Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-6 (2022) 433–438

# A Sensor Watermarking Design for Threat Discrimination

**Kangkang Zhang** [*] **Andreas Kasis** [*] **Marios M. Polycarpou** [*]
**Thomas Parisini** [*,**]

[*] *KIOS Research and Innovation Center of Excellence and Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus (e-mail: zhang.kangkang@ucy.ac.cy; kasis.andreas@ucy.ac.cy; mpolycar@ucy.ac.cy).*

[**] *Dept. of Electrical and Electronic Engineering, Imperial College London, London, SW7 2AZ, UK, and Dept. of Engineering and Architecture, University of Trieste, Trieste, 34127, Italy, (e-mail: t.parisini@gmail.com)*

**Abstract:** This paper proposes a sensor additive switching watermark methodology for detecting physical faults and replay attacks, and then identifying the occurring threat type: either the physical fault or the replay attack. The sensor watermark methodology includes a switching watermark generator in the plant side and a switching watermark remover in the control and monitoring side. The switching protocols of the generator and the remover and their common watermark seeds are specifically designed, guaranteeing the switch synchronization in both the nominal and fault cases, and allowing the switch asynchronization in the replay attack case. The latter is used for discriminating the two considered threat cases. The detectability and discrimination ability of the proposed watermark approach, characterizing the class of attacks and faults that can be detected and discriminated, is rigorously investigated. A simulation example is presented to illustrate the effectiveness of the proposed approach.

## 1. INTRODUCTION

Cyber-physical systems (CPS) integrate techniques from control, computation and communication areas (Cardenas et al., 2008). In addition to traditional physical faults, new security vulnerabilities such as cyber attacks, arise in CPS due to such integration. Considering traditional physical faults and malicious cyber attacks motivates the development of threat diagnostic technologies.

Replay attacks are more easily applicable than other integrity attacks such as zero-dynamic attacks and covert attacks (Smith, 2011) due to their simplicity. The Stuxnet attack on the Iranian nuclear facilities is a replay attack (Dibaji et al., 2019). In a replay attack event, the attacker first obtains access to the communication network and records data from the normal operation and then replays the data to the supervisory system. Compared with other integrity attacks, no system knowledge is required, which reduces the resources required by the attacker and simplifies its implementation. In addition, replay attacks possess high stealthiness since the malicious data are taken from the normal system operation.

Watermark-based active attack detection approaches have been widely studied (Gallo et al., 2021; Weerakkody et al., 2017; Porter et al., 2021). A typical watermarking detection methodology includes a watermark added to the control input of the physical plant, which is compared with the measured watermark in the control and monitoring (CM) side. The difference between the added watermark and the measured one is used to indicate the occurrence of a replay attack. In (Mo and Sinopoli, 2009; Mo et al., 2013), a stochastic zero-mean time-stamped watermark is injected into the control input, resulting in the increase of the replay attack detection rate. Such a typical watermarking detection approach causes control performance degradation since watermarks are injected into the system control input. An online learning watermark is designed in (Liu et al., 2018) to achieve the optimal trade-off between the control performance and the attack detection rate. A sensor additive watermarking detection approach is proposed in (Gallo et al., 2018), in which a watermark is added to the sensor measurements at the physical plant side, and is removed at the CM side. Such a methodology is able to detect replay attacks and at the same time, does not affect the control performance in the nominal case. In addition, a sensor multiplicative switching watermark is proposed in (Ferrari and Teixeira, 2020), in which a switching watermark generator in the physical plant side and a switching watermark remover in the CM side are included. The switches in the generator and the remover are synchronized in the normal operation and thus, do not

[⋆] This work has been supported by: European Union's Horizon 2020 Research and Innovation Program (Grant No. 739551 (KIOS CoE) and No. 891101 (SmarTher Grid)) and from the Republic of Cyprus through the Directorate General for European Programs, Coordination, and Development, the Italian Ministry for Research in the framework of the 2017 Program for Research Projects of National Interest (PRIN) (grant no. 2017YKXYXJ), the National Natural Science Foundation of China (Grant No. 61903188), the Natural Science Foundation of Jiangsu Province (Grant No. BK20190403) and the China Postdoctoral Science Foundation (Grant No. 2019M650114).

10.1016/j.ifacol.2022.07.167

affect the control performance. In addition, the asynchronization of the switches is used to indicate the occurrence of a replay attack. The switch synchronization is guaranteed by some specific information sharing channels in (Gallo et al., 2018) while the switching protocols are used to synchronize the switches in (Ferrari and Teixeira, 2020).

Threat discrimination aims to identify the threat type, namely determine what type of threat (physical faults or attacks) occurs, which helps the operator in making correct decisions and taking suitable remedial actions. In the aforementioned watermark detection approaches, the discrimination issues between general physical faults and replay attacks are overseen. In addition, typical anomaly detectors such as the ones in (Ding, 2008) may be not efficient in distinguishing physical faults or replay attacks. The threat discrimination problem between replay attacks and physical sensor bias faults remains an open problem, and few results have been published. A discrimination methodology for distinguishing between sensor replay attacks and physical faults is proposed in (Zhang et al., 2021), in which only constant sensor bias faults can be identified. This paper considers more general physical faults, including process faults, actuator faults and sensor faults, and aims to discriminate between general physical faults and sensor replay attacks.

