Event-Triggered Watermarking Control to handle Cyber-Physical Integrity Attacks

José Rubio-Hernán[1], Luca De Cicco[2], and Joaquín García-Alfaro[1]

[1] SAMOVAR, Telecom SudParis, CNRS, Université Paris-Saclay, Evry, France
jose.rubio_hernan@telecom-sudparis.com,
joaquin.garcia_alfaro@telecom-sudparis.com
[2] Politecnico di Bari, Dipartimento di Ingegneria Elettrica e dell'Informazione, Italy
luca.decicco@poliba.it

Abstract. The use of control-theoretic solutions to detect attacks against cyber-physical systems is a growing area of research. Traditional literature proposes the use of control strategies to retain, f.i., satisfactory close-loop performance, as well as safety properties, when a communication network connects the distributed components of a physical system (e.g., sensors, actuators, and controllers). However, the adaptation of these strategies to handle security incidents, is an ongoing challenge. In this paper, we analyze the use of a watermark-based detector that handles integrity attacks. We show that (1) the detector is able to work properly under the presence of adversaries using non-parametric methods to escape detection; but (2) it fails at detecting adversaries using parametric identification methods to escape detection. We propose a new strategy that complements the watermark-based detector in order to detect both adversaries. We validate the detection efficiency of the new strategy via simulation.

Keywords: Cyber-Physical Security, Critical Infrastructures, Attack detection, Adversary Model, Networked Control System.

1 Introduction

As an evolution of traditional industrial control systems [9], cyber-physical systems [11] combine feedback control technologies with novel computing and communication capabilities. The recently coined cyber-physical security term refers to mechanisms that address security issues associated to these environments. The use of inadequate cyber-physical security mechanisms can have an adverse effect in critical infrastructures, either national or private ones [6]. These issues place the study of cyber-physical security mechanisms as a hot research topic.

Given the control-theoretic nature of cyber-physical systems, the control community is actively working to adapt traditional control strategies to detect faults and errors, towards detectors of malicious attacks [7, 8, 17]. Motivated by the same objectives, we present in this paper a solution that combines two different control strategies to handle integrity attacks against cyber-physical systems.

The contributions of this paper can be summarized as follows. First, we analyze the effectiveness of a challenge-response detector based on control-theoretic watermarks, under the assumption of integrity cyber-physical attacks. We reexamine the security of an existing contribution by Mo et al. in [13], and revisit its security effectiveness under a new adversarial scenario. We show that under the new assumptions, the original contribution presents some weaknesses. We then propose a new detection strategy that combines event-triggered control strategies with the previous watermark-based detector, in order to cover the new adversaries. Finally, we validate our proposed approach via numerical simulations. Our results show the effectiveness of our novel proposal.

The paper is organized as follows. Section 2 provides the necessary background. Section 3 reviews the watermark-based detector scheme by Mo et al. [13], provides a new adversary model and reexamines the security of the detector under the new adversary model. Section 4 presents the new detection strategy to handle the uncovered limitations, and validates the approach via numerical simulations. Section 5 reviews related work. Section 6 concludes the paper.

2 Background

2.1 Cyber-Physical Attacks

The use of communication networks and IT components in traditional control systems paves the way to new vulnerability issues. Attacks against these setups are named cyber-physical attacks. These attacks target physical processes through the network. In [19], authors propose a taxonomy of cyber-physical attacks based on the resources of the adversaries. Such resources are mainly measured in terms of adversary knowledge (e.g., *a priori* knowledge of the adversary about the system and its security measures). For instance, the knowledge of the adversary about the system is the main resource used to build up complex attacks, and to make them undetectable. Based on the degree of the adversary knowledge, the attacks may succeed at violating system properties, e.g., availability and integrity, as well as at obtaining operational information about the system to make the attacks undetectable.

Based on the adversary knowledge, cyber-physical attacks related to integrity can be classified as: (i) the replay attack where the adversary does not need knowledge about the system model [13]; (ii) injection attack, where the adversary injects false data or deviation of the legitimate data. These attacks are not detected if the data are compatible with the dynamics of the system [19], i.e., the adversary must to know the physical processes; and, (iii) covert attack, where the adversary knows perfectly the cyber-physical system behaviour. This attack is defined in [18] where the authors conclude that it is not possible to be detected.

Let us now present the techniques developed in the literature against these attacks; (a) signal-based detector method [1]; (b) statistical detection method [5]; (c) stationary watermark-based detector method, adapting failure detector mechanisms [13]. In the following sections, we re-examine the watermark-based technique, and the control strategies, in order to propose an improved security technique against integrity attacks. The new detection strategy handles cyber-physical adversaries which are not detected with the aforementioned techniques.

Such cyber-physical adversaries use a parametric technique to obtain the knowledge about the system model.

