Revisiting a Watermark-based Detection Scheme to Handle Cyber-Physical Attacks

José Rubio-Hernán
SAMOVAR,
Telecom SudParis, CNRS,
Université Paris-Saclay,
Evry, France
Jose.Rubio_Hernan@telecom-sudparis.com

Luca De Cicco
Politecnico di Bari
Dipartimento di Ingegneria
Elettrica e dell'Informazione
Bari, Italy
luca.decicco@poliba.it

Joaquín García-Alfaro
SAMOVAR,
Telecom SudParis, CNRS,
Université Paris-Saclay,
Evry, France
Joaquin.Garcia_Alfaro@telecom-sudparis.com

***Abstract*—We address detection of attacks against cyber-physical systems. Cyber-physical systems are industrial control systems upgraded with novel computing, communication and interconnection capabilities. In this paper we reexamine the security of a detection scheme proposed by Mo and Sinopoli (2009) and Mo *et al.* (2015). The approach complements the use of *Kalman filters* and *linear quadratic regulators*, by adding an authentication watermark signal for the detection of integrity attacks. We show that the approach only detects *cyber adversaries*, i.e., attackers with the ability to eavesdrop information from the system, but that do not attempt to acquire any knowledge about the system model itself. The detector fails at covering *cyber-physical adversaries*, i.e., attackers that, in addition to the capabilities of the *cyber adversary*, are also able to infer the system model to evade the detection. We discuss an enhanced scheme, based on a multi-watermark authentication signal, that properly detects the two adversary models.**

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

I. Introduction

In an effort of reducing complexity and costs, traditional industrial control systems are being upgraded with novel computing, communication and interconnection capabilities. Industrial control systems that close the loop through a communication network are hereinafter denoted *cyber-physical systems*. The adoption of new communication capabilities comes at the cost of introducing new security threats that are required to be holistically handled, both in terms of safety and security (in the traditional ICT sense). The recently coined *cyber-physical security* term refers to the mechanisms that address this specific challenge [1].

In this paper, we focus on the adaptation of physical-layer failure detection mechanisms (e.g., systems for the detection of faults and accidents) to handle, as well, attacks (e.g., replay and integrity attacks conducted by malicious adversaries). We reexamine the security of a specific scheme by Mo *et al.* proposed in [2], [3]. This scheme relies on the adaptation of a real-time failure detector based on a *linear time-invariant* model of the system. Built upon *Kalman filters* and *linear-quadratic regulators*, the scheme employs authentication watermarks to protect the integrity of physical measurements communicated over the cyber and physical control domains of a networked control system (NCS). Without the protection of the messages, malicious actions can be conducted to mislead the system towards unauthorized or improper actions and affect the availability of the system services.

The main contributions of this paper are summarized as follows:

- We reexamine the security of the attack detector proposed in [2], [3] under a new adversary model.
- We show security weaknesses in [2], [3] under the new adversary model.
- An enhanced detector approach is presented and validated via numerical simulations.

Section II provides the related work. Section III reviews the detector scheme in [2], [3], provides a new adversary model and reexamines the security of the detector under the new adversary model. Section IV presents a novel multi-watermark detection scheme that handles the uncovered limitations of the previous construction. Section V concludes the paper.

II. Related Work

Security of cyber-physical systems is drawing a great deal of attention recently [1] after the infamous StuxNet malware [4] uncovered the potential of successful security attacks carried out against such systems. Several authors have studied the requirements to take into account the new security issues when designing security mechanisms for cyber-physical systems. In [5], Cardenas *et al.* define the issue of secure control by analyzing separately the problem first from a information security point of view and then by looking at specific control issues. In [6], Cardenas *et al.* also outline for the first time the difference between corporate ICT security and cyber-physical system security.

From a cyber perspective, the protocols for industrial control systems built upon a networked control system must cover regulation rules such as delays and faults [7]. Indeed, most industrial control protocols (e.g., Modbus, DNP3, AGA-12, PROFINET and Ethernet/IP), are not designed to provide security from a traditional information or network perspective. Nevertheless, there are some protocols with security extensions. AGA-12 uses cryptography to add integrity and confidentiality protection, but with high deployment cost [8]. DNP3 has an extension named DNP3-SA (fifth version IEEE-1815-2012), adding new security features to DNP3, ensuring integrity and authentication for the message. Even so, current NCSs use these protocols over TCP/IP or UDP/IP communication (e.g., Modbus and DNP3 over TCP, PROFINET over TCP, Ethernet/IP over TCP or UDP). Over Ethernet network communication, there is traditional ICT security until application layer. In the application layer, we find the NCSs protocols, such as PROFINET which has a new layer, PROFIsafe, but this layer has been designed to ensure safety, hence protecting the PROFINET protocol against malfunction (e.g., transmission errors). It does not ensure security against intentional malicious acts [9]. It is worth noting that most of the protocols of application layer are modifications of serial protocols and do not provide security. So, although transport and network layers can provide some security elements, these mechanism are not sufficient to ensure control-data protection [3]. To solve the problem of control-data protection, cryptography could also be used. However, without underestimating a cryptographic solution, in this paper we revisit and analyze a complementary and alternative solution proposed by Mo et al. [10], to ensure the integrity and the authentication of control-data using also the control domain of the networked-control system. This security solution proves to be useful, e.g., in the case where a cyber adversary bypasses a cryptographic solution, by adding an additional protection layer.

