# Secure Control of Networked Control Systems Using Dynamic Watermarking

Dajun Du, Changda Zhang, Xue Li, Minrui Fei, Taicheng Yang, Huiyu Zhou

Abstract-We here investigate secure control of networked control systems developing a new dynamic watermarking (DW) scheme. Firstly, the weaknesses of the conventional DW scheme are revealed, and the tradeoff between the effectiveness of false data injection attack (FDIA) detection and system performance loss is analysed. Secondly, we propose a new DW scheme, and its attack detection capability is interrogated using the additive distortion power of a closed-loop system. Furthermore, the FDIA detection effectiveness of the closed-loop system is analysed using auto/cross covariance of the signals, where the positive correlation between the FDIA detection effectiveness and the watermarking intensity is measured. Thirdly, the tolerance capacity of FDIA against the closed-loop system is investigated, and theoretical analysis shows that the system performance can be recovered from FDIA using our new DW scheme. Finally, experimental results from a networked inverted pendulum system demonstrate the validity of our proposed scheme.

*Index Terms*—Dynamic watermarking, attack detection effectiveness, system performance, watermarking intensity, tolerance capacity.

### I. INTRODUCTION

With the rapid popularization and application of network technologies, the coming decades may witness the extensive deployment of networked control systems (NCSs). Such systems are often embedded by physical plants and digital devices (e.g., digital filter and controller), which are linked by communication networks. Smart grids [1], Internet of Things [2] and networked robots [3] are examples of NCSs. Yet, nowadays it is easier to access the network by malicious users, and NCSs are vulnerable to cyber attacks such as denial-ofservice attack [4], replay attack [5], [6], and false data injection attack (FDIA) [7]-[9]. Insecure and attacked NCSs may suffer fateful consequences including huge economic losses, just as the attacks on the Iran nuclear plant [5] and Ukraine electric grid [7] demonstrated. In this context, it is not surprising that secure control of NCSs has attracted widespread attention with emerging attack detection [10] and resilient control [11].

Based on the use of the probing signals, attack detection can be roughly classified into passive and active detection. Passive detection does not inject the probing signal into the system,

D. Du, C. Zhang, X. Li, and M. Fei are with Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China.

T. Yang is with Department of Engineering and Design, University of Sussex, Brighton BN1 9QT, U.K.

H. Zhou is with School of Informatics, University of Leicester, Leicester LE1 7RH, U.K.

which is sometimes invalid for some sophisticated attacks such as replay attack [12] and FDIA [13]. To solve the problem, active detection is developed by injecting the probing signal into the system, and watermarking-based detection is a kind of typical active detection. It is motivated by the traditional digital watermarking (i.e., the digital code) that is embedded in electronic documents, which is employed to preserve the valuable information [14]. According to signal sources of generation, watermarking-based detection can be roughly divided into two categories: additive and multiplicative watermarking. A typical example of additive watermarking is termed as physical watermarking or dynamic watermarking (DW). Their basic concept is that the control law is actively encrypted by injecting certain probing signal (e.g., an independent identically distributed [12], [15], [16] or stationary [17] Gaussian signal), and specific tests (e.g.,  $\chi^2$  test [12] or DW tests [18]– [20]) are performed to infer a malicious activity. Physical watermarking [12] is designed particularly for preventing replay attack, which is also driven to investigate the tradeoff between the watermarking intensity and replay attack detection effectiveness. As an evolution of physical watermarking, DW [18] is designed to yield the security property, which has been further improved for general linear time-invariant systems [21] and time-varying systems [22].

However, physical watermarking and DW accompany nonzero system performance loss for attack detection. Indeed, the lower watermarking intensity is cautiously chosen in the system to ensure less system performance loss, where the worse attack detection effectiveness is determined. Against these limitations, multiplicative watermarking has been developed, and a typical example is termed as sensor multiplicative watermarking [23]–[25]. In such watermarking scheme, each sensor output is separately encrypted by an infinite/finite impulse response filter [23], and the encrypted sensor output is decrypted by an equalizing filter [24], [25]. With the advantage of zero system performance loss from the watermarking scheme, sensor multiplicative watermarking is also originally designed for replay attack [23] and later developed for FDIA (e.g., routing attack [24] as well as man-in-the-middle attack [25]). But then, to the best of our knowledge, there is no theoretical guarantee of the security property using DW for sensor multiplicative watermarking. Therefore, to this end, the design of a new comprehensive watermarking scheme combining the security property of DW and the advantage of sensor multiplicative watermarking is developed.

At a high level, resilient control has been investigated in several research works, which can be roughly classified into detection-independent and detection-dependent con-

The work of D. Du, C. Zhang, X. Li, M. Fei, T. Yang, and H. Zhou was supported by the National Science Foundation of China under Grant Nos. 92067106, 61773253, 61633016 and 61533010, 111 Project under Grant No.D18003.

trol. Inspired from fault-tolerant control [26]-[28], detectionindependent control examines the tolerance capacity of attacks against systems; e.g., the characterization of maximum perturbation from FDIA is posed as reachable set computation [29] and an FDIA tolerance principle is established based on adaptively truncating the injection channels of attacks [30]. Note that even though the tolerance capacity of attacks against systems is examined, the system experiences performance loss, which calls for detection-dependent control. The idea of detection-dependent control is that once there is an attack to be alarmed, mechanism-depending detection can be used to recover system performance. For example, modification of the attacked signals [31] and the compensation mechanisms [32]–[34] are conducted, where the malicious system outputs are discarded. However, to the best of our knowledge, few research studies concerns with resilient control based on the watermarking schemes. Therefore, an additional desire is to design a mechanism for the proposed watermarking scheme to recover the system performance under attacks.

Motivated by the above observations, this paper looks at a new DW scheme. Specifically, the following challenges will be addressed:

- (1) What are the security weaknesses of the conventional DW scheme?
- (2) How to design a new DW scheme against the security weaknesses? What are the security property and the attack detection performance of the new DW scheme?
- (3) What is the recovery capability of NCSs based on the new DW scheme?

To deal with these challenges, this paper explores a new DW scheme under cyber attacks. The main contributions of this paper are summarized as follows:

- (1) The security weaknesses of the conventional DW scheme are revealed, and a tradeoff between the FDIA detection effectiveness and system performance loss is explored. In addition, when a low watermarking intensity has to be chosen, there exists an FDIA that cannot be effectively detected and causes instability by the conventional DW scheme.
- (2) A new DW scheme is proposed by integrating the watermarking as symmetric-key encryption and new DW testing and compensation, where attacks can be detected using the additive distortion power of the closed-loop system. Furthermore, FDIA detection effectiveness of the closed-loop system is analysed using auto/cross covariance of signals, and the positive correlation between FDIA detection effectiveness and watermarking intensity is analysed.
- (3) The tolerance capacity of FDIA against the closed-loop system is investigated, and it is shown that system performance can be recovered from FDIA, where the quantitative relationship between the new DW scheme and the system performance is revealed.

To clearly present the contributions of this paper, compared with existing results in the literatures, a comparative analysis is listed in Tab. I. It is shown from Tab. I that the existing results have only solved part of the problems of attack

TABLE I Comparative Analysis between the Contributions of This Paper and the Existing Results in the Literatures

	$ADEW^1$	$SPAW^2$	$ZPLW^3$	$ATCA^4$	$SPR^5$
CDW <sup>6</sup> [12], [15]–[22]	1	~	×	×	×
SMW <sup>7</sup> [23]–[25]	1	×	1	×	×
DIC <sup>8</sup> [29], [30]	×	×	1	1	×
DDC <sup>9</sup> [31]–[34]	×	×	1	1	1
New DW (this paper)	1	1	1	1	~

<sup>1</sup> Attack detection effectiveness enhanced by watermarking.

<sup>2</sup> Security property analysis from CDW.

<sup>3</sup> Zero system performance loss from watermarking.

<sup>4</sup> Attack tolerance capacity analysis. <sup>5</sup> System performance recovery.

<sup>6</sup> Conventional DW. <sup>7</sup> Sensor multiplicative watermarking.

<sup>8</sup> Detection-independent control. <sup>9</sup> Detection-dependent control.

detection and resilient control, but the proposed new DW scheme can generate a comprehensive solution of the problems by enhancing the attack detection effectiveness, developing the security property, guaranteeing zero system performance loss from watermarking, and providing the attack tolerance capacity as well as system performance recovery.

The remainder of this paper is organized as follows. Section II contains problem formulation, focusing on the security property and weaknesses of the conventional DW scheme. Section III presents secure control of NCSs based on a new DW scheme, where the design of the new DW scheme, security property and processes are presented. Experimental results for an inverted pendulum system are given in Section IV, followed by conclusion shown in Section V.

*Notation:*  $\mathbb{E}[\cdot]$  represents the expectation of a vector or matrix,  $A_i$ .  $(A_{\cdot i})$  is used for the sub-matrix of a given matrix A formed from row (column) i,  $tr(\cdot)$  denotes the trace of a matrix,  $\rho(\cdot)$  denotes the spectral radius of a matrix,  $\|\cdot\|$  denotes the Euclidian norm of a vector or the spectral norm of a matrix,  $\|\cdot\|_F$  denotes the Frobenius norm of a matrix,  $|\cdot|$  denotes the absolute value of a real number, and  $\sup \mathbb{R}$  (inf  $\mathbb{R}$ ) denotes the supremum (infimum) of a real set  $\mathbb{R} \subseteq \mathcal{R}$ . The zero vector is denoted as  $0_n \in \mathcal{R}^n$ , whilst  $0_{n \times m} \in \mathcal{R}^{n \times m}$  and  $I_{n \times n} \in \mathcal{R}^{n \times n}$  indicate respectively the zero and identity matrices. For simplicity, the subscripts are often omitted if their dimensions are clear. Table II summarizes the notations most frequently used throughout the remainder of the paper.

### **II. PROBLEM FORMULATION**

### A. Security Property Analysis of the Conventional DW Scheme

The framework of secure control of NCSs based on the conventional DW scheme is illustrated in Fig. 1. The system output y(k) is first sampled and transmitted to the estimator via a network, which may be attacked, notated as  $y_a(k)$ . Note that when there exist cyber attacks,  $y_a(k) \neq y(k)$ ; otherwise  $y_a(k) = y(k)$ . Then, the estimator has the state estimate  $\hat{x}_d(k|k)$  by using  $y_a(k)$ , which is used to compute the controller signal  $u_d(k)$  in the controller node. Furthermore,

Т	≜	Time window size for detection
$\mathcal{W}_d, \mathcal{V}_d$	$\triangleq$	Conventional DW (asymptotic) tests 1, 2
$\varphi_{d,1}, \varphi_{d,2}$	$\triangleq$	Conventional DW statistical tests 1, 2
$\vartheta_{d,1}, \vartheta_{d,2}$	$\triangleq$	Thresholds for DW statistical tests 1, 2
$\mathcal{D}_{r_d}$	$\triangleq$	Additional distortion on residual $r_d$
$x_a$	$\triangleq$	State of the false data injection attack
$J^o$	$\triangleq$	System performance under no attack
$J^{o} _{w_d(k)=0}$	$\triangleq$	$J^o$ under no DW scheme
$\mathcal{W}_i, \mathcal{V}$	$\triangleq$	New DW (asymptotic) tests 1, 2
$\varphi_{1,i}, \varphi_2$	$\triangleq$	New DW statistical tests 1, 2
$\vartheta_{1,i}, \vartheta_2$	$\triangleq$	Thresholds for new DW statistical tests 1, 2
$\mathcal{D}_r$	$\triangleq$	Additional distortion on residual r
$J^{o} _{w_{y}(k)=0}$	$\triangleq$	$J^o$ under no new DW scheme
ε	$\triangleq$	Detection results being 0 or 1
h	$\triangleq$	Delay of healthy system output
e	$\triangleq$	Estimation error

TABLE II TABLE OF NOTATIONS

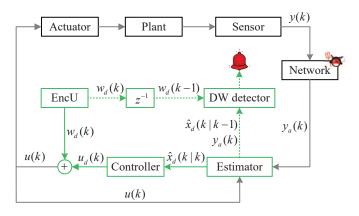


Fig. 1. Framework of secure control of NCSs based on the conventional Dynamic Watermarking (DW) scheme [12], [18].

a watermarking signal  $w_d(k)$  from the watermarking generator EncU is injected into  $u_d(k)$ , resulting in u(k), which is applied to the plant by the actuator. Meanwhile, using  $w_d(k-1)$  from the EncU and  $\hat{x}_d(k|k-1)$  and  $y_a(k)$  from the estimator, the DW detector detects whether or not  $y_a(k)$  is attacked; if an attack takes place, the alarm will be triggered.

To analyse the security property and weaknesses, the discrete-time linear time-invariant plant is considered as follows

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) + \Gamma n(k) \\ y(k) = Cx(k) + v(k) \end{cases}$$
(1)

where  $x(k) \in \mathcal{R}^{m_x}$  is the system state;  $u(k) \in \mathcal{R}$  is the control input;  $y(k) \in \mathcal{R}^{m_y}$  is the system output; the process noise  $n(k) \in \mathcal{R}^{m_n}$  and measurement noise  $v(k) \in \mathcal{R}^{m_y}$  are independent identically distributed (i.i.d.) zero-mean white Gaussians with covariance matrices  $\Sigma_n$  and  $\Sigma_v$  respectively, and mutually independent. A, B,  $\Gamma$ , C are constant matrices with appropriate dimensions.

Next, to estimate the system state for measuring control input and attack detection, the estimator adopts the steadystate Kalman filter by

$$\hat{x}_d(k|k) = \hat{x}_d(k|k-1) + L(y_a(k) - C\hat{x}_d(k|k-1)), \quad (2)$$

$$\hat{x}_d(k+1|k) = A\hat{x}_d(k|k) + Bu(k)$$
 (3)

where  $L = PC^T \Sigma_o^{-1}$  is the steady-state Kalman gain;  $\Sigma_o = CPC^T + \Sigma_v$ , and P is the unique positive definite solution of the Riccati equation [12].

