Detection of Cyber-Attacks: a Multiplicative Watermarking Scheme

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

Abstract This chapter addresses the problem of detecting stealthy data injection attacks on sensor measurements in a networked control system. A multiplicative watermarking scheme is proposed, where each sensor data is post-processed by a time-varying filter called *watermark generator*. At the controller's side, the watermark is removed from each channel by another filter, called the *watermark remover*, thus reconstructing the original signal. The parameters of each remover are matched to those of the corresponding generator, and are supposed to be a shared secret not known by the attacker. The rationale for time-varying watermarks is to allow model-based schemes to detect otherwise stealthy attacks, by constantly introducing mismatches between the actual and the nominal dynamics used by the detector. A specific model-based diagnosis algorithm is designed to this end. Under the proposed watermarking scheme, the robustness and the detectability properties of the model-based detector are analyzed and guidelines for designing the watermarking filters are derived. Distinctive features of the proposed approach, with respect to other solutions like end-to-end encryption, are that the scheme is lightweight enough to be applied also to legacy control systems, its absence of side effects such as delays, and the possibility of utilizing a robust controller to operate the closed-loop system in the event of the transmitter and receiver losing synchronization of their watermarking filters. The results are illustrated through numerical examples.

Riccardo Ferrari
Delft University of Technology, The Netherlands, e-mail: r.ferrari@tudelft.nl

André M. H. Teixeira
Upssala University, Sweden e-mail: andre.teixeira@angstrom.uu.se

1 Introduction

The penetration of information technologies (IT) hardware and software in current networked industrial control systems (ICS) has grown significantly in recent times. This has led ICS to being vulnerable to a steadily increasing number of cyber-threats, as discussed in NCCIC and ICS-CERT (2016); Trend Micro (2018); Gorenc and Sands (2018). Thus, it must not come as a surprise that, in recent years, the control systems community became more and more attentive to the topic of cyber-security, in addition to the established focus on safety (Cárdenas et al, 2008, 2009; Teixeira et al, 2015). A keystone in such endeavour is the introduction of rational adversary models for describing cyber-attack policies, thus differentiating knowledgeable and malicious adversaries with respect to faults. Such adversaries aim at exploiting existing vulnerabilities and limitations in traditional anomaly detection mechanisms, while remaining undetected. The concept of stealthy attacks has been investigated in Pasqualetti et al (2013) and Teixeira et al (2015); Smith (2011), amongst others.

Amongst the proposed approaches to detecting stealthy attacks, Teixeira et al (2012) has shown how they can be detected by taking advantage of mismatches between the system's and the attack policy's initial conditions. Another stream of research considered instead active modifications to the system dynamics, that could expose such otherwise stealthy attacks. For instance, Miao et al (2017) proposed a static multiple-sensors output coding scheme. Nonetheless, both approaches bear some limitations, such as the unrealistic requirement of controlling the plant's initial condition, or the control performances drop caused by active modifications.

Other related approaches found in the literature have been inspired in the concept of watermarking. Watermarking, a classic approach for guaranteeing authenticity in the multimedia industry (Pérez-Freire et al, 2006), has been recently proposed as a way to overcome such drawbacks while making stealthy attacks detectable by existing model-based anomaly detectors. An additive watermarking scheme has been introduced by Mo et al (2015) to detect replay attacks, where colored noise of known covariance is purposely injected in the actuators. A similar, but distributed approach for interconnected microgrids was instead presented in Gallo et al (2018). However, the injection of an additive watermark in the actuators leads to decreased control performances, and does not guarantee against additive stealthy attacks.

As a way to tackle such limitations, in this chapter we further extend the modular multiplicative watermarking scheme proposed in Ferrari and Teixeira (2017a). Such an approach is based on each sensor output being independently pre-processed via a time-varying single-input single-output (SISO) watermark generator before transmission over the control network. A bank of matched watermark removers is included on the controller side, where the original sensors' signals are reconstructed thus preventing any control performances loss (Fig. 1). The approach is independent from the plant's initial condition and does not require extra communication or coordination between multiple sensors.

The proposed solution resembles a channel encryption scheme: indeed, watermarking can be interpreted as a light-weight mechanism enforcing authentication of

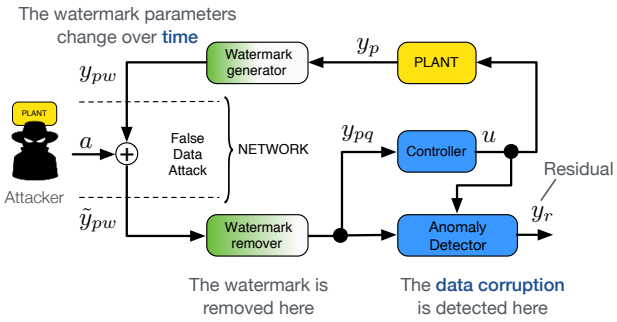


Fig. 1 Scheme of the proposed watermarking scheme under measurement false-data injection attack.

the data and its source, albeit with weaker cryptographic guarantees than strong encryption schemes (Sandberg et al, 2015). For the case of networked control systems, this weakness often translates into a strength. As watermarking requires lighter computational power, it is better suited to meet critical real-time constraints. Furthermore, as authentication and data integrity are in this scenario more important than data confidentiality, the use of strong cryptographic methods may be unwarranted. Additionally, as investigated in this chapter, a robust controller may still be able to stabilize the system when the transmitter and receiver lose synchronization, which is not the case when standard cryptographic schemes are used.

The rationale behind the proposed watermarking scheme is to make stealthy man-in-the-middle attacks detectable, by having them cause an imperfect reconstruction of the sensors' measurements. Such condition would cause a detection by a suitable *anomaly detector* (Ferrari and Teixeira, 2018, 2017a,b). In particular, in this chapter we introduce novel watermark generators and removers, implemented as hybrid switching SISO systems with piece-wise linear dynamics. The design of such switching filters is addressed and it is shown how they can guarantee perfect reconstruction of the plant outputs. Furthermore, their time-varying properties are linked to conditions on the detectability of otherwise stealthy attacks. Stability of the closed-loop system with the proposed watermarking scheme is also analyzed, including the case of constant but mismatched parameter filters at the generator and remover.

The outline of the chapter is as follows. In Section 2, we describe the problem formulation, as well as define stealthy data-injection attacks that are undetectable without watermarking. The design of the switching sensor watermarking scheme is addressed in Section 3, where design guidelines for the watermarking scheme and its synchronization protocol are provided. An application example is provided as well, to illustrate the proposed approach. Detectability properties are investigated in Section 4, while numerical results illustrating the effectiveness of the proposed solutions are reported in Section 5. The paper concludes with final remarks and future work directions in Section 6.

2 Problem formulation

Following the modeling framework for secure control systems presented in Teixeira et al (2015), in this section the control system under attack is described, together with the adversary model and known limitations to its detection. The conceptual structure of the closed-loop system under attack is presented in Fig. 2.

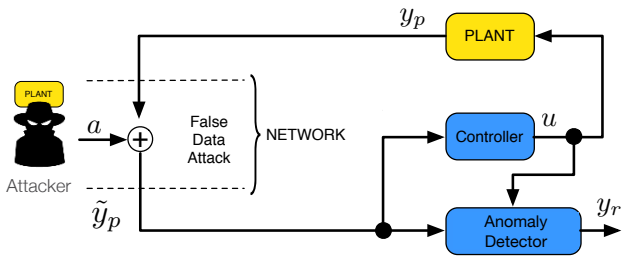


Fig. 2 Scheme of the networked control system under measurement false-data injection attack, without the proposed watermarking and detection architecture.

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 as discrete-time linear systems:

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

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 $y_r[k] \in \mathbb{R}^{n_y}$ is the *residual vector* that is evaluated for detecting anomalies. The variables $\eta[k]$ and $\xi[k]$, finally, denote the unknown process and measurement disturbances.

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

The sensor measurements are exchanged through a communication network, and can thus be targeted by cyber-attacks that manipulate the data arriving at the receiver.

At the plant side, the data transmitted by the sensors is denoted as $y_p[k] \in \mathbb{R}^{n_y}$, while the received sensor data at the detector's side is denoted as $\tilde{y}_p[k] \in \mathbb{R}^{n_y}$.

The operation of the closed-loop system is monitored by the anomaly detector, 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, the following condition holds:

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

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

The main focus of this chapter is to investigate the detection of so-called false data injection cyber-attacks on sensors. This attack scenario, as well as fundamental limitations in its detectability, are described next.

2.1 False-data injection attacks

In the present false-data injection attack scenario, derived from Ferrari and Teixeira (2018), we consider a malicious adversary that is able to access and corrupt the measurements sent to the controller. This attack policy may be modeled as

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

where $a[k]$ is the malicious data corruption added to the measurements. Note that such a scenario may be revised to also include replay attacks as in Ferrari and Teixeira (2017a), which is modeled as $\tilde{y}_p[k] = y_p[k-T]$, where previously recorded data is forwarded again to the controller. Similarly, even routing attacks can be considered as in Ferrari and Teixeira (2017c), which are modeled as $\tilde{y}_p[k] = Ry_p[k]$, where R is a routing matrix. In this chapter, the false data injection attacks are examined in more detail, while brief remarks are given for the case of replay attacks.

The attack is designed by the adversary according to a set of attack goals and constraints, attack resources, and system knowledge (Teixeira et al, 2015). These aspects are further described below.

Attack goals and constraints: The adversary aims at corrupting the sensor data so that the system's operation is disrupted, while remaining undetected by the anomaly detector.

Disruption and disclosure resources: The false-data injection attack on the communication channels requires that the attacker is able to both read the transmitted data and corrupt it. Therefore, we assume that the attacker has the required disclosure resources to eavesdrop on the transmitted data, as well as the disruption resources to corrupt the measurement data received by the controller and anomaly detector.

Model knowledge: Taking a worst-case perspective, the adversary is assumed to have access to the detailed nominal model of the plant, (A_p, B_p, C_p).

Fundamental limitations to detectability: As it is well-known in the literature (Pasqualetti et al, 2013; Teixeira et al, 2015), an attacker with detailed knowledge of the plant may be able to inject false data that mimics the behavior of the plant, and therefore bypass the detection of a linear-time invariant detector. In particular, this chapter discusses the detectability of attacks according to the following definition.

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., $\eta_p[k] = 0$ and $\xi_p[k] = 0$ 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*.

More specifically, we focus on undetectable attacks that are able to produce no visible change to the residual generated by the anomaly detector. To characterize such a class of attacks, the following definition is required.

Definition 2. Consider the system $\Sigma = (A, B, C, D)$ with input $a[k]$ and output $y[k]$, where $B \in \mathbb{R}^{n_x \times n_u}$ 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* of Σ if it satisfies

$$\begin{bmatrix} \lambda I_{n_x} - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} \bar{x} \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \bar{x} \neq 0. \tag{4}$$

Moreover, the input $a[k] = \lambda^{k-k_0} g$ is called an output-zeroing input that, for $x[k_0] = \bar{x}$, yields $y[k] = 0$ for all $k \geq k_0$.

Next we apply the previous definition to the closed-loop system under sensor false-data injection attack (see (1) and (3)), and characterize a specific class of undetectable attacks that complies with the previously described adversary model.