Inspired by the works in (Gallo et al., 2018; Ferrari and Teixeira, 2020), this work proposes a sensor additive switching watermark methodology for detecting replay attacks and general physical faults, and identifying the occurring threat type: either replay attacks or physical faults. The proposed methodology includes a switching watermark generator in the physical plant side, and a switching watermark remover in the CM side that removes the effects on the CPS of the generator. The main contributions of this work are:

- A *two-layer decision strategy* is proposed for threat detection and discrimination, including two adaptive thresholds associated with the nominal and fault scenarios respectively.
- A set of switching protocols and the common seeds of the generator and remover are designed such that the switches in the generator and the remover are synchronized in the nominal and faulty cases. Furthermore, the seeds are designed to allow the switch asynchronization, which enables the detection of the replay attack case.
- The detectability and discrimination ability of the proposed approach are rigorously investigated.

Proofs of the main analytic results are removed due to space restrictions and will be provided in an extended version of this work.

The rest of this paper is organized as follows. In Section 2, we present the problem formulation. In Section 3, the details of the detection and discrimination methodology are presented and in Section 4, a simulation example is given. Finally, conclusions are drawn in Section 5.

*Notation*: The notation $|\cdot|$ represents the absolute value for scalars, and the 2-norm for vectors and matrices. The sets $\mathbb{N}_+$ and $\mathbb{N}_0$ represent natural numbers without zero and natural numbers with zero, respectively. The set $\mathbb{R}$ represents real numbers, and $\mathbb{R}^n$ represents the real vector space with dimension $n$. For $x : [0, +\infty) \to \mathbb{R}^n$, we denote $x^+(t) = \lim_{s\downarrow t} x(s)$ and $x^-(t) = \lim_{s\uparrow t} x(s)$. In addition, the solution of discontinuous systems will be used in this paper. To cope with the discontinuities, a common approach is to decribe the values at the points of discontinuity with a Filippov set valued map (Cortes, 2008). Filippov solutions (see (Filippov, 2013)) are often used for the analysis of discontinuous systems, to enable well-defined solutions to the system. We opt to not make explicit use of such tools within the paper, for simplicity and to avoid a focus diversion from the challenges associated with developing threat detection and isolation approaches. Nevertheless, solutions that include discontinuities should be interpreted as above.

## 2. PROBLEM FORMULATION

In this paper, we consider the CPS depicted in Fig. 1, which consists of a physical plant $\mathcal{P}$, a communication network $\mathcal{N}_s$, a controller $\mathcal{C}$ and an anomaly detector $\mathcal{D}$. We consider two types of threats: 1) sensor replay attacks in $\mathcal{N}_s$ and 2) general physical faults. Under physical faults and replay attacks, the CPS is described by

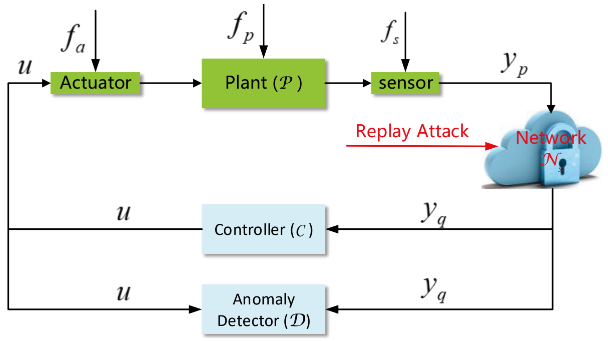

Fig. 1. Diagram of CPS under sensor replay attacks and physical faults.

$$\dot{x}_p = Ax_p + Bu + f_p(t) + f_a(t) + B_d d(t), \tag{1a}$$
$$y_p = Cx_p + f_s(t) + D_d d(t), \tag{1b}$$
$$y_q = y_p + a(t), \tag{1c}$$

where $x_p \in \mathcal{X} \subset \mathbb{R}^{n_p}$ ($\mathcal{X} \triangleq \{|x| \leq \sigma\}$ with $\sigma > 0$) is the state, $u \in \mathbb{R}^{n_u}$ is the control input, $y_p \in \mathbb{R}^{n_y}$ represents the sensor measurements, and $y_q \in \mathbb{R}^{n_y}$ represents the received sensor measurements by the controller $\mathcal{C}$ and the detector $\mathcal{D}$ from $\mathcal{N}_s$. Moreover, $A$, $B$ and $C$, $B_d$ and $D_d$ are known matrices by the defender with appropriate dimensions, and the pair $(A, C)$ is observable. The signal $d \in \mathbb{R}^{n_d}$ represents the lumped disturbances and measurement noises, and satisfies the following assumption.

*Assumption 1.* There exists a constant $\delta_d > 0$ such that

$$|d(t)| \leq \delta_d, \ \forall \ t \geq 0, \tag{2}$$

where $\delta_d$ is known by the defender. ▲

In this work, process faults, actuator faults and sensor faults are considered, which may occur at different time instants. We use $T_f$ to represent the time instants when any fault occurs. In (1), $f_p(t) \in \mathbb{R}^{n_p}$, $f_a(t) \in \mathbb{R}^{n_p}$ and $f_s(t) \in \mathbb{R}^{n_y}$ represent the physical process faults, actuator faults and sensor faults, respectively. In order to simplify the notations, $f_p$, $f_a$ and $f_s$ are jointly denoted as $f$, i.e.,

$f^T(t) \triangleq [f_p^T(t), f_a^T(t), f_s^T(t)]$. Thus, $f_p(t) + f_a(t) = B_f f(t)$ in (1a) with $B_f = [I, I, 0]$, and $f_s = D_f f$ in (1b) with $D_f = [0, 0, I]$. The $f$ satisfies the following assumption.