According to  $\hat{x}_d(k|k)$ , the linear quadratic Gaussian controller is implemented to minimize the objective function Jwith matrices  $Q, R \ge 0$  [12]. It is well known that if there is no attacks (i.e.,  $y_a(k) \equiv y(k)$ ), the solution of the minimization problem on J will lead to a fixed gain controller:

$$u_d(k) = K\hat{x}_d(k|k) \tag{4}$$

where  $K = -(B^T SB + R)^{-1} B^T SA$  is the controller gain, and S is the unique positive definite solution of the Riccati equation [12]. Furthermore,  $u_d(k)$  is injected with a watermarking signal  $w_d(k)$  by EncU, i.e.,

$$u(k) = u_d(k) + w_d(k) \tag{5}$$

where  $w_d(k)$  is drawn from an i.i.d. Gaussian distribution with zero mean and variance  $\sigma_{w_d}^2$ , and  $w_d(k)$  is chosen to be also independent of  $u_d(k)$  [12].

Furthermore, back to the side of the estimator, its signals  $\hat{x}_d(k|k-1)$  and  $y_a(k)$  are transferred to the DW detector to evaluate whether or not the attack takes place. Therefore, using  $w_d(k-1)$ ,  $\hat{x}_d(k|k-1)$  and  $y_a(k)$ , two DW tests [18] are operated in the DW detector, i.e.,

1) DW Test 1: checking whether or not

$$\lim_{T \to \infty} \mathcal{W}_d(T) = 0 \tag{6}$$

where  $\mathcal{W}_d(T) := \frac{1}{T} \sum_{k=1}^T w_d(k-1) Lr_d(k)$ ; T is the time window size;  $r_d(k)$  is the residual, i.e.,  $r_d(k) := y_a(k) - C\hat{x}_d(k|k-1)$ .

2) DW Test 2: checking whether or not

$$\lim_{T \to \infty} \mathcal{V}_d(T) = 0 \tag{7}$$

where  $\mathcal{V}_d(T) := \frac{1}{T} \sum_{k=1}^T Lr_d(k) (Lr_d(k))^T - L\Sigma_o L^T;$  $\Sigma_o$  is the same as the one in (2) and also the covariance matrix of the normal residual.

*Remark 1:* From (6) and (7), it is clear that applying directly DW Tests 1 and 2 in real world is unrealistic owing to the limit  $T \to \infty$ . To solve the problem, two DW statistical tests need to be revisited in a finite T. For instance, two indicators have been defined in [19] as  $\varphi_{d,1}(k) := \|\mathcal{W}_d^k(T)\|_F$ and  $\varphi_{d,2}(k) := |tr(\mathcal{V}_d^k(T))|$ , where  $\mathcal{W}_d^k(T)$  and  $\mathcal{V}_d^k(T)$  are  $\mathcal{W}_d(T)$  and  $\mathcal{V}_d(T)$  calculated at the current time window  $\{k - T + 1, k - T + 2, \dots, k\}$  respectively. Furthermore, to evaluate whether or not the attack takes place, let  $\vartheta_{d,1}, \vartheta_{d,2}$  be the preset thresholds; when an attack takes place, we expect  $\varphi_{d,1}(k) \ge \vartheta_{d,1}$  or  $\varphi_{d,2}(k) \ge \vartheta_{d,2}$  so that the attack alarm can be made.

The above discussion refers to a framework of secure control of NCSs based on the conventional DW scheme described by (1)-(7), then the security property will be analysed. Note that to clearly distinguish between the normal (i.e., attack-free) system where  $y_a(k) \equiv y(k)$  and the system under attacks, we denote  $y^o(k)$ ,  $\hat{x}^o_d(k|k-1)$ ,  $r^o_d(k) = y^o(k) - C\hat{x}^o_d(k|k-1)$ ,  $J^o$  as the attack-free counterpart of  $y_a(k)$ ,  $\hat{x}_d(k|k-1)$ ,  $r_d(k)$ , J. Furthermore, to quantify the additional distortion caused by the attacker on systems, we define

$$\mathcal{D}_{r_d}(k) := Lr_d(k) - Lr_d^o(k). \tag{8}$$

According to  $\{\mathcal{D}_{r_d}\}$ , the *additive distortion power* [18] of systems is defined by

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathcal{D}_{r_d}(k)\|^2.$$
(9)

Using the above definition of additive distortion power (9), the security property of the conventional DW scheme is provided.

Security property. (cf. [18, Th. 5]) If  $y_a(k)$  passes the tests (6) and (7), then

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \|\mathcal{D}_{r_d}(k)\|^2 = 0$$
 (10)

which means that the additive distortion power of the systems that bypass the DW Tests 1 and 2 is restricted to be zero. ■

*Remark 2:* Eq. (10) interprets the theoretical foundation of attack detection. In other words, if an attacker (by, e.g., injecting false data into the network) forces  $y_a(k) \neq y(k)$  and dissatisfies (10), then the attacker will be detected by the conventional DW scheme.

## B. Security Weaknesses Analysis of the Conventional DW Scheme

The above has presented a framework of secure control of NCSs based on the conventional DW scheme and analysed its security property. However, with the FDIA, there remain limitations on attack detection effectiveness, and system performance loss from attacks and watermarking. The details of the FDIA and limitations are analysed below.

Firstly, to analyse limitations of the conventional DW scheme, the following FDIA (cf. [8], [22]) is given by

$$x_a(k+1) = A_a x_a(k), y_a(k) = C x_a(k), k \in [k_0^a, \infty)$$
(11)

where  $x_a(k) \in \mathcal{R}^{m_x}$  is the state of the FDIA;  $k_0^a$  is the initial instant of the FDIA;  $A_a$  is a constant matrix with  $\rho(A_a) < 1$ ; C is the same as the one in (1). Then, under the FDIA (11), the closed-loop system based on the conventional DW scheme described by (1)–(7) can be modelled as

$$\zeta_d(k+1) = \mathcal{A}_0 \zeta_d(k) + \Lambda_d \psi_d(k) \tag{12}$$

where variables  $\zeta_d(k) := [x(k); x_a(k) - \hat{x}_d(k|k-1); x_a(k)],$  $\psi_d(k) := [n(k); v(k); w_d(k)];$  the matrices  $\mathcal{A}_0$  and  $\Lambda_d$  are given by

$$\mathcal{A}_0 := \begin{bmatrix} A & \mathbf{H} \\ 0 & \Xi \end{bmatrix}, \Lambda_d := \begin{bmatrix} \Gamma & 0 & B \\ 0 & 0 & -B \\ 0 & 0 & 0 \end{bmatrix}$$

and  $H := [-BK(I + LC), BK], \Xi := [\Phi_1, \Phi_2; 0, A_a], \Phi_1 := (A + BK)(I - LC), \text{ and } \Phi_2 := A_a - (A + BK).$ 

Remark 3: When  $A_a$  satisfies  $\rho(A_a) < 1$  for the FDIA (11),  $y_a(k)$  looks like normal y(k); otherwise  $\rho(A_a) > 1$  means that  $y_a(k)$  will diverge, and it can easily judge whether the FDIA (11) takes place. Furthermore, it is reasonable to choose  $\rho(A_a) < 1$  for the FDIA (11) because it can be seen from the following limitations 1 and 2 that  $\rho(A_a) < 1$  yields stealthiness and destructiveness of the FDIA (11). An example of the FDIA (11) can be seen in [22, Th. II.3], where  $A_a = A + BK$  and  $\rho(A + BK) < 1$  in (11).

Secondly, to quantify attack detection effectiveness whilst considering the ergodic theorem [35] of stationary processes, the cross covariance of watermarking and residual and the auto covariance of residual are used to approximate the temporal averaging in (6) and (7) respectively, i.e.,

$$\mathbb{E}\left[w_d(k-1)Lr_d(k)\right] \approx \lim_{T \to \infty} \mathcal{W}_d(T), \tag{13}$$

$$\mathbb{E}\left[Lr_d(k)\left(Lr_d(k)\right)^T\right] \approx \lim_{T \to \infty} \mathcal{V}_d(T) + L\Sigma_o L^T.$$
(14)

Finally, according to the defined FDIA (11) and closed-loop system (12) and quantities (13), (14), the following three limitations are revealed.

**Limitation 1–FDIA Detection Effectiveness Limited by Watermarking.** For the system (12), the FDIA (11) will result in

$$\mathbb{E}\left[w_d(k-1)Lr_d(k)\right] = -\sigma_{w_d}^2 LCB,$$
(15a)

$$\mathbb{E}\left[Lr_d(\infty)\left(Lr_d(\infty)\right)^T\right] = LCM_dC^TL^T$$
(15b)

where  $M_d = \Phi_1 M_d \Phi_1^T + \sigma_{w_d}^2 B B^T$ , and  $\Phi_1$  is the same as the one shown in (12). The proof is given in Section I.A of the supplementary materials.

Limitation 2–System Performance Loss from the FDIA. For the system (12), if  $\rho(A) > 1$ , then the FDIA (11) will result in

$$\lim_{k \to \infty} x(k) = \infty.$$
 (16)

The proof is given in Section I.A of the supplementary materials.

*Remark 4:* From limitation 1, it seems that the FDIA (11) cannot bypass DW Tests (6) and (7), provided  $\sigma_{w_d}^2 \neq 0$  in (15a) and (15b). However, the DW statistical tests  $\varphi_{d,1}(k)$  and  $\varphi_{d,2}(k)$  are practically used, where loose (i.e., big) detection thresholds  $\vartheta_{d,1}$ ,  $\vartheta_{d,2}$  can be set. Note that when a small value of  $\sigma_{w_d}^2$  has to be chosen,  $\varphi_{d,1}(k) < \vartheta_{d,1}$  or  $\varphi_{d,2}(k) < \vartheta_{d,2}$  so that the DW statistical tests are bypassed by the FDIA (11) as presented in Section IV. Limitation 2 points out that if the plant (1) is open-loop unstable (i.e.,  $\rho(A) > 1$ ), the system state x(k) in (12) will diverge (or cross the limit in the real world) under the FDIA (11) even though a well-designed controller (4) is used. Therefore, two of the following tasks are to enhance FDIA detection effectiveness and to recovery the system performance.

Limitation 3-System Performance Loss From Watermarking. (cf. [12, Th. 3]) For the system (12), if there is no attacks (i.e.,  $y_a(k) \equiv y(k)$ ), then the attack-free system performance is

$$J^{o} = J^{o}|_{w_{d}(k)=0} + \Delta J^{o} \tag{17}$$

where  $\Delta J^o = \sigma_{w_d}^2 tr(B^T SB + R)$ ;  $J^o|_{w_d(k)=0}$  is the attack-free system performance without the conventional DW scheme;  $\Delta J^o/J^o|_{w_d(k)=0}$  is attack-free system performance loss from watermarking.

*Remark 5:* The tradeoff between FDIA detection effectiveness and system performance loss from watermarking is analysed as follows. Limitation 3 highlights that the inevitable cost paid for the above security property and FDIA detection effectiveness in limitation 1 of the conventional DW scheme is the attack-free system performance loss from watermarking. An intuitive example [12] is detection of replay attacks, while the example for detection of the FDIA (11) is given in Section IV. In the example of [12], one needs to pay the cost of 91% system performance loss from watermarking to collect about 35% detection rate at each step. Therefore, one of the following tasks is to decrease or eliminate the attack-free system performance loss from watermarking.

To deal with the security property and limitations 1-3, a new DW scheme will be designed by considering the following three aspects:

- a) The new DW scheme should develop the security property of the conventional DW scheme;
- b) The new DW scheme should be able to detect effectively the FDIA (11), while the FDIA detection effectiveness should be explored. The system performance loss from watermarking should be zero;
- c) The new DW scheme should be able to recovery system performance, while the relationship between the new DW scheme and its system performance should be explored.

### III. SECURE CONTROL OF NCSS BASED ON A NEW Dynamic Watermarking Scheme

The security property and weaknesses of the conventional DW scheme have been thoroughly analysed. To solve these problems, a new DW scheme integrating watermarking as symmetric-key encryption and new testing and compensation mechanism is firstly designed. Then, we investigate the security property and attempt to overcome the security weaknesses using the new DW scheme.

### A. A New DW Scheme

The framework of secure control of NCSs based on the new DW scheme integrating the watermarking as symmetrickey encryption and new testing and compensation mechanism is shown in Fig. 2. The new DW scheme is analogous to digital watermarking [14] (i.e., the digital code) that is embedded in electronic documents to preserve the valuable information. Therefore, motivated by digital watermarking and watermarking-as-key encryption [22] and symmetric key [36], the system output y(k) is first sampled and encrypted with a watermarking signal  $w_y(k)$  (as a key) by EncY (i.e.,  $y_w^+(k)$ ). After  $y_w^+(k)$  is transmitted over the network, which may be attacked and becomes  $y_a(k)$ , it is decrypted with  $w_u(k)$  by DecY and saved in the buffer (i.e.,  $y_w^-(k)$ ). Furthermore, using  $y_w^-(k)$  from the buffer,  $w_y(k)$  from the DecY and  $\hat{x}(k|k-1)$ from the estimator, the new DW detector checks whether or not  $y_w^-(k)$  is attacked and gives the corresponding detection

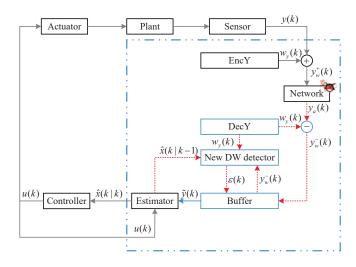


Fig. 2. Framework of secure control of NCSs based on the new DW scheme integrating watermarking as symmetric-key encryption and new testing and compensation mechanism.

results  $\varepsilon(k)$ . According to  $\varepsilon(k)$ , a compensation mechanism is used to update the input of the estimator (i.e.,  $\tilde{y}(k)$ ). Finally, the controller calculates u(k) using the state estimate  $\hat{x}(k|k)$ from the estimator. The following detail analysis is presented.

For the plant (1), to solve the problem of information leakage and improve the confidentiality of signals, y(k) is *encrypted* by EncY in Fig. 2, it follows that

$$y_w^+(k) = y(k) + w_y(k)$$
(18)

where  $w_y(k)$ , as a key, is a watermarking signal drawn from an i.i.d. Gaussian distribution with zero mean and covariance matrix  $\Sigma_{w_y} = diag\{\sigma_{w_{y,1}}^2, \ldots, \sigma_{w_{y,m_y}}^2\}$ , and independent of y(k). Then,  $y_w^+(k)$  is transmitted via the network, if attacked and then becomes  $y_a(k)$ . Note that when there exist cyber attacks,  $y_a(k) \neq y_w^+(k)$ ; otherwise  $y_a(k) = y_w^+(k)$ .