Consider the plant under a sensor data attack, which begins at time $k = k_a$. The respective dynamics and the data received by the controller and anomaly detector are described as

$$\begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] \\ \tilde{y}_p[k] = C_p x_p[k] + a[k]. \end{cases} \tag{5}$$

Based on (1), the trajectories of the closed-loop system under attack, with $\eta[k] = \zeta[k] = 0$, are described by

$$\begin{cases} x[k+1] = Ax[k] + Ba[k] \\ y_r[k] = Cx[k] + Da[k] \end{cases}, \forall k \geq k_a,$$

where $x = \begin{bmatrix} x_p^\top & x_c^\top & x_r^\top \end{bmatrix}$ is the augmented state of the closed-loop system, and the matrices A, B, C, D are defined appropriately from (1).

Observing that $\tilde{y}_p[k]$ serves as input to both the controller and the anomaly detector, we conclude that an output-zeroing attack with respect to $\tilde{y}_p[k]$ would lead to no change on the controller and anomaly detector, and would thus be undetectable. Motivated by this observation, we consider an output-zeroing attack based only on the plant dynamics (5), computed while assuming $u[k] = 0$ and thus captured by the dynamical system $\Sigma = (A_p, 0, C_p, I_{n_y})$. From Def. 2 a zero-dynamics tuple $(\lambda, \bar{x}_a, g)$ of Σ satisfies

$$\begin{bmatrix} \lambda I_{n_x} - A_p & 0 \\ C_p & I_{n_y} \end{bmatrix} \begin{bmatrix} -\bar{x}_a \\ g \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{6}$$

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 $a[k] = \lambda^{k-k_a} C_p \bar{x}_a$. In fact, for the case of sensor attacks, a generic attack signal can be generated by the autonomous system

$$\begin{cases} x_p^a[k+1] = A_p x_p^a[k] \\ a[k] = C_p x_p^a[k] \end{cases}, \forall k \geq k_a, \tag{7}$$

with an arbitrary initial condition $x_p^a[k_a]$ chosen by the adversary.

Let us now look at the effect of such an attack on the closed-loop system (1). Combining (1) with (7), we observe that the compromised measurement output for $k \geq k_a$ is described by $\tilde{y}_p[k] = C_p A_p^{k-k_a} \left(x_p[k_a] + x_p^a[k_a] \right) + \sum_{i=k_a}^{k-k_a-1} C_p A_p^{k-1-k_a-i} B_p u[i]$. Thus, from the output's perspective, the false-data injection attack effectively induces a trajectory identical to that of an impulsive jump of the plant's state at $k = k_a$, from $x_p[k_a]$ to $x_p[k_a] + x_p^a[k_a]$. Given that $\tilde{y}_p[k]$ is the input to the detector and controller, these components will precisely react as if the system experienced the aforementioned impulsive jump. Therefore, the smaller the jump (*i.e.,* $x_p^a[k_a]$), the harder it will be to detect the attack.

As an example, and without loss of generality, let the plant be initialized at the origin $x_p[k_a] = 0$. In this case, the impulsive jump essentially corresponds to a non-zero initial condition. Hence, if the closed-loop system is stable, then the impulsive jump in the attack will result in an asymptotically vanishing transient response, akin to the cases in Teixeira et al (2012).

As discussed above, (6) characterizes a set of undetectable attacks on sensors that essentially mimic a possible trajectory of the system, and thus the anomaly detector cannot distinguish between the attack and a normal transient trajectory. Next we describe a multiplicative watermarking scheme which extends the work in Ferrari and Teixeira (2017a, 2018) and that enables the detection of such attacks, while not affecting the performance of the closed-loop system in normal conditions.

3 Multiplicative Watermarking Scheme

To detect the presence of man-in-the-middle attacks, we consider the watermarking scheme illustrated in Fig. 1, where the following elements are added to the control system: a *Watermark Generator* $\mathcal{W}$ and a *Watermark Remover* $\mathcal{Q}$.

To secure the system against adversaries, the watermark generator and remover should share a private key unknown to the adversary, similar to cryptographic schemes. Furthermore, for increased security, the private key must be updated over time. Essentially, the presence of the private watermarking filter introduces an asymmetry between the adversary's knowledge and the watermarked plant, as illustrated in Fig. 3, which is the key to enable the attack's detection. These aspects will guide the design of the multiplicative watermarking scheme, as described in the remainder of this section.

Fig. 3 The role of multiplicative watermarking in attack detection. The attacker assumes the data being transmitted over the network is produced by the *plant*, of which he/she knows a model. Instead, it is produced by the cascade of the *plant* and of the *watermark generator*. Such asymmetry has a key role in making the attack detectable.

3.1 Watermarking Scheme: a Hybrid System approach

The watermark generator $\mathcal{W}$ and remover $\mathcal{Q}$ are designed as synchronized hybrid discrete-time linear systems, which will both experience discrete jumps at the time indexes contained in the sequence $\mathscr{T} \triangleq \{k^1, \dots, k^N\}$. As anticipated, and as will be detailed later, such switching behaviour is enabling in making stealthy data injection attacks detectable. Between switches, that is for $k^i \leq k < k^{i+1}$, the dynamics of $\mathcal{W}$ and $\mathcal{Q}$ are described by the following state space equations:

$$\begin{aligned} \mathcal{W}: &\begin{cases} x_w[k+1] = A_w(\theta_w[k])x_w[k] + B_w(\theta_w[k])y_p[k] \\ \quad y_w[k] = C_w(\theta_w[k])x_w[k] + D_w(\theta_w[k])y_p[k] \end{cases} \\ \mathcal{Q}: &\begin{cases} x_q[k+1] = A_q(\theta_q[k])x_q[k] + B_q(\theta_q[k])y_w[k] \\ \quad y_q[k] = C_q(\theta_q[k])x_q[k] + D_q(\theta_q[k])y_w[k], \end{cases} \end{aligned} \tag{8}$$

where the vectors $x_w, x_q \in \mathbb{R}^{n_w}$ and $y_w, y_q \in \mathbb{R}^{n_y}$ represent, respectively, the state of the watermark generator $\mathcal{W}$ and of the watermark remover $\mathcal{Q}$ and their output.

Each component, or channel, of the output y_p shall be watermarked independently, thus leading the matrices $A_w, A_q \in \mathbb{R}^{n_w \times n_w}$, $B_w, B_q \in \mathbb{R}^{n_w \times n_y}$, $C_w, C_q \in \mathbb{R}^{n_y \times n_w}$ and $D_w, D_q \in \mathbb{R}^{n_y \times n_y}$ to have a block diagonal structure. This will be denoted as $A_w = \text{blkdiag}(A_w^1, \ldots, A_w^{n_y})$, where the blocks A_w^i will have suitable sizes, and similarly for the other matrices in eq. (8).

The vectors $\theta_w, \theta_q \in \mathbb{R}^{n_\theta}$ denote piece-wise constant parameters affecting the dynamics and constitute the private key used by $\mathcal{W}$ and $\mathcal{Q}$ to generate and remove the watermark. These parameters are updated at switching times, and their values can be defined via two sequences $\Theta_\mathcal{W} \triangleq \{\theta_w[k^1], \ldots, \theta_w[k^N]\}$ and $\Theta_\mathcal{Q} \triangleq \{\theta_q[k^1], \ldots, \theta_q[k^N]\}$, respectively. Moreover, the internal states of the watermark generator and remover are also affected by discrete jumps at switching times. In particular, their values at a switching time $k = k^i$ will not be determined by propagating eq. (8) one time step forward from $k = k^i - 1$, but will be defined via two further sequences: $\mathscr{X}_\mathcal{W}^+ \triangleq \{x_w^+[k^1], \ldots, x_w^+[k^N]\}$ and $\mathscr{X}_\mathcal{Q}^+ \triangleq \{x_q^+[k^1], \ldots, x_q^+[k^N]\}$, respectively. The notation $x_w^+[k]^i$ is introduced, by drawing on the hybrid systems literature (Goebel et al, 2009; Teel and Poveda, 2015), to stress that we denote a value to which the variable x_w is reset after a switch, rather than the value obtained by propagating forward equations (8).

Remark 1 *The sequences $\mathscr{T}$, $\Theta_\mathcal{W}$, $\Theta_\mathcal{Q}$, $\mathscr{X}_\mathcal{W}^+$ and $\mathscr{X}_\mathcal{Q}^+$, which define the switches of $\mathcal{W}$ and $\mathcal{Q}$, can either be assumed to be defined offline a priori, or can be independently computed online by $\mathcal{W}$ and $\mathcal{Q}$ in a way to guarantee synchronicity of the two. Both approaches are acceptable in practice, and are perfectly equivalent with respect to the scope and goals of the present work.*

3.2 Watermarking Scheme Design Principles

In the following, we consider the watermarking scheme's effect on the closed-loop system in the absence of attacks. As opposed to existing additive watermarking approaches such as Mo et al (2015), our aim is to design the watermark generator and remover so that there exists no performance degradation in the absence of attacks. In order to meet such a goal, the watermark generator and remover must essentially act as an encoder and decoder, respectively, where one is the inverse function of the other. Hence, the following design rules are made.

Assumption 2 *The sequences of parameter vectors Θ_w and Θ_q and the dependence of the matrices A_w, B_w, C_w and D_w on θ_w and of the matrices A_q, B_q, C_q and D_q on θ_q are such that, for every instant k:*

1. *$\mathcal{W}$ and $\mathcal{Q}$ are stable and invertible;*
2. *the inverses of $\mathcal{W}$ and $\mathcal{Q}$ are stable;*
3. *$\theta_w = \theta_q$ implies that $\mathcal{Q}$ is the inverse of $\mathcal{W}$.*

Naturally, these conditions are necessary ones, as they ensure that $y_q[k]$ converges to $y_p[k]$ asymptotically. Moreover, these design rules may be trivially met by the

following choice of state-space representation

$$\begin{aligned} &D_q C_w + C_q = 0, \quad B_q D_w = B_w, \quad D_q D_w = 1, \\ &A_q + B_q C_w = A_w, \quad B_q C_w = A_q - B_w C_q, \end{aligned} \tag{9}$$

as concluded by examining the cascade of the generator and remover, and its state-space representation.

However, the above conditions are not sufficient, as they only provide an asymptotic result. An additional condition must be posed on the initialization of the internal states at each switching times, so that no transient behavior is induced by the watermarking scheme.

Assumption 3 *The switching times, and the corresponding jump updates, are designed such that, for every switching instant $k^i \in \mathscr{T}$:*

1. $\theta_w[k^i] = \theta_q[k^i]$;
2. $x_w^+[k^i] = x_q^+[k^i]$.

The first condition ensures that the generator and remover are synchronous and simultaneously update their parameters to the same value. The second condition, together with the state-space description described in (9), guarantees that no transient mismatch occurs between the internal states of the generator and the remover. Consequently, examining the cascade system $\mathfrak{QW}$ under these conditions, one concludes that the relation $y_p[k] = y_q[k]$ holds true for all time instants, which in turn implies that the multiplicative watermarking scheme is transparent under no-attack conditions, and does not affect the closed-loop system operation. For a detailed formal proof the reader is invited to refer to Ferrari and Teixeira (2018).

3.3 Stability Analysis

In this section, we investigate the stability of the closed-loop system with the proposed watermarking scheme when Assumption 3 is not satisfied, and in the absence of attacks. Since the controller design is oblivious to the mismatch between the filters, determining stability of the closed-loop system with mismatched filter parameters is a robust stability problem with multiplicative model uncertainty, where the uncertainty is in fact a hybrid system.

