Detection of sensor data injection attacks with multiplicative watermarking

Teixeira, Andre M.H.; Ferrari, Riccardo

DOI
10.23919/ECC.2018.8550114

Publication date
2018

Document Version
Final published version

Published in
Proceedings of 2018 European Control Conference (ECC2018)

2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

Detection of Sensor Data Injection Attacks with Multiplicative Watermarking

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

Abstract— **In this paper, the problem of detecting stealthy false-data injection attacks on the measurements is considered. We propose a multiplicative watermarking scheme, where each sensor's output is individually fed to a SISO watermark generator whose parameters are supposed to be unknown to the adversary. Under such a scenario, the detectability properties of the attack are analyzed and guidelines for designing the watermarking filters are derived. Fundamental limitations to the case of single-output systems are also uncovered, for which an alternative approach is proposed. The results are illustrated through numerical examples.**

I. Introduction

The topic of cyber-secure control systems has been receiving increasing attention recently. An overview of existing cyber-threats and vulnerabilities in networked control systems is presented in [1]–[3]. Rational adversary models are highlighted as one of the key items in security for control systems, thus making adversaries endowed with intelligence and intent, as opposed to faults. Therefore, these adversaries may exploit existing vulnerabilities and limitations in the traditional anomaly detection mechanisms and remain undetected. In fact, [4] uses such fundamental limitations to characterize a set of stealthy attack policies for networked systems modeled by differential-algebraic equations. Related stealthy attack policies were also considered in [3], [5].

Detectability conditions of stealthy false-data injection attacks to control systems are examined in [6], where it is shown that stealthy attacks may become detectable due to mismatches between the system's and the attack policy's initial conditions. Additionally, modifications to the system dynamics that reveal stealthy attacks were also characterized. Recently, [7] proposed an static output coding scheme combining the outputs of multiple sensors to reveal stealthy data injection attacks on sensors. However, both approaches present certain limitations. On the one hand, the plant's initial conditions cannot be directly controlled, and changing the system dynamics may negatively affect performance. On the other hand, sensor coding schemes require additional communication between sensors and to the controller, and it would not be applicable in single-output systems. These limitations can be tackled by using a multiplicative watermarking scheme, as discussed in this paper.

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

A. Teixeira is with the Division of Signals and Systems, Department of Engineering Sciences, at the Uppsala University, Sweden. andre.teixeira@angstrom.uu.se

R. Ferrari is with the Delft Center for Systems and Controls, at the Delft Technical University, the Netherlands. r.ferrari@tudelft.nl

Watermarking is a well-known solution to the problem of authenticity and integrity verification in the field of multimedia data [8]. An additive watermarking scheme has been proposed by [9] to detect replay attacks, where noise is purposely injected in the system by the actuators to watermark the sensor outputs through known correlations. However, this scheme decreases the performance of the system and fails to detect additive stealthy attacks, drawbacks that can be tackled by employing multiplicative watermarks.

Recently, [10] has proposed the use of an external auxiliary system, with time-varying dynamics unknown to the adversary, whose output is transmitted to the anomaly detector and used to detect the presence of integrity attacks. While sharing similarities with our proposed multiplicative watermarking, the approach in [10] imposes further burdens on the system, such as the communication of the external system's measurement signals and the use of an additional state estimator, which are not required in our watermarking solution. Furthermore, [10] has not addressed possible fundamental limitations to the detection of attacks.

As main contributions of this paper, we consider the modular multiplicative watermarking scheme recently proposed in [11] against replay attacks, where each sensor output is separately watermarked by being fed to a SISO watermark generator and the watermark is latter removed at the controller, therefore not requiring communication between multiple sensors and ensuring a modular architecture. The case of stealthy false-data injection attack to sensor data is analyzed under the proposed multiplicative watermarking scheme, for which fundamental detectability properties are analyzed. In particular, we show how the watermarking scheme can be designed to detect sensor attacks, even for single-output systems, and without affecting the performance of the system in the absence of attacks. The design guidelines of the watermarking filters are independent of the anomaly detection and control schemes, thus ensuring modularity.

The outline of the paper is as follows. In Section II, we describe the problem formulation, as well as the sensor false-data injection attack scenario and recall its detectability properties without watermarking. The sensor watermarking scheme is described in Section III, where the new detectability properties and fundamental limitations are discussed, leading to design guidelines for the watermarking scheme. Numerical examples are presented in Section V, and the paper concludes with final remarks in Section VI.

II. Problem formulation

In this section, we present the control system and describe the main problem at hand. Consider the modeling framework

described in [3], where the control system is composed by a physical plant ($\mathcal{P}$), a feedback controller ($\mathcal{C}$), and an anomaly detector ($\mathcal{R}$). The physical plant, controller, and anomaly detector are modeled in a discrete-time state-space form

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

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 for detecting anomalies. $\eta[k]$ and $\xi[k]$ denote the unknown process and measurement disturbances.

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 $\bar{\xi}[k]$.

