2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Delay-Induced Watermarking for Detection of Replay Attacks in Linear Systems

Christoforos Somarakis[1], Raman Goyal[2], Erfaun Noorani[3], Shantanu Rane[2]

***Abstract*— A state-feedback watermarking signal design for the detection of replay attacks in linear systems is proposed. The control input is augmented with a random time-delayed term of the system state estimate, in order to secure the system against attacks of replay type. We outline the basic analysis of the closed-loop response of the state-feedback watermarking in a LQG controlled system. Our theoretical results are applied on a temperature process control example. While the proposed secure control scheme requires very involved analysis, it, nevertheless, holds promise of being superior to conventional, feed-forward, watermarking schemes, in both its ability to detect attacks as well as the secured system performance.**

I. INTRODUCTION

The widespread integration of the physical and the cyber layers within the Industry 4.0 transformation has fueled research efforts in the field of Cyber-Physical Systems security (CPSs) [1]. The complex nature of CPSs makes such systems vulnerable to malicious attacks. A popular example of system vulnerability in a CPS environment, is this of *replay attack*. In a replay attack, the attacker records the output signal of the real system and then replays it back repeatedly for the controller [2]. In the absence of secure control, the victim stays oblivious to the attack. Figure 1 illustrates a system theory point of view of replay attacks: A master-slave hierarchy of identical systems with the output of the attack system to compromise the real system.

Fast and reliable methods of detecting attacks are, therefore, of deep interest in the field of estimation and control [3], [4]. Our work focuses on a secure method based on *input watermarking*. These are untraceable and unrepeatable signals, e.g. Gaussian noise, that typically get combined with the nominal input signal [5]. Their unique characteristic is that they enable detection of replay attacks whenever certain statistical analysis tools are utilized, such as the χ^2-based ones [6]. The rate at which replay attacks are detected, as well as, the level of performance degradation of the controlled system under watermarking signal inputs, are two metrics that characterize the overall efficiency of watermarking as a secure control method [7]. We present a novel type of watermarking, based on random time-delayed state-feedback. We analyze the resulting dynamics and discuss an example where this design can offer security with comparable, if not improved, detection rate and performance degradation as compared to mainstream watermarking that assume an additive Gaussian signal. The remainder of the paper is organized as follows: Notations are presented in §II. The problem formulation is put in §III. In §IV, we discuss preliminary results and in §V we detail system response and key statistics in presence of a replay attack. In §VI, we discuss the

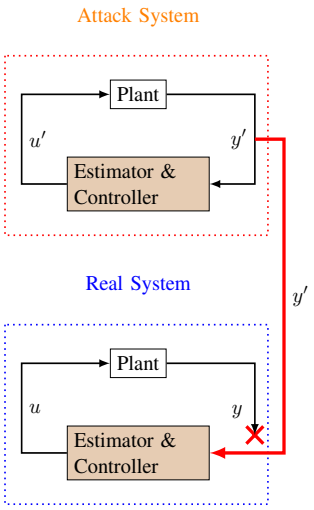

Fig. 1. The replay attack cast as a drive-response system. The attack signal, y', intervenes and compromises the real-time measurements of the real system. Signal attack, y', a recorded part of y, is considered as an output of a copy system running in parallel with the real system.

[1] Christoforos Somarakis is with the Data Science and Applied Mathematics Group, Merck & Co, West Point, PA, USA christoforos.somarakis@merck.com

[2] Raman Goyal, Shantanu Rane are with Palo Alto Research Center - An SRI Company, Palo Alto, CA, USA {rgoyal,srane}@parc.com

[3] Erfaun Noorani is with the Department of Electrical and Computer Engineering, the Institute for System Research (ISR) at the University of Maryland College Park, College Park, MD, USA. E. Noorani is a Clark doctoral Fellow at A. James Clark School of Engineering. enoorani@umd.edu

effect of feedback time-delay control in the system performance. In §VII, we conduct an experimental study of the effectiveness of our watermarking design and benchmark our design against widely known Gaussian watermarking. The numerical example is an Industrial Engineering benchmark of temperature control. Concluding remarks are placed § VIII. Proofs of technical results are omitted due to space limitations.

