, and then designing mechanisms to operate the system in the presence of such attacks. This is by and large the approach that has dominated the literature thus far. In this paper, we pursue an alternate paradigm for securing cyber-physical systems, an "active defense." The technique presented in [18], [19] is the first to our knowledge that looked at this possibility, with the controller commanding the actuator to inject a random signal along with the control signal at each time in order to secure the system against replay attack. Subsequently, [20]–[22] extends this technique to adversaries employing more intelligent attacks.

In this paper, we examine the general approach of "dynamic watermarking," for *arbitrary* attack strategies, with the adversary not being aware of the specific realization of the random excitation. The honest nodes which excite the system with a random excitation check for the consistency of all the reported measurements. In contrast to [18]–[22] that carry out the test of only one residual, effectively the Test 2 described below, our method tests two residuals (Test 1 and Test 2 below). It is then shown that *no attack* by the adversarial nodes can remain undetected if it does *anything* other than add a zero-power signal to the noise already entering the system. To the best of our knowledge, this is the first result to show the resilience of a defense mechanism to *arbitrary* attack models.

In digital information security, a digital watermark is a code that is embedded in an electronic document in such a way that it cannot be removed without destroying the contents of the document. This technique for secure control can be recognized as having parallels with digital watermarking. Specifically, it relies on the fact that the sensor cannot separate the actuator's private excitation from the process noise. There is also the added feature in this scheme that the watermark is not revealed to other parties and only the imprinter of the watermark can detect whether it is present or has been tampered with. This provides the desired type of security, as we establish in the subsequent sections. Further extensions of this approach to more general control systems are elaborated in [23].

IV. Dynamic Watermarking: First-Order Systems

To introduce the concept, we begin by considering the case of a first-order stochastic linear dynamical system from [23]:

$$y[t+1] = ay[t] + bu[t] + w[t], \tag{2}$$

where $y[t] \in \mathbb{R}$ is the output of the system at time t, $u[t] \in \mathbb{R}$ is the input to the system at time t, $a, b \in \mathbb{R}$ are known parameters, and $w[t] \sim \mathcal{N}(0, \sigma_w^2)$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables.

In order to secure the control system, we consider the technique of "dynamic watermarking." The actuator node superimposes on the control law-specified input $u^g[t]$ a random variable $e[t]$ which it draws from a particular distribution $P_e(e)$, independent of the reported measurements $z[t]$ up to time t, and of all past values $e[t-1], e[t-2], ...e[0]$. The distribution P_e is made public, i.e., every node in the system knows the distribution from which the the random sequence $\{e\}$ is drawn. However, the honest actuators are required to hold private and not report the actual realization of the random sequence $\{e[t]\}$ to any other node in the system. Malicious nodes, being Byzantine, can do what they please, as in [6]. For this reason, the sequence $\{e\}$ is termed the actuator node's private excitation. The net input applied to the system, therefore, is given by

$$u[t] = u^g[t] + e[t], \tag{3}$$

A high-level intuition for why private excitation aids and may even be somehow necessary for securing the control system, is as follows. Consider the case where the sensor is malicious and the actuator is not. In the absence of private excitation, the system evolves according to (2), which, incorporating the control law, reduces to

$$y[t+1] = ay[t] + bg_t(z^t) + w[t]. \tag{4}$$

A malicious sensor, in this case, can "simulate" a linear dynamical system, and report its simulated measurements to the actuator. That is, it reports $z[t]$ generated as follows:

$$z[t+1] = az[t] + bg_t(z^t) + w'[t], \tag{5}$$

where $w'[t] \sim \mathcal{N}(0, \sigma_w^2)$ is a sequence of i.i.d. random variables drawn by the sensor, having no relationship to the actual noise sequence $w[t]$, i.e., independent of the physical system. Hence, a malicious sensor need not even observe the output to report measurements, and an honest actuator will never be able to tell the difference between the outputs of the virtual system and that of the real one since $w'[t]$, being drawn from the same distribution as $w[t]$, could have been the actual process noise that affected the system. Therefore, no detection algorithm can expose the maliciousness of the sensor.