Next, to eliminate the side effect of the encrypted signal after transmission over the network,  $y_a(k)$  is *decrypted* by the DecY, i.e.,

$$y_w^-(k) = y_a(k) - w_y(k).$$
 (19)

Note that when there exist cyber attacks,  $y_w^-(k) \neq y(k)$ ; otherwise  $y_w^-(k) = y(k)$ . As a consequence, the watermarking signal  $w_y(k)$  and encryption (18) and decryption (19) comprise the watermarking as symmetric-key encryption.

*Remark 6:* For the decryption in (19), there is a practical key consistency problem: how to produce the same  $w_y(k)$  in (18) and (19) at the same instant k? The watermarking signal  $w_y(k)$  cannot simply be sent via the network because the attacker may observe and distort it *en route*. Alternatively, an online algorithm for producing  $w_y(k)$  in EncY and DecY is designed based on secure mechanism [36] and Marsaglia-Bray method [37], which can guarantee the same  $w_y(k)$  on both sides of the network.

*Remark 7:* Compared with the conventional DW scheme shown in Fig. 1 where the control input is encrypted using the method of (5), the new DW scheme encrypts the system output using the method of (18). When decryption is added to

**Algorithm 1:** Online Algorithm for Producing Same  $w_u(k)$  in EncY and DecY

- Initialization: The same seeds and sequence numbers for EncY and DecY are set, e.g., both seeds are set as 1 and sequence numbers are set as k<sup>+</sup><sub>w</sub> = k<sup>-</sup><sub>w</sub> = 1;
   while k = 1, 2, · · · do
- 3 EncY receives y(k) and then generates  $w_y(k_w^+)$  by using the seed and  $k_w^+$ ;
- 4 EncY encrypts y(k) with  $w_y(k_w^+)$  by using (18); 5  $k_w^+ \leftarrow k_w^+ + 1;$
- 6 DecY receives  $y_a(k)$  and then generates  $w_y(k_w^-)$ by using the seed and  $k_w^-$ ;
- 7 DecY decrypts  $y_a(k)$  with  $w_y(k_w^-)$  by using (19); 8  $k_w^- \leftarrow k_w^- + 1;$
- 8 k 9 end

the conventional DW scheme on the controller side, limitation 3 on system performance loss from watermarking may be overcome, but limitation 1 on FDIA detection effectiveness will still remain. Therefore, the watermarking as symmetric-key encryption (18) and (19), together with the following the new DW Tests, are integrated in the new DW scheme and expected to cope with limitations 1 and 3.

Furthermore,  $y_w^-(k)$  is saved in the buffer, which will be detected to determine whether or not this signal is attacked. Therefore,  $y_w^-(k)$  is transferred to the new DW detector, where the following two designed tests are operated using  $\hat{x}(k|k-1)$  from the estimator and  $w_y(k)$  from the DecY, i.e.,

1) New DW Test 1: checking whether or not

$$\lim_{T \to \infty} \mathcal{W}_i(T) = 0 \tag{20}$$

where  $\mathcal{W}_i(T) := \frac{1}{T} \sum_{k=1}^T w_{y,i}(k) Lr(k), i = 1, \cdots, m_y;$  $w_{y,i}$  is the *i*th component of  $w_y$ ; *T* is the time window size; *L* has been given in (2); due to the decryption (19), the new residual r(k) is defined by  $r(k) := y_w^-(k) - C\hat{x}(k|k-1).$ 

2) New DW Test 2: checking whether or not

$$\lim_{T \to \infty} \mathcal{V}(T) = 0 \tag{21}$$

where  $\mathcal{V}(T) := \frac{1}{T} \sum_{k=1}^{T} Lr(k) (Lr(k))^T - L\Sigma_o L^T; \Sigma_o$  is the same as the one in (2).

*Remark 8:* Compared with the conventional DW Tests (6) and (7), there are two differences in signals for the new DW Tests (20) and (21): the watermarking signal is changed from  $w_d(k-1)$  to  $w_y(k)$ , and the residual signal is changed from  $r_d(k)$  to r(k). Note that the identity is that the tests are formed using the time averaging of products of (i) watermarking signal and residual (i.e., (6) and (20)) and (ii) two identical residual (i.e., (7) and (21)). Therefore, it is expected that the new DW Tests (20) and (21), together with the above watermarking as symmetric-key encryption, can develop the security property of the conventional DW scheme, and overcome limitations 1 and 3.

*Remark 9:* Note that in the new DW Tests 1 and 2, we require  $T \rightarrow \infty$ , which is similar to the conventional DW Tests

1 and 2. To solve the problem, following the definitions made in Remark 1 and [19], the new DW Tests 1, 2 need to be converted to the statistical tests  $\varphi_{1,i}(k), \varphi_2(k)$  for practical applications, i.e.,  $\varphi_{1,i}(k) := \|\mathcal{W}_i^k(T)\|_F, \varphi_2(k) := |tr(\mathcal{V}^k(T))|$ , where  $\mathcal{W}_i^k(T)$  and  $\mathcal{V}^k(T)$  are  $\mathcal{W}_i(T)$  and  $\mathcal{V}(T)$  calculated within the current time window  $\{k - T + 1, k - T + 2, \dots, k\}$ , respectively. Furthermore, let  $\vartheta_{1,i}, \vartheta_2$  be the preset thresholds and if  $y_w^-(k)$  is attacked, we expect  $\varphi_{1,i}(k) \ge \vartheta_{1,i}$  or  $\varphi_2(k) \ge \vartheta_2$  thereby the detection result is set as  $\varepsilon(k) = 0$ ; otherwise,  $\varepsilon(k) = 1$ . The detection results will be used for compensation below.

After  $y_w^-(k)$  has been detected, there may be two cases for the attacked signals:

 If there is no compensation, then y<sub>w</sub><sup>-</sup>(k) will be directly sent to the estimator whether the signal is attacked or not, i.e.,

$$\tilde{y}(k) = y_w^-(k)$$
, no compensation. (22)

2) The above (22) could cause system performance degradation even instability. To improve the system resilience under attacks, a compensation mechanism is employed, where  $y_w^-(k)$  will be sent to the estimator based on the detection results  $\varepsilon(k)$ , i.e.,

$$\tilde{y}(k) = \begin{cases} y_w^-(k), & \text{if } \varepsilon(k) = 1\\ \tilde{y}(k - h'(k)), & \text{if } \varepsilon(k) = 0 \end{cases}$$
(23)

where  $h'(k) := k - \max\{k' | \varepsilon(k') = 1, k' \leq k\} > 0$  is the duration of last successful buffer update. For simplicity, we only consider the situation where the missed detection<sup>1</sup> can be ignored (i.e.,  $y_w^-(k) = y(k)$ , if  $\varepsilon(k) = 1$ ), then (23) becomes

$$\tilde{y}(k) = \begin{cases} y(k), if \ \varepsilon(k) = 1\\ \tilde{y}(k - h'(k)), if \ \varepsilon(k) = 0 \end{cases}$$
(24)

Define

$$h(k) := \begin{cases} 0, if \ \varepsilon(k) = 1\\ h'(k), if \ \varepsilon(k) = 0 \end{cases}$$
(25)

According to (25), (24) can be re-written as

 $\tilde{y}(k) = y(k - h(k)), \text{ with compensation}$  (26)

where h(k) is the delay of healthy system output.

*Remark 10:* Compared with the conventional DW scheme and the scheme (22) without compensation, the compensation (26) based on the detection results  $\varepsilon(k)$  is used to discard the attacked  $y_w^-(k)$  (e.g., injected with false data) and to enable the latest y(k-h(k)). It is expected to overcome the limitation 2, i.e., the recovery of system performance.

Then,  $\hat{x}(k|k)$  can be computed by

$$\hat{x}(k|k) = \hat{x}(k|k-1) + L(\tilde{y}(k) - C\hat{x}(k|k-1)).$$
(27)

Using  $\hat{x}(k|k)$ , since there is no watermarking signal on the controller side, the control input (5) can be re-written as

$$\iota(k) = K\hat{x}(k|k) \tag{28}$$

<sup>1</sup>Missed detection is that even though an attack takes place, the attack alarm is not triggered (i.e.,  $\varepsilon(k) = 1$ ).

where K is the same as (4), which can update  $\hat{x}(k+1|k)$  by

$$\hat{x}(k+1|k) = A\hat{x}(k|k) + Bu(k).$$
 (29)

Finally, under the FDIA (11), to identify the system based on the new DW scheme without or with compensation, we produce the corresponding close-loop system models as follows:

- On the one hand, under the FDIA (11), the system based on the new DW scheme without compensation, described in (1), (18)-(21), (22) and (27)-(29), can be modelled as
  - $\zeta(k+1) = \mathcal{A}_0\zeta(k) + \Lambda_0\psi(k)$ , no compensation (30)

where marks  $\zeta(k) := [x(k); x_a(k) - \hat{x}(k|k-1); x_a(k)],$  $\psi(k) := [n(k); v(k); w_y(k)];$  the matrices  $\mathcal{A}_0$  and  $\Lambda_0$  are given by

$$\mathcal{A}_{0} \text{ in (12), and} \\ \Lambda_{0} := \begin{bmatrix} \Gamma & 0 & BKL \\ 0 & 0 & (A+BK)L \\ 0 & 0 & 0 \end{bmatrix}$$

 On the other hand, under the FDIA (11), the system based on the new DW scheme with compensation, described by (1), (18)-(21), and (26)-(29), can be constructed as

$$\bar{\zeta}(k+1) = A_0 \bar{\zeta}(k) + A_1 E \bar{\zeta}(k-h(k)) + \Gamma_0 \bar{\psi}(k),$$
  
with compensation (31)

where the variables  $\overline{\zeta}(k) := [x(k); \hat{x}(k-1|k-1)],$  $\overline{\psi}(k) := [n(k); v(k-h(k))];$  the matrices  $E, A_0, A_1$  and  $\Gamma_0$  are given by

$$E := [I, 0], \text{ and}$$

$$A_0 := \begin{bmatrix} A & BK(I - LC)(A + BK) \\ 0 & (I - LC)(A + BK) \end{bmatrix},$$

$$A_1 := \begin{bmatrix} BKLC \\ LC \end{bmatrix}, \Gamma_0 := \begin{bmatrix} \Gamma & BKL \\ 0 & L \end{bmatrix}.$$

Remark 11: In the designed secure controller, the attack detection method, instead of the control strategy, usually brings the difference of computational time complexity. Therefore, compared with the tests used in [31]–[34], computational time complexity of the new DW Tests (20) and (21) is analysed as follows. For instance, considering that  $r(k) \in \mathcal{R}^{m_y}$ ,  $x(k) \in \mathcal{R}^{m_x}$  and the finite T, computational time complexity of the residual-based test  $\sum_{s=k-T-1}^{k} r^{T}(s)r(s)$  used in [34] is  $O(m_y^2)$ . However, computational time complexity of the new DW Tests (20) and (21) is  $O(1) + O(m_x^2 + m_x m_y) \approx$  $O(m_x^2 + m_x m_y)$ , where O(1) and  $O(m_x^2 + m_x m_y)$  are spent by watermarking generation, and calculation of  $\mathcal{W}_i(T)$  and  $\mathcal{V}(T)$ , respectively. To clearly compare computational time complexity of the new DW Tests with the residual-based test, a ratio  $\wp := m_y^2/(m_x^2 + m_x m_y)$  is defined and shown in Fig. A.2 of Section II.B in the supplementary materials. It can be seen from Fig. A.2 that if  $m_y < m_x$ , then computational time complexity of the new DW Tests is usually more than that of the residual-based test used in [34], as shown in the experimental results of Section IV.B; if  $m_u$  is much larger than  $m_x$ , computational time complexity of the new DW Tests will be less than that of the residual-based test used in [34], which could take place in the power system.

### B. Security Property Analysis of the New DW Scheme

We have developed a framework for secure control of NCSs based on the new DW scheme, and then the security property of the new DW scheme is analysed using the following Theorem 1.

Theorem 1: For the system (30) or (31), if  $y_w^-(k)$  passes the new DW Tests 1 and 2, then

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \left\| \mathcal{D}_r(k) \right\|^2 = 0 \tag{32}$$

where  $\mathcal{D}_{r}(k) := Lr(k) - Lr^{o}(k)$ ; and  $r^{o}(k) := y_{w}^{o-}(k) - C\hat{x}^{o}(k|k-1), y_{w}^{o-}(k), \hat{x}^{o}(k|k-1)$  are the attack-free (i.e.,  $y_{w}^{-}(k) \equiv y(k)$ ) counterparts of  $r(k), y_{w}^{-}(k), \hat{x}(k|k-1)$ .

*Proof:* The proof is given in Section I.B of the supplementary materials.

*Remark 12:* The security property of the conventional DW scheme indicates that the additive distortion power of the closed-loop systems is restricted to be zero when DW Tests 1 and 2 are bypassed. Theorem 1 reveals that the additive distortion power of the closed-loop systems is restricted to be zero when the new DW Tests 1 and 2 are bypassed. In addition, (32) explains the theoretical basis of attack detection as (10). Therefore, by integrating the watermarking as symmetric-key encryption and the new DW Tests 1 and 2, the new DW scheme develops the security property of the conventional DW scheme.

C. FDIA Detection Effectiveness Analysis of the New DW Scheme

The above discussion is about the security property of the new DW scheme, and FDIA detection effectiveness of the new DW scheme can be evaluated in the following Theorem 2 and Corollary 1.

Theorem 2: For the system (30), if  $\rho(A) > 1$ , then the FDIA (11) will result in

$$\mathbb{E}\left[w_{y,i}(k)Lr(k)\right] = -\sigma_{w_{y,i}}^2 L_{\cdot i},\tag{33a}$$

$$\mathbb{E}\left[Lr(\infty)\left(Lr(\infty)\right)^{T}\right] = L\left(CMC^{T} + \Sigma_{w_{y}}\right)L^{T},\quad(33b)$$

$$\lim_{k \to \infty} x(k) = \infty \tag{33c}$$

where  $M = \Phi_1 M \Phi_1^T + (A + BK) L \Sigma_{w_y} ((A + BK)L)^T$ ,  $\Phi_1$  is the same as the one of (12).

*Proof:* The proof is given in Section I.C of the supplementary materials.

*Corollary 1:* For the system (31), the FDIA (11) will result in (33a).