*Assumption 2.* There exists a constant $\delta_f > 0$ such that

$$|f(t)| \leq \delta_f, \ \forall \ t \geq 0, \tag{3}$$

where $\delta_f$ is known by the defender. ▲

*Remark 1.* It is worth mentioning that given $\delta_d$ in Assumption 1 and $\delta_f$ in Assumption 2, the bound $\sigma$ below (1) can be easily derived based on the knowledge of $A$ and $u$, e.g., a bound on $u$ or the state feedback gain of $u$. Assumption 2 is used for threat discrimination purposes, but is not required for threat detection. Note that the results presented in this paper hold when any upper bound $\delta_f$ to $f(t)$ is known. Hence, a tight upper bound on $|f(t)|$ is not required in the proposed methodology. ∇

The time-dependent variable $a(t) \in \mathbb{R}^{n_y}$ in (1) represents the replay attack signal, which is defined as follows:

$$a(t) \triangleq y_q(t) - y_p(t). \tag{4}$$

In general, replay attacks have recording and replaying procedures (see (Mo and Sinopoli, 2009)). For the considered replay attack in this paper, the adversary first records $y_p$ starting at a time $T_a - T$ and ending at $T_a$, and then replays $y_p$ starting at $T_a$ and ending at $T_a + T$. Based on the above description of the replay attack, the virtual attack signal $a(t)$ can be written as

$$a(t) = \beta(t, T_a, T)(y_p(t - T) - y_p(t)), \tag{5}$$

where $\beta : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies $\beta(t, T_a, T) = 1$ for $t \in [T_a, T_a + T]$ and $\beta(t, T_a, T) = 0$ otherwise, and is used for indicating the occurrence of the replay attack at the time instant $T_a$ and the disappearance at $T_d + T$. The recording time length $T$ satisfies the following assumption.

*Assumption 3.* There exists a constant $\delta_T > 0$ such that

$$T \geq \delta_T, \tag{6}$$

where $\delta_T$ is known by the defender. ▲

The anomaly detector $\mathcal{D}$ in Fig. 1 takes the form of a commonly used detector such as one of the model-based detectors in (Ding, 2008). Specifically, $\mathcal{D}$ has the form:

$$\dot{x}_r = Ax_r + Bu - L(Cx_r - y_q), \tag{7a}$$

$$r = y_q - Cx_r, \tag{7b}$$

where $x_r \in \mathbb{R}^{n_p}$ is the state of the detector and $r \in \mathbb{R}^{n_y}$ is the *residual* used for the anomaly detection. The gain matrix $L \in \mathbb{R}^{n_p \times n_y}$ is chosen such that $A_r \triangleq A - LC$ is a Hurwitz matrix. The estimation error $e \triangleq x_p - x_r$ between the detector $\mathcal{D}$ and the CPS (1), and the residual $r$ can be written as

$$\dot{e} = A_r e + B_d d + B_f f + L(Cx_p - y_q), \tag{8a}$$

$$r = Ce + D_d d + D_f f + a(t). \tag{8b}$$

Under Assumption 1 and given the Hurwitz matrix $A_r$, in the nominal scenario (i.e., $f(t) = 0$ and $a(t) = 0$ identically), the residual $r(t)$ is bounded by a time-varying threshold, i.e., $|r(t)| \leq \bar{r}_1(t)$, where $\bar{r}_1(t)$ is given as follows:

$$\bar{r}_1(t) \triangleq \kappa e^{-\mu t} \bar{e}_0 + \int_0^t \kappa e^{-\mu(t-\tau)} (|B_d - LD_d|\delta_d) d\tau + |D_d|\delta_d. \tag{9}$$

In the above equation, the scalars $\kappa > 0$ and $\mu > 0$ satisfy $|Ce^{A_r t}| \leq \kappa e^{-\mu t}$, and $\bar{e}_0$ satisfies $|x_r(0) - x_p(0)| \leq \bar{e}_0$. Moreover, under Assumptions 1 and 2, in the fault scenario (i.e., $a(t) = 0$ identically), $r(t)$ satisfies $|r(t)| \leq \bar{r}_2(t)$, where $\bar{r}_2(t)$ depends on $\delta_d$ and $\delta_f$, and is given by

$$\bar{r}_2(t) \triangleq \bar{r}_1(t) + \bar{r}_{20}(t), \tag{10a}$$

$$\bar{r}_{20}(t) \triangleq \int_0^t \kappa e^{-\mu(t-\tau)} (|B_f - LD_f|\delta_f) d\tau + |D_f|\delta_f. \tag{10b}$$

In this paper, we consider a *nominal scenario* without attacks and faults, and two threat scenarios: (i) the *replay attack scenario* with a replay attack and (ii) the *fault scenario* with one/multiple physical fault(s). The threat scenarios satisfy the following assumption.

