Detection and isolation of replay attacks through sensor watermarking

Ferrari, Riccardo M.G.; Herdeiro Teixeira, A.M.

DOI
10.1016/j.ifacol.2017.08.1502

Publication date
2017

Document Version
Final published version

Published in
IFAC-PapersOnLine

Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 50-1 (2017) 7363–7368

Detection and Isolation of Replay Attacks through Sensor Watermarking ⋆

Riccardo M.G. Ferrari * **André M.H. Teixeira** **

* *Delft Center for Systems and Controls,*
** *Faculty of Technology, Policy and Management,*
Delft University of Technology, Delft, The Netherlands
(e-mail: {r.ferrari,andre.teixeira}@tudelft.nl*)*

Abstract: This paper addresses the detection and isolation of replay attacks on sensor measurements. As opposed to previously proposed additive watermarking, we propose a multiplicative watermarking scheme, where each sensor's output is separately watermarked by being fed to a SISO watermark generator. Additionally, a set of equalizing filters is placed at the controller's side, which reconstructs the original output signals from the received watermarked data. We show that the proposed scheme has several advantages over existing approaches: it has no detrimental effects on the closed-loop performance in the absence of attacks; it can be designed in a modular fashion, independently of the design of the controller and anomaly detector; it facilitates the detection of replay attacks and the isolation of the time at which the replayed data was recorded. These properties are discussed in detail and the results are illustrated through a numerical example.

Keywords: Fault detection and diagnosis, Control over networks, Channel equalisation

1. INTRODUCTION

Modern control systems are increasingly relying on information and communication technology (ICT) infrastructures to exchange measurement and control signals. However, the increasing use of pervasive and open-standard ICT systems results in control systems becoming increasingly vulnerable to malicious cyberthreats, which may affect the physical processes through the control loop.

The topic of cyber-secure control systems has been receiving increasing attention recently. An overview of existing cyberthreats and vulnerabilities in networked control systems is presented in Cárdenas et al. (2008); Teixeira et al. (2015). Detectability conditions of stealthy false-data injection attacks to control systems are closely examined in Teixeira et al. (2012), where the authors characterized modifications to the system dynamics that reveal stealthy attacks. Recently, Miao et al. (2014) proposed an static output coding scheme combining the outputs of multiple sensors to reveal stealthy data injection attacks on sensors.

Less studied are attacks of multiplicative nature, such as replay (Mo et al., 2015) and routing attacks (Ferrari and Teixeira, 2017). Within this class of attacks, replay attacks have been more extensively analyzed. In Mo et al. (2015), the analysis of detectability conditions for replay attacks shows that, asymptotically, replay attacks are undetectable. To detect replay attacks, the authors proposed a detection scheme through additive watermarking, where noise is purposely injected in the system by the actuators to watermark the sensor outputs through known correlations. However, such additive watermark presents some drawbacks: the performance of the system decreases and the actuators are further burdened with noisy inputs. These two drawbacks can be tackled by employing multiplicative sensor watermarks, akin to the techniques explored in Teixeira et al. (2012); Miao et al. (2014).

As main contributions of this paper, we study the fundamental limitations in detectability of replay attacks and propose tailored detection and isolation schemes to identify these attacks. In particular, to facilitate their detection and identification, we propose a multiplicative sensor watermarking scheme akin to that in Ferrari and Teixeira (2017), where each sensor output is separately watermarked by being fed to a SISO filter with time-varying piecewise constant parameters. Additionally, an equalization filter is incorporated at the controller's side to reconstruct the original plant outputs and ensure the modularity of the scheme, without the need to redesign the controller, and with no detrimental effects on the closed-loop performance.

The outline of the paper is as follows. In Section 2, we present the problem formulation, describe and analyze the detectability of replay attacks without watermarking, and summarize the main elements of the proposed watermarking scheme. The sensor watermarking scheme is described in detail Section 3, where the performance in the absence of attacks and the detectability of replay attacks with the proposed scheme are also analyzed. To detect replay attacks, an observer-based detection scheme with robust adaptive threshold is proposed in Section 4. Numerical results are presented in Section 5, and the paper concludes with final remarks in Section 6.

⋆ This work has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 608224 and from H2020 Programme under grant no. 707546 (SURE).

10.1016/j.ifacol.2017.08.1502

2. PROBLEM FORMULATION

In this section, we present the control system and describe the main problem at hand. Consider the modeling framework described in Teixeira et al. (2015), where the control system is composed by a physical plant ($\mathcal{P}$), a feedback controller ($\mathcal{C}$), and an anomaly detector ($\mathcal{R}$), which are modeled in a discrete-time state-space form as

$$\begin{aligned}\mathcal{P}:&\begin{cases}x_p[k+1]=A_p x_p[k]+B_p u[k]+\eta[k]\\ y_p[k]=C_p x_p[k]+\xi[k]\end{cases}\\ \mathcal{C}:&\begin{cases}x_c[k+1]=A_c x_c[k]+B_c \tilde{y}_p[k]\\ u[k]=C_c x_c[k]+D_c \tilde{y}_p[k]\end{cases}\\ \mathcal{R}:&\begin{cases}x_r[k+1]=A_r x_r[k]+B_r u[k]+K_r \tilde{y}_p[k]\\ y_r[k]=C_r x_r[k]+D_r u[k]+E_r \tilde{y}_p[k]\end{cases}\end{aligned} \tag{1}$$

where $x_p[k]\in\mathbb{R}^{n_p}$, $x_c[k]\in\mathbb{R}^{n_c}$ and $x_r[k]\in\mathbb{R}^{n_r}$ are the state variables, $u[k]\in\mathbb{R}^{n_u}$ is the vector of control actions applied to the process, $y_p[k]\in\mathbb{R}^{n_y}$ is the vector of plant outputs transmitted by the sensors, $\tilde{y}_p\in\mathbb{R}^{n_y}$ is the data received by the detector and controller, and $y_r[k]\in\mathbb{R}^{n_y}$ the residual vector that is used for detecting anomalies. The variables $\eta[k]$ and $\xi[k]$ denote the unknown process and measurement disturbances, respectively.