The line of research that is more closely related to this paper 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, power grids, water distribution systems. Figure 1 shows the way how an adversary conducting a cyber-physical attack can be represented through a block diagram, a representation typically used by the control system community. The $\bigoplus$ symbol in the figure represents a *summing junction*, i.e., a linear element that outputs the sum of a number of input signals. In a nutshell, the figure represents the control loop of a monitored system, and how an adversary succeeds at modifying some of the readings, by recording and replicating previous measurements corresponding to normal operation conditions. Then, the adversary modifies the control input u to affect the system state and disrupt normal operation conditions. If, on one hand, the adversary is not required to have the knowledge of the system process model, on the other hand access to all sensors (i.e., it has access to all components of the vector y) or insecure communication protocols is required to carry out a successful attack. This type of adversary is undetectable with a monitor detector which only verifies faulty measurements.

Several studies have proposed the adaptation of fault detection systems to detect as well attacks. Sophisticated variations of the attack in Figure 1 include (i) bias-injection cyber-physical attacks, in which the new data injected by the adversary corresponds to a bias from the legitimate data, with the aim of leading the system to wrong control decisions (e.g., to cause malfunction in the long-term); and (ii) geometric-injection cyber-physical attacks, in which the bias is gradually injected. The attack may remain undetected when data compatible with the system dynamics are injected, potentially leading the system to irreversible damages. Teixeira *et al.* propose in [11] a mathematical framework to model several attack strategies. They show how to model each considered attack policy within this framework with the goal of evaluating its impact based on the mathematical concept of *safe sets*. An alternative modeling approach is taken by Pasqualetti *et al.* in [12], where the authors propose to employ the theory of geometric control to model cyber-physical systems attacks. Within this framework, a characterization of undetectable and unidentifiable attacks is provided.

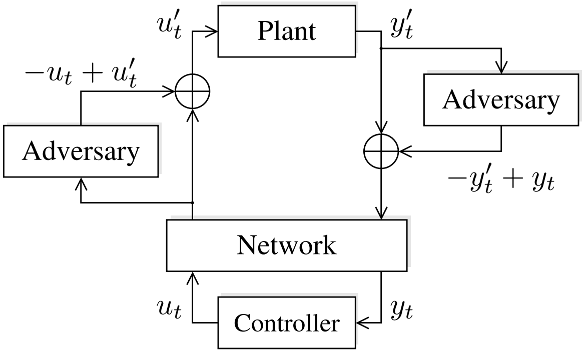


Fig. 1: Representation of a cyber-physical attack against a networked control system.

Techniques to prevent the aforementioned cyber-physical attacks have been proposed in the literature. In [13], Franklin *et al.* describe a signal-based detector method, using discrete wavelet transformations. Brumback and Srinath study in [14] strategies for handling cyber-physical attacks using statistical detection methods. Mo *et al.* propose in [2], [3] the use of watermark-based detection by adapting traditional failure detection mechanisms (e.g., detectors to handle faults and errors). In the following sections, we elaborate further on the watermark-based technique by Mo *et al.*, discuss about some security limitations and propose an improved technique.

III. Watermark-based Attack Detection

In [2], [3], a watermark-based strategy is proposed with the aim of detecting replay and injection attacks against cyber-physical systems. The overall goal of this section is to provide a brief review of the mechanism proposed in [2], [3] and to assess its performance when a new adversary model, that we name *cyber-physical adversary*, is employed. In particular, this section is organized as follows: in Section III-A we provide some necessary definitions and background concerning the class of control systems considered in this paper; Section III-B describes the attack detection scheme proposed in [2], [3]; in Section III-C, we propose the cyber-physical adversary; in Section III-D, we show a practical method that can be employed by the adversary to guess the watermark; finally, we evaluate the performance of the detection scheme when attacked by the cyber-physical adversary via numerical simulations.

A. *Definitions and Background*

We consider plants of industrial control systems that can be mathematically modeled as discrete linear time-invariant (LTI) systems. It is worth mentioning that a mathematical model provides a rigorous way to describe the dynamical behaviour of a given system. Such class of systems can be described as follows:

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

where $x_t \in \mathbb{R}^n$ is the vector of the state variables (or state) at the t-th time step, $u_t \in \mathbb{R}^p$ is the control signal, and $w_t \in \mathbb{R}^n$ is the *process noise* that is assumed to be a zero mean Gaussian white noise with covariance Q, *i.e.* $w_t \sim N(0, Q)$. Moreover, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times p}$ are respectively the *state* matrix and the *input* matrix.

A static relation maps the state x_t to the system output $y_t \in \mathbb{R}^m$:

$$y_t = Cx_t + v_t \tag{2}$$

where $C \in \mathbb{R}^{m\times n}$ is the output matrix. The value of the output vector y_t represents the measurement produced by the sensors that is affected by a noise v_t assumed as a zero mean Gaussian white noise and covariance R, *i.e.* $v_t \sim N(0, R)$.