In the following, we restrict our attention to the inter-switching times (i.e., with constant mismatched parameters), during which the uncertainty behaves as a linear time-invariant system. We start by formulating the nominal system and the uncertainty under analysis.

The key steps in our stability analysis are to first rewrite the closed-loop system with mismatched filters as the nominal closed-loop system (without filter) connected in feedback with a system composed of the mismatched filters. Second, to apply classical robust stability results to the feedback system, in terms of the $\mathcal{H}_\infty$ norm of each sub-system.

The first step is accomplished by rewriting $\tilde{y}_p[k] = y_q[k]$ as $\tilde{y}_p[k] = y_p[k] + \Delta y_q[k]$, where $\Delta y_q[k]$ is described by

$$\mathcal{D}(\theta_w, \theta_q) \begin{cases} \begin{bmatrix} x_w[k+1] \\ x_q[k+1] \end{bmatrix} = \begin{bmatrix} A_w(\theta_w) & 0 \\ B_q(\theta_q)C_w(\theta_w) & A_q(\theta_q) \end{bmatrix} \begin{bmatrix} x_w[k] \\ x_q[k] \end{bmatrix} + \\ \qquad \begin{bmatrix} B_w(\theta_w) \\ B_q(\theta_q)D_w(\theta_w) \end{bmatrix} y_p[k] \\ \Delta y_q[k] = \begin{bmatrix} D_q(\theta_q)C_w(\theta_w) & C_q(\theta_q) \end{bmatrix} \begin{bmatrix} x_w[k] \\ x_q[k] \end{bmatrix} + \\ \qquad \big(D_q(\theta_q)D_w(\theta_w) - I_{n_y}\big)\, y_p[k]. \end{cases} \tag{10}$$

Note that the system $\mathcal{D}(\theta_w, \theta_q)$ has $y_p[k]$ as its input, and $\Delta y_q[k]$ as its output. Furthermore, observe that, under Assumption 3 and the relations in (9), we have $\Delta y_q[k] = 0$ for all k, which corroborates the statements at Section 3.2 that matched watermarking filters do not affect the performance of the closed-loop system.

Next, recalling that $\tilde{y}_p[k] = y_p[k] + \Delta y_p[k]$, we consider the nominal closed-loop system (1) as seen from the input $\Delta y_q[k]$ to the output $y_p[k]$, which is denoted as $\mathcal{S}_{\Delta y_q, y_p}$.

Given the above definitions of $\mathcal{D}(\theta_w, \theta_q)$ and $\mathcal{S}_{\Delta y_q, y_p}$, we are now in place to follow the second step of the robust stability analysis. In fact, note that the perturbed closed-loop system can be described as the nominal closed-loop system, $\mathcal{S}_{\Delta y_q, y_p}$, interconnected through feedback with $\mathcal{D}(\theta_w, \theta_q)$. Defining $\gamma(\Sigma)$ as the $\mathcal{H}_\infty$-norm of a linear system Σ, the following stability result directly follows from classical results on robust stability (Zhou et al, 1996).

Theorem 1 (Ferrari and Teixeira (2020)). *Let the generator $\mathcal{W}$ and the remover $\mathcal{Q}$ be non-synchronized at a switching time instant k^i, and assume no future switching occurs. Then the closed-loop system and watermarking filters are robustly asymptotically stable if $\gamma\big(\mathcal{S}_{\Delta y_q, y_p}\big)\, \gamma\big(\mathcal{D}(\theta_w[k^i], \theta_q[k^i])\big) \leq 1$.*

Although Theorem 1 gives only a sufficient condition, it allows for a simpler design of the filter parameters, by imposing two $\mathcal{H}_\infty$-norm constraints for each pair of filter parameters. The next results formalize this statement.

Corollary 1 (Ferrari and Teixeira (2020)). *Let the generator $\mathcal{W}$ and the remover $\mathcal{Q}$ be non-synchronized at a switching time instant k_i, and assume no future switching occurs. Then the closed-loop system and watermarking filters are robustly asymptotically stable if $\mathcal{W}(z;\theta_i)$, $\mathcal{W}^{-1}(z;\theta_i)$, $\mathcal{W}(z;\theta_j)$, and $\mathcal{W}^{-1}(z;\theta_j)$ are stable for all choice of filter parameters $\theta_i, \theta_j \in \Theta$, and, for all $\theta_i, \theta_j \in \Theta$, $\theta_j \neq \theta_i$, the following frequency domain constraints are satisfied for all $z \in \mathbb{C}$ on the unit circle:*

$$|\,(\mathcal{W}(z;\theta_i) - \mathcal{W}(z;\theta_j))\,| \leq \gamma\big(\mathcal{S}_{\Delta y_q, y_p}\big)^{-1}\, |\mathcal{W}(z;\theta_j)|. \tag{11}$$

Note that these frequency domain inequalities ensuring robust stability could be enforced by requiring different parameters θ_i and θ_j to be sufficiently close, depending on the $\mathcal{H}_\infty$-norm of the nominal closed-loop system. On the other hand, to

enable the detection of the mismatch and replay attacks, one desires that the filter parameters are as different as possible. Therefore one must trade off robust stability and detectability of filter mismatches.

3.4 An application example

In order to illustrate the application of the proposed watermarking technique, we will now introduce an example. We will make use of an unstable LTI plant from the database maintained by Gazdoš et al. (2012) and, in particular, of the fluidized bed system discussed in Kendi and Doyle (1996). The plant is described there by the following unstable second-order transfer function

$$G_p(s) = \frac{1}{(s+0.8695)(s-0.0056)}, \tag{12}$$

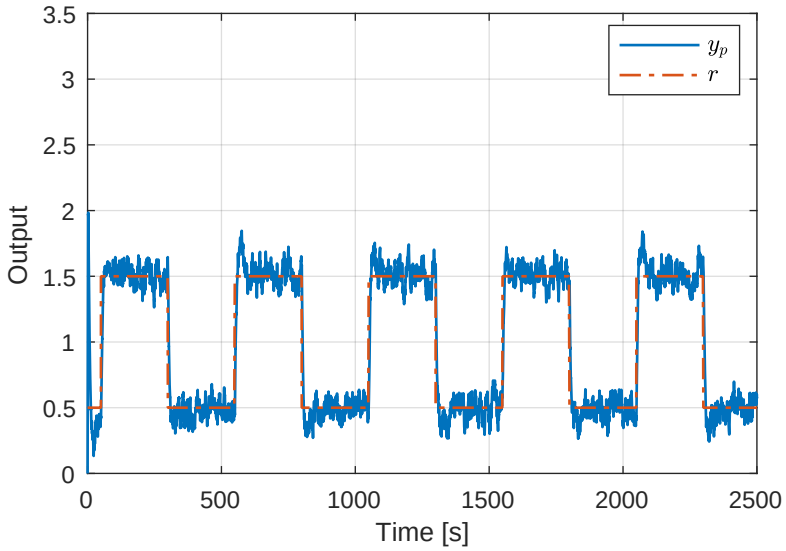
from which, after discretization with a sampling time $T_s = 0.1s$, the following state space realization can be obtained:

$$A_p = \begin{bmatrix} 1.9173 & -0.9172 \\ 1.0000 & 0 \end{bmatrix}, B_p = \begin{bmatrix} 0.125 \\ 0 \end{bmatrix}, C_p = \begin{bmatrix} 0.0389 & 0.0378 \end{bmatrix}. \tag{13}$$

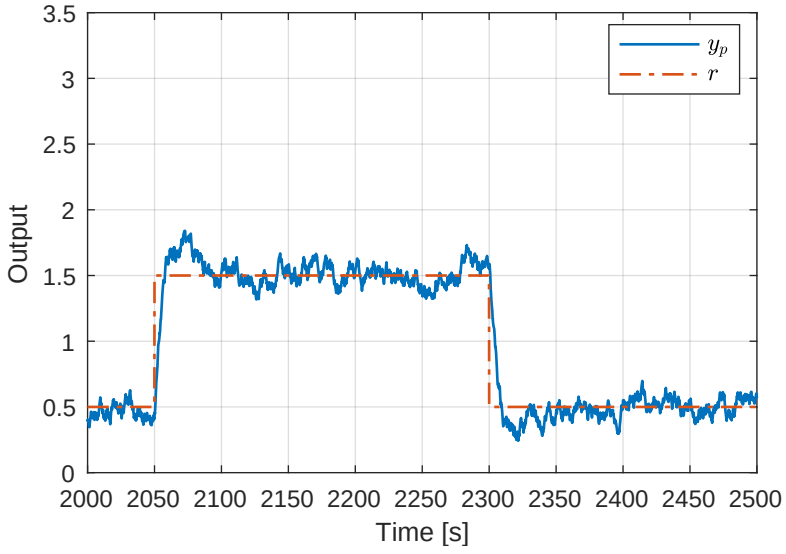
For stabilization and reference tracking, we designed a 1 degrees-of-freedom LQG servo-control law with integral action, leading to a controller the following state-space matrices:

$$\begin{aligned} A_c &= \begin{bmatrix} 1.4898 & -1.2937 & 0.001 \\ 0.6800 & -0.3109 & 0 \\ 0 & 0 & 1.000 \end{bmatrix}, B_c = \begin{bmatrix} -10.4646 \\ -8.2313 \\ 0.1000 \end{bmatrix} \\ C_c &= \begin{bmatrix} -0.1657 & 0.1504 & 0.0078 \end{bmatrix}. \end{aligned} \tag{14}$$

The controller was fed the input $e = \tilde{y}_p - r$, with r being a square wave reference signal switching between the values 0.5 and 1.5 with a period of 500 s and a duty cycle of 50 %. Finally, the uncertainties η_p and ξ_p were set to zero mean Gaussian random variables whose components' absolute values were capped at, respectively, $\bar{\eta}_p = 0.3$ and $\bar{\xi}_p = 0.15$. First of all, we will show the behaviour of the plant when no watermarking mechanism and no attacks are present. The reference signal with the corresponding plant output, the plant input and the tracking error are presented, respectively, in Figures 4, 5 and 6. From these figures, we observe that the controller is following the reference reasonably well, considering the non-negligible uncertainties in the model and measurements. The tracking error, as expected, shows positive and negative peaks corresponding to the reference rising and falling edges, and the control input has a similar behaviour.



(a)



(b)

Fig. 4 The reference signal and the plant output from the example considered throughout this chapter, when no watermark and no attack are present. The second plot shows a zoom in of the first one, during the last repetition of the periodic reference signal.