Assumption 1. The uncertainties represented by η and ξ are unknown, but their norms are upper bounded by some known and bounded sequences $\bar{\eta}[k]$ and $\bar{\xi}[k]$.

Since the sensor measurements, exchanged through a communication network, may have been subject to cyber-attacks, at the plant side, we denote the data transmitted by the sensors as $y_p[k]\in\mathbb{R}^{n_y}$ whereas, at the detector's side, the received sensor data is denoted as $\tilde{y}_p[k]\in\mathbb{R}^{n_y}$.

The anomaly detector is collocated with the controller and it evaluates the behavior of the plant based only on the closed-loop models and the available input and output data $u[k]$ and $\tilde{y}_p[k]$. In particular, given the residue signal y_r, an alarm is triggered to indicate the presence of anomalies if $|y_{r,(i)}[k]|\geq\bar{y}_{r,(i)}[k]$, for at least one time instant k and one component $i\in\{1,\dots,n_y\}$, where $\bar{y}_r\in\mathbb{R}^{n_y}_+$ is a robust detection residual.

Defining $x_{cr}[k]=[x_c[k]^\top\ \ x_r[k]^\top]^\top$, the controller and detector dynamics can be written as

$$\mathcal{F}_{cr}:\begin{cases}x_{cr}[k+1]=A_{cr}x_{cr}[k]+B_{cr}\tilde{y}_p[k]\\ y_r[k]=C_{cr}x_{cr}[k]+D_{cr}\tilde{y}_p[k]\\ u[k]=C_u x_{cr}[k]+D_u\tilde{y}_p[k]\,,\end{cases} \tag{2}$$

where A_{cr}, B_{cr}, C_{cr}, D_{cr}, C_u, and D_u are derived from (1).

The main focus of this paper is to investigate the detection and isolation of cyber replay attacks. This attack scenario, as well a fundamental limitation in their detectability akin to the results of Mo et al. (2015), are described next.

2.1 Replay attack scenario

The replay attack scenario considered in this work is summarized in Figure 1. In this scenario, the adversary first records the measurement signals transmitted by all the sensors starting at time $k_r=k_0-T$, after which the adversary replay the recorded signals starting at time k_0. Denoting the delayed variables with a prime, such as in $x'[k]\triangleq x[k-T]$ for $k\geq k_0$, the sensor measurements under replay attack are given by $\tilde{y}_p[k]=y'_p[k]$, $\forall k\geq k_0$.

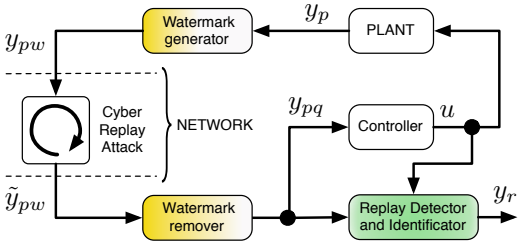

Fig. 1. A block-diagram representation of the setting considered in the present paper, with novel contributions shaded in color.

2.2 Detectability of replay attacks

To analyze the detectability of replay attacks, consider the residual signal at the time in which the measurements were recorded, y'_r, which is described by

$$\begin{aligned}x'_{cr}[k+1]&=A_{cr}x'_{cr}[k]+B_{cr}y'_p[k]\\ y'_r[k]&=C_{cr}x'_{cr}[k]+D_{cr}y'_p[k].\end{aligned} \tag{3}$$

As an inherent limitation in detectability of replay attacks by LTI detectors, the following result provides necessary and sufficient conditions for which, under attack, the residual signal y_r converges to y'_r.

Theorem 1. Suppose that the anomaly detector is an LTI system and assume that $x_{cr}[k_r]\neq x_{cr}[k_0]$. Under a replay attack, the residual signal $y_r[k]$ converges asymptotically to $y'_r[k]$ for arbitrary $x_{cr}[k_0]$ if, and only if, the unstable modes of A_{cr} are unobservable with respect to C_{cr}.

Proof. First, by introducing the notation $\Delta x=x-x'$ and by using (2) with $\tilde{y}_p[k]=y'_p[k]$ and (3), we rewrite the residual as $y_r[k]=y'_r[k]+\Delta y_r[k]$, where $\Delta y_r[k]$ is described by $\Delta y_r[k]=C_{cr}A_{cr}^{k-k_0}\Delta x_{cr}[k_0]$, for $k\geq k_0$. The remainder of the proof directly follows from the PBH observability test (Zhou et al., 1996). ■

If there were no anomalies when $y'_p[k]$ was recorded, an evaluation of y'_r either does not trigger any alarm, or it triggers a false alarm in which no anomaly is present. Therefore, to relate Th. 1 to the undetectability of replay attacks on LTI systems, we make the following assumption.

Assumption 2. The residual y'_r does not trigger any alarm.

2.3 Watermarking and equalization scheme

To allow the presence of replay attacks to be detected, we propose to turn the closed-loop system (1) into a switched system parametrized by a controlled variable $\theta[k]$. Specifically, $\theta[k]$ is defined as a piecewise constant variable $\theta[k]\triangleq\theta_j\in\Theta$, for $k_j\leq k<k_{j+1}$, where $\mathcal{K}_\theta\triangleq\{k_1,\dots,k_j,\dots\}$ denotes the set of switching times and $\Theta\triangleq\{\theta_1,\dots,\theta_M\}$ is the set of possible parameters. Furthermore, we assume that the parameter $\theta[k]$ is only known by the sensors and the anomaly detector and controller. For simplicity of notation, the time argument of $\theta[k]$ is omitted when possible.