On the other hand, in the presence of private excitation $e[t]$ with $e[t] \sim \mathcal{N}(0, \sigma_e^2)$ in (3), the system evolves according to

$$y[t+1] = ay[t] + bu^g[t] + be[t] + w[t]. \tag{6}$$

The output of the system contains a component of $e[t]$, which is known only to the actuator. Hence, the actuator can check if the measurement reported by the sensor has some correlation with its private excitation. In what follows, we prove rigorously that by subjecting the reported measurements to certain tests which essentially check for appropriate correlations with the private excitation, one can constrain the distortion that the sensor can introduce to have zero power on average, if it hopes to remain undetected.

We now design tests that the actuator can perform to check if the sensor is malicious or not. We require the tests to satisfy the following properties:

1) An honest actuator passes the tests almost surely.
2) The tests are implementable by the actuator, based only on the information available to the actuator, viz., $e[t]$ and $z[t]$, and not on information such as $y[t]$ which may not be available to the actuator.

One can verify that the following two tests satisfy the above constraints. They are specified as asymptotic properties to be checked, but can be converted into finite time tests using standard statistical analysis, as described in [23]:

1) **Test 1:** Check if

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(z[k]-az[k-1]-bu^g[k-1])e[k-1]=b\sigma_e^2. \tag{7}$$

2) **Test 2:** Check if

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(z[k]-az[k-1] - bu^g[k-1]-be[k-1])^2=\sigma_w^2. \tag{8}$$

The following theorem proves that subjecting the reported measurements to the above tests ensures that the malicious sensor cannot introduce any distortion of non-zero power without being exposed.

Theorem 1. *Define* $v[t]\coloneqq z[t]-az[t-1]-bu^g[t-1]-be[t-1]-w[t-1]$*, so that* $v\equiv 0$ *for an honest sensor. If the sensor satisfies the above tests (7) and (8), then*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]=0. \tag{9}$$

Proof. Since $\{z\}$ satisfies (7), we have,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(v[k]+be[k-1]+w[k-1])e[k-1]=b\sigma_e^2. \tag{10}$$

Since $e[k]$ and $w[k]$ are uncorrelated, the above becomes

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}e[k-1]v[k]=0. \tag{11}$$

Since the reported sequence also satisfies (8), we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]+2v[k]w[k-1]+w^2[k-1]=\sigma_w^2.$$

Since $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}w^2[k-1]=E\{w^2[k-1]\}=\sigma_w^2$, we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]+\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}2v[k]w[k-1]=0. \tag{12}$$

Equation (11) implies that the sequence $\{v\}$ added by the sensor must be empirically uncorrelated with the actuator's private noise sequence $\{e\}$. In what follows, we show that if $v[k]$ is empirically uncorrelated with $e[k-1]$, then it must also be uncorrelated with $w[k-1]$. Intuitively, since the sensor can observe only the sum $be[k-1]+w[k-1]$ at each time k, and cannot perfectly separate $e[k-1]$ from $w[k-1]$, it may be conjectured that if $v[k]$ is uncorrelated with $e[k-1]$ it must also be uncorrelated with $w[k-1]$. However, it turns out that this entails a more complex proof since the roles of e and w are not symmetric because the former is known to the actuator but the latter is not. The rest of the proof is devoted to establishing this rigorously since the claim follows readily from it, by substituting this in (12).

Define the $\sigma-$algebra $\mathcal{S}_k\coloneqq\sigma(y^{k+1},z^{k+1},e^{k-1})$, and $\widehat{w}[k]\coloneqq E[w[k]\big|\mathcal{S}_k]$. Since the sequence of observations is i.i.d. Gaussian, the conditional mean estimate of $w[k]$ based on observing $be[k]+w[k]$ is [8]

$$\widehat{w}[k]=\frac{\sigma_w^2}{b^2\sigma_e^2+\sigma_w^2}(be[k]+w[k])=\beta(be[k]+w[k]),$$