2.2 Control Strategies

Control theory is a well-known topic, where the evolution of the technology has been the main motivation to create new control policies to manage these systems, keeping the control features. Among these new technologies, we can mention the networked control systems (NCSs), where the loop between the different components of the system is closed through the network. A wide range of research has been reported in the literature focusing on managing these new technologies in order to preserve the control properties of the systems. They have generated new challenges in control/estimation, signal processing, and communication in order to solve the new performance problems as limited power transmission, bandwidth constrains, packet drop, delay or security. The networked control systems have motivated to consider control/estimation and communication in a unified way [10], in order to solve problems as performance or security. Among all control strategies in NCSs, we have focused on the strategies depending on the transmission policy; sampled-data control, or event-triggered control. Into the sampled-data policy, we find mono-frequency sampling, i.e., the same sampling frequency for all the channels, or multi-frequency sampling, i.e., different sampling frequencies depending on the channel (sensor/controller or controller/actuator) [17]. Event-triggered control (ETC) has been also studied depending on the policy to send the events, Periodic event-triggered control (PECT) [8] or stochastic events-triggered schedule [7]. This topic is inline with our research since the security in NCSs includes the management of the control properties through the network to avoid that an external entity, an adversary, has the capacity to control these properties and harm the system.

2.3 Watermark-based Attack Detection

The watermark-based detector is proposed in [13], with the goal of detecting replay attacks against cyber-physical systems. To analyze the watermark-based detector, the authors use an industrial control system modeled mathematically as a discrete linear time-invariant (LTI) system. This mathematical model is used to describe the dynamic behaviour of the system. The system can be represented as follows:

$$x_{t+1} = Ax_t + Bu_t + w_t \tag{2.1}$$

$$y_t = Cx_t + v_t \tag{2.2}$$

where $x_t \in \mathbb{R}^n$ is the state's vector, $u_t \in \mathbb{R}^p$ is the control signal, $y_t \in \mathbb{R}^m$ is the system output, and $w_t \in \mathbb{R}^n$ and v_t are the *process noise* and the measurement noise respectively. The noises are assumed to be a zero mean Gaussian white noise with covariance Q, *i.e.* $w_t \sim N(0, Q)$ and R, *i.e.* $v_t \sim N(0, R)$ respectively. Moreover, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{m\times n}$ are respectively the *state* matrix, the *input* matrix end the *output* matrix.

Let us now define the well-known *Linear Quadratic Gaussian* (LQG) approach used as a control technique in [13]. This technique has two independent components:

1. a *Kalman filter* producing an optimal state estimation $\hat{x}_t$ of the state x:
$$\hat{x}_{t|t-1} = A\hat{x}_{t-1} + Bu_{t-1}$$
$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1}) \tag{2.3}$$
where K_t denotes the Kalman gain, and $\hat{x}_{t|t-1}$ is the *a priori* system state estimation.
2. a *Linear Quadratic Regulator* (LQR) providing the control law u_t.
$$u_t = L\hat{x}_t \tag{2.4}$$
where L denotes the feedback gain of a linear-quadratic regulator.

After describing the model of the plant, hereinafter we present the detection scheme proposed in [13] against replay attacks. The idea is to superpose a watermark signal $\Delta u_t \in \mathbb{R}^p$ to the optimal control law $u_t^\diamond$. The new control input u_t is given by:
$$u_t = u_t^\diamond + \Delta u_t \tag{2.5}$$
Note that the watermark signal is independent from the process noise w_t and the output noise v_t. To detect the adversaries, the watermark-based detector employs a well-known χ^2 detector [3]. The *alarm signal* g_t generates by the detector is defined as:
$$g_t = \sum_{i=t-w+1}^{t} (r_i)^T \mathcal{P}^{-1}(r_i) \tag{2.6}$$
where w is the size of the detection window, $\mathcal{P}$ is the co-variance of input signals from the sensors and $r_t = y_t - C\hat{x}_{t|t-1}$ is the residues generated from the estimator at each t-th time step.

To verify if the system is under attack, g_t is compared with a threshold γ. If g_t is equal or greater than the threshold, $g_t \geq \gamma$, the detector generates an alarm.

3 Watermark-Based Attack Detection against a new Adversary Model

Let us assume the system employs the detector described in Section 2.3, so that the controller superposes its output with an authentication watermark Δu_t. At steady-state, i.e. after the transient has been exhausted, the output of the system can be considered as the sum of its steady-state value and a component that is due to watermark signal that shall be only known by the controller.

Hereinafter we denote the adversary proposed in [13] as a cyber adversary [16]. This attacker has the ability to eavesdrop all the messages sent by the sensors y_t and to inject messages with a signal y'_t to conduct malicious actions without any knowledge about the system model. Let us also define a cyber-physical adversary as the attacker who is able to eavesdrop the message with the intention of improving its knowledge about the system behaviour, in order to conduct malicious actions [16].

Based on the way to model the system's behaviour, two different cyber-physical adversaries can be defined.

Definition 3.1. *An attacker that, only uses the previous input and output of the system to obtain a system behaviour is defined as a non-parametric cyber-physical adversary.*

Remark 1. This adversary can use a Finite Impulse Response (FIR) identification model [20].

Cyber and non-parametric cyber-physical adversaries can be handled using a non-stationary watermark detector scheme [16]. However, if the cyber-physical adversary is able to acquire the parameters of the system, a non-stationary watermark detector scheme is not able to detect the attack.