*Proof:* The proof is omitted.

*Remark 13:* Limitation 1 earlier has revealed the positive correlation between the FDIA detection effectiveness of the conventional DW scheme and the value of  $\sigma_{w_d}^2$ . However, due to limitation 3, a small  $\sigma_{w_d}^2$  needs to be selected for yielding small system performance loss due to watermarking. This makes it possible that the FDIA (11) may bypass the conventional DW scheme. Similarly, Theorem 2 and Corollary 1 highlight that FDIA detection effectiveness of the new DW scheme with or without compensation is positively correlated

with the value of  $\sigma_{w_{y,i}}^2$ . Note that the key difference is that a big  $\sigma_{w_{y,i}}^2$  can be selected in the new DW scheme because  $J^o = J^o|_{w_y(k)=0}$  thanks to the watermarking as symmetric-key encryption (18) and (19), where  $J^o|_{w_y(k)=0}$  is the attack-free system's performance without the new DW scheme. That is, limitation 3 on system performance loss due to watermarking is overcome by the new DW scheme. This makes it possible that the FDIA (11) will be detected by the new DW scheme with an enabled big enough  $\sigma_{w_{y,i}}^2$ . Therefore, by integrating the watermarking as symmetric-key encryption (18) and (19) and the new DW Tests (20) and (21), limitations 1 and 3 are overcome by the new DW scheme.

*Remark 14:* As stated above, the basis of the new DW scheme can be attributed to the overcoming of limitation 3 on system performance loss from watermarking. The conventional DW scheme brings system performance loss from watermarking, because the control signal is only encrypted by the watermarking signal and without decryption. Unlike the conventional DW scheme, the watermarking as symmetric-key encryption (18) and (19) used in the new DW scheme can prevent watermarking signals from exciting the normal system operation, so the limitation 3 can be avoided.

# D. Performance Analysis of NCSs based on the New DW Scheme

Now, we will further analyse why the new DW scheme needs to be employed for the recovery of system performance, and quantify the recovery capability of system performance based on the new DW scheme.

1) Why the New DW Scheme Needs to Be Used for Recovery of System Performance? Theorem 2 has presented that when the compensation is ignored, if  $\rho(A) > 0$  in (1), then the FDIA (11) emerging in  $[k_0^a, \infty)$  will influence the system stability. Therefore, it is necessary to use the new DW scheme against the FDIA (11) emerging in  $[k_0^a, \infty)$  for recovery of system performance. However, if we use the compensation (26) to cope with the FDIA (11) emerging in  $[k_0^a, \infty)$ , h(k)will go infinity and according to time-delay system theory [38], the system will be unstable. The system will lose its stability even though the compensation is used. But then, on the one hand, in practice the operation time of the FDIA (11) must be the union of subsets of  $[k_0^a, \infty)$ ; on the other hand, sometimes the attackers achieve their goal in a short time without worrying about the attack detection. In this situation, is it necessary to use the new DW scheme for the recovery of system performance? To answer this question, based on the switched system theory, under the FDIA (11) with the subsets of  $[k_0^a, \infty)$ , the following stability analysis is presented.

The first task is to construct the corresponding switched system model. The early work (30) has modelled the subsystem based on the new DW scheme without compensation under the FDIA (11), then the attack-free subsystem based on the new DW scheme without compensation can be given by setting  $h(k) \equiv 0$  in (31) and adopting the same variables  $\zeta(k)$  and  $\psi(k)$  in (30), i.e.,

$$\zeta(k+1) = \mathcal{A}_1\zeta(k) + \Lambda_1\psi(k) \tag{34}$$

where the matrices  $\Lambda_1$  and  $\mathcal{A}_1$  are given by

$$\Lambda_{1} := \begin{bmatrix} \Gamma & BKL & 0\\ 0 & -(A+BK)L & 0\\ 0 & 0 & 0 \end{bmatrix},$$
$$\mathcal{A}_{1} := \begin{bmatrix} A+BKLC & -BK(I-LC) & 0\\ -(A+BK)LC & \Phi_{1} & 0\\ 0 & 0 & 0 \end{bmatrix}.$$

Therefore, combining (30) and (34), under the FDIA (11) with the subsets of  $[k_0^a, \infty)$ , the system based on the new DW scheme can be modelled as

$$\zeta(k+1) = \mathcal{A}_{s(k)}\zeta(k) + \Lambda_{s(k)}\psi(k) \tag{35}$$

where s(k) = 0 or 1 is the switching signal: when an attack takes place, there is s(k) = 0; otherwise s(k) = 1. We know  $A_0$  is unstable and  $A_1$  is stable. Then, there exist real numbers  $\lambda_+ > 1$ ,  $0 < \lambda_- < 1$ ,  $g_0$  and  $g_1$  satisfy

$$\left\| \left(\mathcal{A}_{0}\right)^{k} \right\| \leq \left(\lambda_{-}\right)^{g_{0}} \left(\lambda_{+}\right)^{k}, \left\| \left(\mathcal{A}_{1}\right)^{k} \right\| \leq \left(\lambda_{-}\right)^{g_{1}} \left(\lambda_{-}\right)^{k}.$$
 (36)

Meanwhile, we denote the total activation time of the unstable subsystem (or the stable subsystem) by  $\mathcal{T}_0$  (or  $\mathcal{T}_1$ ). We denote the switching of s by  $N_s(0,k) \leq N_0 + \frac{k}{\tau}$  where  $\tau$  is the average dwell time. According to the definitions of attack frequency and duration made in [39], the larger  $\mathcal{T}_0$  and  $\tau$  stand for a longer attack duration and a smaller attack frequency, respectively.

Finally, stability analysis of the switched system (35) is demonstrated in the following Theorem 3.

*Theorem 3:* For the given constant  $\lambda^{\dagger}$ , if the following conditions are satisfied:

$$\mathbb{E}\left[\|\psi(k)\|\right] < \infty,\tag{37}$$

$$\begin{cases} \tau \in \mathbb{R}_+, if \ g \ge 0\\ \tau \ge \tau_{ave} = \frac{g}{\lambda^{\dagger} - \lambda^*}, if \ g < 0 \end{cases}$$
(38)

$$\inf_{k \ge 0} \left[ \mathcal{T}_1 / \mathcal{T}_0 \right] \ge \left( \ln \lambda_+ - \lambda^* \ln \lambda_- \right) / \left( (\lambda^* - 1) \ln \lambda_- \right)$$
(39)

where  $0 < \lambda^{\dagger} < \lambda^* < 1$  and  $g = \min\{g_0, g_1\}$  from (36), then there always exists a finite constant  $\tau$  such that the switched system (35) is uniformly bounded.

*Proof:* The proof is given in Section I.D of the supplementary materials.

*Remark 15:* The tolerance capacity of FDIAs for NCSs based on the new DW scheme without compensation is analysed as follows. Theorem 3 points out that under FDIA (11), the state of the system based on the new DW scheme without compensation is uniformly bounded when the noise is bounded (37), the attack frequency is small (38) and the attack duration is short (39). As shown in Section IV, under the FDIA (11) with a short successive duration, when the initial state of the FDIA (11) is properly chosen, the condition (39) may be violated and the system will lose its stability. Therefore, it is necessary to use the new DW scheme against the FDIA (11) with the subsets of  $[k_0^n, \infty)$ .

2) How Well Does System Performance Recover from the New DW Scheme? We have proved that it is necessary to use the new DW scheme against the FDIA (11) for the recovery

of system performance. Then, the recovery capability of the system based on the new DW scheme is presented in the following, which is described by the quantitative relationship between the system performance and the maximally allowable delay of healthy system output.

Firstly, the system performance is described by the power of the estimation errors, i.e.,

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \|e(k)\|^2$$
(40)

where the estimation errors  $e(k) := x(k) - \hat{x}(k|k)$ .

According to the defined system performance, the relationship between the system performance and the new DW scheme is presented in the following Theorem 4 and Corollary 2.

Theorem 4: Under the zero-initial condition, if given an integer  $\bar{h} \ge h(k)$ , there exists a real number  $\beta$ , real symmetric matrices  $Z_i > 0$  (i = 1, 2, 3), and matrices  $W_i$  (i = 1, 2, ..., 5) satisfying

$$\begin{bmatrix} \Theta & \Omega \\ * & \Upsilon \end{bmatrix} < 0 \tag{41}$$

where

$$\begin{split} \Theta &:= \begin{bmatrix} \Theta_{11} & E^T W_2 - W_1^T & E^T W_3 \\ * & -W_2^T - W_2 - Z_3 & -W_3 \\ * & * & -\beta I \end{bmatrix}, \\ \Theta_{11} &:= W_1^T E + E^T W_1 - Z_1 + E^T Z_3 E, \\ \Omega &:= \\ \begin{bmatrix} \bar{h} W_1^T & \bar{h} (A_0 - I)^T E^T W_4^T & (E - E^c A_0)^T & A_0^T W_5^T \\ \bar{h} W_2^T & \bar{h} A_1^T E^T W_4^T & (-E^c A_1)^T & A_1^T W_5^T \\ \bar{h} W_3^T & \bar{h} \Gamma_0^T E^T W_4^T & (-E^c \Gamma_0)^T & \Gamma_0^T W_5^T \end{bmatrix}, \\ \tilde{\Upsilon} &:= \\ diag \left\{ -\bar{h} Z_2, \bar{h} \left( Z_2 - W_4^T - W_4 \right), -I, Z_1 - W_5^T - W_5 \right\} \end{split}$$

and  $E^c := [0, I]$ , then the system (31) is asymptotically stable and the power of the estimation errors satisfies

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \|e(k)\|^2 = \beta \left( tr\left(\Sigma_n\right) + tr\left(\Sigma_v\right) \right).$$
(42)

*Proof:* The proof is given in Section I.E of the supplementary materials.

Corollary 2: Under the zero-initial condition, if given an integer  $\bar{h} \ge h(k)$ , there exists a real number  $\beta$ , real symmetric matrices  $Z_i > 0$  (i = 1, 2, 3), and matrices  $W_i$  (i = 1, 2, ..., 5) satisfying

$$\begin{bmatrix} \Theta & \Omega C \\ * & \Upsilon \end{bmatrix} < 0 \tag{43}$$

where  $\Theta$ ,  $\Omega$ ,  $\Upsilon$ , and  $E^c$  has been given in Theorem 4,  $\mathcal{C} := diag\{I, I, C^T, I\}$ , then the system (31) is asymptotically stable and the power of the estimation errors satisfies

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \left\| Ce(k) \right\|^2 = \beta \left( tr\left(\Sigma_n\right) + tr\left(\Sigma_v\right) \right).$$
(44)

*Proof:* The proof is omitted.

*Remark 16:* Limitation 2 and Theorem 3 have revealed that the FDIA (11) emerging in  $[k_0^a, \infty)$  or the union of subsets of  $[k_0^a, \infty)$  can destroy system stability without compensation. Theorem 4 reveals that by using the new DW scheme, the system performance can be recovered from unstable mode under the FDIA (11) to stable mode (specifically, lying on noise level (42) or (44) with respect to process and measurement noises), and the relationship between the system performance and the maximally allowable delay  $\bar{h}$  of healthy system output is quantified. Therefore, by collating all the compensations, limitation 2 can be overcome when  $\bar{h}$  is not violated.

Remark 17: Note that over detection<sup>2</sup> could lead to  $h(k) > \bar{h}$ , causing the system to be unstable (i.e., infinite estimation errors). This shows the cost to pay for the *compensation mechanism* (26) against limitation 2 on system performance loss from attacks: i.e., it is highly recommended that a small time window size T should be selected when the proposed compensation mechanism is used in the real world, as done in Section IV.

### **IV. EXPERIMENTS**

The platform of the networked inverted pendulum visual servo system (NIPVSS) [40] is employed to validate the new DW scheme. We first give the structure and corresponding parameters of the platform. Then, the real-time experiments are carried out on the platform.

### A. Experiment Platform Setup

The experimental platform of the NIPVSS based on the new DW scheme is shown in Fig. 3. The single acA640-120gm monochrome camera at 100 Hz @  $640 \times 480$  pixels with light sources acting as the sensor captures the images of the inverted pendulum. Then the images are sent to the computer through the 1 Gbps wired network, which runs Microsoft Visual Studio 2010 in Windows XP with an Intel Core i5 processor (3.2 GHz) and 4 GB RAM. After having received the images, the computer processes the images to obtain the state information of the cart position and the pendulum angle. Furthermore, in the computer, the state information is encrypted and decrypted based on the new DW scheme and Algorithm 1, then the current results are used to formulate the estimator, the new DW detector and the controller. Finally, the control signal is applied to the inverted pendulum by use of the motion control box serving as the actuator.

By use of linearization and setting the sampling period as 10 ms,  $x(k) := \left[\alpha(k); \theta(k); \dot{\alpha}(k); \dot{\theta}(k)\right] \in \mathcal{R}^4$  is set as the state of the cart position, pendulum angle, cart velocity and pendulum angular velocity, and  $u(k) = \ddot{\alpha}(k) \in \mathcal{R}$  is set as the control input. Considering the process noise n(k) with the

<sup>2</sup>Over detection is that after the real attacks stop, attack alarms (i.e.,  $\varepsilon(k) = 0$ ) will continue for some time depending on the value of the window size T.

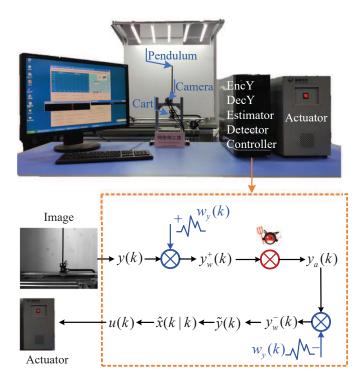


Fig. 3. Experimental platform of the NIPVSS with the new DW scheme.

covariance<sup>3</sup>  $\Sigma_n = diag\{10^{-5}, 10^{-5}\}$ , the discrete-time model of the inverted pendulum as described in (1) is

$$A = \begin{bmatrix} 1 & 0 & 0.0100 & 0 \\ 0 & 1.0015 & 0 & 0.0100 \\ 0 & 0 & 1 & 0 \\ 0 & 0.2945 & 0 & 1.0015 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0.0002 \\ 0.0100 \\ 0.0300 \end{bmatrix},$$
$$\Gamma = [0, 0; 0, 0; 1, 0; 0, 1].$$

By use of the camera as the sensor, the states of the cart position and the pendulum angel can be retrieved, i.e., C = [I, 0] in (1). The measurement noise  $v(k) \in \mathcal{R}^2$  in (1) with the covariance matrix  $\Sigma_v = diag\{2.7 \times 10^{-7}, 5.5 \times 10^{-6}\}$ from the computational error [40] is also introduced, which is analysed in Section II.A of the supplementary materials.