where $\beta\coloneqq\frac{\sigma_w^2}{b^2\sigma_e^2+\sigma_w^2}<1$. This can be written as

$$\widehat{w}[k]=\alpha e[k]+\beta w[k], \tag{13}$$

where $\alpha\coloneqq b\beta$. Let $\widetilde{w}[k]\coloneqq w[k]-\widehat{w}[k]$. Then, $\widetilde{w}[k-1]$ is a martingale difference sequence with respect to the filtration $\{\mathcal{S}_k\}$. This is because $\widetilde{w}[k-1]\in\mathcal{S}_k$ and $E[\widetilde{w}[k]\big|\mathcal{S}_k]=0$. Also, we have $v[k]=z[t]-az[t-1]-(y[t]-ay[t-1])$. Therefore, $v[k]\in\mathcal{S}_{k-1}$. The Martingale Stability Theorem (MST) [24] applies, and we have

$$\sum_{k=1}^{T}v[k]\widetilde{w}[k-1]=o(\sum_{k=1}^{T}v^2[k])+O(1). \tag{14}$$

Hence,

$$\begin{aligned}\sum_{k=1}^{T}v[k]w[k-1]&=\sum_{k=1}^{T}v[k](\widehat{w}[k-1]+\widetilde{w}[k-1])\\&=\sum_{k=1}^{T}v[k]\widehat{w}[k-1]+o(\sum_{k=1}^{T}v^2[k])+O(1).\end{aligned}$$

Substituting for the estimate from (13) in the above equation yields,

$$\begin{aligned}\sum_{k=1}^{T}v[k]w[k-1]=\alpha\sum_{k=1}^{T}v[k]e[k-1]+\beta\sum_{k=1}^{T}v[k]w[k-1]\\+o(\sum_{k=1}^{T}v^2[k])+O(1).\end{aligned}$$

Simplifying,

$$\sum_{k=1}^{T} v[k]w[k-1] = \frac{\alpha}{1-\beta}\sum_{k=1}^{T} v[k]e[k-1] + o(\sum_{k=1}^{T} v^2[k]) + O(1).$$

From (11), we have $\sum_{k=1}^{T} v[t]e[t-1] = o(T)$. It follows that

$$\sum_{k=1}^{T} v[k]w[k-1] = o(\sum_{k=1}^{T} v^2[k]) + o(T) + O(1). \quad (15)$$

So,

$$\sum_{k=1}^{T} v^2[k] + \sum_{k=1}^{T} 2v[k]w[k-1] = (1+o(1))(\sum_{k=1}^{T} v^2[k]) + O(1)$$

Dividing the above equation by T, taking the limit as $T \to \infty$, and invoking (12) completes the proof. □

Remark 1: The only uncertainties in the system are the initial state of the system $y[0]$ and the noise realization $\{w[1], w[2], w[3], \cdots\}$. Since the actuator can compute $z[t+1] - az[t] - bg_t(z^t) - be[t]$, which will be equal to the process noise if the sensor reports measurements truthfully, the sensor reporting the sequence $\{z\}$ is equivalent to it reporting the sequence of process noise. From the definition of $v[t]$, we have

$$z[t+1] - az[t] - bg_t(z^t) - be[t] = w[t] + v[t+1].$$

Note that the above can be computed by the actuator. i.e., it can compute the sequence $\{w + v\}$. Hence, the above theorem essentially states that a malicious sensor cannot distort the noise realization $\{w[1], w[2], w[3], \cdots\}$ beyond adding a zero-power sequence to it. As shown in [23], this leads to preservation of stability and performance for stable systems:

Theorem 2. *Suppose $|a| < 1$, i.e., the system is stable.*

(i) *Define the distortion $d[t] \coloneqq z[t] - y[t]$. Then,*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} d^2[k] = 0.$$

(ii) *If the malicious sensor is to remain undetected, the mean-square performance of $y[t]$ is the same as the reported mean-square performance $z[t]$ that the actuator believes it is:*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} y^2[k] = \lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} z^2[k].$$