Definition 3.2. *An attacker able to estimate the parameters of the system using input and output data to mislead the controller detector is defined as a parametric cyber-physical adversary.*

The signal injected by the parametric cyber-physical adversary cannot be detected by the χ^2 detector (cf. Equation (2.6)), using a non-stationary watermark-based scheme.

Remark 2. This adversary can use an ARX (autoregressive with exogenous input) or an ARMAX (autoregressive-moving average with exogenous input) approach in order to estimate the model of the system [14].

We assume that the main constraint of this adversary is the energy spent to eavesdrop and analyze the communication data, i.e., the number of samples eavesdropped to obtain the system model parameters.

Proof. If the system uses a watermark-based detector, the system control inputs are represented by Equation (2.5), and the outputs are represented by:

$$y_t = C(Ax_t + B(u_t^{\diamond} + \Delta u_t) + w_t) + v_t \tag{3.1}$$

note that the watermark can be defined as an independent and identically distributed Gaussian distribution or a stationary Gaussian distribution. Using the ARX approach we can define the system defined in Equations (2.1) and (2.2) as follows:

$$Y(z) = H(z)U(z) + V(z) \tag{3.2}$$

where $U(z)$ and $Y(z)$ represent the inputs and the outputs of the plant respectively. V(z) represents the external noise which affects the outputs of the plant. And $H(z)$ is another way to describe the model of the system presented in Section 2.3, using frequency domain.

$$H(z) = \frac{Y(z) - V(z)}{U(z)} = \frac{\mathcal{N}(z)}{\mathcal{D}(z)} = \left(\frac{n_0 z^m + n_1 z^{m-1} + ... + n_m}{d_0 z^n + d_1 z^{n-1} + ... + d_n} \right) \tag{3.3}$$

where $\mathcal{N}(z)$ and $\mathcal{D}(z)$ are the polynomial functions which build the model of the system. We prove that under the attacker model of Definition 3.2, the adversary is able to know exactly the watermark signal and thus $\Delta u_t = \Delta u'_t$.

Proposition 1. *A parametric cyber-physical adversary is able to obtain the system model, $H(z)$, and mislead the controller, eavesdropping the control inputs and the measurements of the sensors. The probability to be detected, is equal to the probability to obtain an erroneous model. This probability, is directly proportional to the order of the system, i.e., the order of $\mathcal{D}(z)$, and inversely proportional to the window size to eavesdrop the data channel.*

Proof. If the adversary knows all the control inputs, and the measurements of the sensors, then the model obtained by the adversary can be defined as; $H_{at}(z) = (Y(z) - V(z))/U(z)$. Comparing the adversary model of the system and the real model system, it is straightforward to prove that both system models are equal, $H_{at}(z) = H(z)$. Nevertheless, the adversary has an error that depends on the order selected to create the model and the number of samples eavesdropped to compute the parameters of the model, the window size. Following the Mean Square Error (MSE):

$$MSE = \frac{\mathcal{H}(\zeta)}{\hat{T}} \tag{3.4}$$

where $\mathcal{H}(\zeta)/\hat{T}$ is the error variance, since the system model used in this paper (cf. Section 2.3) contains no bias error [2]. This error is directly proportional to system complexity (flexibility), ζ, and inversely proportional to the samples eavesdropped by the adversary. It is worth to note that the complexity is directly proportional to the system order. Indeed, for a system with a small order is easier to obtain a good approximation model by the adversary.

To summarize, these adversaries look at the real system like a black box. They can increase the order (complexity) of their model to improve the possibility to go into the order's range where the real system could be identified. Nevertheless, they need to use a larger window size to minimize the MSE value. For this reason, the computation cost of the attack increases for a high order of the system, since the adversary needs to increase their order model, as well as, the window size in order to minimize the MSE. It is worth mentioning that the number of samples eavesdropped before the attack, as well as the order system of the adversary, are the main parameters to avoid detection.

3.1 Numerical Validation

In the previous sections we have seen that the watermark detector proposed in [13] and the improvement proposed in [16] are not able to detect parametric cyber-physical adversaries. We have validated both watermark detector against the parametric cyber-physical adversary presented in Definition 3.2. Hereinafter we present only the detection ratio with respect to this adversary using the detector improvement proposed in [16] due to space constraint. Nevertheless, we have obtained the same detection ratio using the detector proposed in [13]. This adversary is able to identify the system model parameters from the input and output plant signals. To validate the watermark detector against the parametric cyber-physical adversary, we define three different use cases:

1. First use-case: the adversary knows only a subset of control inputs and measurements of the sensors. This adversary will be detected by the watermark-based detector proposed in [13].