The sensor measurements are exchanged through a communication network. To model the fact that the sensor measurements 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 if for at least one time instant k

$$\|y_r\|_{p,[k,k+N_r)} \triangleq \sum_{j=k}^{k+N_r-1} \|y_r[j]\|_p \geq \bar{y}_r[k], \tag{2}$$

where $\bar{y}_r[k] \in \mathbb{R}_+^{n_y}$ is a robust detection residual and $1 \leq p < +\infty$ and $N_r \geq 1$ are design parameters.

The main focus of this paper is to investigate the detection of cyber false-data injection attacks on sensors. This attack scenario, as well as a fundamental limitation in their detectability akin to the results of [3], [4], are described next, where the detectability of attacks is discussed according to

Definition 1: Suppose that the closed-loop system is at equilibrium such that $y_r[-1] = 0$, and that there are no unknown disturbances, i.e., $\eta[k] = 0$ and $\xi[k] = 0$ for all k. An anomaly occurring at $k = k_a \geq 0$ is said to be ε-stealthy if $\|y_r\|_{p,[k,k+N_r)} \leq \varepsilon$ for all $k \geq k_a$. In particular, an ε-stealthy anomaly is termed as simply *stealthy*, whereas a 0-stealthy anomaly is named *undetectable*.

A. Measurement false-data injection attack

In the present scenario, a malicious adversary injects false-data into the measurements sent to the controller, which is captured by adding an attack vector $a[k] \in \mathbb{R}^{n_y}$

$$\tilde{y}_p[k] = y_p[k] + a[k], \tag{3}$$

Attack goals and constraints: The adversary aims at disrupting the system's behavior by corrupting the sensor data, while remaining stealthy with respect to the anomaly detector. Such an adversary model may be characterized by the following attack policy [4], [12]:

$$\begin{aligned} x_a[k+1] &= A_p x_a[k] \\ a[k] &= C_p x_a[k] \end{aligned}, \; x_a[k_a] = \bar{x}_a, \tag{4}$$

where $\bar{x}_a \in \mathbb{R}^{n_p}$ is an eigenvector of A_p.

Disruption and disclosure resources: The adversary is assumed to only have disruption resources to corrupt the measurement data.

Model knowledge: In the present scenario, the adversary also has access to the detailed model of the plant, (A_p, C_p), which is used to compute the attack policy.

Attack detectability: To discuss false-data injection attack detectability, the following definition is required.

Definition 2: Consider the system $\Sigma = (A, B, C, D)$ with $B \in \mathbb{R}^{n_x \times n_u}$ and $C \in \mathbb{R}^{n_y \times n_x}$. A tuple $(\lambda, \bar{x}, g) \in \mathbb{C} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u}$, is a *zero dynamics* (ZD) of Σ if it satisfies

$$\begin{bmatrix} \lambda I_{n_x} - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} \bar{x} \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \bar{x} \neq 0. \tag{5}$$

It is well-known that a ZD tuple can generate an input that results in a zero output. More formally, given a system $\Sigma = (A, B, C, D)$ with a ZD tuple $(\lambda, \bar{x}, g)$ and initial condition $x[k_0] = \bar{x}$, an input of the form $u[k] = \lambda^{k-k_0} g$ applied to Σ will result in the output $y[k] = 0$ for all $k \geq k_0$.

Next we apply this result to the closed-loop system under a sensor false-data injection attack (see (1) and (3)). To compute the attack's contribution to the residue output, suppose that $x_c[k_a]$ and $x_r[k_a]$ are both zero. Recalling (1), we observe that the state of the controller and anomaly detector will remain unchanged as long as $\tilde{y}_p[k] = 0$ for all $k \geq k_a$. Hence, the plant under attack, with input $a[k]$ and output $\tilde{y}_p[k]$, is described by the dynamics $(A_p, 0, C_p, I_{n_y})$.

From Def. 2 a ZD tuple $(\lambda, \bar{x}_a, g)$ of Σ satisfies

$$\begin{bmatrix} \lambda I_{n_x} - A_p & 0 \\ C_p & I_{n_y} \end{bmatrix} \begin{bmatrix} -\bar{x}_a \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

from which we conclude that $\bar{x}_a$ is an eigenvector of A_p associated with λ, $g = C_p \bar{x}_a$, and the corresponding attack signal is $a[k] = \lambda^{k-k_a} C_p \bar{x}_a$. Recalling that $A_p \bar{x}_a = \lambda \bar{x}_a$, we conclude that the attack signal generated by (4) does indeed correspond to a ZD input of Σ. Hence, if Σ is initialized at $x_p[k_a] = -\bar{x}_a$, the attack signal (4) yields a zero output, i.e., $\tilde{y}_p[k] = 0$ for $k \geq k_a$, which is undetected by the anomaly detector. The case for initial conditions $x_p[k_a] \neq -\bar{x}_a$ will result in an asymptotically vanishing transient response if the closed-loop system is stable, akin to the cases in [6].

Attack impact: One relevant aspect is the possible impact of the sensors data injection attack to the states of the physical plant. As an ε-stealthy attack may be parameterized by $a[k] = \lambda^{k-k_a} C_p \bar{x}_a$, if $|\lambda| > 1$ then a stabilizing feedback controller will make the plant's states grow unbounded.