(iii) *Suppose the control law is $u(t) = fy(t)$ with $|a + bf| < 1$. The malicious sensor cannot compromise the performance of the system if it is to remain undetected, i.e., the mean-square performance of the system is*

$$\lim_{T\to\infty} \frac{1}{T}\sum_{k=0}^{T-1} y^2[k] = \frac{\sigma_w^2 + b^2\sigma_e^2}{1 - |a+bf|^2}.$$

V. Dynamic Watermarking: General ARMAX Systems

In this section, we extend the results to the more general class of ARMAX systems with colored noise, which are of considerable interest in process control [25]. In the case of an ARMAX system, the actuator's private excitation is generated in such a way that its output spectrum matches that of the colored process noise. This constitutes a form of an Internal Model Principle [26] for dynamic watermarking, and is a phenomenon that does not emerge in the analysis of first-order systems. Moreover, the watermarking mechanism and the results for first-order systems can be recovered as a special case of those for ARMAX systems.

Consider a system governed by (1). The system can equivalently be expressed in the $z-$domain as

$$A(z^{-1})y[t] = B(z^{-1})u[t] + C(z^{-1})w[t], \quad (16)$$

where $A(z^{-1}) \coloneqq 1 + \sum_{i=1}^{m} a_i z^{-i}$, $B(z^{-1}) \coloneqq \sum_{i=1}^{n} b_i z^{-i}$, and $C(z^{-1}) \coloneqq 1 + \sum_{i=1}^{p} c_i z^{-i}$. We assume both $B(z^{-1})$ and $C(z^{-1})$ to be strictly minimum phase.

For ARMAX systems as described above, we do not directly superimpose the private excitation $\{e\}$ on the control law-specified input. Rather, we filter it through $B^{-1}(z^{-1})C(z^{-1})$ before applying it to the input:

$$u[t] = u^g[t] + B^{-1}(z)C(z)e[t]. \quad (17)$$

This yields

$$y[t] = -\sum_{i=1}^{m} a_i y[t-i] + \sum_{i=1}^{n} b_i u^g[t-i] + n[t] + \sum_{i=1}^{p} c_i n[t-i], \quad (18)$$

where $n[t] \coloneqq e[t] + w[t]$.

We now develop tests that the actuator should perform to check for the maliciousness of the sensor. The basic idea behind the tests is to check if the prediction-error has the right statistics. The actuator constructs the prediction-error $\{\widetilde{y}[k]\}$ as follows.

$$\widetilde{y}[k] = 0 \ \ \forall k \in \{-1, -2, ..., -p\}, \quad (19)$$

$$\widehat{y}_{k|k-1} = \sum_{i=1}^{m} a_i z[k-i] + \sum_{i=1}^{n} b_i u^g[k-i] + \sum_{i=1}^{p} c_i[z[k-i] - \widehat{y}_{k-i|k-i-1}], \quad (20)$$

$$\widetilde{y}[k] = z[k] - \widehat{y}_{k|k-1}. \quad (21)$$

For an honest sensor, i.e., if $z[t] \equiv y[t]$, the prediction error sequence $\widetilde{y}[t]$ satisfies $C(z^{-1})(\widetilde{y}[t] - w[t] - e[t]) \equiv 0$, and since $C(z^{-1})$ is strictly minimum phase, it follows that $lim_{t\to\infty}(\widetilde{y}[t] - w[t] - e[t]) = 0$. Thus for such an honest sensor it will turn out that $\frac{1}{T}\sum_{t=1}^{T} \widetilde{y}^2[t] = \sigma_w^2 + \sigma_e^2$. Based on this, the actuator performs the following tests. Note that they satisfy the properties that an honest sensor passes the tests, and that they are implementable by the actuator.

1) **Test 1:** Check if

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\widetilde{y}^2[k]=\sigma_e^2+\sigma_w^2. \tag{22}$$

2) **Test 2:** Check if

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(\widetilde{y}[k]-e[k])^2=\sigma_w^2. \tag{23}$$

The following theorem proves that the above tests suffice to constrain a malicious sensor to adding a distortion that has only zero power.