Proof. Assuming, on the one hand, a system defined as $H(z) = (Y(z)-V(z))/U(z)$, where $U(z) = U_1(z) + U_2(z)$; and, on the other hand, an adversary whose model can be defined as $H_{at_1} = (Y(z) - V(z))/U_1(z)$, since this attacker only knows a subset of inputs $U_1(z)$ [21]. Then, if all the inputs and outputs are correlated, the adversary will be detected by the system, since:

$$H_{at_1} = \frac{Y(z) - V(z)}{U_1(z)} \neq \frac{Y(z) - V(z)}{U(z)} = H(z) \tag{3.5}$$

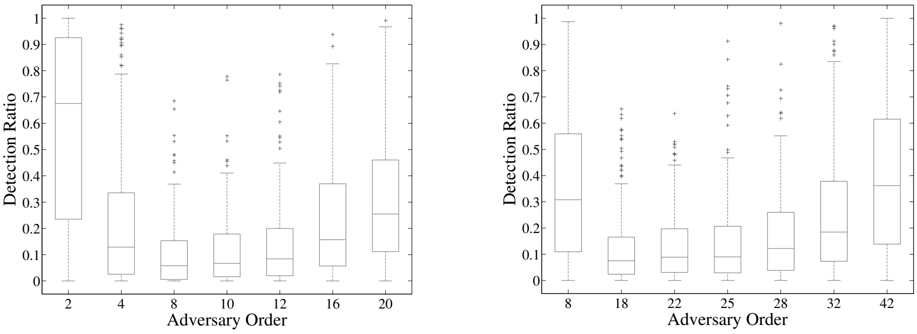
proves that the model used by the adversary, H_{at_1}, is different to the real system model.

2. Second use-case: the adversary has access to all the control inputs and measurements of the sensors. In this case, the parametric cyber-physical adversary could be able to obtain the model of the system with great accuracy. To do so, the adversary has to use the order of the unknown system, p, and to use a large window size, $\hat{T}$, to eavesdrop the data in order to get the correct system model.

Figures 3.1(a) and (b) show the detection ratio of the watermark detector against a parametric cyber-physical adversary. Figure 3.1(a) shows the results of 200 Monte Carlo simulations using systems of order ten, against this adversary. The results present the ratio of detection if the adversary uses a window size equal to 200 and different system orders for the model. If the attacker chooses the correct system order for the model, the ratio of detection is around 7%. Nevertheless, if the adversary order varies in the range $[8, 12]$, the detection ratio is not higher than 10%. Out of this range, the ratio of detection increases drastically. Figure 3.1(b) shows the ratio of detection for 200 Monte Carlo simulations using systems of order 25, against seven different parametric cyber-physical adversaries. The assumed window size is settled to $\hat{T} = 300$. If an adversary uses a model of the system with the correct order, the ratio of detection is around 8%. The range of orders where the ratio of detection does not increase drastically is $[18, 28]$. If an adversary uses an order in this range, the ratio of detection is not higher than 10%. Otherwise, the likelihood to detect the adversary is high.

Figure 3.2 shows the ratio of detection of the same system, against a parametric cyber-physical adversary with different window sizes (125, 150, 200, 250, and 300), and the correct system order. The results confirm that the adversary needs a bigger window size in order to attack a system using a higher order, with a ratio of detection less than 10%. From these results we can conclude that a parametric cyber-physical adversary, who is capable to eavesdrop and analyze a large number of samples from the communication channel, and using an equivalent order system, is capable of evading detection.

3. Third use-case: This is a particular case of the second use-case, where the adversary knows a subset of inputs (control inputs) and outputs (measurements of the sensors). These inputs and outputs are independent of any other inputs and outputs. For this reason, the adversary is able to attack this subset of the system. In this use-case, the adversary has all the knowledge about a subset of the system since it is independent of the other subsets of the same system.



(a) Detection ratio with respect to the adversary order for systems of order 10

(b) Detection ratio with respect to the adversary order for systems of order 25

Fig. 3.1. Detection ratio function with respect to the adversary order. (a) For systems of order 10 against a parametric cyber-physical adversary with a window size equal to 200. And (b) for systems of order 25 against a parametric cyber-physical adversary with a window size equal to 300

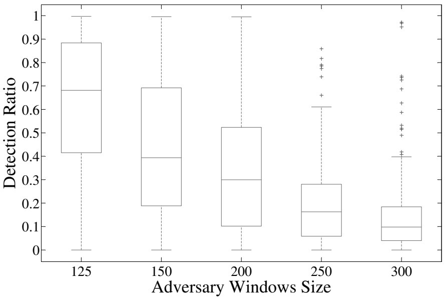


Fig. 3.2. Detection ratio function with respect to the adversary windows size. The order used by the parametric cyber-physical adversary is the correct systems order, $p = 25$

4 PIETC Watermark-Based Detection Strategy

In the previous section we have seen that the watermark-based schemes are able to handle attacks carried out by adversaries with limited knowledge about the system dynamics, f.i., the ones defined in our work as either cyber adversaries or non-parametric cyber-physical adversaries (cf. Definition 3.1). Nevertheless, it fails at detecting those adversaries with enough knowledge about the system dynamics, defined in our work as parametric cyber-physical adversaries (cf. Definition 3.2). In this section we present a new detector scheme, hereinafter denoted as periodic and intermittent event-triggered control watermark detector (PIETC-WD). This new detector aims at detecting the three adversary models defined in our work.

Our scheme consists of a local controller located in the sensors and a remote controller creating a distributed controller. The cooperation between the local and the remote controller allows us to create an intrusion detection policy to capture integrity attacks. The local controllers manage the dynamics of the plant, and the remote controller manages the system closed-loop in order to ensure the system against integrity attacks. Notice that our new scheme requires an additional controller together with the sensors, that must have enough computation power to process data estimations, e.g., to predict errors between environmental and estimated data. The actuators do not require additional computational power. Nevertheless, during the time between two consecutive events, they must keep the last data received from the remote controller.

