Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Enhanced Anomaly Detector for Nonlinear Cyber-Physical Systems against Stealthy Integrity Attacks

Kangkang Zhang * **Marios M. Polycarpou** *
Thomas Parisini *,**

* *KIOS Research and Innovation Center of Excellence and the Dept. of Electrical and Computer Engineering, University of Cyprus, Nicosia, 1678, Cyprus (e-mail: kangzhang359@nuaa.edu.cn; zhang.kangkang@ucy.ac.cy; mpolycar@ucy.ac.cy).*
** *Dept. of Electrical and Electronic Engineering, Imperial College London, London, UK (e-mail: t.parisini@gmail.com)*

Abstract: The detection of stealthy integrity attacks for nonlinear cyber-physical systems is a great challenge for the research community. This paper proposes a backward-in-time detection methodology to enhance the anomaly detector against stealthy integrity attacks for a class of nonlinear cyber-physical systems. It uses the virtual value of the state at a time instant prior to the occurrence time of the attacks for detecting stealthy attacks. The definition of stealthy integrity attacks is formulated in the context of nonlinear plants such that they are undetectable with respect to traditional anomaly detectors. A H_∞ fixed-point smoother is developed for estimating the analytical virtual values of the states at a prior time to the attack occurrence time, and then, the backward-in-time detection schemes are proposed based on the smoother. Based on the prior estimates, attack residual generation and threshold generation schemes are designed. Finally, a simulation is presented to illustrate the effectiveness of the enhanced anomaly detector.

Keywords: Stealthy integrity attacks, nonlinear cyber-physical systems, backward-in-time detection methodology.

1. INTRODUCTION

Cyber-physical systems (CPSs) are attracting more and more research efforts currently owing to their wide applications to critical infrastructure systems, for example, electric power transmission and distribution systems, waste water and gas distribution systems, transportation systems, and so on. However, several malicious attack incidents in modern industrial CPSs have taken place and been reported in recent years. Prompted by increasing safety and security of CPSs specifications, there has been some increasing activities on research dealing with cyber attack detection issues.

In the context of CPSs, cyber attacks are classified into two categories according to (Cardenas et al., 2008) and (Teixeira et al., 2015): denial of service (DOS) attacks and integrity (or deception) attacks. *Integrity attacks* include replay attacks (Mo and Sinopoli, 2009), covert attacks (Smith, 2015), zero-dynamics attacks (Teixeira et al., 2015) and so on, where adversaries compromise the *integrity* of CPSs to keep deceptions when injecting false data. In terms of stealthiness, integrity attacks, if intelligently designed, are stealthy and can bypass classical anomaly detectors such as fault diagnosis schemes (Ding, 2008). Therefore, enhancing the attack detectability of anomaly detectors against stealthy integrity attacks presents a key challenge in cyber-physical security.

Detection of integrity attacks based on dynamic models of CPSs has been previously investigated by the research community. Some model-based attack detection methods such as the physical watermarking (Mo and Sinopoli, 2009) and moving target (Mo and Sinopoli, 2010) have been proposed and developed in the past decade. The watermarking approach is proposed to detect replay attacks in (Mo and Sinopoli, 2009) and (Mo et al., 2013), which is realized by adding watermarks to the control inputs and then detecting them based on the received output measurements from communication networks to determine the occurrences of replay attacks. However, additive watermarks to the control inputs may cause degradation of control performance in CPSs. To deal with this drawback, (Ferrari and Teixeira, 2017) proposes a sensor watermarking method using multiplicative watermarks to detect and isolate replay attacks.

In this paper, a backward-in-time detection (BTD) methodology to enhance the attack detection abilities of the traditional anomaly detectors is proposed, aiming at detecting stealthy integrity attacks for a class of nonlinear CPSs. It uses the virtual value of the state at a time instant

⋆ This work has been supported by the European Union's Horizon 2020 Research and Innovation Program under the grant agreement No. 739551 (KIOS CoE), the National Natural Science Foundation of China (Grants No 61903188), the Natural Science Foundation of Jiangsu Province (Grants No BK20190403) and the China Postdoctoral Science Foundation 2019M660114.

prior to the occurrence time of the attacks for detecting stealthy attacks. Specifically, the stealthy integrity attacks for nonlinear CPSs are formulated. The attack model is then proposed based on geometric control theory and stability theory, and sufficient conditions are rigorously analyzed such that the generated attacks are stealthy to traditional anomaly detectors. A H_∞ fixed-point smoother is then proposed as diagnostic observer for providing an estimate of the analytical value of the state at a particular time instant prior to the attack occurrences time. Furthermore, based on the BTD methodology and using the prior estimates, the corresponding attack residual generation and threshold generation schemes are designed. The detectability is rigorously investigated for characterizing the class of attacks that can be detected.

Notation: For any vectors $x,\ y \in \mathbb{R}^n$, $\mathrm{Co}(x, y)$ denotes the convex hull of the set $\{x, y\}$. For a signal $x(t) \in \mathbb{R}^n$, and 2-norm in a time interval $t \in [t, t + T_w]$ where T_w is the time window is defined as follows:

$$\|x(t)\|_{2,T_w} = \Big(\int_t^{t+T_w} x^T(\tau)x(\tau)d\tau\Big)^{\frac{1}{2}}.$$

2. PROBLEM FORMULATION

Considering the potential of cyber-attacks, a general architecture of CPSs is shown in Fig. 1, which normally consists of a physical plant $\mathcal{P}$, a feedback controller $\mathcal{C}$ and an anomaly detector $\mathcal{D}$, actuator communication network $\mathcal{N}_a$ and sensor communication network $\mathcal{N}_s$ for data transmission between the control and plant sides. During