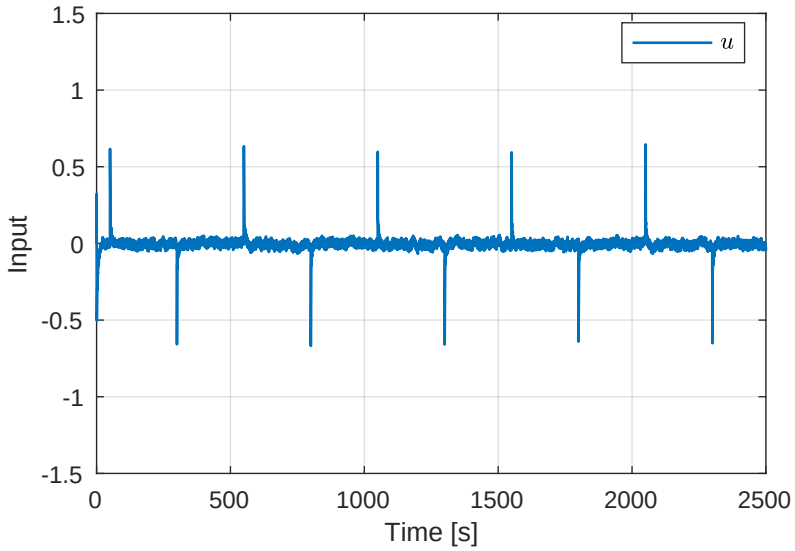


Fig. 5 The plant input from the example considered throughout this chapter, when no watermark and no attack are present.

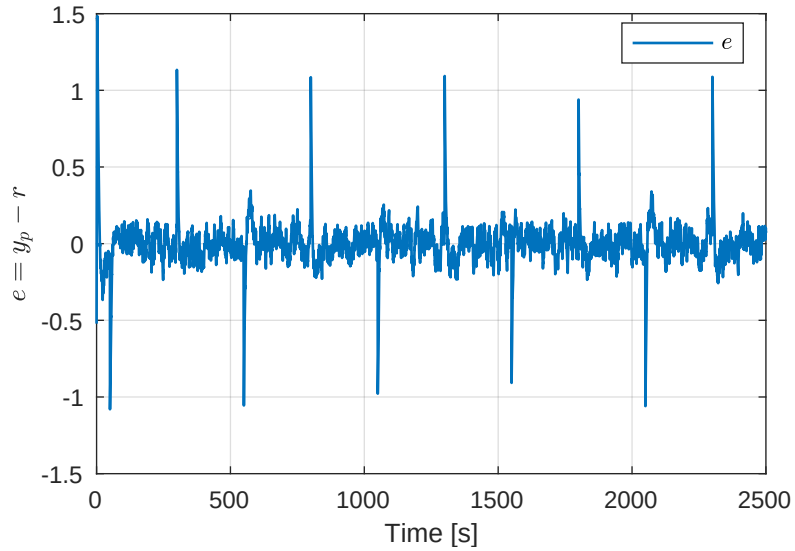


Fig. 6 The tracking error from the example considered throughout this chapter, when no watermark and no attack are present.

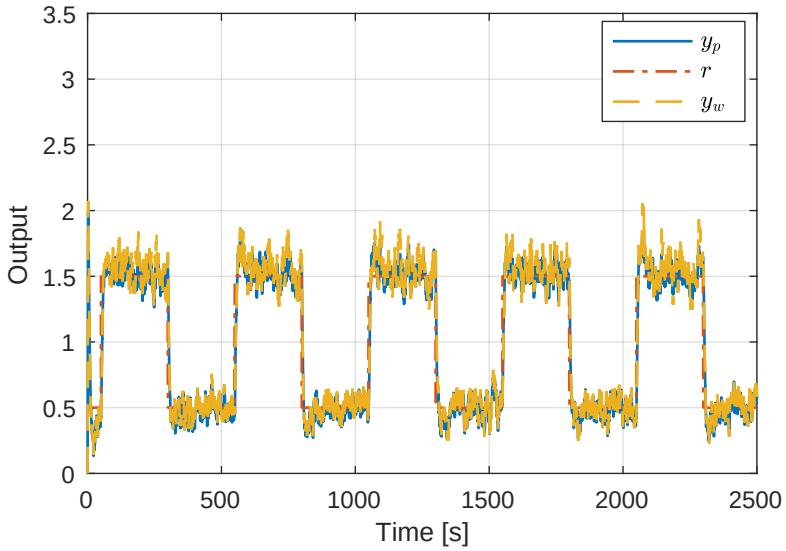
We will now show the effects of the presence of the watermark, still in absence of an attack. To do this, first we will produce a sequence of $N = 7$ random watermark generators and removers, whose parameters will make up the sequences $\Theta_{\mathcal{W}}$ and Θ_{Q}. In particular, the watermark generators will be chosen to be Finite Impulse Response (FIR) filters of order 3 and each filter transfer function will be defined as

$$\mathcal{W}(z) = w_{B,(1)} + w_{B,(2)}z^{-1} + w_{B,(3)}z^{-2} + w_{B,(4)}z^{-3}, \tag{15}$$

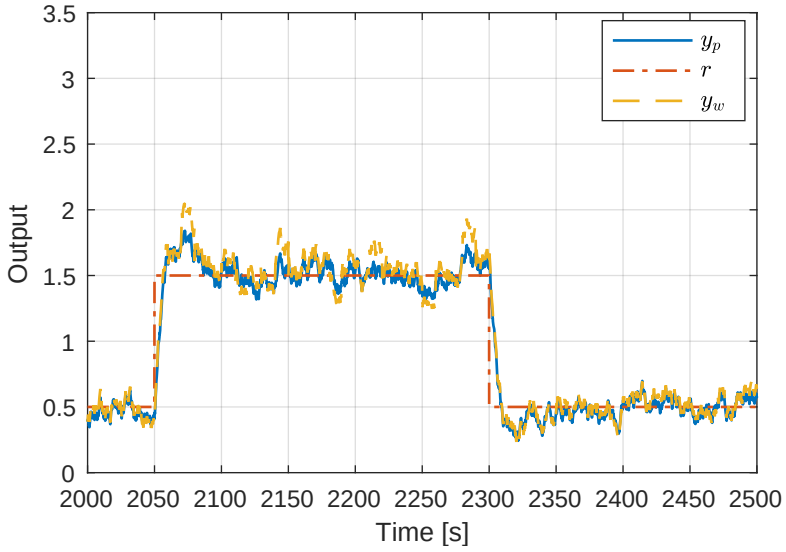
where z^{-1} denotes the unitary delay, $w_{B,(1)} = 1$ and $w_{B,(j)}$ with $j \in \{2, \ldots, 4\}$ are random numbers drawn from the interval $[-w_M \; w_M]$, different for each filter in the sequence. The scalar $w_M \in \mathbb{R}$ will be termed the watermark *magnitude*, and each filter parameter θ_w in the sequence $\Theta_{\mathcal{W}}$ can be interpreted as $\theta_w = w_B \in \mathbb{R}^4$.

We are thus ready to present the results of applying such watermarks to the example, following the closed loop scheme with watermark generation and removal presented in Figure 1. In the present example, as well as in the following simulations throughout this chapter, the watermark parameters are changed every $10\,s$, unless specified otherwise. This means that the switching time instants will be $k^1 = 100$, $k^2 = 200$ and so on, as the sampling time is equal to $0.1\,s$. After every $N = 7$ switching instants, the parameter sequence will cycle back and use again the first watermark parameter, such that during the entire simulation the watermark parameters will keep being switched.

In Figures 7 and 8, we can see that the watermark's presence is barely noticeable when setting its amplitude to $w_M = 5\%$. Most of all, the watermark remover does properly recover the true output such that the control performances will stay exactly the same as in the non-watermarked case.

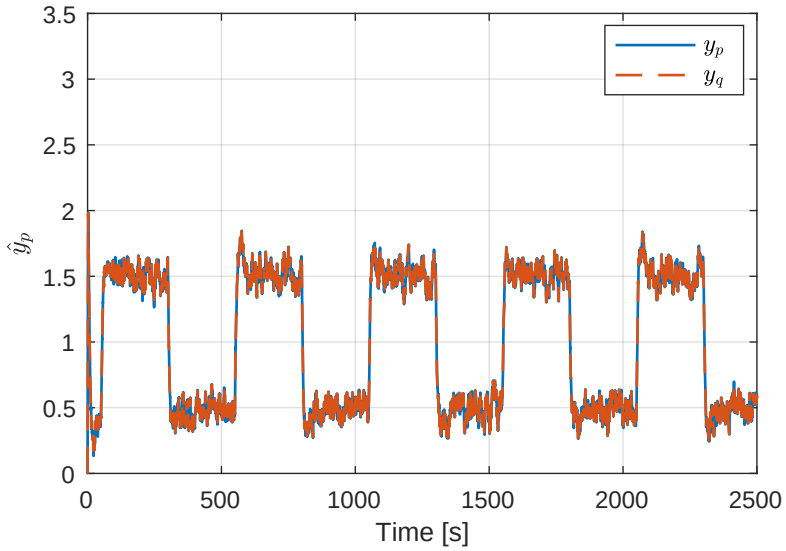


(a)

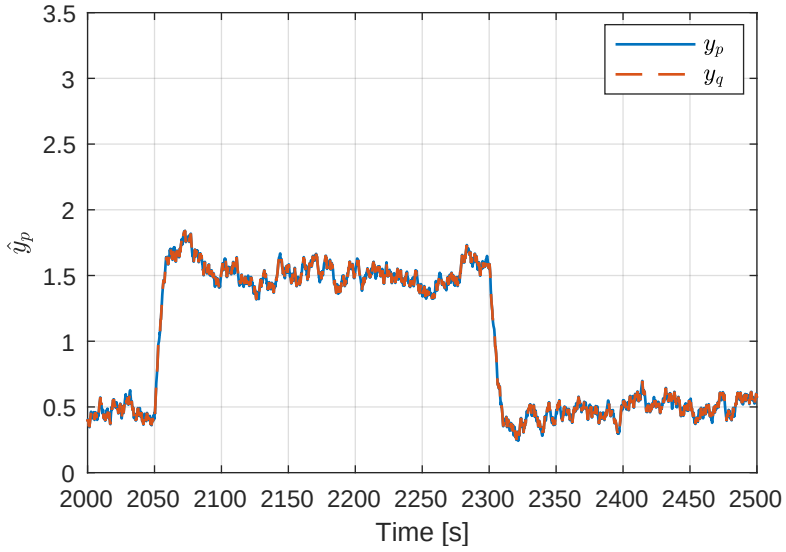


(b)

Fig. 7 The reference signal, the true plant output and the watermarked one from the example considered throughout this chapter, when no attack is present and a watermark of 5% amplitude is present. The second plot shows a zoom in of the first one, during the last repetition of the periodic reference signal.