For such a class of systems, a widely used control technique is the *Linear Quadratic Gaussian* (LQG) approach. The overall goal of an LQG controller is to produce a control law u_t such that a quadratic cost J, that is function of both the state x and the control input u, is minimized:

$$J = \lim_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{n-1}(x_i^T W x_i + u_i^T U u_i)\right] \tag{3}$$

where W and U represent positive definite cost matrices [15].

It is well-known that such a control problem has, under some technical conditions, an optimal solution that, thanks to the separation principle, is made of two components that can be designed independently:

1) a *Kalman filter* that, based on the noisy measurements, produces an optimal state estimation $\hat{x}_t$ of the state x;
2) a *Linear Quadratic Regulator* (LQR) that, based on the state estimation $\hat{x}_t$, provides the control law u_t that solves the LQR problem (3).

Let us briefly illustrate how these two components are designed. The Kalman filter estimates the state as follows:

- Predict (*a priori*) system state $\hat{x}_{t|t-1}$ and covariance:

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1} + Bu_{t-1}$$

$$P_{t|t-1} = AP_{t-1}A^T + Q$$

- Update parameters and (*a posteriori*) system state and covariance:

$$K_t = (P_{t|t-1}C^T)(CP_{t|t-1}C^T + R)^{-1}$$

$$\hat{x}_t = \hat{x}_{t|t-1} + K_t(y_t - C\hat{x}_{t|t-1})$$

$$P_t = (I - K_tC)P_{t|t-1}$$

where K_t and P_t denote, respectively, the Kalman gain and the *a posteriori* error covariance matrix, and I is the identity matrix of appropriate dimensions.

The optimal control law u_t provided by the LQR is a linear controller:

$$u_t = L\hat{x}_t \tag{4}$$

where L denotes the feedback gain of a linear-quadratic regulator (*LQR*) which minimizes the control cost (3) and it is defined as follows (see [2], [3] for further details):

$$L = -(B^TSB + U)^{-1}B^TSA,$$

with S being the matrix that solves the following discrete time algebraic Riccati equation:

$$S = A^TSA + W - A^TSB[B^TSB + U]^{-1}B^TSA.$$

B. *The χ^2 Detector*

This section briefly describes the detection scheme proposed in [2], [3]. The procedure is applicable to discrete LTI plants controlled by a LQG controller as detailed in Section III-A.

Before presenting the detection scheme, we provide a definition of the adversary model considered in [2], [3]:

Definition 1. *An attacker that has the ability to eavesdrop all the messages containing the sensor outputs y_t and to inject messages with a signal y'_t to conduct malicious actions is defined as a cyber adversary.*

Remark. *It is important to notice that the definition given above does not suppose that the attacker possesses (or makes attempts to gather) any knowledge about the system model, reason why we name such attacker a cyber adversary.*

In the following, we will denote with u_t^* the output of the LQR controller given by Equation (4) and with u_t the control input that is sent to the plant (see Equation (1)). The idea is to superpose to the optimal control law u_t^* a watermark signal $\Delta u_t \in \mathbb{R}^p$ that serves as an authentication signal. Thus, the control input u_t is given by:

$$u_t = u_t^* + \Delta u_t \tag{5}$$

The watermark signal is a Gaussian random signal with zero mean that is independent both from the state noise w_t and the measurement noise v_t. Such an authentication watermark is expected to detect replay and integrity attacks modeled by the cyber adversary defined above. Now that the optimal control law u_t^* is equipped with the authentication signal Δu_t, a *detector* – physically co-located with the controller – can be designed having the goal of generating alarms when an attack takes place. Towards this end, [2], [3] propose to employ a χ^2 detector, a well-known category of real-time anomaly detectors classically used for fault detection in control systems [16], for the purpose of attack detection.

An *alarm signal* g_t is computed based on the residues $r_t = y_t - C\hat{x}_{t|t-1}$ generated by the estimator. Then, g_t is compared with a threshold γ to decide whether the system is in a normal state. The threshold is tuned to minimize false alarms [2], [3]. The alarm signal g_t is computed as follows:

$$g_t = \sum_{i=t-w+1}^{t} (y_i - C\hat{x}_{i|i-1})^T \mathcal{P}^{-1}(y_i - C\hat{x}_{i|i-1}) \quad (6)$$

where w is the size of the detection window and $\mathcal{P} = (CPC^T + R)$ is the co-variance of an independent and identically distributed (i.i.d.) Gaussian input signal from the sensors.

The system is considered not under attack if $g_t < \gamma$, otherwise if $g_t \geq \gamma$ the system is considered to be under attack and the detector generates an alarm.

C. Cyber-Physical Adversary

Let us assume the system employs the detector described in Section III-B, 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.

Let us now introduce an enhanced adversary that is aware of the fact that the system employs the χ^2 detector presented above. Since the detector is based on the watermark signal Δu_t, we will show that an adversary that is able to extract a stationary signal Δu_t, from the signal u_t is able to conduct a replay attack while remaining undetected.