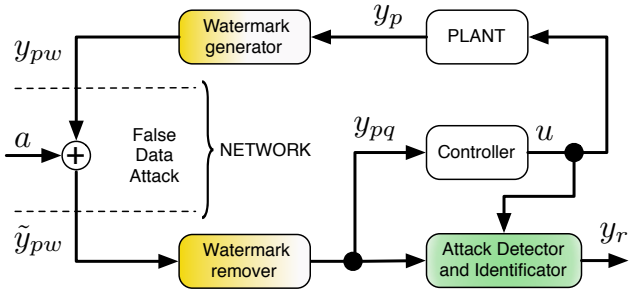

Fig. 1. Scheme of the proposed watermarking scheme under measurement false-data injection attack.

B. Watermarking and equalization scheme

To allow the anomaly detector to detect the presence of false-data injection attacks, we introduce a pre-processing step, denoted as *sensor watermarking* [11], where each sensor processes its measurements through a filter parametrized by θ before transmitting them. 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, the parameter $\theta[k]$ is only known by the sensors and the anomaly detector and controller. For brevity, the time argument of $\theta[k]$ is omitted when possible.

Denoting $\mathcal{W}(\theta)$ as the watermarking filters and $y_{pw}[k]$ as the watermarked sensor outputs to be transmitted, it holds

$$\begin{aligned} \mathcal{P}: &\begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] + \eta[k] \\ y_p[k] = C_p x_p[k] + \xi[k] \end{cases} \\ \mathcal{W}(\theta): &\begin{cases} x_w[k+1] = A_w(\theta) x_w[k] + B_w(\theta) y_p[k] \\ y_{pw}[k] = C_w(\theta) x_w[k] + D_w(\theta) y_p[k]. \end{cases} \end{aligned} \tag{6}$$

At the controller side of the network, the received watermarked data $\tilde{y}_{pw}[k]$ is preprocessed through an equalizer filter parametrized by the very same $\theta[k]$. The objective is to remove the watermark, thus reconstructing in nominal conditions the plant outputs. The equalizer outputs $y_{pq}[k]$ are thus fed to the anomaly detector and controller (Fig. 1). As argued earlier, cyber-attacks can lead to $y_{pw}[k] \neq \tilde{y}_{pw}[k]$.

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

$$\begin{aligned} \mathcal{Q}(\theta): &\begin{cases} x_q[k+1] = A_q(\theta) x_q[k] + B_q(\theta) \tilde{y}_{pw}[k] \\ y_{pq}[k] = C_q(\theta) x_q[k] + D_q(\theta) \tilde{y}_{pw}[k] \end{cases} \\ \mathcal{F}_{cr}: &\begin{cases} x_{cr}[k+1] = A_{cr} x_{cr}[k] + B_{cr} y_{pq}[k] \\ y_r[k] = C_{cr} x_{cr}[k] + D_{cr} y_{pq}[k] \\ u[k] = C_u x_{cr}[k] + D_u y_{pq}[k], \end{cases} \end{aligned} \tag{7}$$

where $x_{cr}[k] = [x_c[k]^\top \ x_r[k]^\top]^\top$, and the matrices A_{cr}, B_{cr}, C_{cr}, D_{cr}, C_u, and D_u are derived from (1).

In the next sections, we derive the conditions under which the attacks are detectable for the disturbance-free case. Then, we identify cases where fundamental limitations still exist, and propose an alternative approach to enforce detection, thus providing guidelines for our watermark scheme design.

III. Sensor Watermarking

Let the watermark generator of the generic ith sensor be implemented through an infinite impulse response (IIR) filter:

$$\begin{aligned} -w^i_{A,(N+1)} y_{pw,(i)}[k] = &\sum_{n=1}^{N} w^i_{A,(N+1-n)} y_{pw,(i)}[k-n] + \\ &\sum_{n=0}^{N} w^i_{B,(N+1-n)} y_{p,(i)}[k-n], \end{aligned} \tag{8}$$

where $w^i_A = [w^i_{A,(1)} \ \dots \ w^i_{A,(N+1)}]^\top \in \mathbb{R}^{N+1}$ and $w^i_B = [w^i_{B,(1)} \ \dots \ w^i_{B,(N+1)}]^\top \in \mathbb{R}^{N+1}$ are the filter parameters, N its order and $w^i_{A,(N+1)} = -1$ by convention. Regarding the watermark remover, a simple approach would be to consider the equalizing filter of the ith measurement as the inverse of the respective watermark filter (see (8) in [11]).

In relation to the watermarking scheme proposed in the previous section, each admissible value of the piecewise constant variable θ is obtained as $\theta_j = \text{col}(\theta^i_j, \ i = 1, \dots n_y)$, with $\theta^i_j = \{w^i_{A,j}, \ w^i_{B,j}\}$ and $w^i_{A,j}$, $w^i_{B,j}$ being a particular choice of filter parameters for the ith measurement. Similarly to the previous section, when no specific jth admissible value is meant, the notation $\theta^i = \{w^i_A, \ w^i_B\}$ is used.