*Assumption 4.* The replay attack scenario and the fault scenario do not occur simultaneously, that is a threat event includes exactly one of the two threat scenarios. ▲

*Remark 2.* Assumption 4 is reasonable in practice. The likelihood that replay attacks and sensor bias faults occur simultaneously is very low in practical scenarios. Therefore, this work is done under Assumption 4. ∇

The *objective* of this paper is to indicate the occurrence of the considered threats, and then identify the threat scenario: either the fault scenario or the replay attack scenario. The detection and discrimination strategy proposed in this paper is based on the anomaly detector $\mathcal{D}$ in (7) and the time-varying thresholds $\bar{r}_1(t)$ in (9) and $\bar{r}_2(t)$ in (10). Specifically, the proposed strategy is referred to as "*two-layer decision strategy*", and is given as follows:

**Layer-1:** If $|r(t_d)| > \bar{r}_1(t_d)$ for some $t_d$, then a threat is detected at $t_d$. Otherwise, no threat is detected;

**Layer-2:** Assume that $|r(t_d)| > \bar{r}_1(t_d)$ for some $t_d$ in **Layer-1**. Then, if $|r(t_a)| > \bar{r}_2(t_a)$ for some $t_a > t_d$, then the attack scenario is identified at $t_a$. Otherwise, a fault scenario is identified.

The first layer is to indicate the occurrence of the threats, and the second layer is activated when any threat is detected in the first layer, indicating the threat discrimination result. In the context of the *two-layer decision strategy*, the main challenge is to drive the residual $|r(t)|$ to exceed the threshold $\bar{r}_2(t)$ in the replay attack scenario but not the fault scenario, such that threat discrimination is achieved. This paper designs a sensor additive switching watermark scheme to solve this problem, which includes a set of suitably designed watermark seeds and switching protocols. The watermark seeds and switching protocols are jointly designed such that the proposed scheme does not affect the behaviour of the closed loop system in the normal and fault scenarios. Moreover, the watermark seeds are designed to drive $|r(t)|$ to exceed the threshold $\bar{r}_2(t)$ in only the replay attack scenario but not the fault scenario, allowing the discrimination. The details of the proposed design are provided in the following section.

## 3. WATERMARK-BASED DETECTION AND DISCRIMINATION METHODOLOGY

In this section, a sensor additive switching watermark methodology, including seeds and switching protocols, is designed. Based on this, the detection and discrimination methodology is developed. The ability of the proposed approach to detect and discriminate between faults and replay attacks is also rigorously investigated. We start by showing the schematic diagram of the developed watermark methodology in Fig. 2, where $y_w$ represents the input of the data transmission network in the physical plant side and $\tilde{y}_w$ is the output of the data transmission network in the CM side. More specifically, the schemes include a switching watermark generator $\mathcal{W}$ and a switching watermark remover $\mathcal{Q}$. The generator $\mathcal{W}$ produces a watermark $w(t)$ by switching the seeds, in which the switches are triggered by a state-dependent triggering protocol (STP) and a time-dependent triggering protocol (TTP). Note that the TTP of $\mathcal{W}$ can generate triggering commands randomly such that the triggering time instants are not predictable to the attacker. The remover $\mathcal{Q}$ generates a switching watermark $q(t)$ to cancel out $w(t)$. Then, the variables $y_w$ and $\tilde{y}_w$ are described by

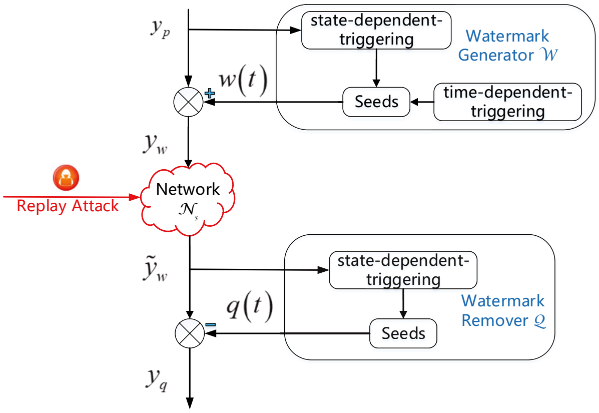
$y_p$
state-dependent-triggering
Watermark Generator $\mathcal{W}$
$w(t)$
+
Seeds
time-dependent-triggering
$y_w$
Network $\mathcal{N}_s$
Replay Attack
$\tilde{y}_w$
state-dependent-triggering
$q(t)$
Watermark Remover $\mathcal{Q}$
−
Seeds
$y_q$

Fig. 2. Schematic diagram of the developed sensor additive switching watermark methodology.

$$y_w(t) = y_p(t) + w(t),\ y_q(t) = \tilde{y}_w(t) - q(t). \qquad (11)$$

In addition, the seeds of $\mathcal{Q}$ are the same as the ones of $\mathcal{W}$. The STP of $\mathcal{Q}$ guarantees that the switches in $\mathcal{Q}$ synchronize with the ones in $\mathcal{W}$ in the normal and faulty cases, i.e., $w(t) = q(t)$ identically, which, from (11), implies that $y_q(t) = y_p(t)$ identically.

### *3.1 Seed Set and Switch Protocol Design*

*Seed Set Design:* Common seeds are used in $\mathcal{W}$ and $\mathcal{Q}$, belonging to a seed set $\mathcal{K}_\Delta$ given by

$$\mathcal{K}_\Delta \triangleq \{\Delta_k \in \mathbb{R}^{n_y} |\ \Delta_{k+1} = \Delta_k + \Delta,\ k \in \mathbb{N}_0\}, \qquad (12)$$

where $\Delta_0 = 0$ and $\Delta \in \mathbb{R}^{n_y}$ is a design parameter to be determined later.