Definition 2. *An attacker that, in addition to the capabilities of the cyber adversary, is also able to eavesdrop the messages containing the output of the controller u_t with the intention of improving its knowledge about the system model is defined as a cyber-physical adversary.*

Let us consider a cyber-physical adversary that wants to carry out a replay attack. Signals u'_t and y'_t are assumed to be respectively the output of the controller and the output of the measurement when a replay attack is taking place. We denote with $\Delta u'_t$ the watermark guessed by the cyber-physical adversary. For the time being, we do not consider the algorithm employed by the adversary to guess the watermark. See Section III-D for such details.

Proposition. *A cyber-physical adversary that is able to exactly estimate the watermark signal injected by the controller cannot be detected by the χ^2 detector (6).*

Proof: We consider an attack is started at time T_0 and we compute the residues r_t for $t \in [T_0, T_0 + T - 1]$:

$$r_t = y'_t - C\hat{x}_{t|t-T} \quad (7)$$

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

$$\begin{aligned} \hat{x}_{t|t-T} &= \hat{x}'_{t|t-T} + \mathcal{A}^{t-T_0}(\hat{x}_{T_0|T_0-1} - \hat{x}'_{T_0|T_0-1}) \\ &+ \sum_{i=0}^{t-T_0-1} (\mathcal{A}^i B(\Delta u_{t-1-i} - \Delta u'_{t-1-i})) \quad (8) \end{aligned}$$

where $\hat{x}'$ is the estimated state when the system is under attack and $\mathcal{A} = (A + BL)(I - KC)$ is a stable matrix [2], [3]. Substitution of (8) in (7) yields:

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

Let us consider separately the three terms in the equation written above: the first term follows the same distribution of $(y_t - C\hat{x}_{t|t-1})$; since $\mathcal{A}$ is asymptotically stable – i.e. all its eigenvalues are inside the open unit disk of the complex plane – the second term converges exponentially fast to zero. In fact, the entries of $\mathcal{A}^{t-T_0}$ converge exponentially fast to zero. Now, if the third term would be equal to zero, the dynamics of r_t would recover the dynamics of the residues when no attack is undergoing and thus, the attack would not be detected. Under the hypothesis of this proposition, the adversary knows exactly the watermark signal and thus $\Delta u_t = \Delta u'_t$ which makes the third term equal to zero and concludes the proof. ■

D. Acquiring the Watermark Signal

Motivated by Proposition III-C, we show now a practical method that can be used to acquire the watermark signal Δu_t. In particular, we propose an adversary that employs an adaptive Least Mean Square (LMS) filter with the purpose of running an online identification of the system model. With the identified model, it is possible to obtain the watermark and, finally, using it to authenticate messages with the aim of driving the system to an undesired state.

We denote with p the LMS filter order and with μ its step size. The step size μ is upper bounded by $2/\lambda_{max}$, where λ_{max} is the maximum eigenvalue of the auto-correlation matrix $R = E[XX^H]$, where X^H is the Hermitian transpose, or conjugate transpose, of X. Observe that if μ is chosen too small, the time to converge to optimal weights tends to be large [17]. The adversary initializes the weight matrix $\mathcal{W}$ to be equal to the zero matrix. Then, the adversary's algorithm shown in Algorithm 1, is run online. It is worth noting that in this algorithm; X is the input signal, e is the error, $\bar{e}$ is its complex conjugate, and d is the desired output signal.

Once the system model has been identified, the adversary is able to extract the watermark and to carry out the replay attack. In particular, the adversary follows the steps described below:

1) *Eavesdropping of u_t and y_t and decomposition*: The adversary captures both the control law u_t and the sensors output y_t to make the decomposition between the information data and the watermark using the LMS filter as a noise cancellation adaptive filter. With this first step, we are able to separate u^*_t and the watermark Δu_t starting from u_t. Notice

Algorithm 1 Cyber-Physical Adversary Algorithm

```
procedure ADVERSARY ALGORITHM
    $k \leftarrow$ length of eavesdropped data
    $p \leftarrow$ filter order
    $j \leftarrow p$
top:
    if $j < k$ then $i \leftarrow 1$.
loop:
    if $i \leqslant p$ then
        $ini \leftarrow j - p + 1$.
        $e(ini) \leftarrow d(ini) - \mathcal{W}^T X[x(ini), ... x(j)]$.
        $W \leftarrow \mathcal{W} + \mu \bar{e}(ini) X[x(ini), ... x(j)]$.
        $j \leftarrow j + i$.
        $i \leftarrow i + 1$.
        goto loop.
        close;
    goto top.
```

that, since the system is linear, it follows from the superposition principle that $y_t = y_t^* + y_t^{\Delta u}$, being y_t^* the output due to u_t^* and $y_t^{\Delta u}$ the output due to the watermark Δu_t.

2) *Acquiring the weight matrix, $\mathcal{W}$*: The adversary uses the LMS adaptive filter described before, as a system identification method.