To carry out with our scheme it is necessary to define communication policies among the sensors, the actuators and the remote controller. We define two communication policies for ensuring the system: (i) *periodic communication policy*, which the communication from the sensors to the remote controller is periodical, with a T_{sc} period, and also from the remote controller to the actuators, with a T_{ca} period; and, (ii) *intermittent communication policy*, which allows for sending data from the sensors to the remote controller if the local controller produces an alarm. Notice that T_{sc} cannot be equal to T_{ca} to avoid that an intermittent communication takes place while the periodic communication is being sent.

Definition 4.1. *Periodic and intermittent event-triggered control watermark detector (PIETC-WD) is a detector strategy with distributed control tasks. On the one hand, the sensors control the system periodically, using their local controllers and a local watermark-based detector [13]. On the other hand, the remote controller uses the estimation error received from each sensor to periodically generate the control inputs. The remote controller also controls the closed-loop communication with an intermittent watermark.*

We provide more information about the controllers and the communication policies in the following subsections.

4.1 Local Controller Design

The local controller is located in the sensors and uses a watermark in order to verify that the dynamics of the system is correct. Each sensor has a local controller with a LQG approach (cf. Section 2.3). We denote the local controller in each sensor by $i \in \{0, 1, ..., N-1\}$, where N is the number of sensors in the system. This controller adds a watermark to the sensor measurement before sending the residue to the remote controller:

$$y_t^{(i)} = y_t^{\diamond(i)} + \Delta y_t^{(i)} \tag{4.1}$$

$$r_t^{(i)} = y_t^{(i)} - C_i \hat{x}_{t|t-1}^{(i)} \tag{4.2}$$

where $y_t^{\diamond(i)}$ is the sensor measurement, $\Delta y_t^{(i)}$ is the watermark added by the local controllers, and $r_t^{(i)}$ is the residue sent to the remote controller to compute the control input $u_t^{(i)}$. Notice that the new sensor measurement $y_t^{(i)}$ is computed after verifying that $y_t^{\diamond(i)}$ is the correct sensor measurement.

4.2 Remote Controller Design

The remote controller receives periodically the residue of each sensor, $r_t^{(i)}$, and computes these residues using the LQG approach (cf. Section 2.3) to obtain the state estimation:

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(r_t) \tag{4.3}$$

where r_t is a vector generated by all the residues of the sensors. We can define the control inputs vector, u_t, as follows:

$$u_t = L(\hat{x}_{t|t-1} + K_t r_t) = L(\hat{x}_{t|t-1} + K_t(r_t^* + \Delta y_t)) \tag{4.4}$$

where r_t^* is the residues' vector before adding the watermark, and Δy_t is the vector generated by all the sensors' watermarks.

The watermark used intermittently by the remote controller is added to the control inputs. The controller adds a watermark with probability β. Denoting $\lambda_t = 1$ or 0 as indication function whether the watermark is added or not, we assume that $\lambda's$ are iid. Bernoulli random variables with $E[\lambda_t] = \beta$.

The intermittence of the watermark communication allows us to define the watermark behaviour as a non-stationary distribution. This watermark, Δu_t (cf. Equation (2.5)), permits us to detect if the closed-loop is being manipulated. It is worth noting that Δu_t is a stochastic signal with the same variance as Δy_t.

4.3 Periodic Communication Policy

The periodic communication policy is managed by the sensors. The sensors add the watermark in the measurements received by the plant and send the residue r_t to the remote controller. The remote controller uses these residues to generate the control inputs sent to the actuators. The actions of these actuators produce change in the state of the plant that are captured by the sensors. If the real state differ from the state estimated by the sensors, then the sensors will switch from periodic communication policy to intermittent communication policy (cf. Section 4.4).

In order to validate the proposal, let us assume that an attack is started at time T_0 and we compute the residue $r_t^{(i)}$ for $t \in [T_0, T_0 + T - 1]$:

$$r_t^{(i)} = y_t'^{(i)} - C_i \hat{x}_{t|t-T}^{(i)} \tag{4.5}$$

where $y_t'^{(i)}$ is the sensor measurement sent to the controller by the adversary. Moreover, it is easy to show that the following holds:

$$\begin{aligned}\hat{x}_{t|t-T}^{(i)} &= \hat{x}_{t|t-T}'^{(i)} + \mathcal{A}_i^{t-T_0}(\hat{x}_{T_0|T_0-1}^{(i)} - \hat{x}_{T_0|T_0-1}'^{(i)}) \\ &\quad + \sum_{j=0}^{t-T_0-1} (\mathcal{A}^j (A_i + B_i L_i) K (\Delta y_{t-1-j}^{(i)} - \Delta y_{t-1-j}'^{(i)}))\end{aligned} \tag{4.6}$$

where $\hat{x}'^{(i)}$ is the local estimated state for each sensor when the system is under attack and $\mathcal{A}_i = (A_i + B_i L_i)(I_i - K_i C_i)$ is a stable matrix [13]. Substitution of (4.6) in (4.5) yields:

$$\begin{aligned}r_t^{(i)} = \underbrace{y_t'^{(i)} - C\hat{x}_{t|t-T}'^{(i)}}_{\text{First term}} - \underbrace{C\mathcal{A}_i^{t-T_0}(\hat{x}_{T_0|T_0-1}^{(i)} - \hat{x}_{T_0|T_0-1}'^{(i)})}_{\text{Second term}} \\ \underbrace{-C \sum_{j=0}^{t-T_0-1} (\mathcal{A}^j (A + BL) K (\Delta y_{t-1-j}^{(i)} - \Delta y_{t-1-j}'^{(i)}))}_{\text{Third term}}\end{aligned}$$