*Switching Protocol Design:* The switching protocols of $\mathcal{W}$ and $\mathcal{Q}$ are designed to guarantee the synchronization of their switches. To this end, we let $t_{w,k}$ and $t_{q,k}$ denote the $k$-th ($k \in \mathbb{N}_0$) switching time instants of $\mathcal{W}$ and $\mathcal{Q}$, respectively, where $t_{w,0} = 0$ and $t_{q,0} = 0$. Then, by using the seed set $\mathcal{K}_\Delta$ in (12), the update laws of the watermarks $w(t)$ and $q(t)$ are designed as follows:

$$w^+(t_{w,k}) = w^-(t_{w,k}) + \Delta,\ \forall\ k \in \mathbb{N}_+, \qquad (13a)$$

$$q^+(t_{q,k}) = q^-(t_{q,k}) + \Delta,\ \forall\ k \in \mathbb{N}_+, \qquad (13b)$$

where $w(t_{w,0}) = \Delta_0$ and $q(t_{q,0}) = \Delta_0$. Recall that the objective of the watermark schemes is to achieve $y_p(t) = y_q(t)$, i.e., $w(t) = q(t)$. By using the common seed set $\mathcal{K}_\Delta$ in (12) and the update laws of $w(t)$ in (13a) and $q(t)$ in (13b), we can deduce that the only requirement for the switching protocols is to achieve *switching time synchronization*, i.e., $t_{w,k} = t_{q,k}$ for any $k \in \mathbb{N}_0$. The specific design for achieving this objective is given below.

First, the STP of the remover $\mathcal{Q}$ is proposed as

$$t_{q,k} \triangleq \inf\left\{t > t_{q,k-1} \,\middle|\, |y_q^-(t) - y_p(t_{q,k-1})| \geq \eta\right\}, \quad \forall\ k \in \mathbb{N}_+, \quad (14)$$

where $\eta > 0$ is predefined by the user.

Note that both the variations of $y_p(t)$ for $t > t_{q,k-1}$ and the switch in $\mathcal{W}$ can trigger the switching condition in (14). Therefore, for the synchronization purposes, the switching protocol of $\mathcal{W}$ has two objectives: **(a)** synchronizing with the spontaneous switches in $\mathcal{Q}$ due to the variations of $y_p(t)$ and **(b)** enforcing a new switch in $\mathcal{Q}$ at any switching time $t_{w,k}$ by using the seed increment $\Delta$. Hence, the switching protocol of $\mathcal{W}$ is proposed as

$$t_{w,k}(t) \triangleq \min\left\{t'_{w,k},\ \hat{t}_{q,k}\right\},\ \forall\ k \in \mathbb{N}_+, \qquad (15)$$

where $t'_{w,k}$ is determined by the TTP of $\mathcal{W}$ that is designed as

$$t'_{w,k} \triangleq t_{w,k-1} + \alpha_k,\ \forall\ k \in \mathbb{N}_+, \qquad (16)$$

with $\alpha_k$ being a random scalar satisfying $0 < \alpha_k < \delta_T$, and being generated at the time instant $t_{w,k-1}$. Moreover, $\hat{t}_{q,k}$ is determined by the STP $\mathcal{W}$ that is designed as

$$\hat{t}_{q,k} \triangleq \inf\left\{t > t_{w,k-1} \,\middle|\, |y_q^-(t) - y_p(t_{w,k-1})| \geq \eta\right\}, \quad \forall\ k \in \mathbb{N}_+. \quad (17)$$

Subsequently, the switching time sequences $\mathcal{K}_w$ of $\mathcal{W}$ and $\mathcal{K}_q$ of $\mathcal{Q}$ can be defined respectively as follows:

$$\mathcal{K}_w \triangleq \{t_{w,k}, k \in \mathbb{N}_0\},\ \mathcal{K}_q \triangleq \{t_{q,k}, k \in \mathbb{N}_0\}. \qquad (18)$$

Then, $\mathcal{K}_w = \mathcal{K}_q$ implies that the switching time instants $t_{w,k}$ of $\mathcal{W}$ and $t_{q,k}$ of $\mathcal{Q}$ are synchronized for $k \in \mathbb{N}_0$.

Note that objective **(a)**, i.e.,synchronizing with the spontaneous switches in $\mathcal{Q}$, is achieved by the design of the STP (17). The approach to achieve objective **(b)**, i.e., a new switch in $\mathcal{Q}$ is enforced by the switch in $\mathcal{W}$, is presented in the following result.