II. Nomenclature

The n-dimensional vector space is denoted by $\mathbb{R}^n$. It consists of vector elements $x \in \mathbb{R}^n$. The $n \times m$ zero matrix is $O_{n\times m}$, while by O_n is the $n \times n$ square zero matrix. The square identity matrix is denoted by I_n. Analysis will leverage four state-space representations: The system state-space embedded in $\mathbb{R}^n$, the system-controller state-space embedded in $\mathbb{R}^{2n}$, the drive-response space embedded in $\mathbb{R}^{4n}$, and the uplifted drive-response space embedded in $\mathbb{R}^{4n(\overline{\tau}+1)}$. Projection operators are denoted by $\mathfrak{P}\big[\,\cdot\,\big] \in \mathbb{R}^n \to \mathbb{R}^m$, for $m < n$. Notation $\mathcal{N}(\mu, \Sigma)$ denotes the n-dimensional normal distribution with mean μ and covariance matrix Σ. Symbol $\mathbb{E}[\cdot]$ is reserved for the expectation operator. $\mathbb{E}_{\sim p}$ denotes expectation with respect to the p random variable. By $\rho(A)$ we denote the spectral radius of matrix A. The Variance $\mathrm{Var}(\cdot)$, Covariance $\mathrm{Cov}(\cdot)$ vectorization, $\mathrm{vec}(\cdot)$ and trace $\mathrm{trace}(\cdot)$ operators are used in the standard manner.

III. Problem Formulation

The plant is a discrete LTI stochastic system. The state estimator is a Kalman filter providing the plant state estimate. The state estimator feedback optimizes a linear quadratic cost by regulating output towards zero. Lastly, another feedback control comprising the watermarking signal is added together with the optimal LQG feedback.

A. Real System Dynamics

1) Plant Equations: The state of plant $x \in \mathbb{R}^{n_x}$ evolves in time $t \in \mathbb{Z}_+$ according to:

$$x_{t+1} = A\,x_t + B\,u_t + w_t, \tag{1}$$

$$y_t = C\,x_t + v_t, \tag{2}$$

where $x_t \in \mathbb{R}^{n_x}$ is the state of the system and $u_t \in \mathbb{R}^{n_u}$ is the control input, at time t. Process noise $w_t \sim \mathcal{N}(\mathbf{0}, \Sigma_W)$, $\forall\, t$ and measurement noises $v_t \sim \mathcal{N}(\mathbf{0}, \Sigma_V)$, $\forall\, t$ follow zero mean Gaussian distribution with respective covariances.

2) Controller Equations: State vector is regulated around a reference value in the mean sense, with a controller that optimizes some cost function.

a) Kalman Filter: For the predicted state estimate we use the notation $\hat{x}_t = x_{t|t-1}$ and for the updated state the notation $\tilde{x}_t = x_{t|t}$. The state dynamics are:

$$x_{t+1|t} = A\,x_{t|t-1} + B\,u_t + L\,(y_t - C\,x_{t|t-1}), \tag{3}$$

where $L := APC^T(CPC^T + \Sigma_V)^{-1}$ is the $n_x \times n_y$ Kalman gain and P is the $n_x \times n_x$ matrix solution of the algebraic Ricatti equation $P = A\big[P - PC^T(CPC^T + \Sigma_V)^{-1}CP\big]A^T + D\Sigma_W D^T$. The Kalman output is

$$x_{t|t} = (I_{n_x} - MC)\,x_{t|t-1} + My_t, \tag{4}$$

with $M = PC^T(CPC^T + \Sigma_V)^{-1}$ the innovation gain.

b) LQG Controller: We implement the feedback

$$u_t^{\mathrm{LQG}} = -K_{\tilde{x}}\,\tilde{x}_t, \tag{5}$$

where gain K_x is designed to minimize the cost:

$$J = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\bigg[\sum_{k=0}^{t-1} x_k^T\,Q\,x_k + u_k^T\,R\,u_k\bigg]. \tag{6}$$

It is well-known, [8], that $K_{\tilde{x}} = (R + B^T\,\Delta\,B)^{-1}B^T\Delta A$ with Δ solves $\Delta = Q + A^T\Delta\,A - A^T\Delta\,B(R + B^T\Delta\,B)^{-1}B^T\Delta\,A$.

c) Watermarking: We utilize a delayed version of state estimate with a time-varying delay with the form

$$u_t^{\mathrm{WM}} = -K_\tau\,\tilde{x}_{t-\tau_t}, \tag{7}$$

where the gain matrix K_τ, and the time-delay τ_t, are design parameters. We consider time-delays as random variables such that for every t, $\tau_t \in \{1, \ldots, \overline{\tau}\}$ with mass probability function p_τ. Time series $\{\tau_t\}_{t\geq 0}$ consists of IID random variables that are also independent of plant process and sensor noises. In conclusion, the control in our problem has the form

$$u_t = u_t^{\mathrm{LQG}} + u_t^{\mathrm{WM}} \tag{8}$$

with u_t^{LQG} as in Eq. (5) and u_t^{WM} as in Eq. (7).