The watermarking filter dynamics for sensor i (8) can be written as $\mathcal{W}(\theta^i)$ in (6), by using the controllable canonical form, where $x^i_w[k] \in \mathbb{R}^N$. Similarly, by using the controllable canonical form and the coordinate transformation matrix $T = w^i_{B,(N+1)} I_N$, the equalizer dynamics can be written as $\mathcal{Q}(\theta^i)$ in (7), where $x^i_q[k] \in \mathbb{R}^N$ and $B^i_q = \begin{bmatrix} 0_{1,N-1} & \dfrac{1}{w^i_{B,(N+1)}} \end{bmatrix}^\top$, $I_N \in \mathbb{R}^{N \times N}$ is the identity matrix and $0_{N,M} \in \mathbb{R}^{N \times M}$ is the null matrix. Inspecting the state-space realizations of $\mathcal{W}(\theta^i)$ and $\mathcal{Q}(\theta^i)$ when the same parameter θ^i is used in both filters, we obtain the following:

$$\begin{aligned} &D^i_q C^i_w + C^i_q = 0, \quad B^i_q D^i_w = B^i_w, \quad D^i_q D^i_w = 1, \\ &A^i_q + B^i_q C^i_w = A^i_q - B^i_w C^i_q = A^i_w. \end{aligned} \tag{9}$$

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

Assumption 2: The watermarking filter $\mathcal{W}(\theta^i)$ and its inverse $\mathcal{Q}(\theta^i)$ are stable for all $\theta^i \in \Theta$. □

For notation simplicity and without loss of generality, when possible we consider the single sensor case, i.e., $n_y = 1$, and therefore omit superscripts. Note that the results extend straightforwardly to the multiple sensor case.

Next, considering the closed-loop system with the proposed watermarking and equalizing filters, we analyze the detectability of stealthy false-data injection attacks. As the aim is to uncover fundamental limitations for arbitrary controllers and anomaly detectors, the core element of the discussion is the cascade of the plant $\mathcal{P}$, the watermarking filter $\mathcal{W}(\theta)$, and the equalizing filter $\mathcal{Q}(\theta)$.

Lemma 1: The open-loop dynamics of the reconstructed output, $y_{pq}[k]$, without disturbances and under a false-data injection attack on the watermarked measurements, $\tilde{y}_{pw}[k] = y_{pw}[k] + a[k]$, can be written as

$$\begin{aligned}\begin{bmatrix} x_p[k+1] \\ x_{wq}[k+1] \end{bmatrix} &= \begin{bmatrix} A_p & 0 \\ 0 & A_q \end{bmatrix} \begin{bmatrix} x_p[k] \\ x_{wq}[k] \end{bmatrix} + \begin{bmatrix} 0 \\ -B_q \end{bmatrix} a[k] \\ y_{pq}[k] &= \begin{bmatrix} C_p & D_q C_w \end{bmatrix} \begin{bmatrix} x_p[k] \\ x_{wq}[k] \end{bmatrix} + D_q a[k]. \end{aligned} \tag{10}$$

Next we discuss the detectability properties of stealthy data injection attacks performed on the system with watermarked sensors, under the following spectral assumptions.

Assumption 3: The matrix A_p has distinct eigenvalues, and the eigenvalues of A_p are not eigenvalues of A_q.

A. Detectability of false-data injection attacks

Here we suppose that the watermark parameters θ are unknown to the attacker and we investigate the detectability of the false-data injection attack $a[k]$ computed according to (4), based only on the plant dynamics. The main result of this section is as follows, where we use the notion of support set of a vector $x \in \mathbb{R}^n$ defined as $\text{supp}(x) \triangleq \{i : x_{(i)} \neq 0\}$.

Theorem 1: Consider the plant with sensor watermarking described in (6), with initial condition $x_{pwq}[0] = [\bar{x}_p^\top \ \bar{x}_w^\top \ \bar{x}_q^\top]^\top$. Suppose the system is under a false-data injection attack on the watermarked measurements, $\tilde{y}_{pw}[k] = y_{pw}[k] + a[k]$, where $a[k]$ is characterized by (4) with $\bar{x}_a$ being an eigenvector of A_p associated with the eigenvalue $\lambda \in \mathbb{C}$. Define the transfer functions $\mathcal{Q}^i(z) \triangleq C_q^i \left(zI_N - A_q^i\right)^{-1} B_q^i + D_q^i$ for all $i = 1, \ldots, n_y$. There exist $\bar{x}_p$, and $\bar{x}_{wq} = \bar{x}_w - \bar{x}_q$ such that the false-data injection attack is 0-stealthy with respect to $y_{pq}[k]$ if, and only if,

$$\mathcal{Q}^i(\lambda) = \mathcal{Q}^j(\lambda), \ \forall\, i, j \in \text{supp}(C_p \bar{x}_a). \tag{11}$$