3) *Computing the attack sensor measurement y_t'*: The adversary attacks the system by sending fake sensor measurements y_t', where $y_t^{\Delta u}$ is computed using the watermark Δu_t as follows:

$$y_t^{\Delta u} = \mathcal{W}^T \Delta u_t$$

and $y_t' = y_{t-1}^* + y_t^{\Delta u}$.

In the remainder of this section, we show via numerical simulations that the detection mechanism proposed in [2], [3] is not sufficiently robust and is not able to detect cyber-physical adversaries (see Section III-C) that are able to identify the system model by eavesdropping the data channel.

In order to simulate the NCS, we have employed a simplified version of the Tennesse Eastman control challenge problem [18] also used as a benchmark in [19]. This system simulates a MIMO system of order $n = 7$ with $p = 4$ inputs and $m = 4$ outputs. In particular the model of the discrete LTI system described by Equations (1)-(2) is defined by the following matrices:

$$\text{A=}\begin{bmatrix} 0.987 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.895 & -0.025 & 0 & 0 & 0 & 0 \\ 0 & 0.036 & 0.999 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -0.008 & 0 & 0 \\ 0 & 0 & 0 & 0.005 & 0.960 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.999 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.990 \end{bmatrix},$$

$$\text{B=}\begin{bmatrix} 0.149 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.071 \\ 0 & 0 & 0 & 0.001 \\ 0.380 & 0 & -0.096 & 0 \\ 1.000 & 0 & -0.096 & 0 \\ 0 & 0.038 & 0 & 0 \\ 0 & 0 & 0 & 0.075 \end{bmatrix},$$

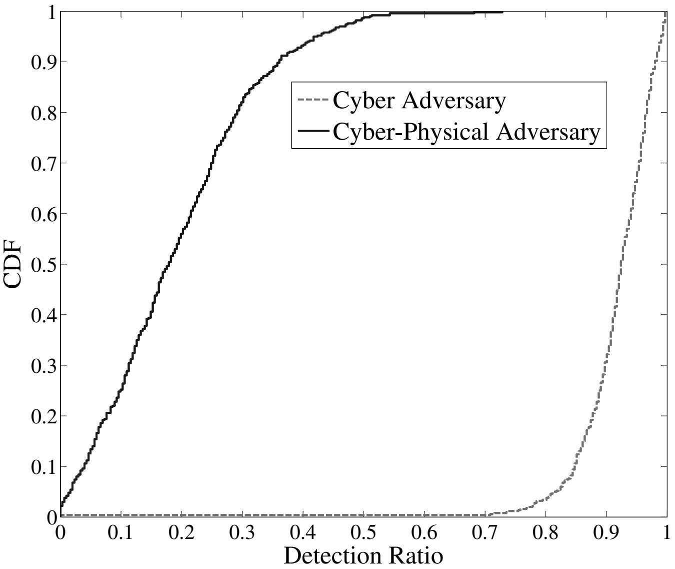


Fig. 2: Cumulative distribution function (CDF) of the detection ratio associated to the χ^2 detector (see Section III-B), obtained by measuring the DR metric (see Equation 9) for 500 simulations (both cyber and cyber-physical adversary cases).

$$\text{C=}\begin{bmatrix} 0.151 & -0.076 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.040 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.133 \end{bmatrix}.$$

Moreover, the co-variance matrices are equal to $Q = 0.01I$ and $R = I$, whereas the cost matrices are $W = 1.5I$ and $U = 10I$.

In order to quantify the detector performance, we define the DR (*Detection Ratio*) metric as follows:

$$DR = \frac{\sum_{t=T_0}^{T_0+T_a} \mathbb{1}_{g_t \geq \gamma}}{T_a} \tag{9}$$

where T_a is the attack duration, and $\mathbb{1}$ is the indicator function whose output is equal to 1 if the Boolean condition given as its argument ($g_t \geq \gamma$) is true; or it is equal to 0 otherwise. In a nutshell, $DR \in [0, 1]$ can be considered as an efficiency index for the detector: DR is equal to one when the attack is always detected; and it is equal to zero when the attack is always undetected.

Figure 2 shows the CDF (Cumulative Distribution Function) of the detection ratio obtained by measuring DR for 500 simulations both in the case of the cyber-physical and the cyber adversary. The figure shows that the detection scheme proposed in [2], [3] is able to provide a median detection ratio that is larger than 0.9 when a cyber adversary attacks the system. However, using a cyber-physical adversary that acquires the watermark, the median detection ratio drops to around 0.2. This quantitatively shows that the detection strategy proposed in [2], [3] is not sufficiently robust for security.

IV. Multi-watermark based Attack Detection

In the previous section we have shown that the watermark-based detection scheme in [2], [3] is able to properly handle attacks carried out by cyber adversaries, but it fails at detecting cyber-physical adversaries. In this section, we propose a detection scheme that extends the one presented in Section III-B and overcomes the limitations shown in the previous section.