B. System-Controller Dynamics

The coupled dynamics enable the augmented state $\boldsymbol{x}_t := \big(x_t^T, \hat{x}_t^T\big)^T \in \mathbb{R}^{2n_x}$, $\boldsymbol{n}_t := \big(w_t^T, v_t^T, v_{t-\tau_t}^T\big)^T \in \mathbb{R}^{n_x+2n_y}$, where $\hat{x}_t := x_{t|t-1}$ from Eq. (3), evolving as

$$\boldsymbol{x}_{t+1} = \mathbf{A}\,\boldsymbol{x}_t + \mathbf{B}\,\boldsymbol{x}_{t-\tau_t} + \mathbf{\Gamma}\,\boldsymbol{n}_t, \quad t > \overline{\tau}, \tag{9}$$

for appropriate $\mathbf{A} \in \mathbb{R}^{2n_x\times 2n_x}$, $\mathbf{B} \in \mathbb{R}^{2n_x\times 2n_x}$, $\mathbf{\Gamma} \in \mathbb{R}^{2n_x\times n_x+2n_y}$.

C. Attack System Dynamics

Replay attacks take place with an attacker compromising the real system sensing. The attack time t' is defined as the first time instance when sensor component, instead of y_t, transmits measurement signal y'_t, for $t \geq t'$. Compromised signal $y' = \{y'_t\}_{t=t_{\text{start}}}^{t_{\text{end}}}$ is assumed to be a signal of the true output Eq. (2) recorded by the attacker at past time, and being replayed during the attack in a loop. Following the line of arguments as in [7], we assume that y'_t is the result of a *virtual* system, i.e. a copy of Eqs. (1) to (4) with control Eq. (8) that is operating in parallel with the real system. The attack system dynamics are

$$\begin{aligned} x'_{t+1} &= A\,x'_t + B\,u'_t + w'_t \\ y'_t &= C\,x'_t + v'_t \\ \hat{x}'_{t+1} &= A\,\hat{x}'_t + B\,u'_t + L\,(y'_t - C\,\hat{x}'_t) \end{aligned} \tag{10}$$

The control law is $u'_t = -K_{\tilde{x}}(I - MC)\,\hat{x}'_t - K_{\tilde{x}}MC\,x'_t - K_{\tilde{x}}M\,v'_t - K_\tau(I - MC)\,\hat{x}'_{t-\tau'_t} - K_\tau MC\,x'_{t-\tau'_t} - K_\tau M\,v'_{t-\tau'_t}$. When the attack moment, t', is large enough, the sources of randomness in the attack system can be assumed independent from the real system's ones.

D. Drive-Response System Under Replay Attack

During a replay attack the real system measures a compromised signal, e.g. a past copy of the real system. This signal could be provided by (10). For implementation purposes, the attack signal occurs as an information leak of measurement for a considerable amount of time, say leaked data set $\{y_t\}_{t=t_1}^{t_2}$, so that replay signal y' is a repeat in a loop. The real system (plant and estimator) operates then according to

$$\begin{aligned} x_{t+1} &= A\,x_t + B\,u_t + w_t \\ y_t &= C\,x_t + v_t \\ \hat{x}_{t+1} &= A\,\hat{x}_t + B\,u_t + L\,(y'_t - C\,\hat{x}_t) \end{aligned} \tag{11}$$

and now the control law (8) takes the form $u_t = -K_{\tilde{x}}(I - MC)\,\hat{x}_t - K_{\tilde{x}}MC\,x'_t - K_{\tilde{x}}M\,v'_t - K_\tau(I - MC)\,\hat{x}_{t-\tau_t} - K_\tau MC\,x'_{t-\tau_t} - K_\tau M\,v'_{t-\tau_t}$. One can cast (10) and (11) as a drive-response system architecture where the attack system (10) drives the dynamics of (11). We construct the drive-response state-space vectors: $\boldsymbol{x}_t^\dagger := \left(x_t^T, \hat{x}_t^T, x_t'^T, \hat{x}_t'^T\right)^T \in \mathbb{R}^{4n_x}$ $\boldsymbol{n}_t^\dagger := \left(w_t^T, w_t'^T, v_t'^T, v'^T_{t-\tau_t}, v'^T_{t-\tau'_t}\right)^T \in \mathbb{R}^{2n_x+3n_y}$, with associated dynamics to be

$$\boldsymbol{x}_{t+1}^\dagger = \mathbf{A}^\ddagger\,\boldsymbol{x}_t^\dagger + \mathbf{B}^\ddagger\,\boldsymbol{x}_{t-\tau_t}^\dagger + \mathbf{C}^\ddagger\,\boldsymbol{x}_{t-\tau'_t}^\dagger + \mathbf{G}^\ddagger\,\boldsymbol{n}_t^\dagger \tag{12}$$

with matrices $\mathbf{A}^\ddagger, \mathbf{B}^\ddagger, \mathbf{C}^\ddagger, \mathbf{G}^\ddagger$ of appropriate form. Note that unless a replay-attack takes place, $\boldsymbol{x}_t^\dagger$ represents the state of two decoupled identical systems.