Proof: Recalling (10) and the attack policy (4), the system under attack can be represented as an autonomous system. Furthermore, the attack is 0-stealthy if and only if the following initial conditions $x_p[0] = \bar{x}_p$, $x_{wq}[0] = \bar{x}_{wq}$, and $x_a[0] = \bar{x}_a$, with $\bar{x}_a$ being an eigenvector of A_p, satisfy the PBH unobservability test [13], which can be written as

$$\begin{bmatrix} \lambda I_{n_x} - A_p & 0 & 0 \\ 0 & \lambda I_N - A_q & B_q C_p \\ 0 & 0 & \lambda I_N - A_p \\ C_p & D_q C_w & D_q C_p \end{bmatrix} \begin{bmatrix} \bar{x}_p \\ \bar{x}_{wq} \\ \bar{x}_a \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{12}$$

for some $\lambda \in \mathbb{C}$. As the first and third equations imply that x_p and x_a are both eigenvectors of A_p for the same eigenvalue λ, we conclude that there exists $\alpha \in \mathbb{C}$ such that $\bar{x}_p = \alpha \bar{x}_a$. Including this change of variable in the former set of equations, together with $D_q C_w = -C_q$, we derive

$$\begin{aligned}\begin{bmatrix} \lambda I_N - A_q & B_q \\ -C_q & \alpha I_{n_y} + D_q \end{bmatrix} \begin{bmatrix} \bar{x}_{wq} \\ C_p \bar{x}_a \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ (\lambda I - A_p)\bar{x}_a &= 0. \end{aligned} \tag{13}$$

The proof concludes by recalling that, from the attack policy, $\bar{x}_a$ satisfies the second equation. By solving for $\bar{x}_{wq}$, the first set of equations in (13) can be rewritten as $\left(C_q \left(\lambda I_N - A_q\right)^{-1} B_q + D_q + \alpha I_{n_y}\right) C_p \bar{x}_a = 0$. Recalling that A_q, B_q, C_q, and D_q are all block-diagonal, representing independent filters, the latter equation can be rewritten as $\mathcal{Q}^i(\lambda) = -\alpha, \ \forall i \in \text{supp}(C_p \bar{x}_a)$, where α is a constant. ■

The latter result characterizes under what conditions data injection attacks, computed based on (A_p, C_p), are 0-stealthy, despite the presence of the watermarking filters. This points to design guidelines that enable detection, by ensuring $\mathcal{Q}^i(\lambda) \neq \mathcal{Q}^j(\lambda)$ for all $i, j \in \text{supp}(C_p \bar{x}_a)$ and for all $\lambda \in \mathbb{C}$ in the spectrum of A_p, where $\bar{x}_a$ is the eigenvector of A_p associated with λ. There are, however, fundamental limitations for single-output systems, as well as for the case of multiple outputs with homogeneous filters for all sensors, as formalized next.

Corollary 1: For single-output systems and for multiple-output systems with homogeneous watermark filters, i.e. $w_A^i = w_A^j$ and $w_B^i = w_B^j$ for all $i \neq j$, there exist $\bar{x}_p$ and $\bar{x}_{wq} = \bar{x}_w - \bar{x}_q$ such that the false-data injection attack is 0-stealthy with respect to $y_{pq}[k]$.

Despite such limitations, there is another degree of freedom that may be leveraged to make the attack ε-stealthy, and therefore detectable, even when (11) is satisfied, such as in the cases of Corollary 1. In fact, note that 0-stealthy attacks also require specific initial conditions of the plant and the watermarking filters, $\bar{x}_p$ and $\bar{x}_{wq}$ respectively. Although $\bar{x}_p$ cannot be directly controlled, $\bar{x}_w$ and $\bar{x}_q$ and thus $\bar{x}_{wq}$ can, as the filters are implemented in digital computers. In particular, as follows from Theorem 2 in [11], resetting $\bar{x}_w$ and $\bar{x}_q$ to the same value such that $\bar{x}_{wq} = 0$ would have no adverse impact on the closed-loop performance.

Theorem 2: Consider the plant with sensor watermarking described in (6), with initial condition $x_{pwq}[0] = [\bar{x}_p^\top \ \bar{x}_w^\top \ \bar{x}_q^\top]^\top$. Suppose the system is under a sensor false-data injection attack on the watermarked measurements, $\tilde{y}_{pw}[k] = y_{pw}[k] + a[k]$, where $a[k]$ is characterized by (4) with $\bar{x}_a$ being an eigenvector of A_p associated with the eigenvalue $\lambda \in \mathbb{C}$. Furthermore, suppose that $\bar{x}_p = \alpha \bar{x}_a$ and $\mathcal{Q}^i(\lambda) = \alpha, \ \forall\, i \in \text{supp}(C_p \bar{x}_a)$, for some $\alpha \neq 0$, and define $\bar{x}_{wq}^a$ such that $\left[\alpha \bar{x}_a^\top \ \bar{x}_{wq}^{a\top} \ \bar{x}_a^\top\right]^\top$ is a solution to (12).

The output $y_{pq}[k]$ under the measurement false-data injection attack is described by the autonomous system

$$\begin{aligned}\Delta x_{wq}[k+1] &= A_q \Delta x_{wq}[k] \\ y_{pq}[k] &= D_q C_w \Delta x_{wq}[k] \end{aligned} \tag{14}$$

with $\Delta x_{wq}[0] = \bar{x}_w - \bar{x}_q - \bar{x}_{wq}^a$. Furthermore, for $\bar{x}_w - \bar{x}_q \neq \bar{x}_{wq}^a$, the false-data injection attack is ε-stealthy with respect to the output $y_{pq}[k]$, for a finite $\varepsilon > 0$.