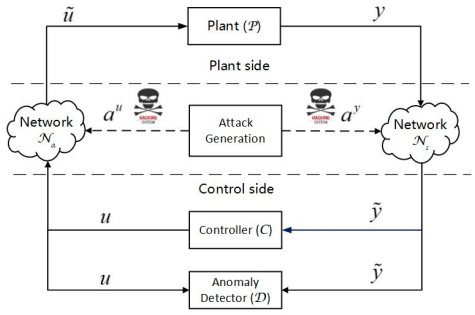

Fig. 1. General architecture of CPSs under possible cyber-attacks.

a cyber-attack event, the attack generation block arises to compromise communication networks $\mathcal{N}_a$ and $\mathcal{N}_s$ by injecting false date a^u and a^y. The data transmission properties of $\mathcal{N}_a$ and $\mathcal{N}_s$ are characterized by

$$\tilde{u} = u + a^u,\ \tilde{y} = y + a^y, \tag{1}$$

where $u,\ \tilde{u} \in \mathbb{R}^m$ represent the transmitted and received actuator data respectively of $\mathcal{N}_a$, while $y,\ \tilde{y} \in \mathbb{R}^q$ are the transmitted and received sensor measurements respectively. The attack signals a^u and a^y are given by $a^u = [a_1^u, \cdots, a_m^u]^T \in \mathbb{R}^m$ and $a^y = [a_1^y, \cdots, a_q^y]^T \in \mathbb{R}^q$ where for $i \in \{1, \cdots, m\}$, ${a_i}^u = 0$ if there is no attack occurring on the ith channel of $\mathcal{N}_a$, and similarly, for $j \in \{1, \cdots, q\}$, ${a_j}^y = 0$ if the jth channel of $\mathcal{N}_s$ is not attacked. Throughout this paper, we suppose that the attacks occur at some unknown time instant T_0. Then, $a^u = 0$ and $a^y = 0$ for $t < T_0$. The dynamics of $\mathcal{P}$, $\mathcal{C}$, $\mathcal{D}$ of the considered CPSs are described by

$$\mathcal{P}: \begin{cases} \dot{x} = Ax + g_p(t, x) + B\tilde{u} + D_1 d, \quad x(0) = x_0, \\ y = Cx + D_2 d, \end{cases} \tag{2}$$

$$\mathcal{C}: \begin{cases} \dot{c} = g_c(t, c, \tilde{y}), \\ u = K_1 c + K_2 \tilde{y}, \end{cases} \tag{3}$$

$$\mathcal{D}: \begin{cases} \dot{x}_r = Ax_r + g_r(t, x_r) + Bu + L\tilde{y}, \\ y_r = Cx_r - \tilde{y}, \\ J_{th} = J(\bar{d}), \end{cases} \tag{4}$$

where $x \in \mathcal{X}_p \subseteq \mathbb{R}^{n_p}$ ($\mathcal{X}_p$ is a neighborhood of the origin) is the state of the physical plant $\mathcal{P}$, $d \in \mathbb{R}^h$ represents the lumped disturbances, and is assumed to satisfy $\sup_{t\in\mathbb{R}_+} \sqrt{d^T(t)d(t)} \le \bar{d}$. The matrices $A \in \mathbb{R}^{n_p \times n_p}$, $B \in \mathbb{R}^{n_p \times m}$, $C \in \mathbb{R}^{q \times n_p}$, $D_1 \in \mathbb{R}^{n_p \times h}$ and $D_2 \in \mathbb{R}^{q \times h}$ are known system matrices. The pair (A, C) is assumed to be observable. The vector function $g_p(\cdot) : \mathbb{R}_+ \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_p}$ is the known nonlinearity of the physical plant $\mathcal{P}$, which is piecewise continuous with respect to (w.r.t.) t and continuously differentiable w.r.t. x. In addition, $g_p(t, x)$ satisfies the following assumption.

Assumption 1. The function $g_p(t, x)$ is locally Lipschitz w.r.t. x for $t \ge 0$, i.e., $\forall\ x, \hat{x} \in \mathcal{X}_p$,

$$\|g_p(t, x) - g_p(t, \hat{x})\| \le l\|x - \hat{x}\|,$$

where l the Lipschitz constant. ∇

The variable $c \in \mathcal{X}_c \subseteq \mathbb{R}^{n_c}$ ($\mathcal{X}_c$ is a neighborhood of the origin) is the state of the output controller $\mathcal{C}$. The vector function $g_c(\cdot) : \mathbb{R}_+ \times \mathbb{R}^{n_c} \times \mathbb{R}^q \to \mathbb{R}^{n_c}$ represents the nonlinear dynamics of $\mathcal{C}$, which is piecewise continuous w.r.t. t and continuously differentiable w.r.t. c and $\tilde{y}$. The matrices $K_1 \in \mathbb{R}^{m \times n_c}$ and $K_2 \in \mathbb{R}^{m \times q}$ are the (known) control gain matrices. For $t \ge 0$, $g_p(t, 0) = 0$, and $g_c(t, 0, 0) = 0$, thereby the origin $x = 0$, $c = 0$ is an equilibrium point at $t = 0$ for the unforced closed-loop system (2)-(3) (i.e., $d = 0$). In addition, the unforced closed-loop system (2)-(3) is assumed to satisfy the following assumption.