In the proposed scheme, we thus introduce a pre-processing step, which we denote as *sensor watermarking*,

where each sensor processes the measurements through a filter parametrized by θ before transmitting the data. Denoting $\mathcal{W}(\theta)$ as the set of watermarking filters, the watermarked sensor outputs to be transmitted, denoted as $y_{pw}[k]$, are described by

$$\mathcal{W}(\theta): \begin{cases} x_w[k+1] = A_w(\theta)x_w[k] + B_w(\theta)y_p[k] \\ y_{pw}[k] = C_w(\theta)x_w[k] + D_w(\theta)y_p[k]. \end{cases} \quad (4)$$

As argued earlier, due the the presence of cyber-attacks and other anomalies, the watermarked data transmitted by the sensors ($y_{pw}[k]$) may differ from the data received at the controller's side ($\tilde{y}_{pw}[k]$). The controller and anomaly detector also apply a pre-processing step, denoted as *equalization*, where the received watermarked data $\tilde{y}_{pw}[k]$ is processed through an equalizing filter parametrized by $\theta[k]$. The objective of the equalization step is to remove the watermark from the received data, $\tilde{y}_{pw}[k]$, thus producing the reconstructed plant outputs $y_{pq}[k]$. As illustrated in Fig. 1, the reconstructed measurements $y_{pq}[k]$ are fed to the anomaly detector and controller.

Denoting $\mathcal{Q}(\theta)$ as the equalizer, the residual and control input are computed from the received data $\tilde{y}_{pw}[k]$ as

$$\begin{aligned} \mathcal{Q}(\theta): &\begin{cases} x_q[k+1] = A_q(\theta)x_q[k] + B_q(\theta)\tilde{y}_{pw}[k] \\ y_{pq}[k] = C_q(\theta)x_q[k] + D_q(\theta)\tilde{y}_{pw}[k], \end{cases} \\ \mathcal{F}_{cr}: &\begin{cases} x_{cr}[k+1] = A_{cr}x_{cr}[k] + B_{cr}y_{pq}[k] \\ y_r[k] = C_{cr}x_{cr}[k] + D_{cr}y_{pq}[k] \\ u[k] = C_u x_{cr}[k] + D_u y_{pq}[k]. \end{cases} \end{aligned} \quad (5)$$

Furthermore, the parameter $\theta[k]$ is changed frequently as to limit the time in which a replay attack may remain undetected, as explained in Sect. 3 and 4. Under a replay attack, the replayed watermarked data is described by

$$\mathcal{W}(\theta'): \begin{cases} x'_w[k+1] = A_w(\theta')x'_w[k] + B_w(\theta')y'_p[k] \\ y'_{pw}[k] = C_w(\theta')x'_w[k] + D_w(\theta')y'_p[k], \end{cases} \quad (6)$$

where the watermarking filter at attack recording time was parametrized by $\theta' = \theta[k-T]$.

To illustrate the reasoning behind the proposed scheme, in the following we describe the scheme in the frequency domain. Denote the nominal transfer function of the residual, without the additional pre-processing filter, as $y_r(z) = F_{cr}(z)y_p(z)$. Let $W(z;\theta)$ and $Q(z;\theta)$ be the transfer function of the pre-processing filters at the sensors and anomaly detector, respectively, which are parametrized by θ. Furthermore, suppose the filters are designed such that

$$Q(z;\theta)W(z;\theta') = I + D(z;\theta,\theta')$$

where, ideally, $D(z;\theta,\theta) = 0$ and $D(z;\theta,\theta')$ is 'large' for $\theta \neq \theta'$. The transfer function of the anomaly detector is given by $y_r(z) = F_{cr}(z)Q(z;\theta)\tilde{y}_{pw}(z)$, whereas the pre-processed measurements are given by $y_{pw}(z) = W(z;\theta)y_p(z)$. Under nominal conditions, i.e. $\tilde{y}_{pw} = y_{pw}$, the residual is then given by

$$y_r(z) = F_{cr}(z)y_p(z) + F_{cr}(z)D(z;\theta,\theta)y_p(z).$$

On the contrary, under a replay attack where $\tilde{y}_{pw}(z) = y'_{pw}(z) = W(z;\theta')y'_p(z)$, we have

$$y_r(z) = F_{cr}(z)y'_p(z) + F_{cr}(z)D(z;\theta,\theta')y'_p(z).$$

Note that, since the filters $W(z;\theta)$ and $Q(z;\theta)$ are designed to ensure $D(z;\theta,\theta) = 0$, the transfer function of the residual (and also the control signal) in the absence of attacks is decoupled from the pre-processing filters. Hence, the anomaly detector and controller can be designed in a modular fashion, independently from the pre-processing filters. To detect replay attacks, a robust threshold is designed so that, in the absence of attacks, the residual evaluation is robust to the unknown disturbances η and ξ. As the term $F_{cr}(z)D(z;\theta,\theta)y_p(z)$ is 0 by design, the modularity also extends itself to the threshold design, which needs not to consider the pre-processing filters.

On the other hand, when a replay attack occurs, the residual will be driven by $F_{cr}(z)D(z;\theta,\theta')y'_p(z)$, which is large by design and, therefore, easily detectable.