Theorem 3. *Define* $v[t] \coloneqq z[t]+\sum_{i=1}^{m}a_i z[t-i]-\sum_{i=1}^{n}b_i u^g[t-i]-n[t]-\sum_{i=1}^{p}c_i n[t-i]$. *If the sensor satisfies (22) and (23), then,*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]=0. \tag{24}$$

Proof. For ease of exposition, we suppose that the initial conditions $n[-p], n[-p+1], ..., n[-1]$ are known and are equal to zero. The proof carries over even if this is not the case.

The prediction can be rewritten as

$$C\widehat{y}_{k|k-1}=(1-A)z[k]+Bu^g[k]+(C-1)z[k]. \tag{25}$$

Rearranging, we get

$$Az[k]=Bu^g[k]+C\widetilde{y}[k]. \tag{26}$$

Substituting for $A(z^{-1})z[k]-B(z^{-1})u^g[k]=v[k]+C(z^{-1})n[k]$, we obtain

$$v[k]+C(z^{-1})n[k]=C(z^{-1})\widetilde{y}[k].$$

Defining $\bar{v}[k] \coloneqq C^{-1}(z^{-1})v[k]$, we get

$$\widetilde{y}[k]=n[k]+\bar{v}[k], \tag{27}$$

where $\bar{v}[k]$ is the result of filtering the distortion sequence $\{v[k]\}$ with $C^{-1}(z^{-1})$. Now, since the reported sequence of measurements satisfies (22), substituting (27) in (22) gives

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}n^2[k]+\bar{v}^2[k]+2n[k]\bar{v}[k]=\sigma_e^2+\sigma_w^2. \tag{28}$$

However, since $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}n^2[k]=\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}w^2[k]+e^2[k]+2w[k]e[k]=\sigma_e^2+\sigma_w^2$, where the last equality follows from the fact that $\{e\}$ and $\{w\}$ are uncorrelated, the above equation reduces to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\bar{v}^2[k]+2n[k]\bar{v}[k]=0. \tag{29}$$

Also, since the reported sequence of measurements satisfy (23), we have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}(\bar{v}[k]+w[k])^2=\sigma_w^2.$$

Using $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}w^2[k]=\sigma_w^2$, the above simplifies to

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\bar{v}^2[k]+2\bar{v}[k]w[k]=0. \tag{30}$$

Substituting above into (29) and noting that $n[t]=e[t]+w[t]$, one arrives at

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\bar{v}[k]e[k]=0. \tag{31}$$

Thus, the filtered distortion $\{\bar{v}\}$ behaves analogously to the way the distortion sequence $\{v\}$ did for the first-order system (compare (30) with (12) and (31) with (11)). Hence, following the same sequence of arguments of the proof for Theorem 1, one arrives at

$$\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}\bar{v}^2[k]=0. \tag{32}$$

Finally, since $v[t]=[C(z^{-1})]\bar{v}[t]$, it follows that $\lim_{T\to\infty}\frac{1}{T}\sum_{k=1}^{T}v^2[k]=0$. □

VI. Statistical Tests

While the tests presented in the previous sections are of an asymptotic nature, as discussed in [23], they can be converted in a straightforward manner into detection thresholds on finite time behavior and thereby implemented to detect malicious activities in finite duration with the usual trade-off between detection and false alarm rates. One such test suitable for the problem at hand is based on the sequential probability ratio test (SPRT) [27], and is described in detail in [23]. Test statistics corresponding to (22) and (23) are formed by computing empirical variances using a moving-average window of length l. Even though the empirical variance can be computed using all past measurements, a moving-average filter is desirable since the adversary may distort measurements in a bursty fashion. The test statistics thus computed can be compared against predetermined thresholds to identify a malicious sensor.

VII. Simulation Results

We now illustrate the above methodology on a second-order ARMAX system with colored noise. Consider a system of the form (1), with the following parameters: $a_0=0.7$, $a_1=0.2$, $b_0=c_0=1$, $b_1=0.5$, $c_1=0.3$, and $\sigma_w^2=1$. The actuator injects private excitation of variance $\sigma_e^2=1$.