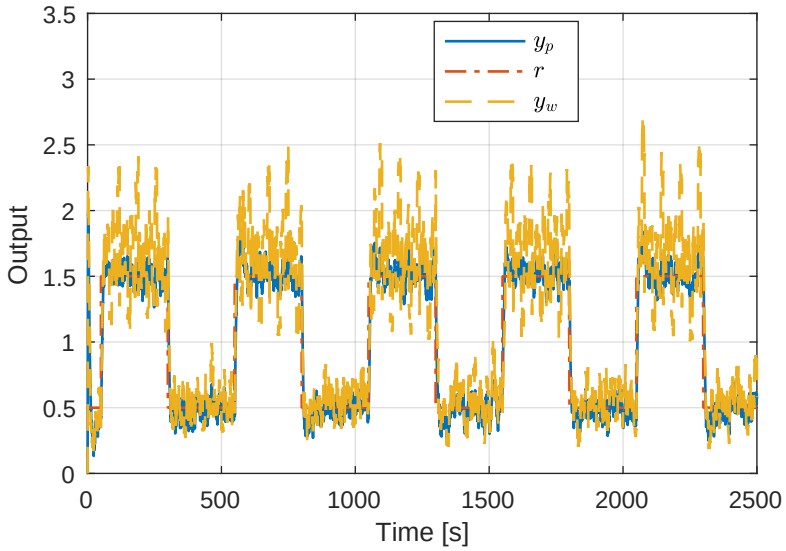


(a)

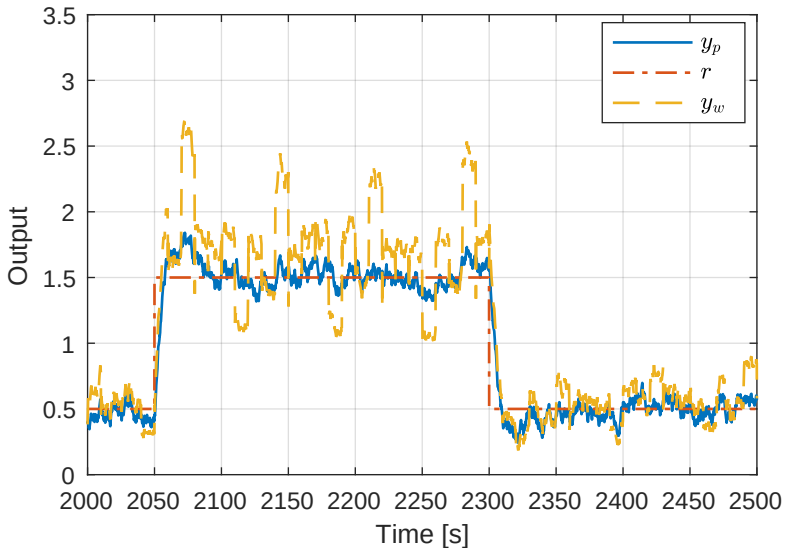


(b)

Fig. 8 The true plant output and the reconstructed one from the example considered throughout this chapter, when no attack is present and a watermark of 5% amplitude is present. The second plot shows a zoom in of the first one, during the last repetition of the periodic reference signal.

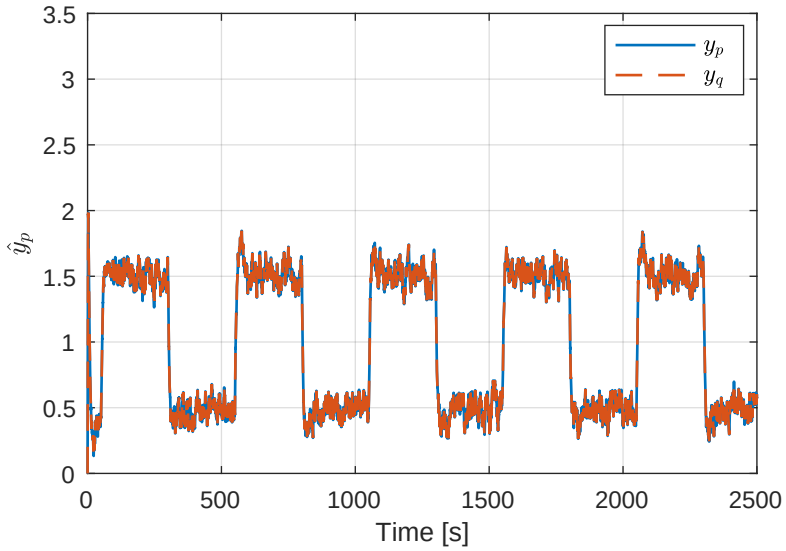


(a)

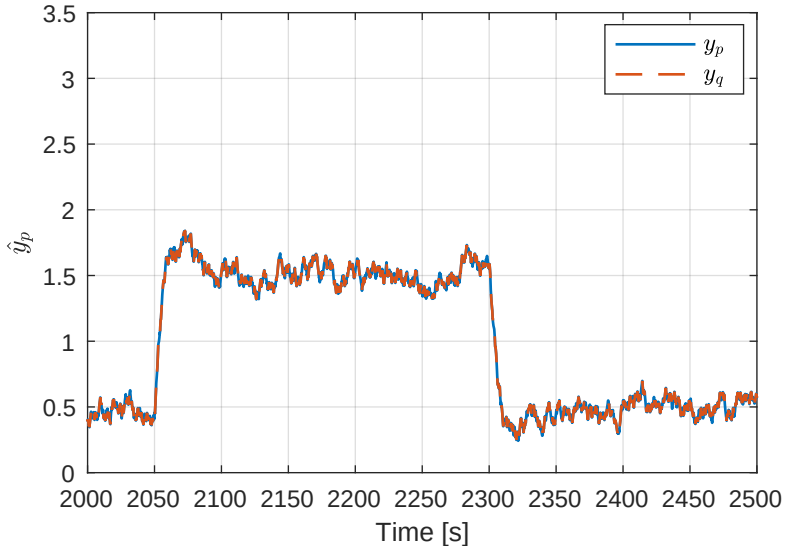


(b)

Fig. 9 The reference signal, the true plant output and the watermarked one from the example considered throughout this chapter, when no attack is present and a watermark of 20% amplitude is present. The second plot shows a zoom in of the first one, during the last repetition of the periodic reference signal.



(a)



(b)

Fig. 10 The true plant output and the reconstructed one from the example considered throughout this chapter, when no attack is present and a watermark of 20% amplitude is present. The second plot shows a zoom in of the first one, during the last repetition of the periodic reference signal.

If we increase the watermark amplitude to $w_M = 20\%$ its presence, as well as the switching times, become clearly apparent (see Fig. 9). Still, the reconstructed output

y_{pq} produced by the watermark remover continues to match exactly the true output y_p, and again the control performances will be unaffected (see Fig. 10).

In the next sections, we derive the conditions under which the attacks are detectable thanks to the multiplicative watermarking scheme. 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.

4 Detection of stealthy false data injection attacks

Having defined all the elements illustrated in Fig. 1, and characterized the essential design rules in normal conditions, the behavior of the watermarking scheme under attack is now examined. Proofs of results are omitted for the sake of brevity, but can be found in Ferrari and Teixeira (2018).

As a first step, we describe the full dynamics of the closed-loop system with watermarking, by having the following equations at the plant's side:

$$\begin{aligned} \mathcal{P}: &\begin{cases} x_p[k+1] = A_p x_p[k] + B_p u[k] + \eta_p[k] \\ y_p[k] = C_p x_p[k] + \xi_p[k] \end{cases} \\ \mathcal{W}(\theta_w): &\begin{cases} x_w[k+1] = A_w(\theta_w) x_w[k] + B_w(\theta_w) y_p[k] \\ y_w[k] = C_w(\theta_w) x_w[k] + D_w(\theta_w) y_p[k]. \end{cases} \end{aligned} \tag{16}$$

The sensors transmit over a network the watermarked data $y_w[k]$, which may be corrupted en-route by an adversary and be replaced by $\tilde{y}_w[k]$.

At the controller side of the network, the residual and control input are computed from the received data $\tilde{y}_w[k]$ as

$$\begin{aligned} \mathcal{Q}(\theta_q): &\begin{cases} x_q[k+1] = A_q(\theta_q) x_q[k] + B_q(\theta_q) \tilde{y}_w[k] \\ y_q[k] = C_q(\theta_q) x_q[k] + D_q(\theta_q) \tilde{y}_w[k] \end{cases} \\ \mathcal{F}_{cr}: &\begin{cases} x_{cr}[k+1] = A_{cr} x_{cr}[k] + B_{cr} y_q[k] \\ y_r[k] = C_{cr} x_{cr}[k] + D_{cr} y_q[k] \\ u[k] = C_u x_{cr}[k] + D_u y_q[k], \end{cases} \end{aligned} \tag{17}$$

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).

As explained earlier, a key assumption in the present work is that the watermark parameters $\theta_w = \theta_q$ are unknown to the attacker. Thus, we investigate the detectability of the false-data injection attack $a[k]$ computed according to (7), based on the attacker knowing only the plant dynamics.

The core of the analysis can be explained as follows. We start by recalling that the adversary knows the plant model, and stages an undetectable attack by mimicking a possible behavior of the plant. However, under the multiplicative watermarking

scheme, we notice that the plant is augmented with the watermark generator, as described in (16). Similarly, as detailed in (17), we rest on the fact that the anomaly detector and the controller are augmented with the watermark remover at their input. Consequently, we can conclude that while the man-in-the-middle attack mimics the plant behavior without watermarking, the anomaly detector instead expects a behavior that is affected by the watermarking generator. This mismatch is what allows for the detection of (previously) undetectable attacks. We will formalize this intuitive explanation in the remaining part of this chapter.

The main result of this section is the following, 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 2. *Consider the plant with sensor watermarking described in (16), 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, $\tilde{y}_w[k] = y_w[k] + a[k]$, where $a[k]$ is characterized by (7) with $\bar{x}_a$ being an eigenvector of A_p associated with the eigenvalue $\lambda \in \mathbb{C}$. Define the channel 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 $y_q[k]$ if, and only if,*

$$\mathcal{Q}^i(\lambda) = \mathcal{Q}^j(\lambda), \ \forall i, j \in \textit{supp}(C_p \bar{x}_a). \tag{18}$$

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 result thus 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 watermarks for all sensors, as formalized next.

Corollary 2. *For single-output systems and for multiple-output systems with homogeneous watermark filters, i.e. $A_w^i = A_w^j$, $B_w^i = B_w^j$, $C_w^i = C_w^j$ and $D_w^i = D_w^j$ 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 $y_q[k]$.*

Despite such limitations, there is another degree of freedom that may be leveraged to make the attack ε-stealthy, and therefore detectable, even when (18) is satisfied, such as in the cases of Corollary 2. 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 Ferrari and Teixeira (2017a), 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 3. *Consider the plant with sensor watermarking described in (16), 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, $\tilde{y}_w[k] = y_w[k] + a[k]$, where*

$a[k]$ is characterized by (7) 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 $\mathcal{Q}^i(\lambda) = \alpha$, $\forall i \in$ supp$(C_p \bar{x}_a)$, for some $\alpha \neq 0$, and define $\bar{x}^a_{wq}$ such that $\left[\alpha \bar{x}_a^\top \; \bar{x}^{a\top}_{wq} \; \bar{x}_a^\top\right]^\top$ satisfy the PBH unobservability test from Zhou et al (1996).

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

$$\begin{aligned} \Delta x_{wq}[k+1] &= A_q \Delta x_{wq}[k] \\ y_q[k] &= D_q C_w \Delta x_{wq}[k] \end{aligned} \tag{19}$$

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

Once the sensor data attack is made detectable through a multiplicative watermarking, the compromised sensors can be isolated through conventional FDI techniques (Hwang et al, 2010) or approaches tailored to detect sparse sensor attacks (Fawzi et al, 2014). For instance, in Ferrari and Teixeira (2018) the following estimator is introduced for attack detection:

$$\hat{\mathcal{P}} : \begin{cases} \hat{x}_p[k+1] = A_p \hat{x}_p[k] + B_p u[k] + K\left(y_q[k] - \hat{y}_p[k]\right) \\ \hat{y}_p[k] = C_p \hat{x}_p[k], \end{cases} \tag{20}$$

where $\hat{x}_p \in \mathbb{R}^{n_p}$ and $\hat{y}_p \in \mathbb{R}^{n_y}$ are the estimates of x_p and y_p, and K is chosen such that $A_r \triangleq A_p - KC_p$ is Schur. By defining $x_r = \hat{x}_p$ and $\varepsilon \triangleq x_p - \hat{x}_p$, in no attack conditions the detection residual $y_r \triangleq y_q - \hat{y}_p$ can be written as the solution to the following dynamical system

$$\begin{cases} \varepsilon[k+1] = A_r \varepsilon[k] - K\xi_p[k] + \eta_p[k] \\ y_r[k] = C_p \varepsilon[k] + \xi_p[k] \end{cases}. \tag{21}$$

The last equation, thanks to the assumed knowledge on the upper bounds of the uncertainties, can be used to compute the detection threshold $\hat{y}_r$.