Let us consider separately the three terms in the equation written above: the first term follows the same distribution of $(y_t - C_i \hat{x}_{t|t-1}^{(i)})$; since $\mathcal{A}_i$ is asymptotically stable – i.e. all its eigenvalues are inside the open unit disk of the complex plane – the second term converges exponentially to zero. In fact, the entries of $\mathcal{A}_i^{t-T_0}$ converge exponentially fast to zero. The third term, under attack, is not equal to zero, since $\Delta y_t^{(i)} \neq \Delta y_t'^{(i)}$, and the adversary is detected; for a cyber adversary viewpoint, the measurements of the sensors change all the time and replay measurements are not accepted; likewise, a cyber-physical adversary is not able to obtain the system model using the methodology proposed in Section 3. The parametric cyber-physical adversary model, using the ARX approach [14], is computed as follows:

$$H_{at_2} = \frac{f(R(z), Y(z)) - V(z)}{U(z)} \tag{4.7}$$

where f is a linear function of the residue $R(z)$, and the output $Y(z)$.

Assuming that the real model is $H = (Y(z) - V(z))/U(z)$, we can see that $H_{at_2} \neq H$, and the adversary is not able to obtain the model of the system.

4.4 Intermittent Communication Policy

The aforementioned periodic communication policy is managed by the sensors. The sensors produce an alarm if $g_t \geq \gamma$. When a sensor produces an alarm, this information is sent immediately to the remote controller. The affected sensor sends the real sensor measurement to the remote controller in order to carry out a second verification. An alarm happens if the control input has been manipulated by an external entity, a problem occurs in the system or the remote controller adds the watermark in the control input.

When the remote controller receives a measurement from a sensor, if a watermark Δu has not been sent, then the remote controller creates an intrusion alarm. Otherwise, if a watermark has been added to the control input, the controller verifies if this alarm is produced by the watermark. If the residue generated between the real measurements of the sensors and the estimation is under the threshold, the remote controller sends the control input generated before adding the watermark. However, if the residue is over the threshold, it means that an external entity is into the closed-loop, and an alarm is activated.

In order to validate our claims, let us assume the following attack in the communication channel between the sensor and the controller after the controller sends a control input with a watermark. It is started at time T_0 and we compute the residues r_t for $t \in [T_0, T_0 + T - 1]$:

$$r_t = y'_t - C\hat{x}_{t|t-T} \tag{4.8}$$

Moreover, it is easy to show that the following holds:

$$\begin{aligned}\hat{x}_{t|t-T} = \hat{x}'_{t|t-T} &+ \mathcal{A}^{t-T_0}(\hat{x}_{T_0|T_0-1} - \hat{x}'_{T_0|T_0-1}) \\ &+ \sum_{j=0}^{t-T_0-1} (\mathcal{A}^j B(\Delta u_{t-1-j} - \Delta u'_{t-1-j}))\end{aligned} \tag{4.9}$$

Substitution of (4.9) in (4.8) yields:

$$\begin{aligned}r_t = \underbrace{y'_t - C\hat{x}'_{t|t-T}}_{\text{First term}} &- \underbrace{C\mathcal{A}^{t-T_0}(\hat{x}_{T_0|T_0-1} - \hat{x}'_{T_0|T_0-1})}_{\text{Second term}} \\ &\underbrace{- C\sum_{j=0}^{t-T_0-1} (\mathcal{A}^j B(\Delta u_{t-1-j} - \Delta u'_{t-1-j}))}_{\text{Third term}}\end{aligned}$$

The first term follows the same distribution of $(y_t - C\hat{x}_{t|t-1})$; the second term converges exponentially to zero. Since the third term is not equal to zero, $\Delta u_t \neq \Delta u'_t$, the adversary is detected; from the cyber adversary viewpoint, the measurements of the sensors change all the time and replay measurements are not accepted; likewise, the cyber-physical adversary is not able to obtain the system model using the methodology proposed in Section 3.

4.5 New Parametric Cyber-Physical Adversary

In this section we present a new parametric cyber-physical adversary with the knowledge about the new detector strategy, in order to evaluate the new detection strategy. This attacker has knowledge about the new communication policies and the existence of the local and the remote watermarks. Nevertheless, the new adversary does not know the watermark co-variances, the controller's parameters used to obtain the correct error between data, and neither the moment when the remote controller forces an intermittent communication.