Assumption 2. In the non-attack cases, there is a region $\mathcal{X}_{pc} \subset \mathbb{R}^{n_p + n_c}$ ($\mathcal{X}_{pc}$ is a neighborhood of the origin) and a continuous differentiable function $V : \mathbb{R}_+ \times \mathcal{X}_{pc} \to \mathbb{R}$ such that on $\mathcal{X}_{pc}$,

$$W_1(c, x) \le V(t, c, x) \le W_2(c, x),$$
$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\mathcal{P}|_{d=0} + \frac{\partial V}{\partial c}\mathcal{C}|_{d=0} \le -W_3(c, x).$$

where W_1, W_2 and W_3 are continuous positive definite functions. ∇

Suppose that $\mathcal{D}$ is one of the commonly used model-based fault detector illustrated in (Ding, 2008). The dynamics in (4) are used for estimating the states of the physical plant $\mathcal{P}$ where $x_r \in \mathbb{R}^{n_p}$ is the estimate of x, the gain L is designed to stabilize an error system (see, e.g., (Zhang et al., 2010)). The variable $y_r \in \mathbb{R}^q$ is the so-called *residual*. Moreover, the scalar J_{th} represents the threshold, and $J : \mathbb{R}_+ \to \mathbb{R}_+$ is a known scalar function of the disturbance bound $\bar{d}$. The decision of the occurrence of any anomalies is made based on the following principle: if there is a time instant $T_d > T_0$ such that $\|y_r(T_d)\|_{2,T_w} > J_{th}$, then an alarm is triggered. Otherwise, the system is considered as healthy and operating in normal condition.

The detection methodology used in $\mathcal{D}$ is called forward-tin-time detection (FTD) since the residual $y_r(t)$ starting

from the time posterior to T_0 is used to detect anomalies. It should be pointed out that most of the fault detectors such as the ones in (Chen and Patton, 2012), (Blanke et al., 2006) and (Ding, 2008) are based on the FTD methodologies. However, they require that the amplitudes of the anomalies are sufficiently large such that $\|y_r(T_d)\|_{2,T_w} > J_{th}$ holds, and thus, are not effective to attacks with sufficiently small amplitude. This motivates to formulate stealthy attacks for nonlinear CPSs. To this end, some notations are given. To distinguish the non-attack cases and attack cases, the superscript n is used for denoting the variables in non-attack case. For example, x^n is used for denoting the plant state. The attack vectors a^u and a^y are represented by a compact vector a, i.e., $a^T = [(a^u)^T, (a^y)^T]$.

Definition 1. The attack $a \neq 0$ is called a stealthy attack w.r.t. $\mathcal{D}$ if in the presence of a at time instant T_0, one has:

(1) $\|y_r(t)\|_{2,T_w} \leq J_{th}$ for $t \geq T_0$,
(2) $\|y_r - y_r^n\| \to 0$ as $t \to \infty$.

Remark 1. In the definition of perfect undetectable attacks in (Pasqualetti et al., 2013), the received sensor measurements are not affected by attacks, that is $\tilde{y} = \tilde{y}^n$ identically. However, perfect undetectable attacks are not easily to be generated, sometimes impossible, for general nonlinear CPSs, and are also not necessary for practical cases. Comparing with perfect undetectable attacks, the conditions in Definition 1 are weaken, which can also guarantee the stealthiness of the attacks to the anomaly detectors using FTD methodologies. ∇

Objective. The objective of this paper is to construct stealthy integrity attack detection schemes for $\mathcal{D}$ to enhance its detectability against stealthy integrity attacks by utilizing signals u and $\tilde{y}$, and the knowledge $(A,\ g_p(\cdot),\ B,\ C,\ D_1,\ D_2)$ of the physical plant $\mathcal{P}$.

3. STEALTHY INTEGRITY ATTACK SCENARIOS

3.1 Preliminaries

Dynamics of Increments Denote the increments of x, $\tilde{y}$, c and u due to attacks by $x^a = x - x^n$, $y^a = y - y^n$, $c^a = c - c^n$ and $u^a = u - u^n$, respectively. Then for $t < T_0$, $x^a = 0$, $y^a = 0$, $c^a = 0$ and $u^a = 0$. Due to the nonlinearities $g_p(t,x)$ and $g_c(t,c,\tilde{y})$, $g_p(t,x) - g_p(t,x^n)$ and $g_c(t,c,\tilde{y}) - g_c(t,c^n,\tilde{y}^n)$ will arise in deriving the dynamics of these increments. Based on the extended differential mean value theorem, there exist matrices $\xi_x = [\xi_{x1}, \cdots, \xi_{xn_p}] \in \mathbb{R}^{n_p \times n_p}$ with $\xi_{xi} \in \mathrm{Co}(x, x^n)$ and

$$G_p(t,\xi_x) = \begin{bmatrix} \frac{\partial g_{p1}}{\partial x_1}(t,\xi_{x1}) & \cdots & \frac{\partial g_{p1}}{\partial x_{n_p}}(t,\xi_{x1}) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{pn_p}}{\partial x_1}(t,\xi_{xn_p}) & \cdots & \frac{\partial g_{pn_p}}{\partial x_{n_p}}(t,\xi_{xn_p}) \end{bmatrix}$$

such that

$$g_p(t,x) - g_p(t,x^n) = G_p(t,\xi_x)(x - x^n) = G_p(t,\xi_x)x^a.$$

Similarly, there exist matrices ξ_c, $\xi_{\tilde{y}}$, $G_{c1}(t,\xi_c,\xi_{\tilde{y}})$ and $G_{c2}(t,\xi_c,\xi_{\tilde{y}})$ with proper dimensions such that

$$g_c(t,c,\tilde{y}) - g_c(t,c^n,\tilde{y}^n) = G_{c1}(t,\xi_c,\xi_{\tilde{y}})c^a + G_{c2}(t,\xi_c,\xi_{\tilde{y}})\tilde{y}^a.$$

Consequently, the dynamical incremental systems $\mathcal{P}^a$ and $\mathcal{C}^a$ can be respectively described by

$$\mathcal{P}^a : \begin{cases} \dot{x}^a = (A + G_p(t,\xi_x))x^a + Bu^a + B_a a, \\ \tilde{y}^a = Cx^a + D_a a, \end{cases}$$

$$\mathcal{C}^a : \begin{cases} \dot{c}^a = G_{c1}(t,\xi_c,\xi_{\tilde{y}})c^a + G_{c2}(t,\xi_c,\xi_{\tilde{y}})\tilde{y}^a, \\ u^a = K_1 c^a + K_2 \tilde{y}^a, \end{cases}$$

where $x^a(T_0) = 0$, $c^a(T_0) = 0$, $B_a = [B, 0]$ and $D_a = [0, I_q]$.