The watermarking  $w_{y,i}(k)$  shown in (18) and (19) is generated by the C++ function gaussrand(mean, std)<sup>4</sup> based on Algorithm 1, where "mean" and "std" are the mean value (i.e. zero) and standard deviation of  $w_{y,i}(k)$ , respectively. The seed of function rand() in gaussrand(mean, std) takes the value 1.

The estimator parameters in (27) and (29) are A, B, C and

$$L = [0.2951, 0; 0, 0.1673; 5.1094, 0; 0, 1.5290].$$

We set  $Q = diag\{10, 10, 10, 10\}$ , R = 1, and the controller gain shown in (28) is

$$K = [2.8889, -36.6415, 4.9141, -7.3267].$$

<sup>3</sup>We cannot determine the real covariance matrix of the process noise. Our solution is to preset the process noise covariance matrix and form the estimator using prior knowledge. If the system operates stably under the process noise and corresponding estimator, then we will use them.

<sup>4</sup>The detailed process of *gaussrand(mean, std)* can be found in https://encyclopedia.thefreedictionary.com/Marsaglia+polar+method.

TABLE III FDIA DETECTION EFFECTIVENESS AND SYSTEM PERFORMANCE LOSS FROM WATERMARKING OF NIPVSSS WITH  $\sigma_{w_d}^2 = \sigma_{w_{y,i}}^2 = 0.0001$ 

Indices ( $\sigma_d^2 = \sigma_{w_{y,i}}^2 = 0.0001$ )	Normal	under FDIA
$\mathrm{CDW}^1$ : $\left\ \mathbb{E}\left[w_d(k-1)Lr_d(k)\right]\right\ _F$	0	3.4459E-8
CDW: $\left  tr\left( \mathbb{E} \left[ Lr_d(\infty) (Lr_d(\infty))^T \right] \right) \right $	2.5660E-5	4.1303E-8
CDW: System performance loss	33.67%	-
New DW: $\ \mathbb{E}[w_{y,1}(k)Lr(k)]\ _F$	0	5.1179E-4
New DW: $\ \mathbb{E}[w_{y,2}(k)Lr(k)]\ _{F}$	0	1.5381E-4
New DW: $\left  tr \left( \mathbb{E} \left[ Lr(\infty)(Lr(\infty))^T \right] \right) \right $	2.5660E-5	3.4152E-1
New DW: System performance loss	0	-
Commention of DW		

<sup>1</sup> Conventional DW.

It should be noted that in the platform, the entry angle is due to the finite field of vision of the camera, and the limit position due to the fixed range of cart movement. Specifically, we have  $|\theta| < 0.8$  rad,  $|\alpha| < 0.3$  m. Once the angle crosses the entry level or the cart position is over the limit, the system will trigger self termination (i.e., the servo is put OFF). Moreover, there is a situation where if the cart velocity or the pendulum angular velocity is too large, the cart position will be put BACK to the original point (i.e., zero) and then the control action is ended. These are shown in real-time experiments below.

### B. FDIA Detection Effectiveness of NIPVSSs Based on the Conventional and New DW Schemes

We select the FDIA (11) emerging at  $k \ge 2$ , where  $A_a = diag\{0.1, 0.1, 0.1, 0.1\}$  and  $x_a(2) = [10^{-7}; 0; 0; 10^{-7}]$ . We set  $\sigma_{w_d}^2 = \sigma_{w_{y,i}}^2 = 0.0001$  for the conventional and new DW schemes. The time window size is set as T = 5, the detection thresholds for the conventional and new DW schemes are set as  $\vartheta_{d,1} = 0.0002$ ,  $\vartheta_{d,2} = 0.0015$ , and  $\vartheta_{1,i} = \vartheta_2 = 0.0007$ , experimentally.

Figs. 4 and 5 present the experiment results of applying the FDIA to NIPVSSs based on the conventional and new DW schemes. When there is FDIA, the states can be separated into two situations. However, the conventional DW scheme fails to detect the FDIA, but the new DW scheme succeeds in detecting the FDAI. Table III shows the FDIA detection effectiveness and system performance loss from watermarking of NIPVSSs based on the conventional and new DW schemes, where it can be seen that compared with the conventional DW scheme, the new DW scheme provides much better FDIA detection effectiveness and zero system performance loss from watermarking.

Computational time complexity of the new DW tests (20) and (21) and the residual-based test [34] is also compared in the experiments of Fig. 5, which is listed in Tab. A.I of Section II.B in the supplementary materials. Tab. A.I shows that the proposed new DW Tests require more time than the residual-based test when  $m_y = 2 < m_x = 4$ . However, the microsecond-level running time of the new DW Tests is negligible in comparison the control cycle (10 ms). Therefore, computational time in the use of the new DW Tests is not significant.

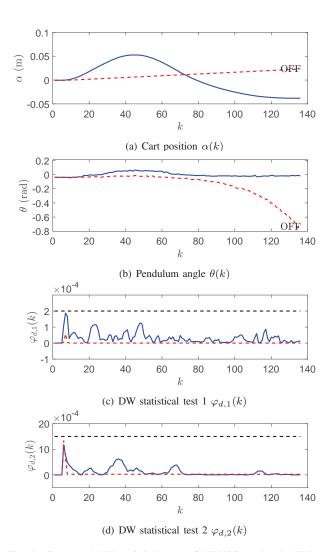
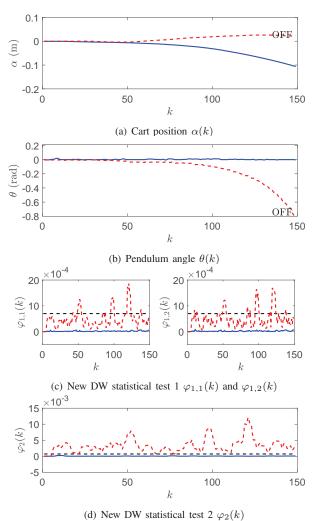


Fig. 4. States and DW statistical tests of NIPVSSs under the FDIA (11) emerging at  $k \ge 2$  based on the conventional DW scheme with  $\sigma_{w_d}^2 = 0.0001$ . Blue line: Normal system. Red line: System under the FDIA. Black line: Detection thresholds. "OFF" denotes that the servo is put off.

To investigate the FDIA under different initial instants, we perform the same experiments of Figs. 4 and 5 again, with the FDIA emerging at  $k \ge 400$ . Experiment results are shown in Figs. A.3 and A.4 of Section II.C in the supplementary materials, which indicates that when the initial instant of the FDIA is close to the initial instant of system operation, the conventional DW scheme with  $\sigma_{w_d}^2 = 0.0001$  cannot detect the FDIA while the new DW scheme with  $\sigma_{w_{y,i}}^2 = 0.0001$  can detect the FDIA; when the initial instant of the FDIA is far from the initial instant of system operation, both the conventional DW scheme with  $\sigma_{w_d}^2 = 0.0001$  and the new DW scheme with  $\sigma_{w_d}^2 = 0.0001$  and the new DW scheme with  $\sigma_{w_d}^2 = 0.0001$  and the new DW scheme with  $\sigma_{w_d}^2 = 0.0001$  and the new DW scheme with  $\sigma_{w_d}^2 = 0.0001$  can detect the FDIA.

Further analysis of FDIA detection effectiveness on the conventional and the new DW schemes with different watermarking density is carried out. Due to space constraints and their similarity to Figs. 4 and 5, the outcomes from further analysis are omitted here. The results show that (i) the conventional DW scheme cannot detect the FDIA even though  $\sigma_{w_d}^2 = 10$  where the system states have severe shaking; (ii) a much larger watermarking density brings better FDIA



ates and the new DW statistical tests of NIPVSSs under th

Fig. 5. States and the new DW statistical tests of NIPVSSs under the FDIA (11) emerging at  $k \ge 2$  based on the new DW scheme with  $\sigma_{w_{y,i}}^2 = 0.0001$ . Blue line: Normal system. Red line: System under the FDIA. Black line: Detection thresholds. "OFF" denotes that the servo is put off.

detection effectiveness by the new DW scheme.

### C. Performance of NIPVSSs Based on the New DW Scheme

To present the performance of NIPVSSs based on the new DW scheme under the FDIA, we select the FDIA (11) taking place at k = 100, 101, 102, 103, whose parameters are set as  $A_a = diag\{0.1, 0.1, 0.1, 0.1\}, x_a(100) = [2; 2; 2; 2]$ . We set  $\sigma_{w_{y,i}}^2 = 0.0001$  for the new DW scheme. The time window size is set as T = 5, and the detection thresholds for the new DW scheme are set as  $\vartheta_{1,i} = \vartheta_2 = 0.0007$ .

1) Performance of NIPVSSs Under No Attacks: We firstly examine the impact of the conventional and new DW schemes with different watermarking densities on the performance of NIPVSSs under no attacks, shown in Fig. A.5 of Section II.D in the supplementary materials. It is shown that when the watermarking density increases to 10, the conventional DW scheme brings 33671.78% system performance loss where the system states have severe shaking, but zero system performance loss is yielded by the new DW scheme. 2) Performance of NIPVSSs Based on the New DW Scheme without Compensation under the FDIA: When the compensation is ignored, the stability analysis is presented in Theorem 3. There always exist non-singular matrices  $V_0$  and  $V_1$  that satisfy

$$\mathcal{A}_0 = V_0 D_0 V_0^{-1}, \mathcal{A}_1 = V_1 D_1 V_1^{-1}$$

where matrices  $D_0 = diag\{\lambda_1(\mathcal{A}_0), \dots, \lambda_8(\mathcal{A}_0)\}$  and  $D_1 = diag\{\lambda_1(\mathcal{A}_1), \dots, \lambda_8(\mathcal{A}_1)\}$ . We select

$$\lambda_{+} = \max \left( \mathcal{M} \left( \lambda_{i}(\mathcal{A}_{0}) \right) \right) = 5.4250, \\ \lambda_{-} = \max \left( \mathcal{M} \left( \lambda_{i}(\mathcal{A}_{1}) \right) \right) = 0.9895, \\ g_{0} = \frac{\ln \left( \mathcal{S}_{\max}(V_{0}) / \mathcal{S}_{\min}(V_{0}) \right)}{\ln \lambda_{-}} = -3879.8947, \\ g_{1} = \frac{\ln \left( \mathcal{S}_{\max}(V_{1}) / \mathcal{S}_{\min}(V_{1}) \right)}{\ln \lambda_{-}} = -614.4731$$

where  $\mathcal{M}(\cdot)$  is the modulus of a complex number,  $\mathcal{S}_{\min}(\cdot)$  $(\mathcal{S}_{\max}(\cdot))$  is the minimal (maximal) singular value of a matrix. Hence  $g = \min\{g_0, g_1\} = g_0$ . Now, by taking  $\lambda^* = 0$ , the infimum of the value of  $\inf_{k\geq 0} [\mathcal{T}_1/\mathcal{T}_0]$  can be obtained as follows

$$\inf_{1>\lambda^* \ge 0} \inf_{k \ge 0} \left[ \mathcal{T}_1 / \mathcal{T}_0 \right]^{\lambda^* = 0} = 159.4495.$$

The experiment is carried out at  $k \in [0, 140]$ , and here  $\mathcal{T}_0 = 4$ ,  $\mathcal{T}_1 = 141 - \mathcal{T}_0 = 137$ . Since  $\mathcal{T}_1/\mathcal{T}_0 = 34.25 < 159.4495$ , the condition (39) is not satisfied. This means that it is possible for the FDIA (11) taking place at k = 100, 101, 102, 103 to influence the stability of the NIPVSS.

Fig. 6 presents the experimental results of applying the FDIA (11) taking place at k = 100, 101, 102, 103 to the NIPVSS without compensation. As shown as the red lines in Fig. 6, the states of the NIPVSS become divergent without compensation.

3) Performance of NIPVSSs Based on the New DW Scheme with Compensation under the FDIA: Fig. 6 also presents the experimental results of applying the FDIA (11) taking place at k = 100, 101, 102, 103 to the NIPVSS with compensation. As shown as the green lines in Fig. 6 (a), (b), the states of the NIPVSS keep stable. Note that the detection indicators  $\varphi_{1,1}(k)$  and  $\varphi_{1,2}(k)$  shown in Fig. 6 (c) validate Corollary 1, i.e., with the compensation used, the new DW scheme has a sufficient ability to detect the FDIA (11).

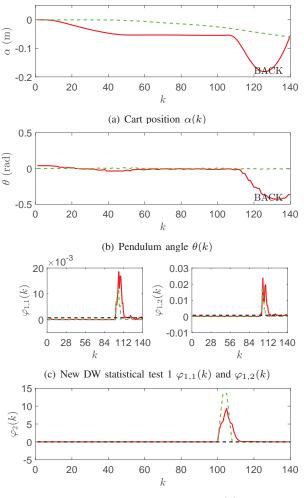
By solving Corollary 2, when the compensation is used, we can obtain  $\bar{h} = 4$  with  $\beta = 2.9800 \times 10^2$ . The corresponding system performance is

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \|Ce(k)\|^2 = 0.0077.$$

To validate Corollary 2, we obtain the system performance in a finite time window T = 5, i.e.,

$$\mathcal{E}_T(k) = \frac{1}{T} \sum_{s=k-T+1}^k \|Ce(s)\|^2.$$

Fig. 7 presents the values of  $\mathcal{E}_T(k)/0.0077$  without and with compensation. It can be seen from Fig. 7 that: 1) the system performance  $\mathcal{E}_T(k)/0.0077$  is recovered with a very big transient value about 8000 (unstable mode) to a very small



(d) New DW statistical test 2  $\varphi_2(k)$ 

Fig. 6. States and the new DW statistical tests of NIPVSSs under the FDIA (11) taking place at k = 100, 101, 102, 103 based on the new DW scheme with  $\sigma_{wy,i}^2 = 0.0001$ . Red line: System without compensation. Green line: System with compensation. Black line: Detection thresholds. "BACK" indicates that the cart position is put back to zero.