*Theorem 1.* Consider the CPS (1), the common seed set $\mathcal{K}_\Delta$ in (12), the watermark update laws in (13), and the switching protocols (14) and (15). Then, in the nominal scenario and the fault scenario, if the seed increment $\Delta$ in (12) satisfies

$$|\Delta| \geq \eta + 2(|C|\sigma + |D_d|\delta_d), \qquad (19)$$

where $\sigma$ is given below (1), then the *switching time synchronization* is guaranteed, i.e., $\mathcal{K}_w = \mathcal{K}_q$. Moreover, $y_p(t_{w,k})$ satisfies

$$y_p(t_{w,k}) = y_q^+(t_{w,k}) = \tilde{y}_w^+(t_{w,k}) - q^+(t_{w,k}),\ \forall\ k \in \mathbb{N}_0. \qquad (20)$$

■

It should be noted that the STP (14) cannot be implemented directly since $y_p(t_{q,k-1})$ used in (14) is not available at the CM side. Similarly, the STP (17) cannot be implemented either due to the unavailabilities of $y_q^-(t)$ at the physical plant side. In order to address these implementation issues, the available resources of $\mathcal{W}$ in between the switches at $t_{w,k-1}$ and $t_{w,k}$ and of $\mathcal{Q}$ in between the switches at $t_{q,k-1}$ and $t_{q,k}$ are given respectively as follows:

$$\mathcal{I}_w(t) \triangleq \{y_p(s)|t_{w,k-1} \leq s \leq t\}, \ \forall \ t > t_{w,k-1}, \quad \text{(21a)}$$

$$\mathcal{I}_q(t) \triangleq \{\tilde{y}_w(s)|t_{q,k-1} \leq s \leq t\}, \ \forall \ t > t_{q,k-1}. \quad \text{(21b)}$$

Then, implementable equivalent forms of the STPs (14) and (17) are specified in the following lemma.

*Lemma 1.* The STPs (14) and (17) with the available resources $\mathcal{I}_q(t)$ and $\mathcal{I}_w(t)$ respectively can be equivalently written as

$$t_{q,k} = \inf\left\{t > t_{q,k-1} \left||\tilde{y}_w^-(t) - \tilde{y}_w(t_{q,k-1})| \geq \eta\right.\right\}, \quad \forall \ k \in \mathbb{N}_+, \quad \text{(22a)}$$

$$\hat{t}_{q,k} = \inf\left\{t > t_{w,k-1} \left||y_p^-(t) - y_p(t_{w,k-1})| \geq \eta\right.\right\}, \quad \forall \ k \in \mathbb{N}_+. \quad \text{(22b)}$$

■

### 3.2 Detectability and Discrimination Ability Analysis

The detectability and discrimination ability of the developed watermark methodology are analyzed in this section, to characterize the physical faults and replay attacks that can be detected and discriminated. To this end, we suppose that the replay attack occurs at $T_a \in [t_{q,k_a-1}, t_{q,k_a})$ with $t_{q,k_a} \in \mathcal{K}_q$, $k_a \geq 2$ and $k_a \in \mathbb{N}_+$, and the recording procedure is initiated at $T_a - T \in [t_{q,k_a-k_0-1}, t_{q,k_a-k_0})$ with $k_0 \in \mathbb{N}_+$ and $k_0 \leq k_a - 1$. Then, a lemma associated with the switches in $\mathcal{Q}$ under the replay attack is presented.

*Lemma 2.* Consider the seed set $\mathcal{K}_\Delta$ in (12) with $\Delta$ satisfying (19) in Theorem 1, and the STP (14). Under Assumption 3, a switch occurs in $\mathcal{Q}$ at the attack time $T_a$. Moreover, switches occur in $\mathcal{Q}$ at $t_{q,k_a-k_0+i} + T$ for $0 \leq i \leq k_0 - 1$ and $i \in \mathbb{N}_0$. ■

Based on Lemma 2, in the presence of the replay attack at $T_a$, the switching time sequence of $\mathcal{Q}$ can be described by

$$\mathcal{K}_{qa} \triangleq \{t_{qa,0}, \cdots, t_{qa,k_a}, t_{qa,k_a+1}, t_{qa,k_a+2}, \cdots\}, \quad \text{(23)}$$

where $t_{qa,i} = t_{q,i}$ for $i \leq k_a - 1$ and $i \in \mathbb{N}_0$, $t_{qa,k_a} = T_a$, and $t_{qa,k_a+i} = t_{q,k_a-k_0+i-1} + T$ for $i \in \mathbb{N}_+$. The above allows to deduce our main results associated with the detectability and discrimination ability of the presented approach, which are given below.

*Theorem 2.* Consider the CPS (1) with the disturbance $d$ satisfying Assumption 1, the detector $\mathcal{D}$ given in (7) and the thresholds $\bar{r}_1(t)$ in (9) and $\bar{r}_2(t)$ in (10). Consider also the seed set $\mathcal{K}_\Delta$ in (12) with $\Delta$ satisfying (19), and the switching protocols in (14) and (15). Then, under Assumption 4, we have the following results:

**(i)** In the fault scenario with the fault satisfying Assumption 2, $|r(t)| \leq \bar{r}_2(t)$ for all $t \geq T_f$, and there exists $t > T_f$ such that $|r(t)| > \bar{r}_1(t)$, if $f(t)$ satisfies

$$|f(t)| > \frac{2\bar{r}_1(t)}{\inf_{\omega \in \mathbb{R}} G_f(j\omega)}, \ \forall \ t \geq 0, \quad \text{(24)}$$

where $G_f(j\omega) = C(j\omega I - A_r)^{-1}(B_f + LD_f) + D_f$.