Geometric Control Foundation Consider the certain LTI part of $\mathcal{P}^a$ as a new system $\mathcal{P}_l$:

$$\mathcal{P}_l : \begin{cases} \dot{x}_l = Ax_l + B_a a,\ x_l(T_0) = x_{l0}, \\ y_l = Cx_l + D_a a, \end{cases}$$

where $x_l \in \mathbb{R}^{n_p}$ is the system state, $y_l \in \mathbb{R}^q$ is the output and $a \in \mathbb{R}^{m+q}$ is the control input. Based on (Trentelman et al., 2012), there is a weakly unobservable subspace $\mathcal{V}(\mathcal{P}_l)$ for $\mathcal{P}_l$ such that $\mathcal{V}(\mathcal{P}_l)$ is the largest subspace satisfying that there exists a linear map $F_a : \mathbb{R}^{m+q} \to \mathbb{R}^{n_p}$ such that

$$(A + B_a F_a)\mathcal{V}(\mathcal{P}_l) \subset \mathcal{V}(\mathcal{P}_l),\ (C + D_a F_a)\mathcal{V}(\mathcal{P}_l) = 0. \quad (5)$$

In addition, utilizing any $F_a \in \underline{F}_a(\mathcal{V}(\mathcal{P}_l))$ ($\underline{F}_a(\mathcal{V}(\mathcal{P}_l))$ is the set of all linear maps satisfying (5)) and any L_a satisfying

$$\mathrm{Im}L_a = \ker D_a \ \cap\ B_a^{-1}\mathcal{V}(\mathcal{P}_l), \quad (6)$$

we can define an input a as follows:

$$a := F_a x_l + L_a \omega(t), \quad (7)$$

where $\omega(t)$ is any vector function with proper dimensions. It has been proved in (Trentelman et al., 2012) that for the initial condition $x_{l0} \in \mathcal{V}(\mathcal{P}_l)$, the output $y_l(t)$ resulting from a and x_{l0} is identically zero if and only if a is designed as (7). Moreover, for any subspace $\mathcal{K}$ of $\mathbb{R}^{n_p}$, we can define a largest controlled invariant subspace for $\mathcal{P}_l$ contained in $\mathcal{K}$ as follows: $\mathcal{I}(\mathcal{K}) := \{x_{l0} \in \mathbb{R}^{n_p} \mid$ there exists an input function a such that $x_l(t) \in \mathcal{K}$ for all $t \geq T_0\}$.

3.2 Attack Model

By replicating $\mathcal{P}_l$ and based on (7), the attack model is proposed as

$$\mathcal{G} : \begin{cases} \dot{z} = (A + B_a F_a)z + B_a L_a \omega(t),\ z(T_0) = z_0, \\ a = F_a z + L_a \omega(t), \end{cases} \quad (8)$$

where $F_a \in \underline{F}_a(\mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x))))$, $\mathrm{Im}L_a = \ker(D_a) \cap B_a^{-1}(\mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x))))$, and $\omega(t)$ is any vector function with proper dimensions. Now, we are ready to present the stealthy property in the following lemma.

Lemma 1. Under Assumptions 1 and 2, the attacks generated by the attack model $\mathcal{G}$ are stealthy attacks defined in Definition 1 if F_a and L_a satisfy

$$F_a \in \underline{F}_a(\mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x)))), \quad (9)$$

$$\mathrm{Im}L_a = B_a^{-1}(\mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x)))), \quad (10)$$

and z_0 satisfies

$$z_0 \in \mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x))) \cap \Omega_0, \quad (11)$$

where Ω_0 is defined in (15). □

Proof. By introducing the error variable $e = x^a - z$, $\tilde{y}^a$ in $\mathcal{P}^a$ can be written as $\tilde{y}^a = \tilde{y}_1^a + \tilde{y}_2^a$ where

$$\tilde{y}_1^a = Ce,$$
$$\tilde{y}_2^a = (C + D_a F_a)z + B_a L_a \omega.$$

Thus, a feasible solution to guarantee $\tilde{y}^a$ to satisfy Definition 1 is that $\tilde{y}_2^a$ is identically zero and $\tilde{y}_1^a$ asymptotically

converges to zero. According to the aforementioned *geometric control theory*, by selecting

$$z_0 \in \mathcal{V}(\mathcal{P}_l) \cap \mathcal{I}(\ker(G_p(t,\xi_x))),$$

$\tilde{y}_2^a = 0$ and $G_p(t,\xi_x)z = 0$ as well. Subsequently, the dynamics $(\mathcal{P}^a, \mathcal{C}^a, \mathcal{G})$ can be characterized by the following reduced-order dynamics in the coordinates (e, c^a):

$$\dot{e} = (A + G_p(t,\xi_x))e + Bu^a, \tag{12}$$
$$\dot{c}^a = G_{c1}(t,\xi_c,\xi_{\tilde{y}})c^a + G_{c2}(t,\xi_c,\xi_{\tilde{y}})\tilde{y}^a, \tag{13}$$
$$u^a = K_1 c^a + K_2 Ce,\ \tilde{y}^a = Ce, \tag{14}$$

where $e(T_0) = -z_0$ and $c^a(T_0) = 0$. From Assumption 2, and based on Theorem 4.9 in (Khalil, 2002), every trajectory starting from $ec^a(T_0)$ where $ec^a(T_0) = [e^T(T_0), c^{aT}(T_0)]^T$, satisfies $\|ec^a(t)\| \leq \beta(\|ec^a(T_0)\|, t - T_0)$ where $\beta(\cdot,\cdot)$ is a class $\mathcal{KL}$ function. Consequently, it follows from (14), that for $t \geq T_0$, one gets

$$\begin{aligned}\|\tilde{y}^a(t)\| \leq& \|C\|\beta(\|ec^a(T_0)\|, t - T_0),\\ \|u^a(t)\| \leq& \|[K_1, K_2C]\|\beta(\|ec^a(T_0)\|, t - T_0),\end{aligned}$$

which indicates that as $t \to \infty$, $\tilde{y}^a(t) \to 0$ and $u^a(t) \to 0$.