3. SENSOR WATERMARKING

Let the watermark generator at each sensor be implemented through an infinite impulse response (IIR) filter of order N. For the ith measurement, the watermark generator is described by the difference equation

$$y_{pw,(i)}[k] = \sum_{n=1}^{N} w^i_{A,(n)} y_{pw,(i)}[k-n] + \sum_{n=0}^{N} w^i_{B,(n)} y_{p,(i)}[k-n], \quad (7)$$

where $w^i_A = [w^i_{A,(1)} \ \dots \ w^i_{A,(N)}]^\top \in \mathbb{R}^N$ and $w^i_B = [w^i_{B,(0)} \ \dots \ w^i_{B,(N)}]^\top \in \mathbb{R}^{N+1}$ are the filter parameters.

Regarding the equalizing filters at the detector's side, their aim is to compute $y_{pq}[k]$, which is a reconstruction of the signals $y_p[k]$ given the received watermarked measurements $\tilde{y}_{pw}[k]$. A simple approach would be to consider the equalizing filter of the ith measurement as the inverse of the respective watermark filter, namely

$$\begin{aligned} y_{pq,(i)}[k] = &\frac{1}{w^i_{B,(0)}} \left(\sum_{n=1}^{N} -w^i_{B,(n)} y_{pq,(i)}[k-n] + \right. \\ &\left. \tilde{y}_{pw,(i)}[k] + \sum_{n=1}^{N} -w^i_{A,(n)} \tilde{y}_{pw,(i)}[k-n] \right). \end{aligned} \quad (8)$$

For notation simplicity and without loss of generality, we suppose that there is only one sensor, i.e., $n_y = 1$, and therefore omit the superscript in the parameters and use the notation $w_A = w^i_A$ and $w_B = w^i_B$. Recall that choosing $w_A = 0$ retrieves a finite impulse response (FIR) filter.

In relation to the replay attack detection scheme proposed in the previous section, each admissible value of the piecewise constant variable $\theta[k]$ is denoted as a particular choice of filter parameters, e.g., $\theta_j = \{w_{A,j}, w_{B,j}\}$.

The watermarking filter dynamics (7) can be written as $\mathcal{W}(\theta)$ in (4), by using the controllable canonical form, where $x_w[k] \in \mathbb{R}^N$ and the matrices are given by

$$A_w(\theta) = \begin{bmatrix} 0_{N-1,1} & I_{N-1} \\ & w_A^\top \end{bmatrix}, \ B_w = \begin{bmatrix} 0_{N-1,1} \\ 1 \end{bmatrix},$$

$C_w(\theta) = \begin{bmatrix} \dots & w_{B,(n)} + w_{B,(0)}w_{A,(n)} & \dots \end{bmatrix}$, for $n = 1, \dots, N$, and $D_w(\theta) = w_{B,(0)}$, where I_N is the identity matrix of order N and $0_{N,M} \in \mathbb{R}^{N\times M}$ is a null matrix.

Similarly, by using the controllable canonical form and the coordinate transformation matrix $T = w_{B,(0)}I$, the equalizer dynamics (8) can be written as $\mathcal{Q}(\theta)$ in (5), where $x_q[k] \in \mathbb{R}^N$ and the matrices are given by $D_q(\theta) = \dfrac{1}{w_{B,(0)}}$,

$$A_q(\theta) = \begin{bmatrix} 0_{N-1,1} & I_{N-1} \\ \dfrac{-1}{w_{B,(0)}} w_B^\top \end{bmatrix}, \; B_q = \begin{bmatrix} 0_{N-1,1} \\ \dfrac{1}{w_{B,(0)}} \end{bmatrix},$$

$$C_q(\theta) = \begin{bmatrix} \cdots & -w_{A,(n)} - \dfrac{w_{B,(n)}}{w_{B,(0)}} & \cdots \end{bmatrix}, \text{ for } n = 1, \ldots, N.$$

In the remainder of the paper, we follow the aforementioned scheme and design the filters so that they are stable.

Assumption 3. The watermarking filter $\mathcal{W}(\theta)$ and its inverse $\mathcal{Q}(\theta)$ are stable for all $\theta \in \Theta$. □

Note that the latter assumption holds when the watermark generator $\mathcal{W}$ is designed as a FIR filter of order N (with N poles at the origin) that has exactly N zeros, all inside the unit circle, which in turn leads to the following assumption.

Assumption 4. The watermarking filter $\mathcal{W}(\theta)$ is an FIR filter with $w_A = 0$ and $w_B = \theta$ for all $\theta \in \Theta$. □

Next, considering the closed-loop system with the proposed watermarking and equalizing filters, we discuss the closed-loop performance in the absence of attacks, followed by an analysis of the detectability of replay attacks. The core element of both discussions is the cascade of the watermarking filter $\mathcal{W}(\theta_2)$ and equalizing filter $\mathcal{Q}(\theta_1)$, which we denote as $\mathcal{QW}(\theta_1, \theta_2)$. By defining $x_{qw} \triangleq \left[x_w^\top,\ x_q^\top\right]^\top$, the cascade system $\mathcal{QW}(\theta_1, \theta_2)$ is described by

$$\begin{aligned} x_{qw}[k+1] &= A_{qw}(\theta_1, \theta_2) x_{qw}[k] + B_{qw}(\theta_1, \theta_2) y_p[k] \\ y_{pq}[k] &= C_{qw}(\theta_1, \theta_2) x_{qw}[k] + D_{qw}(\theta_1, \theta_2) y_p[k], \end{aligned} \tag{9}$$

where

$$\begin{aligned} D_{qw}(\theta_1, \theta_2) &= D_q(\theta_2) D_w(\theta_1), \\ A_{qw}(\theta_1, \theta_2) &= \begin{bmatrix} A_w(\theta_2) & 0 \\ B_q(\theta_1) C_w(\theta_2) & A_q(\theta_1) \end{bmatrix}, \\ B_{qw}(\theta_1, \theta_2) &= \begin{bmatrix} B_w(\theta_2) \\ B_q(\theta_1) D_w(\theta_2) \end{bmatrix}, \end{aligned}$$

and $C_{qw}(\theta_1, \theta_2) = [D_q(\theta_1) C_w(\theta_2) \;\; C_q(\theta_1)]$.