The new adversary could be able to detect the correlation model between the inputs and the outputs of the plant. This adversary can force the sensors' intermittent communication with malfunction control inputs, and mislead the controller with replay error data to obtain the model. Nevertheless, this adversary is not able to know when the communication is periodic or intermittent, since the attacker does not know when the remote control sends the watermark added to the control inputs which generates the intermittent communication. The intermittent communication does not change the communication between the remote controller and the actuators, but produces an intermittent communication between the sensors and the remote controller, necessary to verify the closed-loop.

Briefly, the new adversary is able to attack the integrity of the system. Nevertheless using the PIETC-WD strategy, the adversary is detected by the controllers of the sensors. The remote controller detects the attack when the remote controller verifies the behaviour of the closed-loop. The adversary cannot avoid the alarm in the sensors (local controller). Nevertheless, the attacker can cut off the communication between the sensors and the remote control misleading the remote controller with correct residues (e.g. replay residues). Moreover, in order to avoid the alarm in the remote controller, the adversary can switch between sending the measurements of the sensors or the residues, but the adversary has a great probability to be detected. We validate the PIETC-WD strategy against the new parametric cyber-physical adversary in the next section.

4.6 Numerical Validation

This section validates through numerical simulation the PIETC-WD strategy proposed in previous sections. We validate this strategy using a use case of a chemical plant. This plant has multiple sensors with local controllers, actuators and a remote controller, which manage all the measurements of the sensors and actuators. The sensors used in this use case send information about pressure, temperature, and density. This information is produced when there is an alarm, and also periodically to indicate the behaviour of the system to the controller. This plant has to be controlled periodically since, if during ten consecutive periodical samples, the system receives wrong or malicious control inputs able to disrupt the system, a critical state might be reached.

To avoid that an adversary gets the system into a critical state, we use our detector strategy (PIETC-WD), with a policy for the remote controller's watermark defined as follows:

- The controller's watermark uses a policy based on a probability to add the watermark in a specific window of samples. In this use case, the windows of samples is assumed equal to five. For each sequence of five control input samples, the probability to add the watermark at each sample is $\beta = 50\%$.

The system is able to produce $2^5 = 32$ different sequences with the same probability to be generated, $\theta = 1/2^5$. Nevertheless, if among these five samples, the system does not send any watermark, three more samples are used to add a watermark to the control input until a new control sequence starts. These three samples added to the original control sequence add $2^3 = 8$ more sequences where the five first samples have not watermark, and the three last samples have the following probability to add the watermark:

- The probability to add the watermark in the sixth sample is 60%.
- The probability to add the watermark in the seventh sample is 50% if the watermark is added in the sixth sample. Otherwise, if the watermark is not added, the probability is 60%.
- The probability to add the watermark in the eighth sample is 50%, if the watermark is added in the sixth or seventh sample. Otherwise, the probability is 60%.

Figure 4.1 shows the results of 200 Monte Carlo simulations using the above use case and controller's watermark policy, against the cyber and the cyber-physical adversary. These results present that the ratio of detection is around 97% against the new parametric cyber-physical adversary and more than 99% against the other cyber and cyber-physical adversaries using the PIETC-WD strategy with a correct policy for the remote controller's watermark.

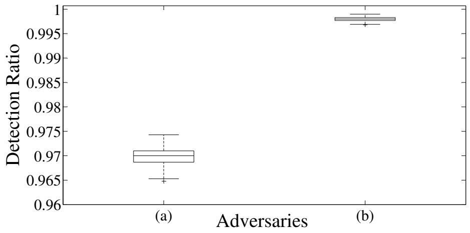


Fig. 4.1. Detection ratio function with respect to the PIETC-WD strategy with a defined controller's watermark policy; (a) against the new parametric cyber-physical adversary; and (b) against cyber or other cyber-physical adversaries

5 Related Work

Security of cyber-physical systems (CPS) is drawing a great deal of attention recently [4]. Solutions focusing on control approaches for the detection of cyber-physical attacks is the research axis more closely related to this paper. This axis is the one that explicitly considers the interconnection between cyber and physical control domains in networked control systems. Recently, the control system community started to study security of cyber-physical systems both under the methodological point of view and from a more technological standpoint by looking at particular problems arising in, e.g., smart grids. Concerning the methodological aspects, several studies have proposed to adapt classical frameworks to handle security issues in networked control systems.

Among cyber-physical attacks handled in the literature, replay attack is the only attack that the adversary is able to carry out without knowledge about system model. To carry out the rest of the attacks, it is necessary some system knowledge. For example, to execute a dynamic false-data injection attack handled by Mo *et al.* [12], the adversary has to have a perfect knowledge of the plant's behaviour, or to execute a covert attack, handled by Smith *et al.* [18], is necessary knowledge of the plant's and controller's behaviour. Otherwise, the adversaries defined in this paper are able to obtain the knowledge of the plant's behaviour in order to attack the system. Concerning the detection mechanism, one line of research has considered the adaptation of fault detection systems to detect a class of attacks [13, 15, 19]. In particular, Mo *et al.* show in [13] that it is possible to detect replay attacks by properly watermarking control inputs. Teixeira *et al.* propose in [19] a mathematical framework to model several attack strategies. An alternative modeling approach is taken by Pasqualetti *et al.* in [15], where the authors propose to employ the theory of geometric control to model cyber-physical systems attacks. In this paper we focus on the interconnection between control strategies and a watermark detector to handle the integrity attacks.