Let $x_r^a = x_r - x_r^n$ and $y_r^a = y_r - y_r^n$ denote the increments of x_r and y_r respectively due to attacks. Then, it follows from (4) that

$$\mathcal{D}^a: \begin{cases}\dot{x}_r^a = A_r x_r^a + g_p(t,x_r) - g_p(t,x_r^n) + Bu^a + L\tilde{y}^a,\\ y_r^a = Cx_r^a + \tilde{y}^a,\end{cases}$$

where $x_r^a(T_0) = 0$, and L is supposed to be designed such that for $t > 0$ and $x_r^a \in \mathcal{X}_{cp}$, there exists a continuously positive definite differentiable function $V_D : \mathbb{R}_+ \times \mathcal{X}_{cp} \to \mathbb{R}$ such that $\frac{\partial V_D}{\partial t} + \frac{\partial V_D}{\partial x_r^a}(A_r x_r^a + g_p(t,x_r) - g_p(t,x_r^n)) \leq -\alpha(x_r^a)$ where $\alpha(\cdot)$ is a continuous positive definite function on $\mathcal{X}_{cp}$. Then, there exists a function $\mu(\|u^a\|, \|\tilde{y}^a\|)$ where $\mu(\|u^a\|, \|\tilde{y}^a\|) \to 0$ as $\|u^a\| \to 0$ and $\|\tilde{y}^a\| \to 0$, such that for all $\|x_r^a\| \geq \mu(\|u^a\|, \|\tilde{y}^a\|)$, $\frac{\partial V_D}{\partial t} + \frac{\partial V_D}{\partial x_r^a}(A_r x_r^a + g_p(t,x_r) - g_p(t,x_r^n) + Bu^a + L\tilde{y}^a) < 0$. Based on Theorem 4.18 in (Khalil, 2002), the solution $x_r^a(t)$ to $\mathcal{D}^a$ satisfies

$$\|x_r^a(t)\| \leq \beta(0, t - T_0) + \rho(\mu(\|u^a\|, \|\tilde{y}^a\|)), \forall\ t \geq T_0$$

where $\rho(\cdot)$ is a class $\mathcal{K}$ function. Furthermore, based on the relation that $y_r^a = Cx_r^a + \tilde{y}^a$, $y_r^a(t)$ satisfies $\|y_r^a(t)\| \leq \delta(z_0, t),\ \forall\ t \geq T_0$, where

$$\begin{aligned}\delta(z_0,t) =& \|C\|(\beta(0, t - T_0) + \rho(\mu(\|u^a\|, \|\tilde{y}^a\|))\\ &+ \beta(\|ec^a(T_0)\|, t - T_0).\end{aligned}$$

Thus, since $\mu(\|u^a\|, \|\tilde{y}^a\|) \to 0$ as $t \to \infty$, $\delta(z_0, t) \to 0$ as $t \to \infty$, and further, $\|y_r^a(t)\| \to 0$ as $t \to \infty$.

In addition, based on the superposition relation $y_r = y_r^n + y_r^a$, for $t \geq T_0$,

$$\begin{aligned}\|y_r(t)\|_{2,T_w} &\leq \|y_r^n(t)\|_{2,T_w} + \|y_r^a(t)\|_{2,T_w}\\ &\leq \|y_r^n(t)\|_{2,T_w} + \int_t^{t+T_w} \delta^2(z_0,\tau)d\tau.\end{aligned}$$

Thus, if $z_0 \in \Omega_0$ where

$$\Omega_0 = \{z_0 | \int_t^{t+T_w} \delta^2(z_0,\tau)d\tau \leq J_{th} - \|y_r^n\|_{2,T_w}\}, \tag{15}$$

then $\|y_r\|_{2,T_w} \leq J_{th}$ for $t \geq T_0$.

Consequently, Definition 1 is satisfied, and the result follows. □

From Lemma 1, the stealthy attacks can not be detected by the traditional anomaly detectors using FTD methodologies. Attack detection schemes based on BTD methodologies will be designed in the following section for detecting these stealthy attacks.

4. ENHANCED ANOMALY DETECTOR USING BTD METHODOLOGY

4.1 Diagnostic H_∞ Fixed-point Smoother

Let $\breve{x}(T_b)$ be the unknown analytical value of the state of the plant at $T_b < T_0$. Then $\breve{x}(T_b) \neq x^n(T_b)$ due to attacks. The aim of the fixed-point smoother is to constructing a procedure for providing an estimate for $\breve{x}(T_b)$. Firstly, the unified dynamics of the underlying CPS for both the non-attack cases and attack cases should be determined. By introducing X and $\tilde{Y}$ defined as follows:

$$X = \begin{cases} x^n,\ t < T_0,\\ x - z,\ t \geq T_0,\end{cases} \quad \tilde{Y} = \begin{cases}\tilde{y}^n,\ t < T_0,\\ \tilde{y},\ t \geq T_0,\end{cases}$$

the unified dynamics of $\mathcal{P}$, $\mathcal{N}_a$ and $\mathcal{N}_s$ in the non-attack cases and in the presence of the attack a generated by $\mathcal{G}$ under the condition of Lemma 1 can be described by

$$\dot{X} = AX + g_p(t,X) + B\tilde{u} + D_1 d, \tag{16}$$
$$\tilde{Y} = CX + D_2 d. \tag{17}$$