A. The Proposed Multi-watermark Signal

The goal of the new detection scheme is to increase the difficulty in retrieving the authentication watermark Δu_t from the control signal u_t, so that the probability of detecting an attack from a cyber-physical adversary can be increased. We assume that the NCS employs exactly the same LQG controller and the same detection strategy presented in Section III. The only difference in the proposed detection scheme is the way the watermark signal Δu_t is generated. The control input u_t, as in the case of the detection scheme presented in Section III-B, is computed as the superposition of the optimal control signal u_t^* produced by the LQR controller and the multi-watermark signal Δu_t. The idea is to construct the watermark signal by switching between N different and independent processes with different co-variance and average (offsets). More precisely, the non-stationary watermark, Δu_t, is obtained by periodically switching, with a period T, between N signals $\Delta u_t^{(i)}$, with $i \in \mathcal{I} = \{0, 1, \dots, N-1\}$, extracted by different stochastic processes. Hence, the watermark signal Δu_t can be formalized as follows:

$$\Delta u_t = \Delta u_t^{(s(t,T))} \tag{10}$$

where $s : \mathbb{N} \times \mathbb{R} \to \mathcal{I}$ is a static function that maps the time sample t and the switching period T to an element of the index set $\mathcal{I}$, defined as follows:

$$s(t,T) = \left\lfloor \frac{1}{T} \mod (t, NT) \right\rfloor \tag{11}$$

where $\mod (x,y)$ is the modulo operator and $\lfloor\cdot\rfloor$ is the floor function.

By using the proposed watermark (see Equation 10), we now have an adaptive protection mechanism with two main configurable parameters: the number of distributions N and the switching frequency $f = 1/T$. It is worth to notice that the original watermark signal described in Section III is recovered when $f \to 0$ and when $\Delta u_t^{(0)}$ being a stationary zero mean Gaussian process.

B. Validation

This section validates through numerical simulations the detection scheme proposed in Section IV-A. In particular, we aim at showing that the proposed watermark signal is able to detect cyber-physical adversaries (see Section III-C) with a higher detection ratio with respect to the one obtained with the watermark proposed in [2], [3]. Towards this end, we start with a system described by the following matrices:

$$A = \begin{bmatrix} 0.5 & 0.8 \\ 0 & 0.8 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{12}$$

and co-variance matrices equal to $Q = 0.8I$ and $R = I$. The positive definite cost matrices W and U are both equal to the identity matrix. The simulation uses a Simulink NCS model under a cyber-physical adversary starting the attack at $t =$ 700s. We have used $N = 3$ different distributions switched at random: a Gaussian, a Rician and a Rayleigh distribution. Table I shows the co-variance and offset configured in the simulations for each distribution.

TABLE I: Sample parameters used in the multi-watermark Simulink model.

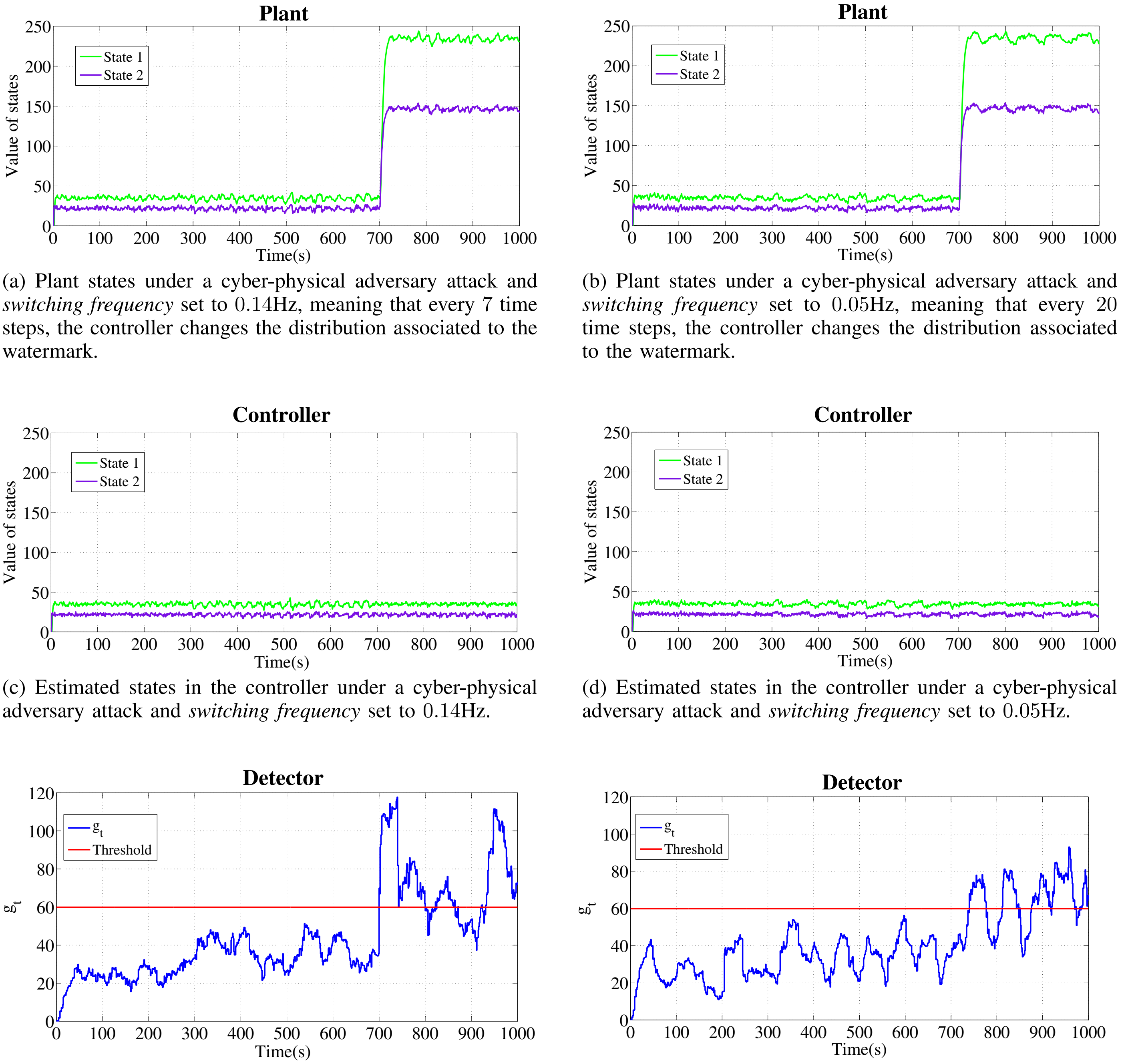
Distribution	**Co-variance**	**Offset**
Gaussian	2.560	0.0
Rician	1.113	2.401
Rayleigh	0.576	1.500