E. Uplifted Drive-Response Level Dynamics

The core difficulty in dynamics (12) is the presence of time-delays that also vary with time hindering an explicit expression of solution to be used in the attack detection. One way to deal with this challenge is by grouping together $\boldsymbol{x}_t^{t-\overline{\tau}}$. The uplifted vectors are $\mathbb{X}_t = \left[\boldsymbol{x}_t^\dagger, \boldsymbol{x}_{t-1}^\dagger, \ldots, \boldsymbol{x}_{t-\overline{\tau}}^\dagger\right] \in \mathbb{R}^{4n_x(\overline{\tau}+1)}$ and $\mathbb{N}_t \leftrightarrow \boldsymbol{n}_t^\dagger$. The dynamics in the uplifted space read as:

$$\mathbb{X}_{t+1} = \mathcal{A}_t\,\mathbb{X}_t + \mathcal{G}\,\mathbb{N}_t \tag{13}$$

with $\mathcal{A}_t = \mathcal{A}_{\tau_t,\tau'_t}$ a time-dependent (random) matrix, and $\mathcal{G}$ a constant matrix. For every $t > 0$, $\mathcal{A}_t = \mathcal{A}_{\tau_t,\tau'_t}$ is sampled according to probability mass function $\{p_\tau p_{\tau'}\}_1^{\overline{\tau}}$ with $\overline{\mathcal{A}} := \mathbb{E}[\mathcal{A}_t] = \mathbb{E}\left[\mathcal{A}_{\tau_t,\tau'_t}\right] = \sum_{\tau,\tau'=1}^{\overline{\tau}} \mathcal{A}_{\tau,\tau'}\,p_\tau\,p_{\tau'}$ to be the expected value of $\mathcal{A}$. Turning system Eq. (9) into Eq. (13) suppresses time-delays at the expense of embedding the solution in a high-dimensional space evolving with time-dependent (random) dynamics. The vector $\mathcal{G}\,\mathbb{N}_t$ involves random variables distributed according to $\{p_\tau\}$, $\{p'_\tau\}$ as well as normal distribution. Solution of Eq. (13) can be represented for $t > \overline{\tau}$ as

$$\mathbb{X}_t = \mathcal{A}_{t:0}\,\mathbb{X}_0 + \sum_{k=0}^{t-1} \mathcal{A}_{t:k+1}\,\mathcal{G}\,\mathbb{N}_k \tag{14}$$

where $\mathcal{A}_{t_2:t_1} := \mathcal{A}_{t_2-1}\cdots\mathcal{A}_{t_1}$ for $t_2 > t_1$ is the transition matrix, with the standard properties: $\mathcal{A}_{t_2,t_1} = \mathcal{A}_{t_2,t'}\mathcal{A}_{t',t_1}, \forall\, t_2 \geq t' \geq t_1$, and $\mathcal{A}_{t:t} = I_{4n_x(\overline{\tau}+1)}$. Evidently, $\mathcal{A}_{t_2:t_1}$ is a product of $t_2 - t_1$ independent and identically distributed matrix-valued random variables each of which encapsulates the (mutually independent) random variables τ_t and τ'_t from the real and the attack system accordingly.

IV. Preliminaries

The stability of the free dynamics of Eq. (9) characterize the long-term properties of the corresponding state matrices in (12) and (13).

A. Stability on the System-Controller Level Dynamics

Watermarking signals are expected to impact negatively the overall system performance. The proposed watermarking as a feedback loop with time-delays may also destabilize the real system. It is therefore critical to design feedback gain matrices K_τ for u_t^{WM} in (7) and distributions $\{1, \ldots, \overline{\tau}\}$ such that $\mathbf{x}_t := \mathbb{E}_{\sim w,v}[\boldsymbol{x}_t]$ satisfies $\lim_{t\to+\infty} \mathbf{x}_t := \lim_{t\to+\infty} \mathbb{E}_{\sim w,v}[\boldsymbol{x}_t] = 0$. The result of this section states sufficient conditions for existence of such K_τ and $\overline{\tau}$. There are a few approaches we can adopt towards this end. Without

loss of generality, we assume $\tau_t \in \{1, \dots, \overline{\tau}\}$ to be some deterministic function of time taking values in $\{1, \dots, \overline{\tau}\}$. These dynamics satisfy:

$$\mathbf{x}_{t+1} = \mathbf{A}\,\mathbf{x}_t + \mathbf{B}\,\mathbf{x}_{t-\tau_t}. \tag{15}$$