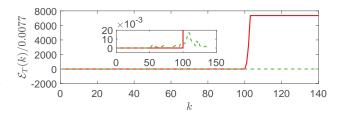


Fig. 7. Values of  $\mathcal{E}_T(k)/0.0077$  without or with compensation. Red line: without compensation. Green line: with compensation.

transient value about 0.02 (stable mode), and 2) the value of  $\mathcal{E}_T(k)/0.0077$  with compensation is less than 0.02, which means that the real system performance is far better than the theoretical system performance.

To further investigate the compensated variables based on the new DW scheme, we define the detection indictor on

$$\tilde{\varphi}_2(k) := \tilde{y}(k) - C\hat{x}(k|k-1) \text{ as follows}$$
$$\tilde{\varphi}_2(k) := \left| tr\left(\frac{1}{T} \sum_{p=k-T+1}^k L\tilde{r}(p)(L\tilde{r}(p))^T - L\Sigma_o L^T\right) \right|.$$

Fig. A.6 of the supplementary materials presents the values of  $\tilde{\varphi}_2(k)$ , where it can be seen that the detection indictor  $\tilde{\varphi}_2(k)$  is less than the detection threshold by use of compensation.

### V. CONCLUSION

A new DW scheme was proposed for NCSs secure control. Firstly, the security weaknesses of the conventional DW scheme was analysed. Secondly, to overcome the security weakness, a framework of NCSs secure control based on the new DW scheme was designed, which integrated the watermarking as symmetric-key encryption and new DW tests and compensation mechanism. Then, by using the additive distortion power of the closed-loop system, whether attacks can be detected was analysed. Furthermore, the positive correlation between the FDIA detection effectiveness and the watermarking intensity was explored by using the cross covariance of watermarking and residuals and auto covariance of residuals, where zero performance loss from watermarking was yielded thanks to watermarking as symmetric-key encryption. Thirdly, the tolerance capacity of the FDIA against the system was discussed and it was shown that the system performance can be recovered from the FDIA, where the quantitative relationship between the new DW scheme and system performance was studied. Finally, the proposed scheme was applied to a real inverted pendulum system, and experimental results demonstrated its superior performance.

To deal with the weaknesses of the conventional DW scheme, we have studied the secure control of NCSs by a new developed DW scheme. However, there exist some NCSs with many subsystems (e.g., networked multi-agent systems [41], [42]), where information exchange is performed among subsystems via communication networks. If the communication networks between subsystems are attacked, the performance of the subsystems or the whole system goes down. Therefore, it is interesting to investigate the secure control of such NCSs based on the proposed DW scheme.

#### REFERENCES

- M. Masera, E. F. Bompard, F. Profumo, and N. Hadjsaid, "Smart (electricity) grids for smart cities: Assessing roles and societal impacts," *Proc. IEEE*, vol. 106, no. 4, pp. 613–625, 2018.
- [2] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, "IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8182– 8201, 2019.
- [3] W. He, H. Huang, and S. S. Ge, "Adaptive neural network control of a robotic manipulator with time-varying output constraints," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3136–3147, 2017.
- [4] X. Wang, J. H. Park, H. Liu, and X. Zhang, "Cooperative outputfeedback secure control of distributed linear cyber-physical systems resist intermittent DoS attacks," *IEEE Trans. Cybern.*, 2020, to be published, doi: 10.1109/TCYB.2020.3034374.
- [5] B. Chen, D. W. C. Ho, G. Hu, and L. Yu, "Secure fusion estimation for bandwidth constrained cyber-physical systems under replay attacks," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1862–1876, 2018.

- [6] J. Liu, Y. Wang, J. Cao, D. Yue, and X. Xie, "Secure adaptiveevent-triggered filter design with input constraint and hybrid cyber attack," *IEEE Trans. Cybern.*, 2020, to be published, doi: 10.1109/TCYB.2020.3003752.
- [7] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 Ukraine blackout: Implications for false data injection attacks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, 2017.
- [8] Z. Pang, G. Liu, D. Zhou, F. Hou, and D. Sun, "Two-channel false data injection attacks against output tracking control of networked systems," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3242–3251, 2016.
- [9] D. Ye and T. Y. Zhang, "Summation detector for false data-injection attack in cyber-physical systems," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2338–2345, 2020.
- [10] S. Tan, J. M. Guerrero, P. Xie, R. Han, and J. C. Vasquez, "Brief survey on attack detection methods for cyber-physical systems," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5329–5339, 2020.
- [11] X.-M. Zhang, Q.-L. Han, X. Ge, and L. Ding, "Resilient control design based on a sampled-data model for a class of networked control systems under denial-of-service attacks," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3616–3626, 2020.
- [12] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in 47th Annu. Allerton Conf. Commun., Control, Comput., 2009, pp. 911–918.
- [13] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyberattack on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 4–13, 2017.
- [14] M. A. Qadir and I. Ahmad, "Digital text watermarking: Secure content delivery and data hiding in digital documents," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 21, no. 11, pp. 18–21, 2006.
- [15] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *53rd IEEE Conf. Decis. Control*, 2014, pp. 3757–3764.
- [16] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, 2014.
- [17] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Syst. Mag.*, vol. 35, no. 1, pp. 93–109, 2015.
- [18] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber-physical systems," *Proc. IEEE*, vol. 105, no. 2, pp. 219–240, 2017.
- [19] T. Huang, B. Satchidanandan, P. R. Kumar, and L. Xie, "An online detection framework for cyber attacks on automatic generation control," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6816–6827, 2018.
- [20] B. Satchidanandan and P. R. Kumar, "On the design of securityguaranteeing dynamic watermarks," *IEEE Control Syst. Lett.*, vol. 4, no. 2, pp. 307–312, 2020.
- [21] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general LTI systems," in *IEEE 56th Annu. Conf. Decis. Control*, 2017, pp. 1834–1839.
- [22] M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, "Detecting generalized replay attacks via time-varying dynamic watermarking," *IEEE Trans. Autom. Control*, 2020, to be published, doi: 10.1109/TAC.2020.3022756.
- [23] R. M. G. Ferrari and A. M. H. Teixeira, "Detection and isolation of routing attacks through sensor watermarking," in *Am. Control Conf.*, 2017, pp. 5436–5442.
- [24] A. M. H. Teixeira and R. M. G. Ferrari, "Detection of sensor data injection attacks with multiplicative watermarking," in *Eur. Control Conf.*, 2018, pp. 338–343.
- [25] R. M. G. Ferrari and A. M. H. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Trans. Autom. Control*, 2020, to be published, doi: 10.1109/TAC.2020.3013850.
- [26] H. Yang, Q.-L. Han, X. Ge, L. Ding, Y. Xu, B. Jiang, and D. Zhou, "Fault-tolerant cooperative control of multiagent systems: A survey of trends and methodologies," *IEEE Trans. Ind. Inf.*, vol. 16, no. 1, pp. 4–17, 2020.
- [27] Y. Zou and K. Xia, "Robust fault-tolerant control for underactuated takeoff and landing UAVs," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 5, pp. 3545–3555, 2020.
- [28] J.-W. Zhu, C.-Y. Gu, S. X. Ding, W.-A. Zhang, X. Wang, and L. Yu, "A new observer-based cooperative fault-tolerant tracking control method with application to networked multiaxis motion control system," *IEEE Trans. Ind. Electron.*, vol. 68, no. 8, pp. 7422–7432, 2021.
- [29] Y. Mo and B. Sinopoli, "On the performance degradation of cyberphysical systems under stealthy integrity attacks," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2618–2624, 2016.

- [30] L. An and G. Yang, "Secure state estimation against sparse sensor attacks with adaptive switching mechanism," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2596–2603, 2018.
- [31] E. Mousavinejad, F. Yang, Q.-L. Han, X. Ge, and L. Vlacic, "Distributed cyber attacks detection and recovery mechanism for vehicle platooning," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3821–3834, 2020.
- [32] D. Du, X. Li, W. Li, R. Chen, M. Fei, and L. Wu, "ADMM-based distributed state estimation of smart grid under data deception and denial of service attacks," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 8, pp. 1698–1711, 2019.
- [33] D. Du, C. Zhang, H. Wang, X. Li, H. Hu, and T. Yang, "Stability analysis of token-based wireless networked control systems under deception attacks," *Inf. Sci.*, vol. 459, pp. 168–182, 2018.
- [34] Y. Guan and X. Ge, "Distributed attack detection and secure estimation of networked cyber-physical systems against false data injection attacks and jamming attacks," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 4, no. 1, pp. 48–59, 2018.
- [35] R. Torres, Z. Lizarazo, and E. Torres, "Fractional sampling theorem for  $\alpha$ -bandlimited random signals and its relation to the von Neumann ergodic theorem," *IEEE Trans. Signal Process.*, vol. 62, no. 14, pp. 3695–3705, 2014.
- [36] J. Katz and Y. Lindell, *Introduction to Modern Cryptography*, 2nd ed., ser. Chapman & Hall/CRC Cryptography and Network Security Series. Chapman and Hall/CRC, 2014.
- [37] G. Marsaglia and T. A. Bray, "A convenient method for generating normal variables," *SIAM Rev.*, vol. 6, no. 3, pp. 260–264, 1964.
- [38] X.-M. Zhang, Q.-L. Han, A. Seuret, F. Gouaisbaut, and Y. He, "Overview of recent advances in stability of linear systems with timevarying delays," *IET Control Theory Appl.*, vol. 13, no. 1, pp. 1–16, 2019.
- [39] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denialof-service," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2930–2944, 2015.
- [40] D. Du, C. Zhang, Y. Song, H. Zhou, X. Li, M. Fei, and W. Li, "Realtime  $H_{\infty}$  control of networked inverted pendulum visual servo systems," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 5113–5126, 2020.
- [41] D. Ding, Q.-L. Han, Z. Wang, and X. Ge, "Recursive filtering of distributed cyber-physical systems with attack detection," *IEEE Trans. Syst., Man, Cybern., Syst.*, 2020, to be published, doi: 10.1109/TSMC.2019.2960541.
- [42] D. Ding, Z. Wang, and Q.-L. Han, "Neural-network-based consensus control for multiagent systems with input constraints: The eventtriggered case," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3719–3730, 2020.

# Supplementary Materials—Secure Control of Networked Control Systems Using Dynamic Watermarking

Dajun Du, Changda Zhang, Xue Li, Minrui Fei, Taicheng Yang, Huiyu Zhou

Abstract—This is the supplementary document of the paper entitled "Secure Control of Networked Control Systems Using Dynamic Watermarking" submitted to IEEE Transactions on Cybernetics. Section I provides the proofs of limitations and theorems. Supplements of experimental results on the measurement noise and system performance with different watermarking densities under no attacks are given in Section II.

### I. PROOFS OF LIMITATIONS AND THEOREMS

### A. Proof of Limitations 1 and 2

Considering the system (12) in the main text, to analyse the ability of the conventional DW scheme for FDIA detection, we firstly define  $\xi_d(k) := x_a(k) - \hat{x}_d(k|k-1)$ . The dynamics of  $\xi_d(k)$  can be given by

$$\begin{cases} \xi_d(k) = \Phi_1 \xi_d(k-1) + \Phi_2 x_a(k-1) - B w_d(k-1) \\ r_d(k) = C \xi_d(k) \end{cases}$$
(A.1)

where  $\Phi_1$ ,  $\Phi_2$  are the same as (12) in the main text. Then, using (A.1), the variable  $\xi_d(k)$  can be computed recursively from 0 to k as follows

$$\xi_d(k) = \Phi_1^k \xi_d(0) + \sum_{p=1}^k \left( \Phi_1^{p-1} \Phi_2 x_a(k-p) - \Phi_1^{p-1} B w_d(k-p) \right).$$
(A.2)

Considering (A.1) and  $w_d(k-1) \perp \xi_d(0)$ ,  $w_d(k-1) \perp x_a(j)$ for  $\forall j$ , and  $w_d(k-1) \perp w_d(j)$  for  $j \neq k-1$  in (A.2), it shows that

$$\mathbb{E}\left[w_d(k-1)Lr_d(k)\right] = -\mathbb{E}\left[w_d(k-1)LCBw_d(k-1)\right].$$
(A.3)

The above equation yields (15a) in the main text.

Secondly, we focus on the steady-state behavior of (A.1). Note that  $\lim_{k\to\infty} x_a(k) = 0$  of the FDIA (11) in the main text; taking the limitation as  $k \to \infty$  for (A.1) yields

$$\begin{cases} \lim_{k \to \infty} \xi_d(k) = \lim_{k \to \infty} \Phi_1 \xi_d(k-1) - Bw_d(k-1) \\ \lim_{k \to \infty} r_d(k) = \lim_{k \to \infty} C\xi_d(k). \end{cases}$$
(A.4)

The work of D. Du, C. Zhang, X. Li, M. Fei, T. Yang, and H. Zhou was supported by the National Science Foundation of China under Grant Nos. 92067106, 61773253, 61633016 and 61533010, 111 Project under Grant No.D18003.

D. Du, C. Zhang, X. Li, and M. Fei are with Shanghai Key Laboratory of Power Station Automation Technology, School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China.

T. Yang is with Department of Engineering and Design, University of Sussex, Brighton BN1 9QT, U.K.

H. Zhou is with School of Informatics, University of Leicester, Leicester LE1 7RH, U.K.

Denote  $\mathbb{E}\left[\xi_d(\infty)\xi_d^T(\infty)\right] := M_d$  and consider  $w_d(k-1)\perp\xi_d(k-1)$ , then multiplying the both sides of the equal sign by the corresponding transposed vector shown in (A.4) and taking the expectation yields (15b) in the main text.

Up to now, limitation 1 has been proved. The following is the proof of limitation 2.

To investigate the action of the system state x(k), in the main text, specifically,  $\zeta_d(k) = [x(k); \xi_d(k); x_a(k)]$ , and the matrices  $\mathcal{A}_0 = [A, \mathrm{H}; 0, \Xi]$  and  $\Xi = [\Phi_1, \Phi_2; 0, A_a]$  in (12). Since  $\rho(A_a) < 1$  and  $\rho(\Phi_1) < 1$ , it follows  $\rho(\Xi) < 1$ , i.e.,  $[\xi_d(k); x_a(k)]$  is exponentially bounded in the mean-square sense. Inequality  $\rho(A) > 1$  results in  $\zeta_d(\infty) \to \infty$ . Therefore, it is clear that  $x(\infty) \to \infty$ , i.e., (16) in the main text. It completes the proof.