3.1 Performance in the absence of replay attacks

Although the main aim of the proposed scheme is to detect replay attacks, it is important that it does not decrease the nominal performance in the absence of attacks. To analyze the impact of the watermarking scheme in the absence of attacks, we compare the nominal system (1) and the watermarked system without attacks described by (4) and (5), where $\mathcal{W}(\theta)$ and $\mathcal{Q}(\theta)$ are matched w.r.t. θ. As a first step, we have the following result.

Lemma 1. Consider the pair of filters $\mathcal{W}(\theta)$ and $\mathcal{Q}(\theta)$, where $\mathcal{Q}(\theta)$ is the stable inverse of the FIR filter $\mathcal{W}(\theta)$. The output of the cascade $\mathcal{QW}(\theta, \theta)$ is given by

$$y_{pq}[k] = y_p[k] - C_q(\theta) A_q(\theta)^{k-k_0} (x_w[k_0] - x_q[k_0]). \tag{10}$$

Furthermore, $y_{pq}[k]$ converges asymptotically to $y_p[k]$.

Proof. The proof follows from (9) for $w_{A,1} = w_{A,2} = 0$ and $w_{B,1} = w_{B,2} = w_B$. Using the transformation $\bar{x}_{qw} = T x_{qw}$ such that $\bar{x}_{qw} = \left[x_w^\top \;\; (x_w - x_q)^\top\right]^\top$, we obtain

$$\bar{A}_{qw}(\theta) = \begin{bmatrix} A_w(\theta) & 0 \\ 0 & A_q(\theta) \end{bmatrix}, \; \bar{B}_{qw}(\theta) = \begin{bmatrix} B_w(\theta) \\ 0 \end{bmatrix}, \; \bar{D}_{qw}(\theta) = 1,$$

and $\bar{C}_{qw}(\theta) = [0 \;\; -C_q(\theta)]$.Observing that, for all $k \geq 0$, $\bar{C}_{qw}(\theta) \bar{A}_{qw}(\theta)^k \bar{B}_{qw}(\theta) = 0$ and $\bar{C}_{qw}(\theta) \bar{A}_{qw}(\theta) \bar{x}_{qw}[k] = -C_q(\theta) A_q(\theta)(x_w[k] - x_q[k])$, the output of $\mathcal{QW}(\theta, \theta)$ can be written as $y_{pq}[k] = y_p[k] - C_q(\theta) A_q(\theta)^{k-k_0} (x_w[k_0] - x_q[k_0])$. Recalling that $A_q(\theta)$ is Schur concludes the proof. ■

Next, we analyze the performance of the closed-loop system with the proposed scheme in the absence of attacks.

Theorem 2. Consider the closed-loop system with watermarked sensors described by (4) and (5). Furthermore, suppose that $\theta[k]$ is updated at times $k \in \mathcal{K}_\theta$. In the absence of replay attacks (i.e., $\tilde{y}_{pw} = y_{pw}$ and $\tilde{y}_p = y_p$), the performance of the closed-loop system with the matched filters $\mathcal{Q}(\theta)$ and $\mathcal{W}(\theta)$ is the same as the performance of the nominal closed-loop system (1) if, and only if, the states of $\mathcal{Q}(\theta)$ and $\mathcal{W}(\theta)$ are such that $x_q[k] = x_w[k]$ for all $k \in \mathcal{K}_\theta$.

Proof. Lemma 1 states that $y_{pq}[k] = y_p[k]$ if, and only if, $x_q[k_\theta] = x_w[k_\theta]$ for all $k_\theta \in \mathcal{K}_\theta$, which implies that (1) and the closed-loop system described by (4) and (5) have identical state trajectories. ■

Remark 1. By imposing that, at switching times, the watermarking and equalizing filters set their states to 0, our proposed scheme does not reduce the performance of the nominal system in the absence of attacks, thus ensuring the modularity of the scheme by decoupling the design of the controller and detector from that of the filters, as opposed to the scheme proposed in Mo et al. (2015).

3.2 Detectability of replay attacks with sensor watermarking

As the main step to analyze the detectability of replay attacks under the proposed watermarking scheme, we derive the following result.

Lemma 2. Consider the pair of filters $\mathcal{W}(\theta_2)$ and $\mathcal{Q}(\theta_1)$, where $\mathcal{Q}(\theta_1)$ is the stable inverse of the FIR filter $\mathcal{W}(\theta_1)$. For $\theta_1 \neq \theta_2$, the cascade $\mathcal{QW}(\theta_1, \theta_2)$ has a minimal realization of order N, which has the same poles as $\mathcal{Q}(\theta_1)$.

Lemma 2 implies that the cascade system $\mathcal{QW}(\theta_1, \theta_2)$ has a non-trivial transfer function (i.e., different from 1) for $\theta_1 \neq \theta_2$. Furthermore, the cascade $\mathcal{QW}(\theta_1, \theta_2)$ can be written as $\mathcal{QW}(\theta_1, \theta_2) = I + \mathcal{D}(\theta_1, \theta_2)$, where $\mathcal{D}(\theta_1, \theta_2) = (A_{qw}(\theta_1), B_{qw}, C_{qw}(\theta_1, \theta_2), D_{qw}(\theta_1, \theta_2) - 1)$ is the system describing the signal $\Delta y_p[k] \triangleq y_{pq}[k] - y_p[k]$ with y_p as input. Thus, we have the following intermediate result.