6 Conclusion

In this paper, we have addressed security issues in cyber-physical system. We have focused on designing a robust distributed control strategy (PIETC-WD strategy), in order to detect parametric cyber-physical adversaries. This adversary is able to acquire the knowledge of the system needed to compromise the control inputs and the measurements of the sensors to attack the system.

We have reviewed the watermark-based detector proposed in [13]. We have shown that the detector fails at properly handling attacks carried out by parametric cyber-physical adversaries. In particular, we have shown that an adversary that learns about the system model is able to model the watermark from the control signal and succeeds at attacking the system without being detected. We have also shown that the watermark-based detector works against parametric cyber-physical adversary who knows only a set of control inputs, [21]. Nevertheless, if the adversary knows all the control inputs and sensor measurements of the system, and uses the correct orders range with a window size sufficiently long, the watermark-based detector fails.

Finally, we have presented and validated our strategy (PIETC-WD). This strategy is capable to detect cyber and cyber-physical adversaries with a great detection ratio, even if the adversary finds the correct model of the system.

Acknowledgements. The authors acknowledge support from the Cyber CNI Chair of Institut Mines-Télécom. The chair is held by Télécom Bretagne and supported by Airbus Defence and Space, Amossys, EDF, Orange, La Poste, Nokia, Société Générale and the Regional Council of Brittany. It has been acknowledged by the Center of excellence in Cybersecurity.

References

[1] A. Arvani and V. S. Rao. Detection and protection against intrusions on smart grid systems. *International Journal of Cyber-Security and Digital Forensics*, 2014.

[2] M. Barenthin Syberg. Complexity issues, validation and input design for control in system identification. 2008.

[3] B. Brumback and M. Srinath. A chi-square test for fault-detection in kalman filters. *Automatic Control, IEEE Transactions on*, 32(6):552–554, 1987.

[4] D. Corman, V. Pillitteri, S. Tousley, and U. Tehranipoor, Lindqvist. NITRD Cyber-Physical Security Panel. 35th Symposium on Security and Privacy, 2014.

[5] V. L. Do, L. Fillatre, and I. Nikiforov. A statistical method for detecting cyber/-physical attacks on scada systems. In *Control Applications*. IEEE, 2014.

[6] N. Falliere, L. O. Murchu, and E. Chien. W32. stuxnet dossier. *White paper, Symantec Corp., Security Response*, 5, 2011.

[7] D. Han, Y. Mo, J. Wu, S. Weerakkody, B. Sinopoli, and L. Shi. Stochastic event-triggered sensor schedule for remote state estimation. *IEEE Transactions on Automatic Control*, 60(10):2661–2675, 2015.

[8] W. Heemels, M. Donkers, and A. R. Teel. Periodic event-triggered control for linear systems. *Automatic Control, IEEE Transactions on*, 58(4):847–861, 2013.

[9] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu. A survey of recent results in networked control systems. *IEEE*, 95(1):138, 2007.

[10] Y. Ke-You and X. Li-Hua. Survey of recent progress in networked control systems. *Acta Automatica Sinica*, 39(2):101–117, 2013.

[11] K.-D. Kim and P. R. Kumar. Cyber-physical systems: A perspective at the centennial. *IEEE*, 100(Special Centennial Issue):1287–1308, 2012.

[12] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli. False data injection attacks against state estimation in wireless sensor networks. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 5967–5972. IEEE, 2010.

[13] Y. Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *Control Systems, IEEE*, 35(1):93–109, 2015.

[14] H. Natke. System identification: Torsten söderström and petre stoica. *Automatica*, 28(5):1069–1071, 1992.

[15] F. Pasqualetti, F. Dorfler, and F. Bullo. Cyber-physical security via geometric control: Distributed monitoring and malicious attacks. In *Decision and Control, 2012 IEEE 51st Annual Conference on*, pages 3418–3425. IEEE, 2012.

[16] J. Rubio-Hernán, L. De Cicco, and J. García-Alfaro. Revisiting a watermark-based detection scheme to handle cyber-physical attacks. In *11th International Conference on Availability, Reliability and Security*, Salzburg, Austria, 2016. IEEE.

[17] J. Salt, V. Casanova, A. Cuenca, and R. Pizá. Sistemas de control basados en red modelado y diseño de estructuras de control. *Revista Iberoamericana de Automática e Informática Industrial RIAI*, 5(3):5–20, 2008.

[18] R. Smith. Covert misappropriation of networked control systems: Presenting a feedback structure. *Control Systems, IEEE*, 35(1):82–92, 2015.

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

[20] S. Tripathi and M. A. Ikbal. Step size optimization of lms algorithm using aunt colony optimization & its comparison with particle swarm optimization algorithm in system identification. 2015.

[21] S. Weerakkody, Y. Mo, and B. Sinopoli. Detecting integrity attacks on control systems using robust physical watermarking. In *Decision and Control*, 2014.