The LQG control makes $\mathbf{A}$ Hurwitz, so that the Lyapunov equation $\mathbf{A}^T\mathbf{H}\,\mathbf{A} - \mathbf{H} = -\mathbf{C}$ has a unique positive definite solution $\mathbf{H} = \mathbf{H}^T$, for any positive definite $\mathbf{C} = \mathbf{C}^T$. Let $c > 0$ be the smallest eigenvalue of $\mathbf{C}$ and $0 < \underline{\eta} \leq \overline{\eta}$ the largest and the smallest eigenvalues of $\mathbf{H}$. Define

$$\alpha := \begin{cases} \frac{\left|\overline{\eta} - c + |\mathbf{A}^T\mathbf{H}\mathbf{B}|\right|}{\overline{\eta}}, & c > |\mathbf{A}^T\mathbf{H}\mathbf{B}| \\ \frac{\left|\underline{\eta} - c + |\mathbf{A}^T\mathbf{H}\mathbf{B}|\right|}{\underline{\eta}}, & c \leq |\mathbf{A}^T\mathbf{H}\mathbf{B}| \end{cases}, \tag{16}$$

$$\beta := \underline{\eta}^{-1}\big(|\mathbf{A}^T\mathbf{H}\mathbf{B}| + |\mathbf{B}^T\mathbf{H}\mathbf{B}|\big). \tag{17}$$

Theorem IV.1. *Let Eq.* (15) *with* $\tau_t : [0, \infty) \to \mathcal{T}$ *for* $\mathcal{T} \subset \mathbb{N}$*, and* $\overline{\tau} = \max_{\tau \in \mathcal{T}}\{\tau\} < \infty$*, and its solution* $\mathbf{x}$*. Assume that* $\alpha + \beta < 1$ *for* α, β *as in Eq.* (16) *and Eq.* (17)*. Then* $\mathbf{x}$ *converges to zero exponentially fast.*

The probabilistic counterpart of the long term behavior of $\mathbf{x}_t$ follows naturally. Under conditions of Theorem IV.1, $\{\mathbf{x}_t\}_t$ converges to 0, almost surely and in probability. Furthermore, $\|\mathbf{x}_t\|$ is bounded hence $\{\mathbf{x}_t\}_t$ converges with respect to the L^r norm for every $r > 0$.

B. Stability on the Drive-Response Level Dynamics

In the event of an attack, the virtual system remains unaffected as it drives the real system. The latter's internal stability gets compromised, leading to performance degradation or instability. The case of interest is the former one.

Assumption IV.1. The drive-response system under attack, is internally asymptotically stable.

C. Stability on the Uplifted Level Dynamics

Theorem IV.1 together with Assumption IV.1 leads on conclusions about the long term behavior of $\mathcal{A}_t$ and the quantity $\mathbb{A} := \mathbb{E}_{\sim\tau,\tau'}\big[\mathcal{A}_{\tau,\tau'} \otimes \mathcal{A}_{\tau,\tau'}\big]$ that will come of use in the next section.

Proposition IV.1. The following conditions hold true:

1) $\lim_{t\to+\infty} \mathcal{A}_{t:t_0} = O_{4n_x(\overline{\tau}+1)}$ almost surely.
2) $\rho\big(\mathbb{A}\big) < 1$.

D. The Asymptotic Auto-Covariance

A central quantity to this work is this of the asymptotic covariance matrix of $\boldsymbol{x}_t$ which we can express via $\mathbb{X}_t$ and appropriate projection. Let $\mathcal{C} = \lim_{t\to+\infty} \mathrm{Cov}(\mathbb{X}_t, \mathbb{X}_t)$ which is equal to

$$\mathcal{C} = \lim_{t\to+\infty} \mathbb{E}\big[\mathbb{X}_t\mathbb{X}_t^T\big] \tag{18}$$

The derivation of a formula of $\mathcal{C}$ involves manipulation of (14), i.e. calculating expectations of random products with statistically dependent quantities. The technical details are overwhelmingly complicated, though calculation remains tractable. For our purposes we only state the formula of $\mathcal{C}$ which is:

$$\mathcal{C} = \mathrm{vec}^{-1}\left(\big(I_{16n_x^2(\overline{\tau}+1)^2} - \mathbb{A}\big)^{-1}\boldsymbol{\omega}\right) \tag{19}$$