Lemma 3. Consider a replay attack that has recorded measurement data $y_{pw}[k]$ from time $k_r = k_0 - T$ to $k_f = k_0 - T_f$, and let $\theta[k] = \theta'$ for $k_r \leq k \leq k_f$. Suppose the recorded data $y'_{pw}[k]$ is replayed as $\tilde{y}_{pw}[k]$ from time k_0, let $\theta[k] = \theta$ for $k \geq k_0$, and redefine $\Delta y_p[k] \triangleq y_{pq}[k] - y'_p[k]$. During the replay attack, the residual output y_r is driven by the replayed data y'_p as described by

$$\begin{aligned} x_{qw}[k+1] &= A_{qw}(\theta, \theta') x_{qw}[k] + B_{qw}(\theta, \theta') y'_p[k] \\ \Delta y_p[k] &= C_{qw}(\theta, \theta') x_{qw}[k] + (D_{qw}(\theta, \theta') - 1) y'_p[k], \\ x_{cr}[k+1] &= A_{cr} x_{cr}[k] + B_{cr} y'_p[k] + B_{cr} \Delta y_p[k] \\ y_r[k] &= C_{cr} x_{cr}[k] + D_{cr} y'_p[k] + D_{cr} \Delta y_p[k], \end{aligned} \tag{11}$$

where $x_{qw}[k] = \left[x'_w[k]^\top \; x_q[k]^\top\right]^\top$.

Proof. The proof follows directly from (6) and (5). ■

We now present the main result of this section regarding the detectability of replay attacks under the proposed watermarking scheme.

Theorem 3. Consider a replay attack that has recorded data from time $k_r = k_0 - T$ to $k_f = k_0 - T_f$, and let $\theta[k] = \theta'$ for $k_r \leq k \leq k_f$. Suppose the recorded data is replayed from time k_0 and let $\theta[k] = \theta$ for $k \geq k_0$. During the replay attack, y_r converges asymptotically to y'_r for any y'_p if and only if $\theta = \theta'$.

Proof. The main step of the proof is to use the notation $\Delta x[k] = x[k] - x'[k]$ and Lemma 3 to conclude that the residual y_r can be rewritten as

$$\begin{aligned}\Delta x_{cr}[k+1] &= A_{cr}\Delta x_{cr}[k] + B_{cr}\Delta y_p[k]\\ y_r[k] &= y'_r[k] + C_{cr}\Delta x_{cr}[k] + D_{cr}\Delta y_p[k],\end{aligned}$$

where $\Delta y_p[k]$ is the output of $\mathcal{D}(\theta, \theta')$ as given by (11). Sufficiency readily follows from Lemma 1, which states that $\mathcal{D}(\theta, \theta')$ is an autonomous system for $\theta = \theta'$. Thus, we conclude that the effect of $\Delta y_p[k]$ decays asymptotically to zero regardless of y'_p and, from Th. 1, we have that y_r converges asymptotically to y'_r. Regarding the necessity, suppose that $\theta \neq \theta'$. Then, as per Lemma 2, $\mathcal{D}(\theta, \theta')$ is a forced system whose output $\Delta y_p[k]$ depends non-trivially on y'_p. Therefore, there exists a signal y'_p for which the effect of $\Delta y_p[k]$ on $y_r[k]$ does not decay to zero. ■

Th. 3 indicates that, when $\theta \neq \theta'$, the undetectability of the replay attack is not guaranteed *a priori*, since it depends on the exogenous input y'_p. Next, we design an anomaly detector and a robust threshold to evaluate the residual and detect replay attacks.

4. DETECTION AND ISOLATION OF REPLAY ATTACKS

We now leverage the proposed watermarking scheme to first detect replay attacks by means of an observer and a robust threshold, and then isolate the recording time of replayed data, by identifying its watermark parameter θ'.

Assumption 5. No replay attacks are present for $0 \leq k < k_0$, with k_0 being the attack start time. Moreover, the variables x_p, x_{pw} and u remain bounded before the occurrence of an attack, i.e., there exist some stability regions $\mathscr{S} = \mathscr{S}^{x_p} \times \mathscr{S}^{x_{pw}} \times \mathscr{S}^{u} \subset \mathbb{R}^{n_p} \times \mathbb{R}^{n_{pw}} \times \mathbb{R}^{m}$, such that $(x_p, x_{pw}, u) \in \mathscr{S}, \forall\, k \leq k_0$. □

Assumption 6. (A_p, C_p) is a detectable pair. □

4.1 Detection of Replay Attacks

The detector $\mathcal{R}$ in (1) will be implemented as the following observer (Ferrari et al., 2008),

$$\hat{\mathcal{P}} : \begin{cases} \hat{x}_p[k+1] = A_p\hat{x}_p[k] + B_p u[k] + K\left(y_{pq}[k] - \hat{y}_p[k]\right)\\ \hat{y}_p[k] = C_p\hat{x}_p[k], \end{cases} \tag{12}$$

where $\hat{x}_p$ and $\hat{y}_p$ of suitable size are dynamic estimates of x_p and y_p and the output error gain matrix K is chosen such that $A_r \triangleq A_p - KC_p$ is Schur. By defining the output residual as $y_r \triangleq y_{pq} - \hat{y}_p$, this corresponds to choosing $x_r = \hat{x}_p$, $A_r = A_p - KC_p$, $B_r = B_p$, $K_r = K$, $C_r = -C_p$, $D_r = 0$, $E_r = I_{n_y}$ in the definition of $\mathcal{R}$ in (1), and feeding it the reconstructed output y_{pq}. In the absence of attacks and assuming the watermarking filter and equalizer are initialized according to Remark 1, the estimation errors $\epsilon \triangleq x_p - \hat{x}_p$ dynamics follows from (1), (10) and (12)

$$\begin{cases} \epsilon[k+1] = A_r\epsilon[k] - K\xi[k] + \eta[k]\\ y_r[k] = C_p\epsilon[k] + \xi[k] \end{cases},$$

leading to the following solution for the output residual

$$y_r[k] = C_p\left[\sum_{h=0}^{k-1}(A_r)^{k-1-h}\left(\eta[h] - K\xi[h]\right) + (A_r)^k\epsilon[0]\right] + \xi[k]$$