In order to design the fixed-pointed smoother, the dynamical variable $\phi(t)$ satisfying

$$\dot{\phi}(t) = 0,\ \phi(T_b) = \breve{x}(T_b) \tag{18}$$

is also introduced for $t \geq T_b$. The output $\breve{y}(T_b)$ ($\breve{y}(T_b)$ is the unknown output corresponding to $\breve{x}(T_b)$) can be expressed in terms of ϕ as

$$\breve{y}(T_b|t) = C\phi(t). \tag{19}$$

Let $\hat{X}$, $\hat{\phi}$ and $\hat{\breve{y}}(T_b|t)$ denote the estimates of X, ϕ and $\breve{y}(T_b|t)$, respectively. The H_∞ diagnostic smoother over the time interval $[T_b, T_s]$ ($T_s > T_0$) is proposed as follows:

$$\mathcal{S}\begin{cases}\dot{\hat{X}} = A\hat{X} + g_p\left(t,\hat{X}\right) + B\tilde{u} - K_X(t)\left(\tilde{Y} - C\hat{X}\right),\\ \dot{\hat{\phi}} = -K_\phi(t)\left(\tilde{Y} - C\hat{X}\right),\\ \hat{\breve{y}}(T_b|t) = C\hat{\phi}(t),\end{cases}$$

where the time-varying matrices $K_X(t)$ and $K_\phi(t)$ are respectively designed as

$$K_X(t) = -\pi_1(t)C^T R^{-1},\ K_\phi(t) = -\pi_2^T(t)C^T R^{-1},$$

where $R = D_2 D_2^T$, and $\pi_1(t)$ and $\pi_2(t)$ are respectively determined by

$$\begin{aligned}\dot{\pi}_1 =& \pi_1 A^T + A\pi_1 - \pi_1\left(C^T R^{-1}C - \epsilon^{-2}l^2 I\right)\pi_1\\ &+\pi_2\left(\epsilon^{-2}l^2 I + \gamma^{-2}C^T C\right)\pi_2 + \epsilon^2 I + D_1 D_1^T,\end{aligned} \tag{20}$$

$$\begin{aligned}\dot{\pi}_2 =& A\pi_2 - \pi_1\left(C^T R^{-1}C - \epsilon^{-2}l^2 I\right)\pi_2\\ &+\pi_2\left(\epsilon^{-2}l^2 I + \gamma^{-2}C^T C\right)\pi_3,\end{aligned} \tag{21}$$

where ϵ is any nonzero scalar, γ is the value of the given H_∞ performance index (further discussed later), and π_3 is determined by

$$\begin{aligned}\dot{\pi}_3 =& -\pi_2^T\left(C^T R^{-1}C - \epsilon^{-2}l^2 I\right)\pi_2\\ &+\pi_3\left(\epsilon^{-2}l^2 I + \gamma^{-2}C^T C\right)\pi_3 + \epsilon^2 I.\end{aligned} \tag{22}$$

Moreover, the values of π_1, π_2 and π_3 at starting time instant T_b satisfy

$$\pi_1(T_b) = \pi_2(T_b),\ \begin{bmatrix}\pi_1(T_b) & \pi_2(T_b)\\ \pi_2^T(T_b) & \pi_3(T_b)\end{bmatrix} = \Theta_0^{-1},$$

where $\Theta_0 = \Theta_0^T > 0$ is a given matrix. Hence, the following proposition is ready to be presented.

Proposition 1. Under Assumption 1, for the given scalar $\gamma > 0$ and any nonzero scalar ϵ, if there exists a solution π_1, π_2 and π_3 to the dynamics (20)-(22), then the smoother $\mathcal{S}$ can guarantee that

$$\frac{\int_{T_b}^{T_s} \left\| \breve{y}\left(T_b|t\right) - \hat{\breve{y}}\left(T_b|t\right) \right\|^2 d\tau}{\int_{T_b}^{T_s} \left\| d \right\|^2 dt + \left\| e(T_b) \right\|_{\Theta_0}^2} \leq \gamma^2, \tag{23}$$

where $e^T(T_b) = [(X - \hat{X})^T(T_b), (\phi - \hat{\phi})^T(T_b)]$. □

Proof. The proof is omitted due to space limitations.

4.2 Residual and Threshold Generation

The residual is proposed as follows:

$$r_a\left(T_b|t\right) \triangleq Cx_r^n(T_b) - \hat{\breve{y}}(T_b|t), \tag{24}$$

and the type of norm $\| \cdot \|_{2,T_w}$ is used as the evaluation function, that is $J_a\left(T_b|t\right) = \left\| r_a\left(T_b|t\right) \right\|_{2,T_w}$. The threshold is proposed as follows:

$$J_{ath} \triangleq J_{th}(\bar{d}) + \|D_2\| T_w \bar{d} + \gamma \sqrt{k_0 + (T_s - T_b)\bar{d}^2}. \tag{25}$$

The decision principle is given as follows:

- if $J_a(T_b|t) > J_{ath}$, then alarms are triggered,
- else, no alarm.