**(ii)** In the replay attack scenario with the attack satisfying Assumption 3, there exists $t > T_a$ such that $|r(t)| > \bar{r}_2(t)$, if the seed increment $\Delta$ satisfies

$$|\Delta| \geq \frac{2\bar{r}_1(t) + r_{20}(t)}{2\inf_{\omega \in \mathbb{R}} G_a(j\omega)}, \ \forall \ t \geq 0, \quad \text{(25)}$$

where $G_a(j\omega) = C(j\omega I - A_r)^{-1}L + I$ and $r_{20}(t)$ is given in (10b). ■

In Theorem 2(i), condition (24) characterizes the faults that can be detected by the developed watermark methodology. Theorem 2(ii) shows that the replay attacks satisfying Assumption 3 can be detected by our watermark methodology if $\Delta$ is designed to satisfy (12). In addition, based on the *two-layer decision strategy*, the fault scenario with the fault satisfying Assumption 2 can be identified if the fault $f$ satisfies (24), and the attack scenario with the replay attacks satisfying Assumption 3 can be identified if $\Delta$ is designed to satisfy (12).

Based on Theorems 1 and 2, $\Delta$ should be designed to satisfy the following requirement:

$$|\Delta| \geq \max\left\{\eta + 2(|C|\sigma + |D_d|\delta_d), \frac{2\bar{r}_1(t) + r_{20}(t)}{2\inf_{\omega \in \mathbb{R}} G_a(j\omega)}\right\}, \quad \text{(26)}$$

where $\eta$ is given in (14), $\sigma$ is given below (1), $\bar{r}_1(t)$ and $r_{20}(t)$ are given in (9) and (10b) respectively, and $G_a(j\omega)$ is given below (25).

## 4. SIMULATION

In this section, a simulation based on a system described by (1) is presented. The system matrices are given as follows:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \ B = \begin{bmatrix} -1.67 & 0 & 0 \\ 0 & -1.93 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

$$B_d = \begin{bmatrix} 0.1 & 0 \\ 0.12 & 0 \\ 0.21 & 0 \end{bmatrix}, \ D_d = \begin{bmatrix} 0 & 0.1 \\ 0 & 0.4 \end{bmatrix},$$

where the pair $(C, A)$ is observable. The bound $\sigma$ of the system state is set as 20. The disturbance $d$ satisfies $|d(t)| \leq 5$ and the fault $f$ satisfies $|f(t)| \leq 10$. Thus, Assumptions 1 and 2 are satisfied with $\delta_d = 5$ and $\delta_f = 10$. Regarding the replay attack, the recording time length $T$ is not smaller than 4s, which indicates that Assumption 3 is satisfied with $\delta_T = 4$. Moreover, the gain $L$ of the detector $\mathcal{D}$ in (8) is given by

$$L = \begin{bmatrix} 0.3021 & 0.9998 \\ -0.0622 & -1.4971 \\ 0.0464 & 0.5577 \end{bmatrix}.$$

The thresholds $\bar{r}_1(t)$ and $\bar{r}_{20}(t)$ are given as follows: $\bar{r}_1(t) = 21.4845 - 14.4229e^{-0.2t}$ and $\bar{r}_{20}(t) = 23.0401 - 20.8040e^{-0.2t}$.

We proceed with the design of the sensor additive switching watermark methodology. Regarding the switching protocols, the user-defined parameter $\eta$ of STPs (14) and (17) is set as $\eta = 10$. In addition, the seed increment $\Delta$ of the seed set $\mathcal{K}_\Delta$ in (12) is calculated following (26). Given $L$, we can calculate $\inf_\omega G_a(j\omega) \geq 1$ for $\omega \in \mathbb{R}$, and then based on (26), we can choose $\Delta = [60, 60]^T$.

For the simulation setup, the disturbance $d(t)$ and the physical fault $f(t)$ occurring at $T_f = 10$s are given as follows: $d(t) = [\sin(50t), d_1(t)]^T$ and $f(t) = [1 + 2\sin(15t), 2 + 4\sin(25t), 0.5 + 4\sin(35t)]^T$, where $d_1(t)$ is randomly selected at each time instant from the uniform distribution $[-2, 2]$. In addition, in terms of the replay attack, the recording procedure initiates at $T_a - T = 19$s and ends at $T_a = 25$s, and then the replaying procedure starts at $T_a$. Hence, $T = 6$s.

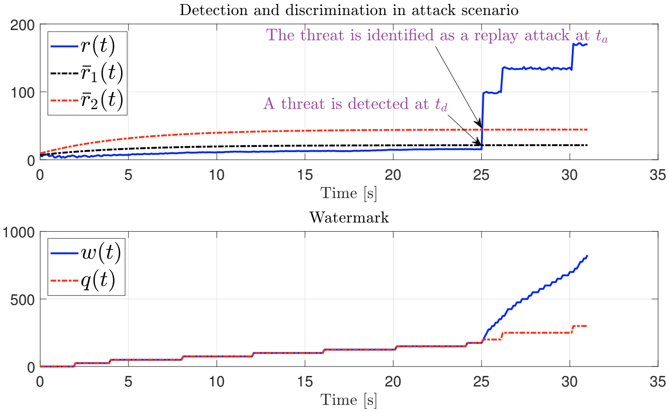

Fig. 3. Detection and discrimination results, and the watermarks in the attack scenario.

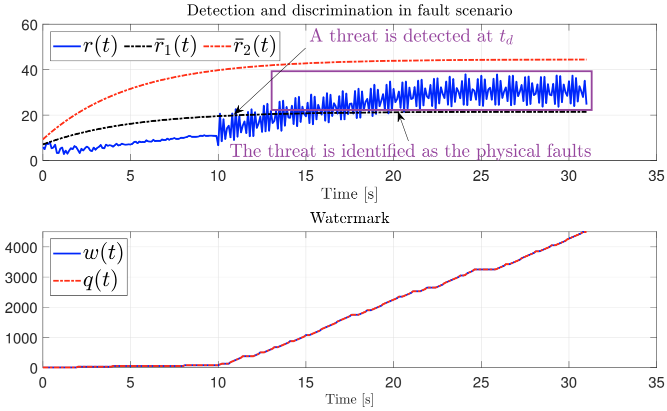

Fig. 4. Detection and discrimination results, and the watermarks in the fault scenario.