When, instead, an attack is present the detector will use the attacked signal $\tilde{y}_q$ for implementing the detection estimator in eq. (20). This will lead to the residual being instead the solution of this system:

$$\begin{cases} \tilde{\varepsilon}[k+1] = A_r \tilde{\varepsilon}[k] - K(\xi_p[k] + \delta_a[k]) + \eta_p[k] \\ y_r[k] = C_p \tilde{\varepsilon}[k] + \xi_p[k] + \delta_a[k] \end{cases}, \tag{22}$$

where the term δ_a, called *attack mismatch*, can be obtained from the following

Lemma 1 (Ferrari and Teixeira (2018)). *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*

$$\begin{aligned} \begin{bmatrix} x_q[k+1] \\ x_a[k+1] \end{bmatrix} &= \begin{bmatrix} A_q & B_q C_q \\ 0 & A_p \end{bmatrix} \begin{bmatrix} x_q[k] \\ x_a[k] \end{bmatrix} \\ \delta_a[k] &= \begin{bmatrix} C_q & (D_q - I)C_p \end{bmatrix} \begin{bmatrix} x_q[k] \\ x_a[k] \end{bmatrix}, \end{aligned} \tag{23}$$

for all $k \geq k^*$, with $x_q[k^*] = 0$ and $x_a[k^*] = \lambda^{k^*-k_a}\bar{x}_a$ being the values at which x_q and x_a have been reset to at the last watermark switch.

The importance of the term δ_a is that it can drive the residual to larger values than those it would have because of the presence of the uncertainties ξ_p and η_p alone, thus possibly allowing for detection. It has been shown in Ferrari and Teixeira (2018), furthermore, that frequently switching the watermark parameters will help detection by continuously resetting δ_a dynamics.

In the following section, a numerical study is presented, which illustrates the problem of detecting an attack in a single output system, and shows how the use of a switched watermark can solve such challenge.

5 Numerical study

The numerical study uses the same example introduced in Section 3.4 to show the effects of a stealthy false-data injection attack, and how the combined use of switching watermarks and of the detection observer introduced in the previous section can lead to a successful detection.

The attack is defined as $a[k] = C_p A_p^{k-k_a}\bar{x}_a = \lambda^{k-k_a} C_p \bar{x}_a$, where $\lambda = 1.0006$ is the plant's unstable eigenvalue and $\bar{x}_a = -10^{-4} \times \begin{bmatrix} 0.7073 & 0.7069 \end{bmatrix}$ is an initial condition aligned with the corresponding eigenvector. The attack starts at time $T_a = k_a \cdot T_s = 30\,\mathrm{s}$ and, as it can be seen from Fig. 11, its magnitude begins to be comparable to the reference signal at about $2000\,s$.

When no watermarking is used, the exponentially increasing attack signal causes the true plant output y_p to diverge, while the received output $\tilde{y}_p$ appears to follow the square wave reference faithfully (see Fig. 12). The input signal does not show any anomalous behaviour (see Fig. 13) and the residual, too, does not reveal any sign of the attack as it is well below the threshold (see Fig. 14). Indeed, when only the detection observer introduced in the previous section is used and no watermarking is present, this attack is 0-stealthy.

The addition of a watermark with amplitude $w_M = 5\%$, but with constant parameters that are not switched, does not lead to detection as shown in Figs. 15 and 16. This corresponds to the case encompassed by Corollary 2, but it can fortunately be avoided by introducing switching parameters. By choosing the reset states of $\mathcal{W}$ and $\mathcal{Q}$ according to Theorem 3 the attack can be made ε-stealthy only and, as such, detectable.

Indeed, by looking at Figs. 17 and 18, we can see that the watermark switching will introduce significant peaks in both the reconstructed output y_q and the resid-

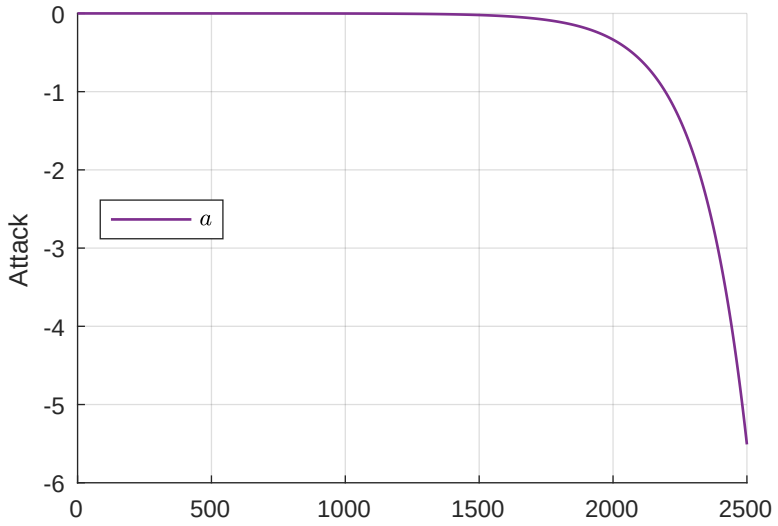


Fig. 11 The attack signal used in the numerical study.

ual y_r, whose amplitude increase with the attack magnitude, ultimately leading to detection.

Finally, in case the watermark amplitude is raised to $w_M = 20\%$, the peaks in both the reconstructed output y_q and the residual y_r are even larger than in the previous case, leading to detection at an earlier time instant (Figs. 19 and 20).

Finally, the detection capabilities of the proposed watermarking scheme for the different cases presented here will be quantitatively presented in Table 5. In particular, the detection time, the ratio between the residual and the threshold at detection and the attack amplitude at detection will be used as indexes for defining the scheme performance. From such results, it can be concluded that a switching watermark with large amplitude will lead to better detection performances, although the large amplitude will cause the watermark to be apparent to an adversary that is eavesdropping the signal y_w.

6 Conclusions

Inspired in authentication techniques with weak cryptographic guarantees, we have proposed a multiplicative watermarking scheme for networked control systems. In this scheme, each sensor's output is individually fed to a switching SISO watermark generator, which produces the watermarked data that is transmitted through the possibly unsecured communication network. At the controller's side, the watermark

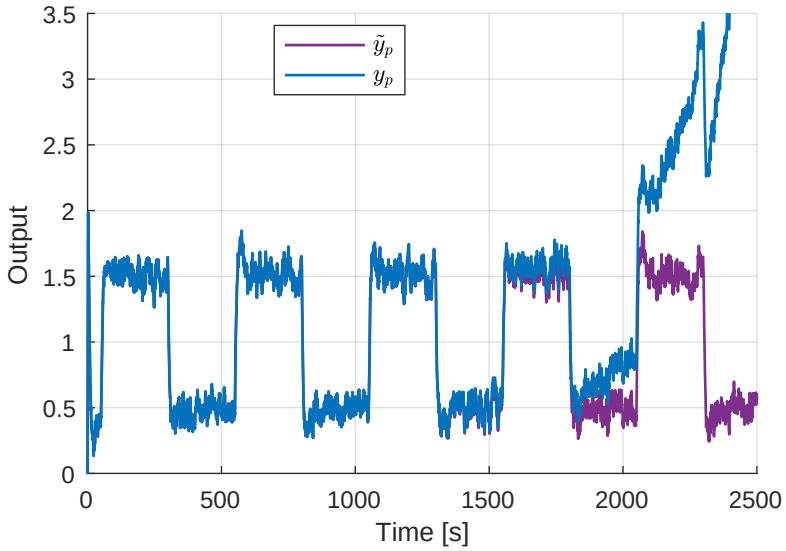


Fig. 12 The true plant output and the received one during an attack, when no watermark is in place.

Table 1 Performance of different watermarking strategies.

index	none	small		large	
		sw.	no sw.	sw.	no sw.
$k_d \cdot T_s$	N/A	2010 s	N/A	1870 s	N/A
$\frac{\|y_r[k_d]\|}{\bar{y}_r[k_d]}$	N/A	1.01	N/A	1.31	N/A
$\left\|\frac{a[k_d]}{y_p[k_d]}\right\|$	N/A	0.43	N/A	0.30	N/A

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; "small", amplitude is $w_M = 5\%$; "large", amplitude is $w_M = 20\%$; "sw.", parameters switched every 10 s; "no sw.", fixed parameters. "N/A" indicates that no detection occurred during simulation time.

remover reconstructs the original measurement data. This approach, combined with a model-based anomaly detector, is shown to lead to detection of otherwise stealthy false-data injection attacks. In particular, the periodic switching of the watermark generator and remover parameters are key to a successful detection. An application example, as well as theoretical results guaranteeing the absence of control performance losses and characterizing the scheme's detectability condition, are provided. Finally, simulation results illustrate the proposed approach and give insight into the correlation between the watermark magnitude and the attack detection performances. In the future, an extension to the case of nonlinear plant dynamics, as well as nonlinear watermarks, could significantly augment the scheme applicability as

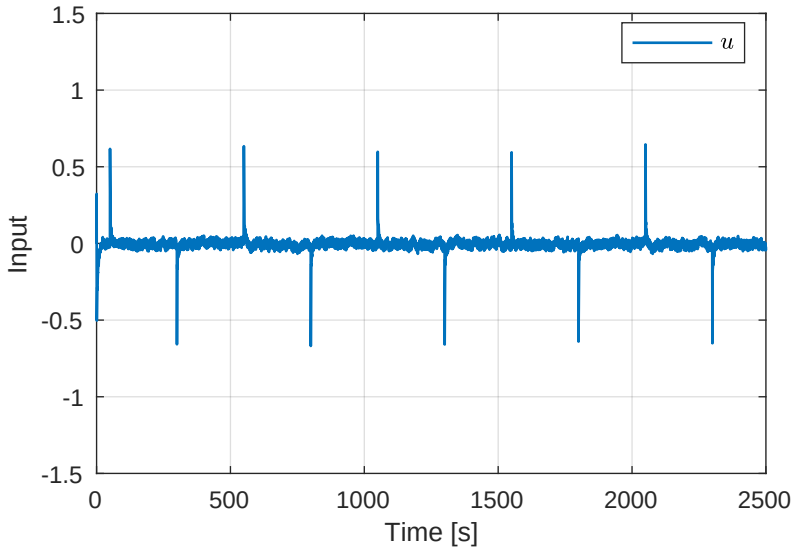


Fig. 13 The plant input during an attack, when no watermark is in place. The input signal in this case is indistinguishable from the case when no attack is present (Fig. 5).

well as its resilience against advanced adversaries that may try to reverse-engineer the watermarking scheme.

Acknowledgements This work is financed by the Swedish Foundation for Strategic Research, the Swedish Research Council under the grant 2018-04396, the European Union Seventh Framework Programme (FP7/2007-2013) under grant no. 608224, and by EU H2020 Programme under grant no. 707546 (SURE) .