To validate the proposed attack detection scheme, we compare the system dynamics considering two different switching frequencies. We have simulated a high frequency switching watermark configured to switch each 7 time samples, and a low frequency switching configured to switch each 20 time samples. Figures 3(a) and (c) show the plant dynamics and the dynamics of the states estimated by the controller in the case of a switching frequency watermark configured to 7 time samples and a cyber-physical adversary attack. Figure 3(a) shows that the adversary is able to drive the states to an undesired value. Nevertheless, the controller misled by the adversary, does not perceive such situation (see Figure 3(c)). Figures 3(b) and (d) show the plant dynamics and the dynamics of the states estimated by the controller when the watermark is switched each 20 time samples. The dynamics show exactly the same behavior described above.

Figures 3(e) and 3(f) show the dynamics of the alarm signal g_t produced by the detector, respectively in the case of high and low switching frequency. Notice that switching the watermark distributions at a high frequency provides better detection performances compared to the case of a low switching frequency.

To quantify the effectiveness of the proposed detection scheme, we compute the detection ratio DR as a function of the switching frequency. In particular, for each considered frequency f we run 500 Monte Carlo simulations (with randomly generated system parameters) both in the case of the cyber-physical and the cyber adversary, and we compute the CDF of the detection ratio.

Let us now confront the performance obtained with the detection strategy based on multiple watermark signals proposed in this paper with that proposed in [2], [3] in both the case of a cyber-physical and a cyber adversary. In the case of the proposed multi-watermark strategy we consider two switching frequencies $f_L = 0.05$Hz (switching watermark each 20 time steps) and $f_H = 0.14$Hz (switching watermark each 7 time steps). The results of this comparison are shown in Figure 4. Let us focus on the detection strategy proposed in [2], [3]: as shown before, the detector is able to consistently detect cyber attack but it performs poorly when a cyber-physical adversary attacks the system. On the other hand, the proposed detection strategy based on multiple watermarks is able to provide a higher detection ratio: in particular we notice that the detector employing a higher switching frequency f_H provides better performances with respect to the case of using the lower switching frequency f_L.



(a) Plant states under a cyber-physical adversary attack and *switching frequency* set to 0.14Hz, meaning that every 7 time steps, the controller changes the distribution associated to the watermark.

(b) Plant states under a cyber-physical adversary attack and *switching frequency* set to 0.05Hz, meaning that every 20 time steps, the controller changes the distribution associated to the watermark.

(c) Estimated states in the controller under a cyber-physical adversary attack and *switching frequency* set to 0.14Hz.

(d) Estimated states in the controller under a cyber-physical adversary attack and *switching frequency* set to 0.05Hz.

(e) Detector results, *switching frequency* set to 0.14Hz.

(f) Detector results, *switching frequency* set to 0.05Hz.

Fig. 3: Simulation results where *State 1* and *State 2* are the temperatures of two different chemical processes. Attacks start at $t = 700s$. (a),(b) The dynamics of the states vector in the plant under a cyber-physical adversary attack and switching frequency configured with two different configurations (0.14Hz and 0.05Hz). (c),(d) The dynamics of the states vector estimated in the controller, under the same scenarios. (e),(f) The dynamics of the alarm signal g_t produced by the multi-watermark based detector, under the same scenarios.