For attack detection, the following threshold shall be used

$$\begin{aligned}\bar{y}_{r,(i)}[k] \triangleq \alpha^i\Bigg[\sum_{h=0}^{k-1}\left(\delta^i\right)^{k-1-h}\big(\bar{\eta}[h] +\\ \|K\|\bar{\xi}[h]\big) + \left(\delta^i\right)^k\bar{x}_r[0]\Bigg] + \bar{\xi}[k]\end{aligned} \tag{13}$$

where α^i and δ^i are two constants such that $\|C_{p,(i)}(A_r)^k\| \leq \alpha^i\left(\delta^i\right)^k \leq \|C_{p,(i)}\|\cdot\|\left(A_r\right)^k\|$ with $C_{p,(i)}$ being the i–th row of matrix C_p (see (Ferrari et al., 2008) and (Dowler, 2013, Th. 3.5)). Furthermore, $\bar{\eta}$, $\bar{x}_r[0]$ and $\bar{\xi}$ are upper bounds on the norms of, respectively, η, $x_r[0]$ and ξ, which can be computed via Assumption 1 and 5.

Theorem 4. (Attack Detectability). If there exists a time index $k_d > k_0$ and a component $i \in \{1, \ldots, n_y\}$ such that during a cyber replay attack the following inequality holds

$$\begin{aligned}&\left|C_{p,(i)}\left[\sum_{h=k_0}^{k_d-1}(A_r)^{k_d-1-h}\left(B_p\Delta u[h] - K\Delta y_p[h]\right)\right] + \Delta y_p[k]\right|\\ &\quad> 2\alpha^i\sum_{h=0}^{k_d-1}\left(\delta^i\right)^{k_d-1-h}\left(\bar{\eta}[h] + \|K\|\bar{\xi}[h]\right) +\\ &\qquad\left(\delta^i\right)^{k_d-k_0}\left(\alpha^i\bar{x}_r[k_0] + \bar{y}_{r,(i)}[k_0]\right) + 2\bar{\xi}[k_d]\end{aligned}$$

where $\bar{y}_{r,(i)}[k_0] \triangleq \max_{x_p \in \mathscr{S}^{x_p}} |y_{r,(i)}[k_0]|$ and $\Delta u \triangleq u' - u$ is the difference between delayed and actual input, then the attack will be detected at the time instant k_d.

Proof. During a replay attack, $\mathcal{Q}$, $\mathcal{R}$ and $\mathcal{C}$ are disconnected from $\mathcal{P}$, and instead are fed $y'_p[k]$. By redefining ϵ as $\epsilon \triangleq x'_p - \hat{x}_p$ and remembering Lemma 3, the solution for y_r can be computed by subtracting (12) from (6):

$$\begin{aligned}y_r[k] = C_p\Bigg[\sum_{h=k_0}^{k-1}(A_r)^{k-1-h}\big(B_p\Delta u[h] - K(\xi'[h] + \Delta y_p[h])\\ +\eta'[h]\big) + (A_r)^k\epsilon[k_0]\Bigg] + \Delta y_p[k] + \xi'[k]\,.\end{aligned}$$

The proof follows from Ferrari et al. (2008, Th. 3.1). ■

4.2 Isolation and Identification of Replay Attacks

Only after a successful detection, a bank of $|\Theta| = M$ filters is activated in order to isolate the replay attack, and to identify the replayed data parameter θ' which provides information as to when the data was recorded.

Each filter j is parametrized by $\theta_j \in \Theta$ and designed as

$$\mathcal{Q}(\theta_j) : \begin{cases} x_{q,j}[k+1] = A_q(\theta_j)x_{q,j}[k] + B_q(\theta_j)\tilde{y}_{pw}[k]\\ y_{pq,j}[k] = C_q(\theta_j)x_{q,j}[k] + D_q(\theta_j)\tilde{y}_{pw}[k], \end{cases}$$

$$\mathcal{F}_{cr,j} : \begin{cases} x_{cr,j}[k+1] = A_{cr}x_{cr,j}[k] + B_{cr}y_{pq,j}[k]\\ y_{r,j}[k] = C_{cr}x_{cr,j}[k] + D_{cr}y_{pq,j}[k]. \end{cases}$$

The isolation and identification logic relies on Th. 3: only for the filter j with $\theta_j = \theta'$, $y_{r,j}[k]$ will converge

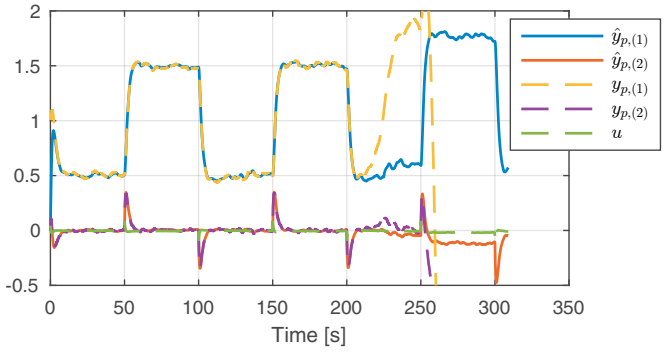

Fig. 2. Estimated true plant outputs produced by the detector (solid lines), and true plant outputs and input (dashed lines).