The detection and discrimination results in the attack scenario and the fault scenario are illustrated in Figs. 3 and 4, respectively. It is shown in Fig. 3 that the residual $r(t)$ exceeds the threshold $\bar{r}_1(t)$ at $t_d$ and exceeds the threshold $\bar{r}_2(t)$ at $t_a$. Thus, based on the two-layer decision strategy, a threat is detected at $t_d$, and the threat scenario is identified as a replay attack at $t_f$. The watermarks $w(t)$ and $q(t)$ in Fig. 3 show that the switches in the generator $\mathcal{W}$ and the remover $\mathcal{Q}$ are synchronized before $T_a$, and occur at $t_{w,0} = 0$s, $t_{w,1} = 2$s, $t_{w,2} = 4$s, $t_{w,3} = 8$s, $t_{w,4} = 12$s, $t_{w,5} = 16$s, $t_{w,6} = 20$s and $t_{w,7} = 24$s. In addition, at $T_a$, the switch synchronization is lost. The watermark $q(t)$ has switches at $T_a$, $t_{w,6} + T = 26$s and $t_{w,7} + T = 30$s, which verifies Lemma 2.

As shown in Fig. 4, the residual $r(t)$ exceeds $\bar{r}_1(t)$ at $t_d$, and remains below $\bar{r}_2(t)$ for all the simulation time. Thus, based on the two-layer decision strategy, a threat is detected at $t_d$ and is identified as a physical fault in the simulation time duration. The switches of $w(t)$ and $q(t)$ in Fig. 4 remain synchronized, which verifies Theorem 1.

## 5. CONCLUSIONS

A sensor additive switching watermark design has been proposed for detecting physical faults and replay attacks, and then identifying the occurring threat type. The switching protocols and the common seeds were designed such that in both the nominal and the fault cases, the switches in the generator and the remover are synchronized. Furthermore, in the replay attack case, the switch synchronization is lost and the detection residual exceeds a detection threshold. The detectability and discrimination ability of the proposed approach was also rigorously investigated. The simulation was presented to verify the developed detection and discrimination approach.

## REFERENCES

Cardenas, A., Amin, S., and Sastry, S. (2008). Secure control: Towards survivable cyber-physical systems. In *2008 The 28th International Conference on Distributed Computing Systems Workshops*, 495–500. IEEE.

Cortes, J. (2008). Discontinuous dynamical systems. *IEEE Control systems magazine*, 28(3), 36–73.

Dibaji, S., Pirani, M., Flamholz, D., Annaswamy, A., Johansson, K., and Chakrabortty, A. (2019). A systems and control perspective of CPS security. *Annual Reviews in Control*, 47, 394–411.

Ding, S. (2008). *Model-based fault diagnosis techniques: design schemes, algorithms, and tools.* Springer Science & Business Media.

Ferrari, R. and Teixeira, A. (2020). A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks. *IEEE Transactions on Automatic Control*, 66(6), 2558–2573.

Filippov, A. (2013). *Differential equations with discontinuous righthand sides: control systems.* Springer Science & Business Media.

Gallo, A., Anand, S., Teixeira, A., and Ferrari, R. (2021). Design of multiplicative watermarking against covert attacks. *arXiv preprint arXiv:2110.00555.*

Gallo, A., Turan, M., Boem, F., Ferrari-Trecate, G., and Parisini, T. (2018). Distributed watermarking for secure control of microgrids under replay attacks. *IFAC-PapersOnLine*, 51(23), 182–187.

Liu, H., Yan, J., Mo, Y., and Johansson, K. (2018). An on-line design of physical watermarks. In *2018 IEEE Conference on Decision and Control (CDC)*, 440–445. IEEE.

Mo, Y., Chabukswar, R., and Sinopoli, B. (2013). Detecting integrity attacks on scada systems. *IEEE Transactions on Control Systems Technology*, 22(4), 1396–1407.

Mo, Y. and Sinopoli, B. (2009). Secure control against replay attacks. In *2009 47th annual Allerton conference on communication, control, and computing (Allerton)*, 911–918. IEEE.

Porter, M., Hespanhol, P., Aswani, A., Johnson-Roberson, M., and Vasudevan, R. (2021). Detecting generalized replay attacks via time-varying dynamic watermarking. *IEEE Transactions on Automatic Control*, 66(8), 3502–3517.

Smith, R. (2011). A decoupled feedback structure for covertly appropriating networked control systems. *IFAC Proceedings Volumes*, 44(1), 90–95.

Weerakkody, S., Ozel, O., and Sinopoli, B. (2017). A bernoulli-gaussian physical watermark for detecting integrity attacks in control systems. In *55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 966–973. IEEE.

Zhang, K., Keliris, C., Parisini, T., and Polycarpou, M. (2021). Identification of sensor replay attacks and physical faults for cyber-physical systems. *IEEE Control Systems Letters*, 6, 1178–1183.