where

$$\begin{aligned}\boldsymbol{\omega} = {} & \mathrm{vec}\left(\mathcal{G}\mathcal{Q}\mathcal{G}^T\right) + \mathbb{E}_{\sim\tau,\tau'}\Big[\mathcal{A}_{\tau,\tau'} \otimes \mathcal{G}\mathcal{S}\mathcal{G}^T\,\mathrm{vec}(\overline{\mathcal{A}}^T)^{\tau-1}\Big] \\ & + \mathbb{E}_{\sim\tau,\tau'}\Big[\mathcal{G}\mathcal{S}\mathcal{G}^T \otimes \mathcal{A}_{\tau,\tau'}\,\mathrm{vec}\left(\overline{\mathcal{A}}^{\tau-1}\right)\Big] + \\ & \sum_{l=1}^{\overline{\tau}} \mathbb{E}_{\sim\tau,\tau'}\Big[\left(\mathcal{A}_{\tau,\tau'} \otimes \mathcal{G}\,\mathcal{R}_{l,\tau,\tau'}\mathcal{G}^T\right)\mathrm{vec}\left((\overline{\mathcal{A}}^T)^{l-1}\right) \\ & + \left(\mathcal{G}\left(\mathcal{R}_{l,\tau,\tau'}\right)^T\mathcal{G}^T \otimes \mathcal{A}_{\tau,\tau'}\right)\mathrm{vec}\left(\overline{\mathcal{A}}^{l-1}\right)\Big].\end{aligned}$$

and matrices $\mathcal{R}_{(\cdot)}$, $\mathcal{Q}$ and $\mathcal{S}$ stated in Appendix §.

V. Replay Attacks & Detectability Analysis

Our hypothesis is that the intrusion signal is a past copy of the original output. A loop transmission of such signal to the real system constitutes a stealthy and successful attack, that can be detected by a feedback controller Eq. (8) that includes the term Eq. (7).

A. The χ^2 Fault Detector

The availabel data to examine are the observed output Eq. (2) and the Kalman estimate Eq. (3). The χ^2 detector is a standard tool in system diagnostics that leverages these data to detect data incoherencies that can be measured from the residuals $y - C\hat{x}$, [6]. The next result characterizes the limit distribution of the residuals.

Theorem V.1. *[6] For system Eqs.* (1) *and* (2) *with filter Eqs.* (3) *and* (4) *and LQG controller Eq.* (5)*, the residues* $y_t - C\hat{x}_{t|t-1}$ *are IID Gaussian distributed for every* t*, with* $\lim_t y_t - C\hat{x}_{t|t-1} \sim \mathcal{N}(0, \Sigma_R := CPC^T + \Sigma_V)$.

Given a time detection window T, the χ^2 detector within T is defined as

$$g_\kappa(T) = \sum_{t=\kappa}^{\kappa+T} (y_t - C\hat{x}_t)^T\Sigma_R^{-1}(y_t - C\hat{x}_t) \tag{20}$$

In view of Theorem V.1, $g_\kappa(T)$ follows a χ^2 distribution with Tn_y degrees of freedom [9]. Manipulation of (20)

with (9) leads to a crucial observation: in the absence of replay attacks the distribution of residual is agnostic to watermarking statistics (i.e. randomness induced by u_t^{WM}). Thus Theorem V.1 is valid when the system is not under attack. In the remainder of the section we will explore the effect of Eq. (5) on $\mathbb{E}[g_\kappa(T)]$.

B. Detection Analysis

During attacks, the compromised readings $y'(t)$, modify the χ^2 detector to:

$$g'_\kappa(T) = \sum_{t=\kappa}^{\kappa+T} (y'_t - C\hat{x}_t)^T \Sigma_R^{-1} (y'_t - C\hat{x}_t). \qquad (21)$$

Note that for $\kappa >> 1$ the expected value of Eq. (20) or Eq. (21) for is calculated by considering the expected value of the sum elements when $t \to +\infty$. In particular for Eq. (21) we observe that

$$\lim_{t\to+\infty} \mathbb{E}\left[(y'_t - C\hat{x}_t)^T \Sigma_R^{-1} (y'_t - C\hat{x}_t)\right]$$
$$= \text{trace}\left[\big(\lim_{t\to\infty} \text{Var}\,(x'_t - \hat{x}_t) + \Sigma_V\big) C^T \Sigma_R^{-1} C\right]$$

Following [7], we write the residual element of time t from (21) as $x'_t - \hat{x}_t = \big(x'_t - \hat{x}'_t\big) + \big(\hat{x}'_t - \hat{x}_t\big)$. for which the variance yields

$$\begin{aligned}\text{Var}\,\big(x'_t - \hat{x}_t\big) = &\text{Var}\,\big(x'_t - \hat{x}'_t\big) + \text{Var}\,\big(\hat{x}'_t - \hat{x}_t\big) + \\ &\text{Cov}\,\big(x'_t - \hat{x}'_t, \hat{x}'_t - \hat{x}_t\big) + \\ &\text{Cov}\,\big(\hat{x}'_t - \hat{x}_t, x'_t - \hat{x}'_t\big)\end{aligned}$$