References

Gazdoš et al F (2012) Database of unstable systems [Online], Available: http://www.unstable-systems.cz, 19/03/2020

Cárdenas AA, Amin S, Sastry SS (2008) Secure control: Towards survivable cyber-physical systems. In: First Int. Work. Cyber-Physical Syst.

Cárdenas AA, Amin S, Sinopoli B, Giani A, Perrig A, Sastry SS (2009) Challenges for securing cyber physical systems. In: Work. Futur. Dir. Cyber-physical Syst. Secur., U.S. DHS

Fawzi H, Tabuada P, Diggavi S (2014) Secure estimation and control for cyber-physical systems under adversarial attacks. IEEE Trans on Autom Control 59(6):1454–1467

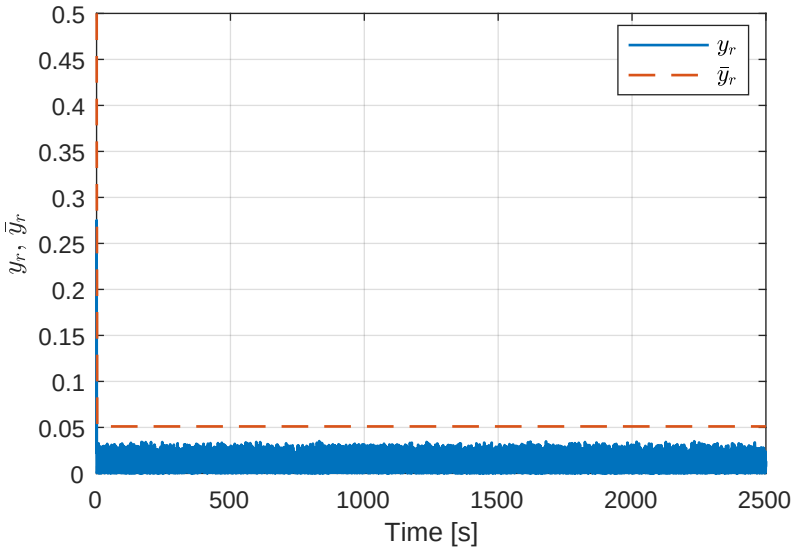


Fig. 14 The detection residual and threshold during an attack, when no watermark is in place. The residual shows no sign of the presence of the attack, and is always well below the threshold.

Ferrari RMG, Teixeira AMH (2017a) Detection and isolation of replay attacks through sensor watermarking. In: Procs. of 20th IFAC World Congress, Toulouse (France) July 9 - 14, 2017, IFAC

Ferrari RMG, Teixeira AMH (2017b) Detection and isolation of routing attacks through sensor watermarking. In: 2017 American Control Conference (ACC), pp 5436–5442

Ferrari RMG, Teixeira AMH (2017c) Detection and isolation of routing attacks through sensor watermarking. In: Proc. Am. Control Conf., IEEE, pp 5436–5442, DOI 10.23919/ACC.2017.7963800

Ferrari RMG, Teixeira AMH (2018) Detection of sensor data injection attacks with multiplicative watermarking. In: Procs. of European Control Conference (ECC 2018), Limassol (Cyprus) June 12 - 15, 2018

Ferrari RMG, Teixeira AMH (2020) A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. IEEE Trans Autom Control Conditionally accepted.

Gallo AJ, Turan MS, Boem F, Ferrari-Trecate G, Parisini T (2018) Distributed watermarking for secure control of microgrids under replay attacks. IFAC-PapersOnLine 51(23):182 – 187, 7th IFAC Workshop on Distributed Estimation and Control in Networked Systems NECSYS 2018

Goebel R, Sanfelice RG, Teel AR (2009) Hybrid dynamical systems. IEEE Control Systems 29(2):28–93

Gorenc B, Sands F (2018) The state of SCADA HMI vulnerabilities. URL https://documents.trendmicro.com/assets/wp/wp-hacker-

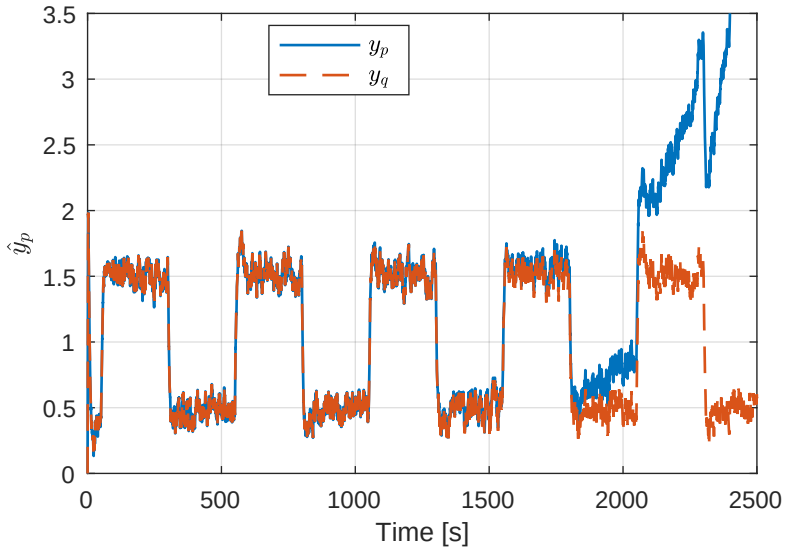


Fig. 15 The true plant output and the one reconstructed by $\mathcal{Q}$ during an attack, when a watermark with magnitude $w_M = 5\%$ is in place, but the watermark parameters are kept constant. The reconstructed output y_{pq} is indistinguishable from the attacked output $\tilde{y}_p$ received in the case where no watermark is present (Fig. 12), and from the reconstructed output in case no attack is present (Fig. 8). As y_{pq} is the signal used by the attack detector, it is by no surprise that no detection is possible in this case either.

machine-interface.pdf

Hwang I, Kim S, Kim Y, Eng C (2010) A survey of fault detection, isolation, and reconfiguration methods. Ieee Trans Control Syst Technol 18(3):18, DOI 10.1109/TCST.2009.2026285

Kendi TA, Doyle FJ (1996) Nonlinear control of a fluidized bed reactor using approximate feedback linearization. Industrial & Engineering Chemistry Research 35(3):746–757

Miao F, Zhu Q, Pajic M, Pappas GJ (2017) Coding schemes for securing cyber-physical systems against stealthy data injection attacks. IEEE Trans on Contr of Network Sys 4(1)

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

NCCIC, ICS-CERT (2016) ICS-CERT year in review 2016. URL https://ics-cert.us-cert.gov/sites/default/files/Annual_Reports/Year_in_Review_FY2016_Final_S508C.pdf

Pasqualetti F, Dorfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. IEEE Trans Automat Contr 58(11):2715–2729

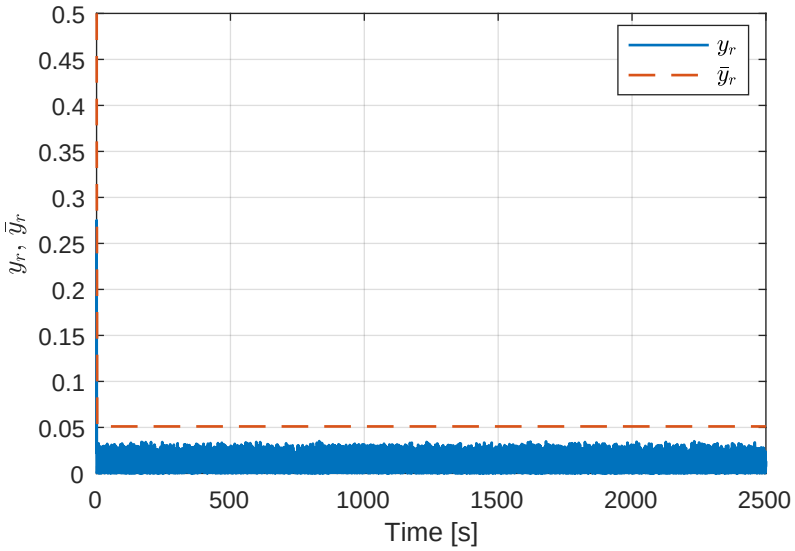


Fig. 16 The detection residual and threshold during an attack, when a watermark with magnitude $w_M = 5\%$ is in place, but the watermark parameters are kept constant. The residual behaves as in the case without watermark (Fig. 14) and is always well below the threshold.

Pérez-Freire L, Comesaña P, Troncoso-Pastoriza JR, Pérez-González F (2006) Trans. on Data Hiding and Multim. Security I, Springer Berlin Heidelberg, chap Watermarking Security: A Survey

Sandberg H, Amin S, Johansson KH (2015) Cyberphysical security in networked control systems: An introduction to the issue. IEEE Control Systems Magazine 35(1):20–23

Smith RS (2011) A decoupled feedback structure for covertly appropriating networked control systems. In: IFAC Proc. Vol., vol 18, pp 90–95

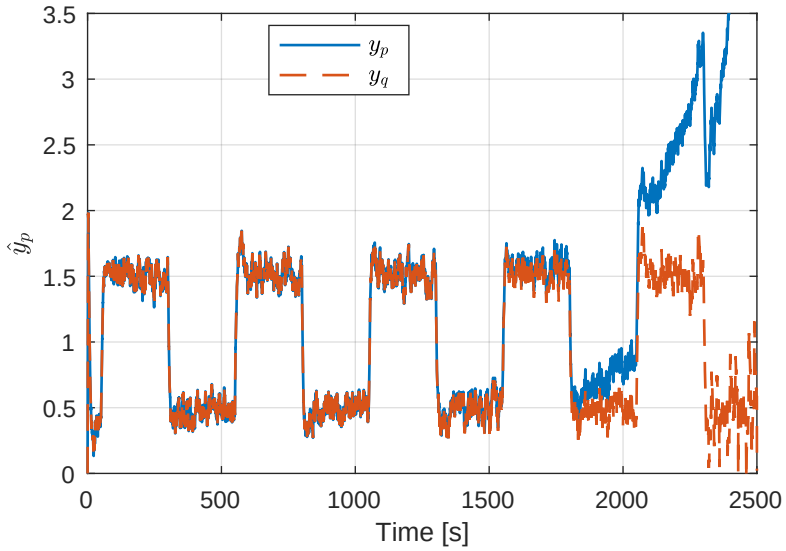
Teel AR, Poveda JI (2015) A hybrid systems approach to global synchronization and coordination of multi-agent sampled-data systems. IFAC-PapersOnLine 48(27):123–128

Teixeira A, Shames I, Sandberg H, Johansson KH (2012) Revealing stealthy attacks in control systems. In: 50th Annu. Allert. Conf. Commun. Control. Comput., DOI 10.1109/Allerton.2012.6483441