### B. Proof of Theorem 1

Since  $y_w^-(k)$  passes test (20) in the main text, it follows that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} w_{y,i}(k) \left( \mathcal{D}_r(k) + Lr^o(k) \right) = 0.$$
 (A.5)

Substituting  $\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} w_{y,i}(k) Lr^o(k) = 0$  into (A.5) yields

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} w_{y,i}(k) \mathcal{D}_r(k) = 0.$$
 (A.6)

Since  $y_w^-(k)$  also passes test (21) in the main text, it follows that

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \left( \mathcal{D}_r(k) + Lr^o(k) \right) \left( \mathcal{D}_r(k) + Lr^o(k) \right)^T$$

$$= L \Sigma_o L^T.$$
(A.7)

Substituting  $\lim_{T\to\infty} \frac{1}{T} \sum_{k=0}^{T-1} Lr^o(k) (Lr^o(k))^T = L\Sigma_o L^T$  into (A.7) yields

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathcal{D}_r(k) \mathcal{D}_r^T(k) + 2\mathcal{D}_r(k) (Lr^o(k))^T = 0 \quad (A.8)$$

which can be re-written as the component-wise form

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=0}^{T-1} \mathcal{D}_{r,i}(k) \mathcal{D}_{r,j}(k) + \mathcal{D}_{r,i}(k) \left( L_j \cdot r^o(k) \right) + \mathcal{D}_{r,j}(k) \left( L_i \cdot r^o(k) \right) = 0$$
(A.9)

where  $\mathcal{D}_{r,i}(k)$  is the *i*th component of  $\mathcal{D}_r(k)$ .

According to decryption (19) and state estimate calculation (25), (27) in the main text, the decryption and state estimate calculation of the attack-free system can be given as

$$y_w^{o-}(k) = y_w^{o+}(k) - w_y(k), \tag{A.10}$$

$$Lr^{o}(k) = -C \left(A + BK\right) \hat{x}^{o}(k-1|k-1) + y_{w}^{o-}(k) \quad (A.11)$$

where  $y_w^{o+}(k)$  is the attack-free counterpart of  $y_w^+(k)$ . Substituting (A.10) into (A.11) yields

$$r_{L,i}^{o}(k) = -(C(A+BK))_{i}\hat{x}^{o}(k-1|k-1) + y_{w,i}^{o+}(k) - w_{y,i}(k)$$
(A.12)

where the variables  $r_{L,i}^{o}(k) := L_{i} \cdot r_{o}(k)$ , and  $y_{w,i}^{o+}(k)$  is the *i*th component of  $y_{w}^{o+}(k)$ . From (A.10)-(A.12) and let  $\mathcal{F}_{k} := \sigma\left((\hat{x}^{o})_{k-2}, (y_{w,i}^{o+})_{k-1}, (w_{y,i})_{k-1}\right)$ , a Markov chain can be formed by

$$\begin{pmatrix} (\hat{x}^{o})_{k-2}, (y_{w,i}^{o+})_{k-1}, (w_{y,i})_{k-1} \end{pmatrix} \to \\ (\hat{x}^{o}(k-1|k-1), y_{w,i}^{o+}(k)) \to r_{L,i}^{o}(k)$$

where the variables

$$(\hat{x}^{o})_{k-2} := \{ \hat{x}^{o}(k-2|, k-2), \hat{x}^{o}(k-3|, k-3), \dots, \hat{x}^{o}(0|0) \}$$

and  $(y_{w,i}^{o+})_{k-1}$ ,  $(w_{y,i})_{k-1}$  are well defined likewise. According to the above Markov chain, it can be given by

$$\hat{r}_{L,i}^{o}(k) = \mathbb{E}\left[r_{L,i}^{o}(k) \left| r_{L,i}^{o}(k) + w_{y,i}(k) \right.\right] \\ = \eta\left(r_{L,i}^{o}(k) + w_{y,i}(k)\right)$$
(A.13)

where variable  $\hat{r}_{L,i}^{o}(k) := \mathbb{E}\left[r_{L,i}^{o}(k) | \mathcal{F}_{k}\right]$ , and the constant  $\eta := \left(L_{i} \cdot \Sigma_{o} L_{i}^{T}\right)^{2} / \left(\left(L_{i} \cdot \Sigma_{o} L_{i}^{T}\right)^{2} + \sigma_{w_{y,i}}^{2}\right)$ . Let  $\tilde{r}_{L,i}^{o}(k) := r_{L,i}^{o}(k) - \hat{r}_{L,i}^{o}(k)$  and the following holds

$$\tilde{r}_{L,i}^{o}(k-1) \in \mathcal{F}_k, \mathbb{E}\left[\tilde{r}_{L,i}^{o}(k) \left| \mathcal{F}_k \right.\right] = 0.$$
(A.14)

From (A.14), it is clear that  $\{\tilde{r}_{L,i}^o(k)\}$  is a Martingale difference sequence. Applying the Martingale stability theorem [S1, Le. 2] to  $\tilde{r}_{L,i}^o(k)$  yields

$$\sum_{k=1}^{T} \mathcal{D}_{r,i}(k) \tilde{r}_{L,i}^{o}(k) = o\left(\sum_{k=1}^{T} \mathcal{D}_{r,i}^{2}(k)\right) + O(1) \quad (A.15)$$

where  $o(\cdot)$  is the infinitesimal of higher order over an infinitesimal, specially, o(1) is the infinitesimal of higher order over any constant; O(1) is the bounded quantity. Substituting (A.13) and (A.15) into (A.8) yields

$$\sum_{k=1}^{T} \mathcal{D}_{r,i}(k) r_{L,i}^{o}(k) = \sum_{k=1}^{T} \mathcal{D}_{r,i}(k) \tilde{r}_{L,i}^{o}(k) + \mathcal{D}_{r,i}(k) \hat{r}_{L,i}^{o}(k)$$
$$= o\left(\sum_{k=1}^{T} \mathcal{D}_{r,i}^{2}(k)\right) + O(1) + \eta \sum_{k=1}^{T} \mathcal{D}_{r,i}(k) r_{L,i}^{o}(k)$$
$$+ \eta \sum_{k=1}^{T} \mathcal{D}_{r,i}(k) w_{y,i}(k).$$
(A.16)

Substituting (A.6) into (A.16), it follows that

$$\sum_{k=1}^{T} \mathcal{D}_{r,i}(k) r_{L,i}^{o}(k) = o\left(\sum_{k=1}^{T} \mathcal{D}_{r,i}^{2}(k)\right) + O(1).$$
(A.17)

Therefore, from (A.17), the following holds

$$\sum_{k=1}^{T} \mathcal{D}_{r,i}^{2}(k) + 2\mathcal{D}_{r,i}(k)r_{L,i}^{o}(k) =$$

$$(1+o(1))\sum_{k=1}^{T} \mathcal{D}_{r,i}^{2}(k) + O(1).$$
(A.18)

Dividing the above equation by T, taking the limit as  $T \to \infty$ , and invoking (A.9), we have:

$$\lim_{T \to \infty} \frac{1}{T} \sum_{k=1}^{T} \mathcal{D}_{r,i}^2(k) = 0, i = 1, 2, \dots, m_x.$$
(A.19)

Summing (A.19) from i = 1 to  $i = m_x$  and considering the definition of norm  $\|\mathcal{D}_r(k)\|$  yields (30) in the main text. It completes the proof.

### C. Proof of Theorem 2

Considering system (28) in the main text, to analyse the ability of the new DW scheme in FDIA detection, we firstly define  $\wp(k) := x_a(k) - \hat{x}(k|k-1)$ . The dynamics of  $\wp(k)$  can be given by

$$\begin{cases} \wp(k+1) = \Phi_1 \wp(k) + \Phi_2 x_a(k) + (A + BK) L w_y(k) \\ r(k) = C \wp(k) - w_y(k) \end{cases}$$
(A.20)

where  $\Phi_1$ ,  $\Phi_2$  are the same as (12) in the main text. Then, using (A.20), the variable  $\wp(k)$  can be computed recursively from 0 to k as follows

$$\wp(k) = \Phi_1^k \wp(0) + \sum_{p=1}^k \Phi_1^{p-1} \Phi_2 x_a(k-p)$$

$$+ \sum_{p=1}^k \Phi_1^{p-1} (A+BK) L w_y(k-p).$$
(A.21)

For (A.21), considering  $w_{y,i}(k) \perp \wp(0)$ ,  $w_{y,i}(k) \perp x_a(p)$  for  $\forall p$ ,  $w_{y,i}(k) \perp w_{y,i}(p)$  for  $p \neq k$ , and  $w_{y,i}(k) \perp w_{y,j}(p)$  for  $\forall p$ , the following holds

$$\mathbb{E}\left[w_{y,i}(k)\wp(k)\right] = 0,\tag{A.22}$$

$$\mathbb{E}\left[w_{y,i}(k)Lr(k)\right] = -\mathbb{E}\left[w_{y,i}(k)L_{\cdot i}w_{y,i}(k)\right].$$
 (A.23)

Equation (A.23) yields (31a) in the main text.

Secondly, we focus on the steady-state behavior of (A.20). Note that  $\lim_{k \to a} x_a(k) = 0$  of the FDIA (11) in the main text; taking the limitation as  $k \to \infty$  for (A.20) produces

$$\begin{cases} \lim_{k \to \infty} \wp(k+1) = \lim_{k \to \infty} \Phi_1 \wp(k) + (A + BK) L w_y(k) \\ \lim_{k \to \infty} r(k) = \lim_{k \to \infty} C \wp(k) - w_y(k). \end{cases}$$
(A.24)

Denoting  $\mathbb{E}\left[\wp(\infty)\wp^T(\infty)\right] = M$  and considering (A.22), we multiply both sides of the equal sign by the corresponding transposed vector shown in (A.24) and take the expectation, yielding (31b) in the main text.

Thirdly, to investigate the action of the system state x(k), let us recall the main text, or more specifically, the variable  $\zeta(k) = [x(k); \wp(k); x_a(k)]$ , and the matrices  $\mathcal{A}_0 =$  $[A, H; 0, \Xi]$  and  $\Xi = [\Phi_1, \Phi_2; 0, A_a]$  in (28). Since  $\rho(A_a) < 1$ 

and  $\rho(\Phi_1) < 1$ , it arrives at  $\rho(\Xi) < 1$ , i.e.,  $[\xi_d(k); x_a(k)]$ is exponentially bounded in the mean-square sense. Inequality  $\rho(A) > 1$  yields  $\zeta_d(\infty) \to \infty$ . Therefore, it is clear that  $x(\infty) \to \infty$ , i.e., (31c) in the main text. It completes the proof.

### D. Proof of Theorem 3

For the switched system (33) in the main text, we first compute recursively  $\zeta(k)$  from  $t_0$  to k as follows

$$\zeta(k) = \mathcal{A}_{q_{i+1}}^{(k-t_i)} \mathcal{A}_{q_i}^{(t_i-t_{i-1})} \cdots \mathcal{A}_{q_1}^{(t_1-t_0)} \zeta(t_0) + \varpi(k)$$
(A.25)

where  $q_i = 0, 1$ ;  $t_i$  is the switching instant; and

$$\varpi(k) = \sum_{j=1}^{i-1} \sum_{p=t_{j-1}}^{t_j-1} \mathcal{A}_{q_{i+1}}^{k-t_i} \cdots \mathcal{A}_{q_{j+1}}^{t_j-t_{j-1}} \mathcal{A}_{q_j}^{t_j-p-1} \Lambda_{q_j} \psi(p) \\
+ \sum_{p=t_{i-1}}^{t_i-1} \mathcal{A}_{q_{i+1}}^{(k-t_i)} \mathcal{A}_{q_i}^{t_i-p-1} \Lambda_{q_i} \psi(p) + \sum_{p=t_i}^{k-1} \mathcal{A}_{q_{i+1}}^{k-p-1} \Lambda_{q_{i+1}} \psi(p).$$
(A.26)

Taking the norm of the both sides of the equal sign in (A.25), using norm triangle equality and taking the expectation yields

$$\mathbb{E}\left[\left\|\zeta(k)\right\|\right] \leqslant \mathbb{E}\left[\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_1}^{(t_1-t_0)}\zeta(t_0)\right\|\right] + \mathbb{E}\left[\left\|\varpi(k)\right\|\right].$$
(A.27)

Secondly, according to the norm consistency principle and (34) in the main text and the definition of  $\mathcal{T}_0$ ,  $\mathcal{T}_1$ ,  $N_s(0,k)$ , the first term on the right hand of (A.27) becomes

$$\mathbb{E}\left[\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_1}^{(t_1-t_0)}\zeta(t_0)\right\|\right] \leqslant \lambda_{-}^{g(i+1)}\lambda_{+}^{\mathcal{T}_0}\lambda_{-}^{\mathcal{T}_1}\mathbb{E}\left[\left\|\zeta(t_0)\right\|\right] \\
= \lambda_{-}^g\lambda_{-}^{gN_s(t_0,k)}\lambda_{+}^{\mathcal{T}_0}\lambda_{-}^{\mathcal{T}_1}\mathbb{E}\left[\left\|\zeta(t_0)\right\|\right]. \tag{A.28}$$

Substituting (37) in the main text into (A.28) yields

$$\mathbb{E}\left[\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_1}^{(t_1-t_0)}\zeta(t_0)\right\|\right] \leqslant \lambda_{-}^g \lambda_{-}^{gN_s(t_0,k)+\lambda^*k} \mathbb{E}\left[\left\|\zeta(t_0)\right\|\right].$$
(A.29)

For (A.29), there are two cases according to the value of g: 1) When  $g \ge 0$ , in (A.29), it follows that

$$\lim_{k \to \infty} \lambda_{-}^{g} \lambda_{-}^{gN_s(t_0,k) + \lambda^* k} = 0$$
 (A.30)

for any  $N_s(t_0, k)$ , i.e., for any  $\tau$  in (36) in the main text. Furthermore, by using the condition  $0 < \lambda^{\dagger} < \lambda^* < 1$ , (A.29) becomes

$$\mathbb{E}\left[\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_1}^{(t_1-t_0)}\zeta(t_0)\right\|\right] \leqslant$$

$$\lambda_{-}^g \lambda_{-}^{gN_s(t_0,k)+\lambda^{\dagger}k} \mathbb{E}\left[\left\|\zeta(t_0)\right\|\right].$$
(A.31)