The first term comes exclusively from the virtual system and from the discussion above it can be deduced that $\text{trace}\left[\big(\text{Var}(x'_t - \hat{x}'_t) + \Sigma_V\big) C^T \Sigma_R^{-1} C\right] \to Tn_y$. The rest of the terms can be represented with the use of (19) and projection operators $\mathfrak{P}\,\mathbb{X}_t := \hat{x}'_t - \hat{x}_t$, $\mathfrak{Q}\,\mathbb{X}_t := x'_t - \hat{x}'_t$. Straightforward algebra yields

$$\begin{aligned}\lim_{\kappa\to+\infty} \mathbb{E}\big[g'_\kappa(T)\big] &= Tn_y + \\ &+ T\,\text{trace}\left[\left(\mathfrak{P}\mathcal{C}\mathfrak{P}^T + 2\mathfrak{P}\mathcal{C}\mathfrak{Q}^T\right) C^T \Sigma_R^{-1} C\right].\end{aligned} \qquad (22)$$

C. Practical Considerations

Formula (22) demonstrates that, when system transitions from clean to attacked, the χ^2-detector gets triggered so that $\lim_{\kappa\to+\infty} \mathbb{E}\big[g'_\kappa(T)\big] > \lim_{\kappa\to+\infty} \mathbb{E}\big[g_\kappa(T)\big] = Tn_y$, and (22) quantifies the difference. A strong enough deviation of $\mathbb{E}[g_\kappa(T)]$ should be a reliable attack detector. The practical way that $g_\kappa(T)$ is utilized for fault detection is by comparing the sample mean against a threshold $\psi := \mathcal{O}(Tn_y)$. The intuition is that in the absence of an attack the residuals' statistics are as in Theorem V.1 and η can be the cutoff of false positive attack. In the attack event, the compromised signal y_t, denoted as y'_t, gets compared against $C\hat{x}$ must yield different statistics from Theorem V.1.

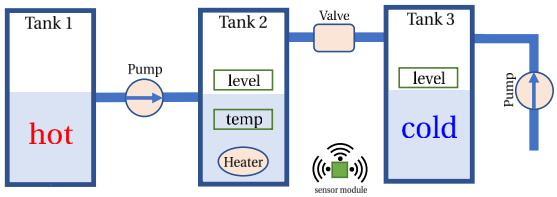

Fig. 2. The chemical process with four actuators (hot, cold pumps, valve, and heater) that control the level and temperature of tank 2 and level of tank 3.

VI. Performance Implications

The feedback controller is designed to minimize Eq. (6). An additive input control signal will yields in higher cost, i.e. lower system performance. The effect of (7) in the value of (6) occurs as an extra term to the optimal control cost value. Indeed a standard Dynamic Programming argument shows that (6) under control (8), denoted as J_{WM}, can be cast as

$$J_{\text{WM}} = J_* + \text{trace}\left[K_\tau^T (B^T \Delta_0 B + R) K_\tau\right]. \qquad (23)$$

where J_* is the optimal cost value of (6) with (5) only, and $\Delta_o = \Delta_o^T$ the solution of the Ricatti equation

$$\Delta_o = Q + A^T \Delta_o A - A^T \Delta_o B (R + B^T \Delta_o B)^{-1} B^T \Delta_o A.$$

The steps to derive (23) are similar to cost expressions derived in [7].

VII. Simulation Example

We consider a chemical process shown in Figure 2 linearized along the lines of [10]. The control inputs are two flow pumps, as illustrated in Figure 2. The three states $x = (x_1, x_2, x_3)^T$ are the level of water in tanks 2 and 3 and the temperature of water in tank 2. The process and measurement noises are zero-mean Gaussian with covariances $W{=}0.5\, I_{n_x}$ and $V{=}0.1\, I_{n_y}$, respectively. The dynamics of this three-tank system are

$$A{=}\begin{bmatrix} 0.96 & 0 & 0 \\ 0.04 & 0.97 & 0 \\ -0.04 & 0 & 0.9 \end{bmatrix}, B{=}\begin{bmatrix} 8.8 & -2.3 & 0 & 0 \\ 0.2 & 2.2 & 4.9 & 0 \\ -0.21 & -2.2 & 1.9 & 21 \end{bmatrix}$$

and also $C = I_{n_x}$. We choose the time window as $T = 85$, threshold $\psi = 110$, the time-delays take values in $[50, 200]$ and time-delay feedback matrix $K_\tau = 0.0713 I_{n_x}$, designed to meet Theorem IV.1.

The performance metric to measure the system response is considered as the LQG cost with the weighting matrices: $Q_s = \text{diag}[0.3, 0.3, 2.4]$ for state cost, and $R = \text{diag}[1, 1, 1, 1]$ for input cost.