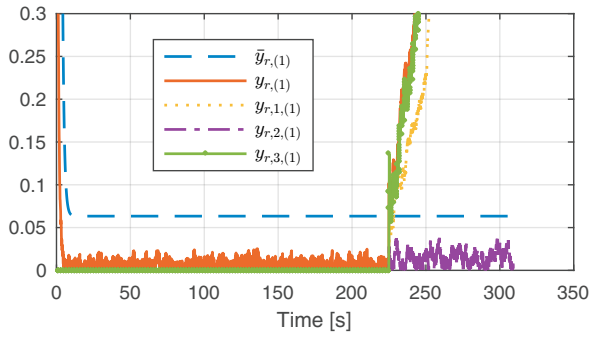

Fig. 3. First components of detection and isolation the residuals and thresholds. Between 210 and 225s no detection is possible as watermark parameters are still holding the value they had during attack recording.

asymptotically to $y'_r[k]$. Therefore, under Assumption 2, the replay attack is said to be isolated at a time $k_j^* > k_0$, which means that $\theta'[k_j^*] = \theta_j$, if the following rules are satisfied, with $j \in \{1, \dots, M\}$ and $i \in \{1, \dots, n_y\}$:

$$\forall k, k_j^* \geq k > k_0 \colon \max_i \left\{ |y_{r,j,(i)}[k]| - \bar{y}_{(i)}[k] \right\} \leq 0,$$

$$\forall l \neq j, \exists k_l^*, k_j^* \geq k_l^* > k_0 \colon \max_i \left\{ |y_{r,l,(i)}[k_l^*]| - \bar{y}_{(i)}[k_l^*] \right\} > 0$$

where $\bar{y}_{r,(i)}[k]$ is the threshold in (13). Furthermore, the time at which the data was recorded, k_r, can be isolated as belonging to the time-interval $[k_j,\ k_{j+1})$, where we recall that $k_j \in \mathcal{K}_\theta$ is such that $\theta[k] = \theta_j$ for $k_j \leq k < k_{j+1}$.

5. NUMERICAL EXAMPLE

As a numerical example, we consider $\mathcal{P}$ to be an unstable discrete time LTI system with $n_p = 2$, $n_u = 1$, $n_y = 2$

$$A_p = \begin{bmatrix} 1 & 0.1 \\ 0.035 & 0.99 \end{bmatrix}, \; B_p = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; C_p = I_2,$$

with I_2 being the 2×2 identity matrix, and $T_s = 0.1$ s the time step. The controller $\mathcal{C}$ is defined by $A_c = I_2$, $B_c = 0.1 \cdot I_2$, $C_c = [\,0.01 \;\; 0.022\,]$, $D_c = [\,0.0875 \;\; 0.1980\,]$ and is fed the error term $e \triangleq r - y_{pq}$, with $r_{(1)}$ a square wave reference varying between 0.5 and 1.5 with a period of 100 s, while $r_{(2)}$ is a null one. Finally, the model and measurement uncertainties are two pairs of random variables uniformly distributed in the intervals $[-0.003\ 0.003]$ and $[-0.006\ 0.006]$, respectively. The coefficients of $\mathcal{W}$ are generated as $w_B^\top = [1,\, 0,\, 0,\, 0] + \omega$, where ω is a random variable uniformly distributed in the box $[-0.1\ 0.1]^4$ and updated at time instants $\mathcal{T}_\theta = \{0,\, 100,\, 225\}$ s. At time $T_0 = 210$ s a replay attack starts, using data recorded from time $T_r = 110$ s onwards.

The attack effects on the estimated plant output $\hat{y}_p$ and the true one y_p are visible in Fig. 2: due to the attack bringing the plant in open loop, y_p quickly diverges, with only a minor deviation on $\hat{y}_p$. Finally, the analysis of the residual y_r and its threshold $\bar{y}_r$ (Fig. 3) shows that the attack is detected and isolated right after $k_3 T_s = 225$ s, when new watermark parameters are generated.

6. CONCLUSIONS

In this work, we proposed a multiplicative sensor watermarking scheme, where each output is separately fed to a SISO watermark generator. As opposed to previous additive watermarking schemes, no additional burden is put on actuators and the closed-loop performance is preserved, thanks to the inclusion of a watermark removing functionality. Analytical results, including attack detectability conditions for the proposed scheme, were derived. Finally, the effectiveness of the proposed approach was illustrated through a numerical study, where a replay attack was detected and identified even during steady-state. Future works will be directed on studying the resilience of the watermarking scheme itself, and on developing nonlinear watermark generators.

REFERENCES

Cárdenas, A.A., Amin, S., and Sastry, S.S. (2008). Secure control: Towards survivable cyber-physical systems. In *First Int. Workshop on Cyber-Physical Systems*.

Dowler, D.A. (2013). *Bounding the Norm of Matrix Powers*. Master's thesis, Brigham Young University-Provo.

Ferrari, R.M., Parisini, T., and Polycarpou, M. (2008). A robust fault detection and isolation scheme for a class of uncertain input-output discrete-time nonlinear systems. In *American Control Conference, 2008*, 2804–2809.

Ferrari, R.M. and Teixeira, A.M. (2017). Detection and isolation of routing attacks through sensor watermarking. In *American Control Conference, 2017*.

Miao, F., Zhu, Q., Pajic, M., and Pappas, G.J. (2014). Coding sensor outputs for injection attacks detection. In *2014 Conf. on Decision and Control (CDC)*.

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

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2012). Revealing stealthy attacks in control systems. In *50th Annu. Allerton Conf. on Comm., Control, and Comp.*

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2015). A secure control framework for resource-limited adversaries. *Automatica*, 51(1), 135–148.

Zhou, K., Doyle, J.C., and Glover, K. (1996). *Robust and Optimal Control*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.