Thus, the detection time T_d can be defined as

$$T_d \triangleq \inf\left\{ t > T_b \,|\, J_a(T_b|t) > J_{ath} \right\}. \tag{26}$$

Subsequently, we have the following theorem:

Theorem 1. (Robustness). For the CPS $(\mathcal{P}, \mathcal{C}, \mathcal{D}, \mathcal{N}_a, \mathcal{N}_s)$ satisfying Assumptions 1 and 2, and the stealthy attacks determined by $\mathcal{G}$ and Lemma 1, the detection decision scheme, characterized by the smoother $\mathcal{S}$, residual (24) and threshold (25), guarantees that there will be no false alarm before the occurrence of the stealthy attacks (i.e., for $t \leq T_0$). □

Proof. From (17) and (19), the residual $r_a\left(T_b|t\right)$ can be split into $r_a\left(T_b|t\right) = y_r^n(T_b) + C\breve{x}^a(T_b) + \breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t) + D_2 d(T_b)$ where $\breve{x}^a(T_b) = \breve{x}(T_b) - x^n(T_b)$. Based on the triangle inequality of vector norms, $J_a\left(T_b|t\right)$ satisfies

$$\begin{aligned} J_a\left(T_b|t\right) \leq & \|y_r^n(T_b)\|_{2,T_w} + \|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_w} & (27)\\ & + \|D_2 d(T_b)\|_{2,T_w} + \|C\breve{x}^a(T_b)\|_{2,T_w}. & (28) \end{aligned}$$

In order to tolerate limit for disturbances under attack free operation conditions, the threshold J_{ath} can be chosen as

$$\begin{aligned} J_{ath} = & \sup_{\breve{x}^a(T_b)=0,\ t \geq T_b} J_a\left(T_b|t\right) = \|y_r^n(T_b)\|_{2,T_w} \\ & + \|D_2 d(T_b)\|_{2,T_w} + \sup(\|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_w}). \end{aligned}$$

Due to the fact that $T_s \geq T_w$, $\|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_w} \leq \|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_s}$. In addition, for any $\hat{x}(T_b)$ and $\hat{\phi}(T_b)$, there always exists a constant k_0 such that $\|e(T_b)\|_{\Theta_0}^2 \leq k_0$. For the bounded disturbance, $\int_{T_b}^{T_s} d^T d \; d\tau \leq (T_s - T_b)\bar{d}^2$. Thus, it follows from Proposition 1 that $\|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_w} \leq \gamma\sqrt{k_0 + (T_s - T_b)\bar{d}^2}$. It straightforwardly follows from the anomaly detector $\mathcal{D}$ that $\|y_r^n(T_b)\|_{2,T_w} \leq J_{th}(\bar{d})$. For the bounded disturbance d, $\|D_2 d(T_b)\|_{2,T_w} \leq \|D_2\| T_w \bar{d}$. Hence, the threshold (25) follows. □

4.3 Detectability Analysis

To characterize the class of stealthy attacks that can be detected, the following theorem is presented.

Theorem 2. (Detectability) For the CPS $(\mathcal{P}, \mathcal{C}, \mathcal{D}, \mathcal{N}_a, \mathcal{N}_s)$ satisfying Assumptions 1 and 2, and the attack detection decision scheme characterized by the smoother $\mathcal{S}$, residual (24) and threshold (25), if

$$\|C\breve{x}^a(T_b)\|_{2,T_w} \geq 2J_{ath}, \tag{29}$$

then, the attack a generated by $\mathcal{G}$ under the condition in Lemma 1, can be detected at T_d (i.e., $J_a(T_b|T_d) > J_{ath}$). □

Proof. In order to detect the attacks, $J_a(T_b|t)$ should be larger than J_{ath} at time instant T_d, which requires

$$\|r_a\left(T_b|T_d\right)\|_{2,T_w} \geq J_{ath}. \tag{30}$$

Due to (27), and $\|y_r^n(T_b)\|_{2,T_w} \leq J_{th}(\bar{d})$, $\|D_2 d(T_b)\|_{2,T_w} \leq \|D_2\| T_w \bar{d}$ and $\|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_w} \leq \|\breve{y}(T_b|t) - \hat{\breve{y}}(T_b|t)\|_{2,T_s}$, a sufficient condition to guarantee (30) can be proposed as $\|C\breve{x}^a(T_b)\|_{2,T_w} \geq J_{ath} + J_{th}(\bar{d}) + \|D_2\| T_w \bar{d} + \gamma\sqrt{k_0 + (T_s - T_b)\bar{d}^2}$ which, based on (25), is equivalent to the inequality (29). Hence, the result follows. □

5. SIMULATION

The knowledge of $\mathcal{P}$ is given as follows:

$$A = \begin{bmatrix} -0.975 & 0 & 0.042 & 0 \\ 0 & -0.977 & 0 & 0.044 \\ 0 & 0 & -0.958 & 0 \\ 0 & 0 & 0.2 & 0.36 \end{bmatrix},\ g_p = \begin{bmatrix} 0.5\sin(0.2x_1) \\ 0 \\ 0 \\ 0 \end{bmatrix},$$
$$B = \begin{bmatrix} 0.0515 & 0.0016 \\ 0.0019 & 0.0447 \\ 0 & 0.0737 \\ 0.0850 & 0 \end{bmatrix},\ D_1 = \begin{bmatrix} 0.1 & 0.1 \\ 0.1 & 0.2 \\ 0.2 & 0 \\ 0 & 0.1 \end{bmatrix},$$
$$C = \begin{bmatrix} 0 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & -0.1 \end{bmatrix}, D_2 = \begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.25 \end{bmatrix},$$

where $x = [x_1, x_2, x_3, x_4]^T$ is the state, and the pair (A, C) is observable. The $\mathcal{C}$ is a static output feedback controller where $g_c = 0$, $K_1 = 0$, and

$$K_2 = \begin{bmatrix} -31.6353 & 153.9421 \\ 25.3556 & -212.9760 \end{bmatrix}.$$

The anomaly detector $\mathcal{D}$ is a typical fault detector designed based on (Ding, 2008).

Based on the given g_p, it can be calculated that the Lipschitz constant $l = 0.1$ and

$$G\left(\xi\right) = \begin{bmatrix} 0.1\cos(0.2x_1) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then, using the *Geometric Approach Toolbox* provided in (Basile and Marro, 1992), the corresponding weakly unobservable subspace $\mathcal{V}(\mathcal{P}_l)$, F_a and L_a can be respectively calculated as: $L_a = [0.0311, -0.9995, 0, 0]^T$,

$$\mathcal{V}(\mathcal{P}_l) = \mathrm{Im}\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ F_a = \begin{bmatrix} 0 & 0.0572 & -0.0299 & 0.0838 \\ 0 & -1.8403 & 0.9377 & -2.6973 \\ 0 & 0 & -0.0002 & 0 \\ 0 & -0.0002 & 0 & 0.0001 \end{bmatrix} \times 10^3.$$

Thus, the attack model $\mathcal{G}$ can be constructed based on (8), and z_0 can also be determined based on (11).