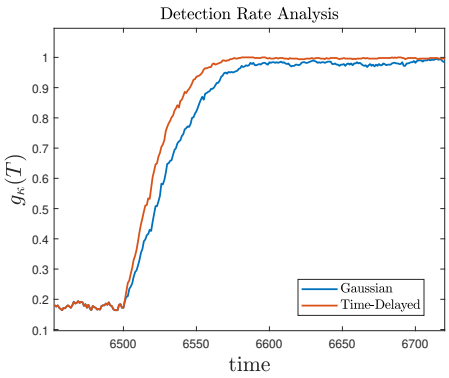

Fig. 3. Comparison of the detection rate of a replay attack between additive Gaussian and delay-induced feedback watermarking. Attack event starts at time $t' = 6500$.

We compare detection rates between the time-delayed watermarking, and an additive Gaussian watermarking signal. The parameters for the Gaussian watermarking signal are chosen to be zero mean with a stationary covariance of $\Sigma_{GW} = 0.015 I_{n_x}$. We ensure a fair comparison by making the detection rate comparable to that of the time-delayed watermarking. Figure 3 shows the results for the detection rate for the attacked system for both Gaussian watermarking and the time-delay watermarking. In the simulations, we record the output of the system from time $t_{start} = 6000$ to $t_{end} = 6300$, and the replay the output values starting from time $t' = 6500$. The steps change of the calculated χ^2 measure $g_\kappa(T)$ past t' highlights that some sort of attack is taking place. The time-delayed watermarking signal results in a faster and overall better detection rate. Finally, the system performance without watermarking (optimal) is 0.7907, with Additive Gaussian is 1.0415 and with Time-Delayed watermarking is 0.8712.

VIII. Discussion

We presented a new type of watermarked security in linear systems. We leveraged a type of random time-delayed feedback for detection of replay attacks. The theoretical framework poses several challenges and key points to pay attention to for the effective synthesis of watermarking signals. The working hypothesis that motivated this effort relies on the advantages of state-feedback control for fault and attack detection in the closed-loop system performance, over additive noise.

Appendix

$$\mathcal{Q} = \begin{bmatrix} \Sigma_W & O & O & O \\ O^T & \Sigma_V & O_{n_v} & O_{n_v} \\ O^T & O_{n_v} & \Sigma_V & \left(\sum_{\tau=1}^{\overline{\tau}} p_\tau^2\right)\Sigma_V \\ O^T & O_{n_v} & \left(\sum_{\tau=1}^{\overline{\tau}} p_\tau^2\right)\Sigma_V & \Sigma_V \end{bmatrix}$$

$$\mathcal{R}_{l,\tau,\tau'} = \begin{bmatrix} O_{n_w} & O & O & O \\ O^T & O_{n_v} & O_{n_v} & O_{n_v} \\ O^T & O_{n_v} & p_{l+\tau}\Sigma_V & p_{l+\tau}\Sigma_V \\ O^T & O_{n_v} & p_{l+\tau'}\Sigma_V & p_{l+\tau'}\Sigma_V \end{bmatrix}$$

$$\mathcal{S} = \begin{bmatrix} O_{n_w} & O & O & O \\ O^T & O_{n_v} & O_{n_v} & O_{n_v} \\ O^T & \Sigma_V & O_{n_v} & O_{n_v} \\ O^T & \Sigma_V & O_{n_v} & O_{n_v} \end{bmatrix}$$

with $O = O_{n_w \times n_v}$.

References

[1] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part i: Quantitative model-based methods," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 293–311, 2003.

[2] H. Liu, Y. Mo, and K. H. Johansson, "Active detection against replay attack: A survey on watermark design for cyber-physical systems," in *Safety, Security and Privacy for Cyber-Physical Systems*. Springer, 2021, pp. 145–171.

[3] A. B. Sharma, L. Golubchik, and R. Govindan, "Sensor faults: Detection methods and prevalence in real-world datasets," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 3, 2010.

[4] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *2008 The 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

[5] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "On the use of watermark-based schemes to detect cyber-physical attacks," *EURASIP Journal on Information Security*, vol. 2017, no. 1, pp. 1–25, 2017.

[6] R. K. Mehra and J. Peschon, "An innovations approach to fault detection and diagnosis in dynamic systems," *Automatica*, vol. 7, no. 5, pp. 637–640, 1971.

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

[8] K. Åström, *Introduction to Stochastic Control Theory*, ser. Dover Books on Electrical Engineering. Dover Publications, 2006.

[9] L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, ser. Addison-Wesley series in electrical and computer engineering. Addison-Wesley Publishing Company, 1991.

[10] J. Milošević, H. Sandberg, and K. H. Johansson, "Estimating the impact of cyber-attack strategies for stochastic networked control systems," *IEEE Transactions on Control of Network Systems*, vol. 7, no. 2, pp. 747–757, 2019.