2) When g < 0, by using (36) in the main text, it has  $gN_s(t_0, k) + \lambda^* k \ge \iota + \lambda^{\dagger} k$ , i.e.,

$$N_s(t_0, k) \leqslant \frac{\iota}{g} + \frac{k - t_0}{\tau_{ave}} \tag{A.32}$$

where  $\iota$  is any constant and  $\tau_{ave} = \frac{g}{\lambda^{\dagger} - \lambda^{*}}$ . Then, substituting (A.32) into (A.29) yields

$$\mathbb{E}\left[\left\|\mathcal{A}_{q_{i+1}}^{(k-t_{i})}\cdots\mathcal{A}_{q_{1}}^{(t_{1}-t_{0})}\zeta(t_{0})\right\|\right] \leqslant \lambda_{-}^{\iota+g}\lambda_{-}^{\lambda^{\dagger}(k-t_{0})}\mathbb{E}\left[\left\|\zeta(t_{0})\right\|\right].$$
(A.33)

Thirdly, by using the norm triangle equality and norm consistency principle, the second term on the right hand of (A.27) becomes

$$\mathbb{E}\left[\|\varpi(k)\|\right] \leqslant \\
\mathbb{E}\left[\sum_{j=1}^{i-1}\sum_{p=t_{j-1}}^{t_{j-1}} \left\|\mathcal{A}_{q_{i+1}}^{k-t_{i}}\cdots\mathcal{A}_{q_{j+1}}^{t_{j}-t_{j-1}}\mathcal{A}_{q_{j}}^{t_{j}-p-1}\right\| \left\|\Lambda_{q_{j}}\right\| \left\|\psi(p)\right\|\right] \\
+ \mathbb{E}\left[\sum_{p=t_{i-1}}^{t_{i-1}} \left\|\mathcal{A}_{q_{i+1}}^{(k-t_{i})}\mathcal{A}_{q_{i}}^{t_{i}-p-1}\right\| \left\|\Lambda_{q_{i}}\right\| \left\|\psi(p)\right\|\right] \\
+ \mathbb{E}\left[\sum_{p=t_{i}}^{k-1} \left\|\mathcal{A}_{q_{i+1}}^{k-p-1}\right\| \left\|\Lambda_{q_{i+1}}\right\| \left\|\psi(p)\right\|\right].$$
(A.34)

When g < 0, it is clear that in (A.34),

$$\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_2}^{(t_2-t_1)}\mathcal{A}_{q_1}^{t_1-t_0}\right\| \leqslant \lambda_{-}^{\iota+g}\lambda_{-}^{\lambda^{\dagger}(k-t_0)}, \qquad (A.35)$$

$$\left\|\mathcal{A}_{q_{i+1}}^{(k-t_i)}\cdots\mathcal{A}_{q_2}^{(t_2-t_1)}\mathcal{A}_{q_j}^{t_j-p-1}\right\| \leqslant \lambda_{-}^{\iota+g}\lambda_{-}^{\lambda^{\dagger}(k-p-1)}.$$
 (A.36)

Substituting (A.35), (A.36), and (35) in the main text (i.e.,  $\mathbb{E}[||\psi(p)||] \leq \psi_u < \infty$ ) into (A.34) yields

$$\mathbb{E}\left[\left\|\varpi(k)\right\|\right] \leqslant \mathbb{E}\left[\sum_{p=t_0}^{k-1} \mu \lambda_{-}^{\iota+g} \lambda_{-}^{\lambda^{\dagger}(k-p-1)} \left\|\psi(p)\right\|\right]$$

$$\leqslant \frac{\mu \lambda_{-}^{\iota+g} \psi_u}{1-\lambda_{-}^{\lambda^{\dagger}}} \lambda_{-}^{\lambda^{\dagger}(k-t_0)} \stackrel{k=t_0}{\leqslant} \frac{\mu \lambda_{-}^{\iota+g} \psi_u}{1-\lambda_{-}^{\lambda^{\dagger}}} < \infty$$
(A.37)

where  $\mu = \max \{ \|\Lambda_0\|, \|\Lambda_1\| \}.$ 

Finally, substituting (A.33) and (A.37) into (A.27) yields

$$\mathbb{E}\left[\left\|\zeta(k)\right\|\right] \leqslant \lambda_{-}^{\iota+g} \lambda_{-}^{\lambda^{\dagger}(k-t_{0})} \mathbb{E}\left[\left\|\zeta(t_{0})\right\|\right] + \frac{\mu \lambda_{-}^{\iota+g} \psi_{u}}{1 - \lambda_{-}^{\lambda^{\dagger}}} \quad (A.38)$$

which is similar to the situation when  $g \ge 0$ . It completes the proof.

### E. Proof of Theorem 4

To analyse the stability and performance of system (29) in the main text, the following Lyapunov-Krasovskii functional candidate is chosen

$$V(k) = \bar{\zeta}^{T}(k)Z_{1}\bar{\zeta}(k) + \sum_{q=-\bar{h}+1}^{0} \sum_{p=k-1+q}^{k-1} \xi^{T}(p)E^{T}Z_{2}E\xi(p) + \sum_{p=k-h(k)}^{k-1} \bar{\zeta}(p)E^{T}Z_{3}E\bar{\zeta}(p)$$
(A.39)

where  $\xi(k) := \overline{\zeta}(k+1) - \overline{\zeta}(k)$ . Conducting difference for (A.39) yields

$$\Delta V(k) = \bar{\zeta}^{T}(k+1)Z_{1}\bar{\zeta}(k+1) - \bar{\zeta}^{T}(k)Z_{1}\bar{\zeta}(k) + \bar{h}\xi^{T}(k)E^{T}Z_{2}E\xi(k) + \bar{\zeta}^{T}(k)E^{T}Z_{3}E\bar{\zeta}(k) - \bar{\zeta}^{T}(k-h(k))E^{T}Z_{3}E\bar{\zeta}(k-h(k)) - \sum_{p=k-\bar{h}}^{k-1}\xi^{T}(p)E^{T}Z_{2}E\xi(p)$$
(A.40)

where  $\Delta V(k) := V(k+1) - V(k)$ .

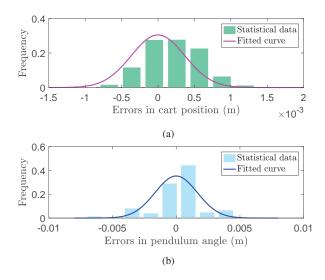


Fig. A.1. Fitted curves on errors of the cart position and pendulum angle.

For convenience, we here define an augmented variable as  $\aleph(k) := [\bar{\zeta}(k); E\bar{\zeta}(k-h(k)); \bar{\psi}(k)].$ 

According to the defined  $\aleph(k)$  and [S2, Le. 1], (A.40) becomes

$$\Delta V(k) + e^T(k)e(k) - \beta \tilde{\psi}^T(k)\tilde{\psi}(k) \leqslant \aleph^T(k)\Pi\aleph(k)$$
 (A.41)

where  $\Pi = \begin{bmatrix} \Theta & \tilde{\Omega} \\ * & \tilde{\Upsilon} \end{bmatrix}$  and

$$\begin{split} \tilde{\Omega} &= \left[ \begin{array}{ccc} hW_1^T & h(A_0 - I)^T E^T & (E - E^c A_0)^T & A_0^T \\ hW_2^T & hA_1^T E^T & (-E^c A_1)^T & A_1^T \\ hW_3^T & h\Gamma_0^T E^T & (-E^c \Gamma_0)^T & \Gamma_0^T \\ \end{array} \right],\\ \tilde{\Upsilon} &= diag \left\{ -hZ_2, -hZ_2^{-1}, -I, -Z_1^{-1} \right\}. \end{split}$$

By introducing two matrices  $W_4$  and  $W_5$  to deal with the nonlinear terms  $Z_2^{-1}$  and  $Z_1^{-1}$  [S2],  $\Pi < 0$  yields (39) in the main text.

As  $\Pi < 0$ , taking the sum of (A.41) from 0 to T - 1 and using the zero-initial condition (i.e., V(0) = 0) gives

$$\sum_{k=0}^{T-1} e^T(k)e(k) - \beta \bar{\psi}^T(k)\bar{\psi}(k) \leqslant -V(T) \leqslant 0.$$
 (A.42)

Dividing (A.42) by T and seeking the limitation as  $T \to \infty$  yield (40) in the main text. It completes the proof.

### **II. SUPPLEMENT ON EXPERIMENTAL RESULTS**

### A. Statistical Analysis of Measurement Noise

Using the data from [39] in the main text and the toolbox *cftool* in MATLAB, the fitted curves on the errors of the cart position and the pendulum angle are illustrated in Fig. A.1. For the fitted curves, it is calculated that the variances of  $v_1$  and  $v_2$  are  $2.7 \times 10^{-7}$  and  $5.5 \times 10^{-6}$  with root mean squared errors 0.0572 and 0.0740, respectively.

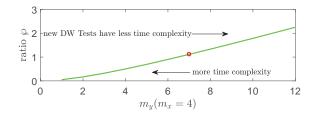


Fig. A.2. The ratio  $\wp$  with  $m_x = 4$ .

TABLE A.I COMPUTATIONAL TIME COMPLEXITY COMPARISON BETWEEN THE RESIDUAL-BASED BEST AND THE NEW DW TESTS.

Test	Computational Time complexity	Running time for test	Run time of experiments in Fig. 5
Residual-based test [34] New DW Tests	$O(m_y^2)$ $O(m_x^2 + m_x m_y)$	6.237 μs 34.158 μs	1.48s 1.48s

### B. Case Comparison of Computational Time Complexity

Fig. A.2 presents the ratio  $\wp$  with  $m_x = 4$  between the computational time complexity of residual-based test and new DW Tests. Tab. A.I lists computational time complexity of the residual-based test and the new DW tests compared in experiments of Fig. 5 in the main text. Note that the estimator and controller are periodic, i.e., control calculation and attack detection are performed every 10 ms: 1) the experiments of Fig. 5 run 148 steps (where 10 ms every step) and consume 1.48 s; 2) the tests are only calculated every 10 ms and the summations are listed in Tab. A.I.

### C. Conventional and New DW Schemes for FDIA with Different Initial Instants

Figs. A.3 and A.4 present experimental results of applying the FDIA at  $k \ge 400$  to NIPVSSs based on the conventional and the new DW schemes with  $\sigma_{w_d}^2 = \sigma_{w_{y,i}}^2 = 0.0001$ , respectively.

## D. States of NIPVSSs with Different Watermarking Densities under no Attacks

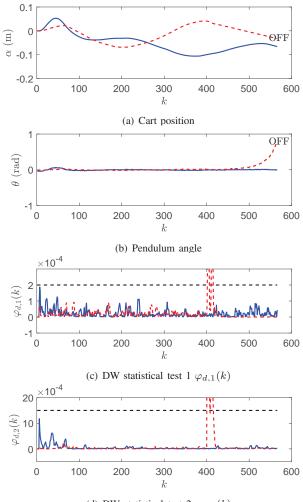
The cart position and pendulum angle are illustrated in Fig. A.5, where  $\sigma_{w_d}^2 = \sigma_{w_{y,i}}^2 = 0.0001$  for (a) and (b), and  $\sigma_{w_d}^2 = \sigma_{w_{y,i}}^2 = 10$  for (c) and (d).

### E. Analysis of the Compensated Variable

Fig. A.6 presents the value of compensated detection indicator  $\tilde{\varphi}_2(k)$ .

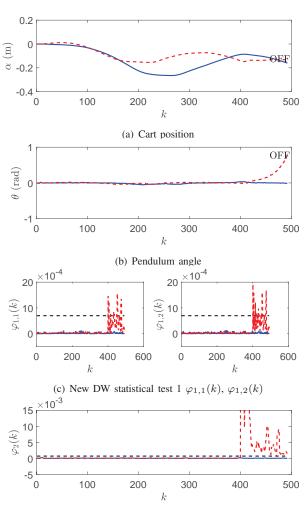
#### REFERENCES

- [S1] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *Ann. Statist.*, vol. 10, no. 1, pp. 154-166, 1982.
- [S2] X. Zhang and Q.-L. Han, "Delay-dependent robust H<sub>∞</sub> filtering for uncertain discrete-time systems with time-varying delay based on a finite sum inequality," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 12, pp. 1466-1470, 2006.



(d) DW statistical test 2  $\varphi_{d,2}(k)$ 

Fig. A.3. States and the DW statistical tests of NIPVSSs under the FDIA emerging at  $k \ge 400$  based on the conventional DW scheme with  $\sigma_{w_d}^2 = 0.0001$ . Blue line: Normal system. Red line: System under FDIA. Black line: Detection thresholds. "OFF" denotes that the servo is put off.



(d) New DW statistical test 2  $\varphi_2(k)$ 

Fig. A.4. States and the new DW statistical tests of NIPVSSs under the FDIA emerging at  $k \ge 400$  based on the new DW scheme with  $\sigma_{w_{y,i}}^2 = 0.0001$ . Blue line: Normal system. Red line: System under FDIA. Black line: Detection thresholds. "OFF" denotes that the servo is put off.

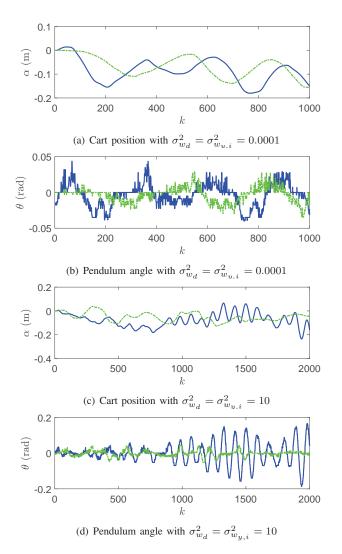


Fig. A.5. State of the NIPVSS under no attack with different watermark densities. Blue line: System based on conventional DW scheme. Green line: System based on new DW scheme.

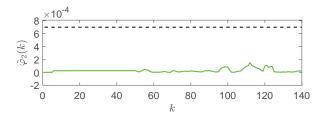


Fig. A.6. Values of  $\tilde{\varphi}_2(k)$  on  $\tilde{r}(k)$ . Green line:  $\tilde{\varphi}_2(k)$ . Black line: Detection threshold.