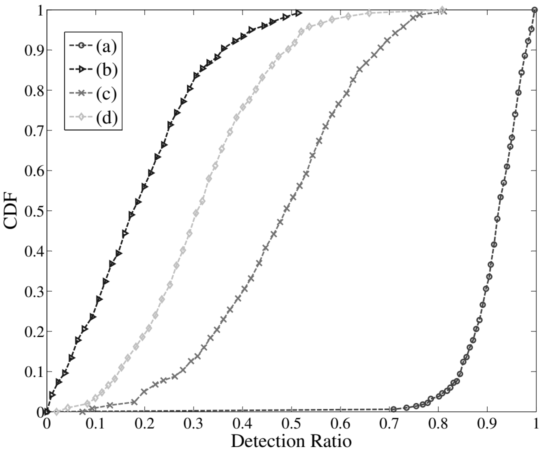


Fig. 4: Confronting the performance of the two detectors. (a) χ^2 detector in [2], [3], and cyber adversary. (b) χ^2 detector and cyber-physical adversary. (c) Multi-watermark detector with switching frequency set to 0.14Hz, and cyber-physical adversary. (d) Multi-watermark detector with switching frequency set to 0.05Hz, and cyber-physical adversary.

V. Conclusion

We have addressed security issues in industrial control systems. We have focused on the adaptation of failure detection mechanisms. The goal is to handle, in addition to faults and errors, the detection of cyber-physical attacks. Cyber-physical attacks refer to malicious activities conducted over industrial control systems with upgraded computing, communication and interconnection capabilities. In other words, they refer to threats against industrial environments that close their loops through networked control systems.

We have revisited a watermark-based attack detection scheme. The approach relies on the adaptation of a failure detector, by adding a complementary authentication watermark signal for the detection of the malicious activities. The approach only requires to inject the watermark from the system controller. The monitored system continues to work regardless of the added watermark signal. This way, the strategy is free from desynchronization. Nevertheless, we have shown that the detection strategy is not sufficiently robust from a security standpoint. Indeed, we have quantitatively shown that the approach only detects *cyber adversaries*, i.e., attackers with the ability to eavesdrop information from the system, but that do not attempt to get any knowledge about the system model itself. We have validated that the detector fails at covering *cyber-physical adversaries*, i.e., attackers that, in addition to the capabilities of the cyber adversary, are also able to infer the system model to evade the detection.

We have then presented a multi-watermark based adaptive detection scheme with two main configurable parameters: number of distributions and switching frequency. The novel multi-watermark proposal succeeds at properly detecting both cyber and cyber-physical adversaries under the assumption that the watermark distributions change frequently. The rationale is that, even under the presence of adversaries with knowledge about the system dynamics, the detector succeeds at reducing their chances of acquiring the authentication watermark and bypass the detector. Numerical simulations validate the detection performance of the new construction.

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] D. Corman, V. Pillitteri, S. Tousley, and U. Tehranipoor, Lindqvist, "NITRD Cyber-Physical Security Panel," 35th IEEE Symposium on Security and Privacy, IEEE SP 2014, San Jose, CA, USA, May 18-21.

[2] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 2009, pp. 911–918.

[3] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *Control Systems, IEEE*, vol. 35, no. 1, pp. 93–109, 2015.

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

[5] 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, pp. 495–500.

[6] 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.

[7] IEC, "Functional safety of electrical/electronic/programmable electronic safety related systems," *IEC 61508*, 2000.

[8] M. Clayton, "Cybersecurity: How US utilities passed up chance to protect their networks," http://www.csmonitor.com/USA/2012/0517/Cybersecurity-How-US-utilities-passed-up-chance-to-protect-their-networks, 2012, [Online; accessed 10-April-2014].

[9] J. Åkerberg and M. Björkman, "Exploring network security in profisafe," in *Computer Safety, Reliability, and Security*. Springer, 2009, pp. 67–80.

[10] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *Control Systems Technology, IEEE Transactions on*, vol. 22, no. 4, pp. 1396–1407, 2014.

[11] 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.

[12] F. Pasqualetti, F. Dorfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. IEEE, 2012, pp. 3418–3425.

[13] A. Arvani and V. S. Rao, "Detection and protection against intrusions on smart grid systems," *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, vol. 3, no. 1, pp. 38–48, 2014.

[14] V. L. Do, L. Fillatre, and I. Nikiforov, "A statistical method for detecting cyber/physical attacks on scada systems," in *Control Applications (CCA), 2014 IEEE Conference on*. IEEE, 2014, pp. 364–369.

[15] G. F. Franklin, J. D. Powell, and M. L. Workman, *Digital control of dynamic systems*. Addison-wesley Menlo Park, 1998, vol. 3.

[16] B. Brumback and M. Srinath, "A chi-square test for fault-detection in kalman filters," *Automatic Control, IEEE Transactions on*, vol. 32, no. 6, pp. 552–554, 1987.

[17] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson Jr, "Stationary and nonstationary learning characteristics of the lms adaptive filter," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1151–1162, 1976.

[18] N. L. Ricker, "Model predictive control of a continuous, nonlinear, two-phase reactor," *Journal of Process Control*, vol. 3, no. 2, pp. 109–123, 1993.

[19] R. Chabukswar, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on scada systems," in *the Proceedings of the 18th IFAC World Congress*, 2011.