As for the diagnostic H_∞ fixed-point smoother, the preset parameters are given as follows: $\gamma = 3.5$, $\epsilon = 0.2$, $T_b = 0$, $\pi_1(T_b) = 9.8124I$, $\pi_2(T_b) = \pi_1(T_b)$ and $\pi_3(T_b) = 19.6248I$. Moreover, z_0 is chosen as $z_0 = [0, 2, -1, 5]$ and $\omega(t) = 500\sin(2t)$. For simulation purpose, the disturbance d is given by $d = [0.3\cos(20t + 0.2), 0.6\sin(10t + 0.3)]^T$. The attack is performed at $T_0 = 25$. The time responses of attack a and $\tilde{y}^a$ are shown in Figs. 2 and 3, where the

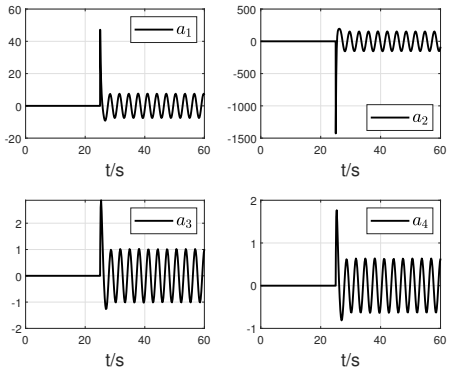

Fig. 2. Time responses of the attack a.

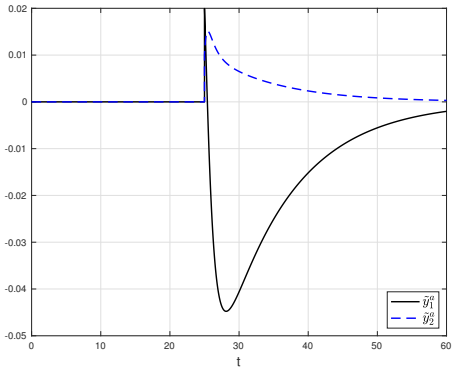

Fig. 3. Time responses of the increment $\tilde{y}^a$ due to the attack.

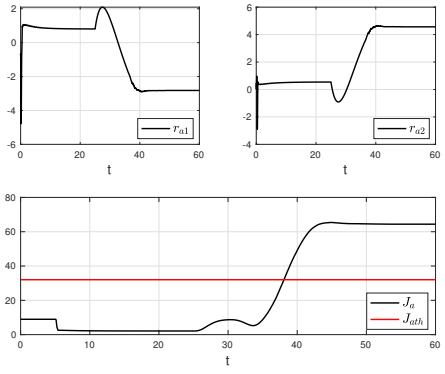

Fig. 4. Time responses of residual r_a, evaluation J_a and threshold J_{ath}.

increments $\tilde{y}_1^a$ and $\tilde{y}_2^a$ are small enough, and converge to zeros asymptotically.

In the residual generation and threshold generation schemes, the time window T_w is chosen as 5, $\bar{d} = 0.6708$ and $k_0 = 11$. In addition, $J_{th}(\bar{d})$ is chosen as $J_{th}(\bar{d}) = 10$. Thus, the threshold J_{ath} can be calculated as $J_{ath} = 32.0139$. The attack detection result is shown in Fig. 4. It can be seen that the attack a is detected at time instant T_d.

6. CONCLUSIONS

In this paper, stealthy integrity attacks have been redefined in the context of nonlinear CPSs. The BTD methodology has been proposed, and the H_∞ fixed-point smoother and corresponding residual generation and threshold generation schemes have also been designed for implementing the BTD methodology. Simulation results have been presented at last to verify the effectiveness of the enhanced anomaly detector. In our future work, the distinguishment issues between faults and attacks for more general nonlinear (not limited to Lipschitz nonlinearities) CPSs will be considered base on the BTD methodology developed in this paper.

REFERENCES

Basile, G. and Marro, G. (1992). *Controlled and conditioned invariants in linear system theory*. Prentice Hall, New Jersey.

Blanke, M., Kinnaert, M., Lunze, J., Staroswiecki, M., and Schröder, J. (2006). *Diagnosis and fault-tolerant control*. Springer Science & Business Media.

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.

Chen, J. and Patton, R. (2012). *Robust model-based fault diagnosis for dynamic systems*. Springer Science & Business Media.

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

Ferrari, R. and Teixeira, A. (2017). Detection and isolation of replay attacks through sensor watermarking. *IFAC-PapersOnLine*, 50(1), 7363–7368.

Khalil, H. (2002). *Nonlinear Systems*. Prentice Hall, New Jersey.

Mo, Y. and Sinopoli, B. (2010). False data injection attacks in cyber physical systems. In *First Workshop on Secure Control Systems*.

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.

Pasqualetti, F., Dörfler, F., and Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE transactions on automatic control*, 58(11), 2715–2729.

Smith, R. (2015). Covert misappropriation of networked control systems: Presenting a feedback structure. *IEEE Control Systems Magazine*, 35(1), 82–92.

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

Trentelman, H., Stoorvogel, A., and Hautus, M. (2012). *Control theory for linear systems*. Springer Science & Business Media.

Zhang, X., Polycarpou, M., and Parisini, T. (2010). Fault diagnosis of a class of nonlinear uncertain systems with lipschitz nonlinearities using adaptive estimation. *Automatica*, 46(2), 290–299.