In our simulation, the actuator implements tests (23) and (22) over a finite duration by computing the sample variances over a finite window of 500 most recent samples.

The system is assumed to operate under normal conditions up to time epoch 4500, at which time the adversary initiates the attack. The adversary implements the attack described by (5) in Section-IV. Note that this attack can never be detected by the actuator in the absence of dynamic watermarking, since the actual sequence reported by the sensor could well have been the result of the process noise affecting the system.

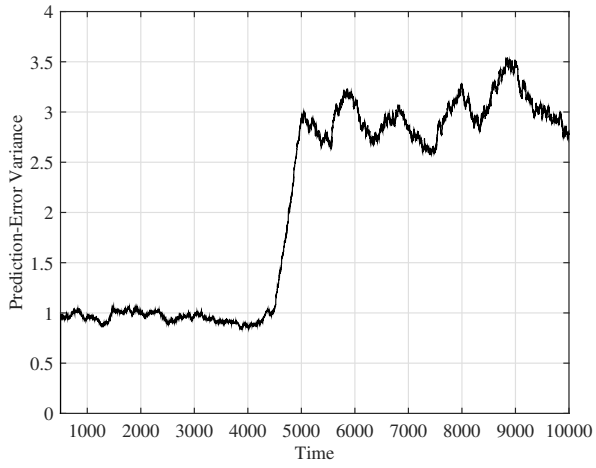

Fig. 2. Variance of Prediction-Error.

Fig. 2 plots the results of the actuator's Test 2 in the presence of dynamic watermarking. Until the time when the adversary initiates the attack, the sample variance (23) remains at 1, which is the nominal variance of the prediction error (since $\sigma_w^2 = 1$). However, once the attack is initiated, the Test-2 statistic increases steadily. Desirable trade-off between detection and false alarm rates can be obtained by setting appropriate detection thresholds.

VIII. Conclusion

In this paper, we have developed provably secure mechanisms for active defense of networked cyber-physical systems. A two-layer approach for securing networked CPS is presented, viz., securing the cyber layer and securing the physical layer. For the security of the physical layer, the method of dynamic watermarking has been presented for ARMAX systems which are of interest in process control. By employing dynamic watermarking, malicious nodes in the system that wish to avoid detection are confined to adding distortion that can only have zero power to the process noise.

References

[1] K.-D. Kim and P. R. Kumar, "Cyber-Physical Systems: A Perspective at the Centennial," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1287–1308, 2012.

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

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

[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] A. A. Cárdenas, S. Amin, and S. Sastry, "Research Challenges for the Security of Control Systems," in *Proceedings of the 3rd Conference on Hot Topics in Security*, ser. HOTSEC'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 6:1–6:6. [Online]. Available: http://dl.acm.org/citation.cfm?id=1496671.1496677

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

[9] 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.

[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] 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.

[12] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure Estimation and Control for Cyber-Physical Systems under Adversarial Attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[13] ——, "Secure State-Estimation for Dynamical Systems under Active Adversaries," in *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011.

[14] A. Teixeira, I. Shames, H. Sandberg, and K. Johansson, "Revealing Stealthy Attacks in Control Systems," in *50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2012.

[15] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in Stochastic Control Systems: Fundamental Limitations and Performance Bounds," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 195–200.

[16] R. Zhang and P. Venkitasubramaniam, "Stealthy Control Signal Attacks in Scalar LQG Systems," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 240–244.

[17] 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

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

[19] 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.

[20] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False Data Injection Attacks Against State Estimation in Wireless Sensor Networks," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010.

[21] 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.

[22] 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.

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

[24] 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.

[25] K. J. Åstrom, *Introduction to Stochastic Control*. New York: Academic Press, 1970, 1970.

[26] B. A. Francis and W. M. Wonham, "The Internal Model Principle of Control Theory," *Automatica*, vol. 12, no. 5, pp. 457–465, 1976.

[27] A. Wald, "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.