Proof: The proof is omitted. ■

In the next section, we further explore the influence of resetting the watermarking filters states on attack detectability.

IV. Detection of False Data Attacks

We now introduce the details of the attack detector $\mathcal{R}$ and provide a practical and sufficient detectability condition. Ass. 5 and 6 from [11, Sect. 4] will be require, and similarly

the detector will be built on top of the following estimator

$$\hat{\mathcal{P}}: \begin{cases} \hat{x}_p[k+1] = A_p\hat{x}_p[k] + B_p u[k] + K\left(y_{pq}[k] - \hat{y}_p[k]\right) \\ \hat{y}_p[k] = C_p\hat{x}_p[k], \end{cases} \tag{15}$$

where $\hat{x}_p \in \mathbb{R}^{n_p}$ and $\hat{y}_p \in \mathbb{R}^{n_y}$ are meant to estimate of x_p and y_p, and K is chosen such that $A_r \triangleq A_p - KC_p$ is Schur. By setting $x_r = \hat{x}_p$ and $\epsilon \triangleq x_p - \hat{x}_p$, when no attack is present the detection residual $y_r \triangleq y_{pq} - \hat{y}_p$ dynamics are

$$\begin{cases} \epsilon[k+1] = A_r\epsilon[k] - K\xi[k] + \eta[k] \\ y_r[k] = C_p\epsilon[k] + \xi[k] \end{cases}, \tag{16}$$

and the detection threshold ith component is computed as

$$\bar{y}_{r,(i)}[k] \triangleq \alpha^i \left[\sum_{h=0}^{k-1} \left(\beta^i\right)^{k-1-h} \left(\bar{\eta}[h] + \|K\|\bar{\xi}[h]\right) + \left(\beta^i\right)^k \bar{\epsilon}[0]\right] + \bar{\xi}[k], \tag{17}$$

assuming an horizon $N_r = 1$ and the 1-norm, and where α^i and β^i are two constants such that $\|C_{p,(i)}(A_r)^k\| \leq \alpha^i \left(\beta^i\right)^k \leq \|C_{p,(i)}\| \cdot \|\left(A_r\right)^k\|$ with $C_{p,(i)}$ being the i–th row of matrix C_p Furthermore, $\bar{\eta}$, $\bar{\epsilon}[0]$ and $\bar{\xi}$ are upper bounds on the norms of, respectively, η, $\epsilon[0]$ and ξ (see [11]). To understand the effect of a sensor false data attack on y_r let us first consider the case where no watermarking is in place. By adding (4) to (1) it is easy to see that the attacked output $\tilde{y}_{pq} = \tilde{y}_p = y_p + a$ can be generated by the following system

$$\begin{cases} \tilde{x}_p[k+1] = A_p\tilde{x}_p[k] + B_p u[k] + \eta[k] \\ \tilde{y}_p[k] = C_p\tilde{x}_p[k] + \xi[k], \end{cases} \tag{18}$$

where it holds $\tilde{x}_p[k] = x_p[k] + x_a[k] = x_p[k] + \lambda^{k-k_a}\bar{x}_a$, with k_a the attack start time. From this it follows that by feeding $\tilde{y}_{pq}$ to the estimator (15), its state estimate $\tilde{x}_p$ will converge to $\tilde{x}_p$ instead than to x_p. Consequently, the detection residual dynamics under attack will be described by (16), with $\epsilon[k] = \tilde{x}_p - \hat{x}_p$, which translates into the stealthiness of the attack.

During an attack, the detector is fed the output $\tilde{y}_{pq} = y_{pq} + a_q$, where a_q is obtained by processing the attack signal $a[k]$ through the watermark remover. Hence, the output $\tilde{y}_{pq}$ can be written as $\tilde{y}_{pq}[k] = C_p\tilde{x}_p[k] + \xi[k] + \delta_a[k]$, where $\delta_a[k]$ is defined as follows.

Lemma 2: Define $k^* \triangleq \max_i\{k_i \,|\, k_i \leq k, i \in \mathbb{N}\}$ as the last watermark switching instant before the current time k, and suppose that $k^* \geq k_a$. The term $\delta_a[k]$ can be written as the output of the following autonomous system

$$\begin{aligned} \begin{bmatrix} x_q[k+1] \\ x_a[k+1] \end{bmatrix} &= \begin{bmatrix} A_q & B_qC_q \\ 0 & A_p \end{bmatrix} \begin{bmatrix} x_q[k] \\ x_a[k] \end{bmatrix} \\ \delta_a[k] &= \begin{bmatrix} C_q & (D_q - I)C_p \end{bmatrix} \begin{bmatrix} x_q[k] \\ x_a[k] \end{bmatrix}, \end{aligned} \tag{19}$$

for all $k \geq k^*$, with $x_q[k^*] = 0$ and $x_a[k^*] = \lambda^{k^*-k_a}\bar{x}_a$.