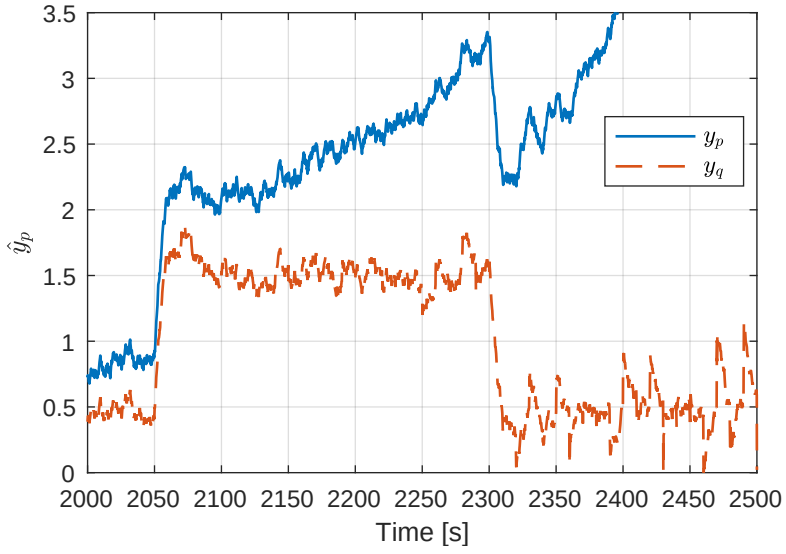
Teixeira A, Shames I, Sandberg H, Johansson KH (2015) A secure control framework for resource-limited adversaries. Automatica 51(1):135–148

Trend Micro (2018) Unseen threats, imminent losses 2018 midyear security roundup. URL https://documents.trendmicro.com/assets/rpt/rpt-2018-Midyear-Security-Roundup-unseen-threats-imminent-losses.pdf

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

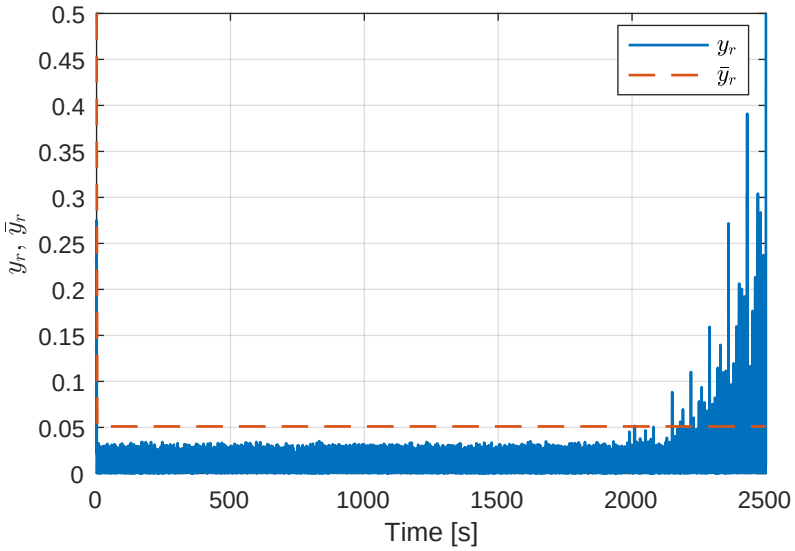


(a)

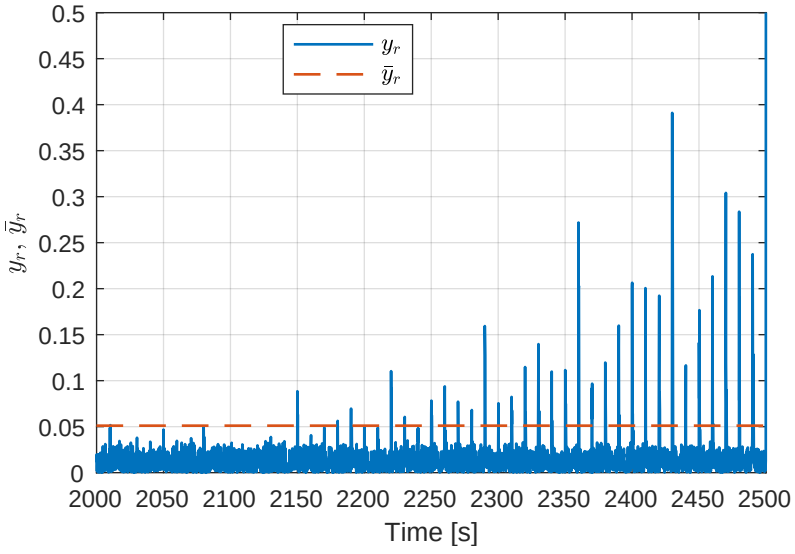


(b)

Fig. 17 (a) The true plant output and the one reconstructed by Ω during an attack, when a watermark with magnitude $w_M = 5\%$ is in place and the watermark parameters are switched every 10 s. (b) At a closer look, after about 2000 s the reconstructed output y_{pq} shows some noticeable differences from the non-attacked case. In particular, peaks in correspondence to the watermark switches, whose amplitude increases along the amplitude of the attack.

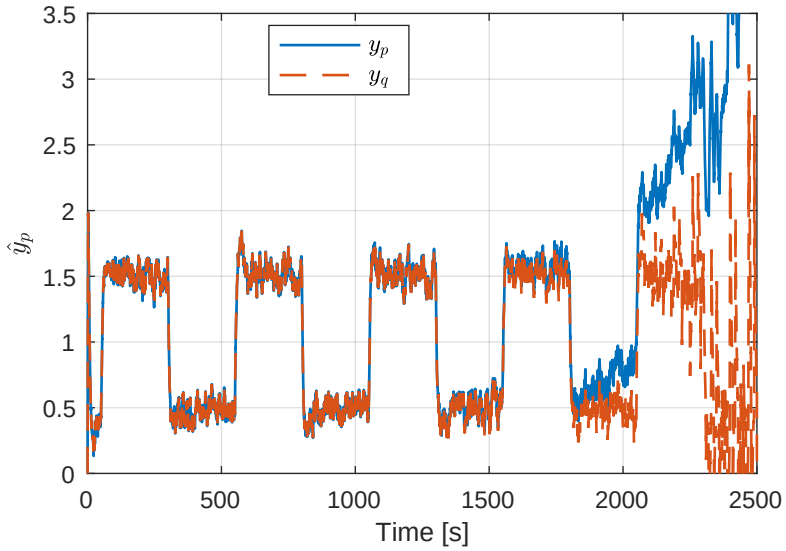


(a)

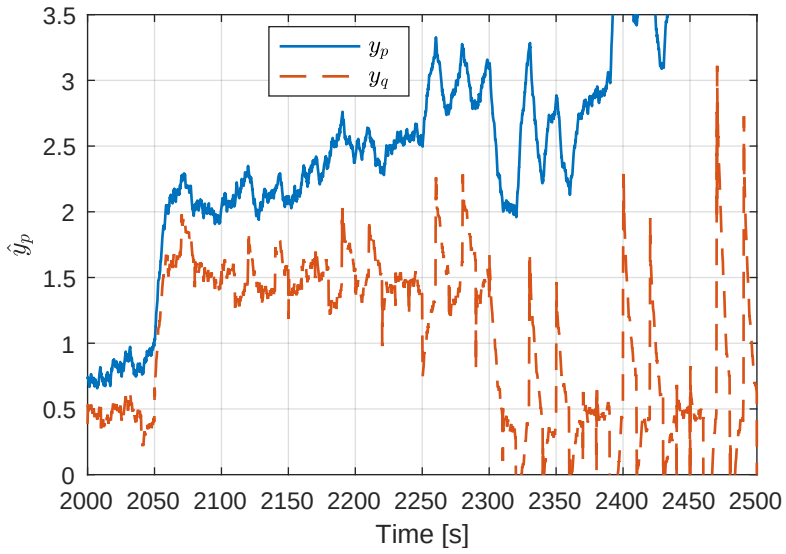


(b)

Fig. 18 (a) The detection residual and threshold during an attack, when a watermark with magnitude $w_M = 5\%$ is in place and the watermark parameters are switched every 10 s. (b) As we could have expected, after about 2000 s the residual is experiencing peaks of increasing magnitude synchronized with the watermark switches, which ultimately lead to detection.

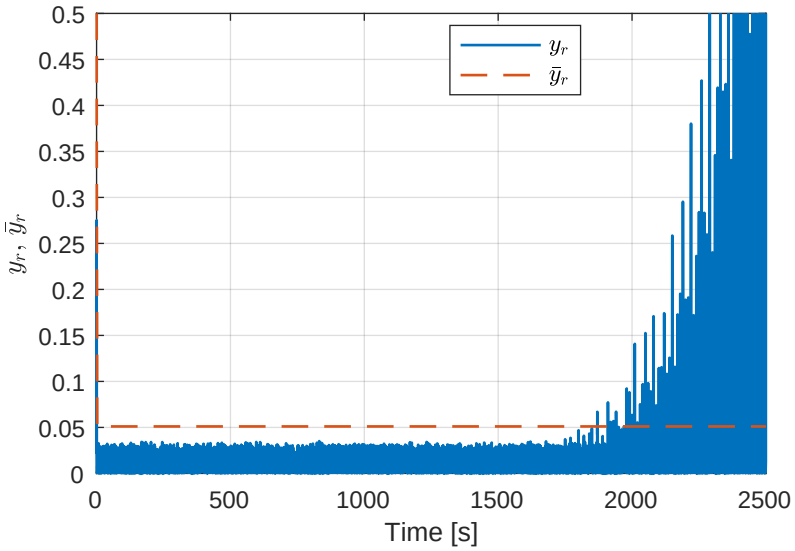


(a)

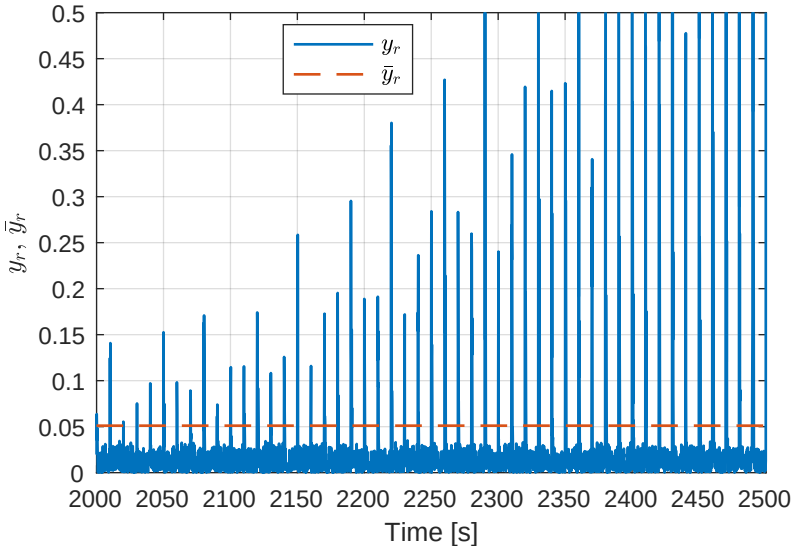


(b)

Fig. 19 (a) The true plant output and the one reconstructed by Ω during an attack, when a watermark with magnitude $w_M = 20\%$ is in place and the watermark parameters are switched every 10 s. (b) With respect to the case with amplitude $w_M = 5\%$ now the difference in the reconstructed output y_{pq} shows even larger differences from the non-attacked case.



(a)



(b)

Fig. 20 (a) The detection residual and threshold during an attack, when a watermark with magnitude $w_M = 20\%$ is in place and the watermark parameters are switched every 10 s. (b) In this case, the residual is showing even larger peaks with respect to the case with $w_M = 5\%$, which largely surpass the threshold already at 2000 s, leading to an earlier to detection.