Given the above characterization of the output, the residual generated by the detector satisfies the following dynamics

$$\begin{cases} \tilde{\epsilon}[k+1] = A_r\tilde{\epsilon}[k] - K(\xi[k] + \delta_a[k]) + \eta[k] \\ y_r[k] = C_p\tilde{\epsilon}[k] + \xi[k] + \delta_a[k] \end{cases}, \tag{20}$$

The following sufficient detectability condition holds:

Theorem 3 (Attack Detectability): If there exists a time index $k_d > k_a$ and a component $i \in \{1, \ldots, n_y\}$ such that during a sensor false data attack the following inequality holds

$$\begin{aligned} &\left| C_{p,(i)} \sum_{h=k_a}^{k_d-1} (A_r)^{k_d-1-h} K\delta_a[h] + \delta_{a,(i)}[k_d] \right| \\ &\quad > 2\alpha^i \sum_{h=0}^{k_d-1} \left(\beta^i\right)^{k_d-1-h} \left(\bar{\eta}[h] + \|K\|\bar{\xi}[h]\right) + \\ &\qquad \left(\beta^i\right)^{k_d} \left(\alpha^i\bar{\epsilon}[0] + \bar{y}_{r,(i)}[0]\right) + 2\bar{\xi}[k_d] \end{aligned}$$

where $\bar{y}_{r,(i)}0] \triangleq \max_{x_p \in \mathscr{S}^{x_p}} |y_{r,(i)}[0]|$ and α^i and β^i are two constants such that $\|C_{p,(i)}\left(A_r\right)^k\| \leq \alpha^i \left(\beta^i\right)^k \leq \|C_{p,(i)}\| \cdot \|\left(A_r\right)^k\|$ with $C_{p,(i)}$ being the i–th row of matrix C_p, then the attack will be detected at the time instant k_d.

Remark 1: The term δ_a is due to the attack being fed through the equalizer, and explains why watermarking can improve detectability. Furthermore, the switching of watermark parameters at instants k_i will abruptly reset δ_a to $(D_q - I)C_p\lambda^{k^*-k_a}\bar{x}_a$, thus possibly easing detection.

However, as suggested by Theorem 2, in the case of homogeneous watermarking filters, the effect of the resetting watermarking filters vanishes asymptotically and, therefore, one expects that the left-hand-side term of the detectability condition in Theorem 3 converges to zero as $k - k^*$ tends to infinity. This behavior is formalized by the next results.

Theorem 4: Suppose that the filters $\mathcal{Q}$ satisfy $\mathcal{Q}^i(\lambda) = \mathcal{Q}^j(\lambda)$ for all $i, j \in \text{supp}(C_p\bar{x}_a)$ and let $k^* \geq k_a$. Define the term $\Delta y_{r,(i)}[k] \triangleq C_{p,(i)} \sum_{h=k_a}^{k-1} (A_r)^{k-1-h} K\delta_a[h] + \delta_{a,(i)}[k]$. As $k - k^*$ tends to infinity, $|\Delta y_{r,(i)}|$ asymptotically converges to 0, for all $i = 1, \ldots, n_y$.

Proof: The proof is omitted. ∎

Theorem 4 illustrates how the limitations uncovered in Corollary 1 affect detectability. Furthermore, it points that the reset of the watermarking filters' initial conditions should be performed regularly, as to limit $k - k^*$ and thus enforcing $\delta_a[k]$ to be in a transient regime where detection is possible.

V. Numerical Examples

Let us consider $\mathcal{P}$ to be an unstable discrete-time LTI system with $n_p = 2$, $n_u = 1$, $n_y = 2$ and matrices

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

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 $A_c = I_2$, $B_c = 0.1 \cdot I_2$, $C_c = \begin{bmatrix} 0.01 & 0.022 \end{bmatrix}$, $D_c = \begin{bmatrix} 0.0875 & 0.1980 \end{bmatrix}$ 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, and $r_{(2)}$ a null one. 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.

TABLE I

PERFORMANCE OF DIFFERENT WATERMARKING STRATEGIES.

index	none	homogeneous		heterogeneous	
		sw.	no sw.	sw.	no sw.
$k_d \cdot T_s$	N/A	140 s	N/A	130 s	145.7
$\frac{\lvert y_{r,(i_d)}[k_d]\rvert}{\bar{y}_{r,(i_d)}[k_d]}$	N/A	1.33	N/A	1.36	1.04
$\left\lvert\frac{a_{(i_d)}[k_d]}{y_{p,(i_d)}[k_d]}\right\rvert$	N/A	0.44	N/A	0.15	0.69

Performance is measured through three indexes: the detection time instant (the smaller, the better), the ratio of the residual and the threshold at detection (the larger, the better) and the ratio of the attack signal to the output at detection (the smaller, the better). Nomenclature: "none", no watermark in place; "homogeneous", same filter parameters w_A and w_B are used for all output components; "heterogeneous", different parameters used; "sw.", parameters switched every 10 s; "no sw.", fixed parameters. The index i_d refers to the component for which the residual first crosses the threshold. "N/A" signals no detection occurred during simulation time.

At time $T_a = k_a \cdot T_s = 75\,\text{s}$, a measurement false-data injection attack described by $a[k] = C_p A_p^{k-k_a} x_a = \lambda^{k-k_a} C_p x_a$, with $x_a = -10^{-4}[-0.9898 \;\; -0.1422]^\top$ and $\lambda = 1.0144$, starts to excite the plant unstable mode.

When no watermarking is used (Fig. 2), the exponentially increasing attack signal being causes the true plant output y_p to quickly diverge, while the estimated output $\hat{y}_p$ appears to follow the square wave reference faithfully. The residual and threshold, too, do not reveal any sign of the attack.

The cases where heterogeneous or homogeneous (see Corollary 1) filters are used, and the sub-cases of parameters being switched every $\tau_{\text{switch}} = 10\,\text{s}$ or being fixed, are compared in Tab. I. The watermark generators consist of third order FIR filters, with $w^i_{A,(N+1)} = 1$, $w^i_{A,(j)} = 0$ for $j = 1, \ldots, 3$, and $w_B^{i\top} = [1,\, 0,\, 0,\, 0] + \omega^i$, ω^i being a random variable uniformly distributed in $[-0.1 \;\; 0.1]^4$. As we expected, best results are obtained with switched heterogeneous filters. Detection is obtained also in the switched homogeneous case, where the effect of the initial condition mismatch $\Delta x_{wq}[k_{\theta_i}] = -\bar{x}^a_{wq}[k_{\theta_i}]$, is exponentially increasing due to the exponential attack signal $a[k]$ (see Fig. 3).

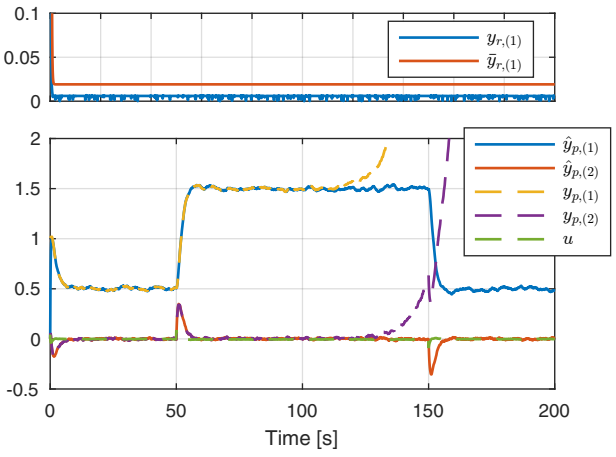

Fig. 2. Results when no watermark is present. Upper: Residual and threshold for first output. Lower: estimated true plant outputs produced by the detector (solid lines), and true plant outputs and input (dashed lines).

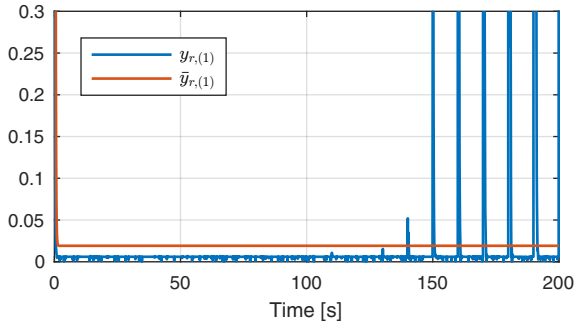

Fig. 3. First components of the detection residuals and thresholds for the switched homogeneous case. Residual spikes correspond to switching times.

VI. CONCLUSIONS

A multiplicative sensor watermarking scheme, where each sensor's output is separately watermarked by a SISO watermark generator, was proposed. As opposed to input watermarking schemes, no additional burden is put on physical actuators. Furthermore, stealthy false-data injection attacks become detectable due to the presence of the watermarking filters. Fundamental limitations for the case of single-output systems are also uncovered, which are overcome by regularly resetting the states of the watermarking filters.

REFERENCES

[1] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Secure control: Towards survivable cyber-physical systems," in 1^{st} *Int. Workshop on Cyber-Physical Syst.*, June 2008.

[2] A. A. Cárdenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. S. Sastry, "Challenges for securing cyber physical systems," in *Workshop on Future Dir. in Cyber-physical Syst. Security*. U.S. DHS, July 2009.

[3] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. 1, pp. 135–148, 2015.

[4] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. on Autom. Contr.*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[5] R. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," in *18th IFAC World Congress*, 2011.

[6] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.

[7] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Trans. on Contr. of Network Sys.*, vol. 4, no. 1, 2017.

[8] L. Pérez-Freire, P. Comesaña, J. R. Troncoso-Pastoriza, and F. Pérez-González, *Trans. on Data Hiding and Multim. Security I.* Springer Berlin Heidelberg, 2006, ch. Watermarking Security: A Survey.

[9] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *Contr. Syst., IEEE*, vol. 35, 2015.

[10] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *Proc. of the 54th IEEE Conf. on Decision and Control (CDC)*, Osaka, Japan, Dec. 2015.

[11] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," in *Proc. of 20th IFAC World Congress*, Toulouse, France, July 2017.

[12] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. on Autom. Control*, vol. 59, no. 6, pp. 1454–1467, June 2014.

[13] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1996.