A Bernoulli-Gaussian Physical Watermark for Detecting Integrity Attacks in Control Systems

Sean Weerakkody Omur Ozel Bruno Sinopoli

Abstract— **We examine the merit of Bernoulli packet drops in actively detecting integrity attacks on control systems. The aim is to detect an adversary who delivers fake sensor measurements to a system operator in order to conceal their effect on the plant. Physical watermarks, or noisy additive Gaussian inputs, have been previously used to detect several classes of integrity attacks in control systems. In this paper, we consider the analysis and design of Gaussian physical watermarks in the presence of packet drops at the control input. On one hand, this enables analysis in a more general network setting. On the other hand, we observe that in certain cases, Bernoulli packet drops can improve detection performance relative to a purely Gaussian watermark. This motivates the joint design of a Bernoulli-Gaussian watermark which incorporates both an additive Gaussian input and a Bernoulli drop process. We characterize the effect of such a watermark on system performance as well as attack detectability in two separate design scenarios. Here, we consider a correlation detector for attack recognition. We then propose efficiently solvable optimization problems to intelligently select parameters of the Gaussian input and the Bernoulli drop process while addressing security and performance trade-offs. Finally, we provide numerical results which illustrate that a watermark with packet drops can indeed outperform a Gaussian watermark.**

I. Introduction

The security of cyber-physical systems (CPS) has become a critical issue [1]. Since CPS such as the smart grid, waste management systems, water distribution systems, transportation systems, and smart buildings are linked to critical infrastructures, it is imperative that they operate securely. Unfortunately, attacks have occurred against CPS. This includes Stuxnet [2], which targeted uranium enrichment facilities in Iran, the Maroochy Shire incident [3], an attack by a malicious insider on a sewage management system, and the Ukraine power attack [4], a hack resulting in widespread blackouts in Ukraine. The threat does not appear to be over as the growing connectivity and heterogeneity of our system architectures provide new attack surfaces for adversaries.

We focus on detecting integrity attacks in control systems in the presence of packet drops at the control input. In an integrity attack, an adversary modifies inputs and sensor measurements in a control system. The goal of such an attacker may be to achieve some economic benefit or cause physical damage to a system. An attacker can potentially maximize his impact by hiding his presence from the operator. Remaining stealthy allows an attacker to affect the system for long periods of time without defender interference. The adversary can avoid detection by intelligently modifying the sensor measurements to fool detectors. One example is a replay attack, as used in Stuxnet, where an attacker replaces true outputs with a previously recorded sequence of measurements [5].

We consider the use of physical watermarking to detect integrity attacks. A physical watermark is a noisy Gaussian signal added on top of an optimal control input to authenticate a system's dynamics. Physical watermarking is a method of active detection, where a defender alters his strategy to recognize attacks. These methods are necessary when standard fault detection methods provably fail [6], [7]. Recent work has investigated physical watermarking. In [5], [8]–[10], the design of watermarks against replay attacks was examined. Additionally, [11] and [12] design asymptotic detectors in systems implementing physical watermarking to ensure zero additive distortion power is introduced into sensor measurements. Additionally, in a scalar setting, [13] demonstrates the optimality of Gaussian watermarks against Gaussian attackers and vice versa. [14] evaluates the use of non-stationary watermarks to hamper system identification. Finally, [15] considers watermarks to thwart adversaries who have access to a subset of inputs and model knowledge.

However, prior work fails to consider the scenario where there exists packet drops in the network. In this paper, we generalize the design of the Gaussian physical watermark by incorporating Bernoulli drops at the control inputs. This enables the operator to account for imperfect networks when designing a Gaussian watermark for secure detection. We also argue that using Bernoulli drops together with a Gaussian watermark can improve detection. This motivates the analysis and design of a joint Bernoulli-Gaussian watermark. In our preliminary work [16], we proposed using packet drops in a setting without Gaussian watermarks to detect replay attacks. This article extends these results by providing a rigorous mathematical setting to jointly design parameters of both a Gaussian watermark and Bernoulli drop process.

We investigate two types of watermark design: 1) a watermark with an independent and identically distributed (IID) Gaussian additive input multiplied by a Markovian Bernoulli drop process at the control input and 2) a watermark with a stationary Gaussian additive input generated by a hidden Markov model (HMM) multiplied by an IID Bernoulli drop

S Weerakkody, O. Ozel, and B. Sinopoli are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA 15213. Email: {sweerakk, oozel}@andrew.cmu.edu, brunos@ece.cmu.edu

S. Weerakkody is supported in part by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. The work by S. Weerakkody, O. Ozel, and B. Sinopoli is supported in part by the Department of Energy under Award Number DE-OE0000779 and by the National Science Foundation under award number CCF 1646526.

process at the control input. We incorporate a correlation detector [17], [8] to recognize integrity attacks and characterize adversarial scenarios where the Bernoulli-Gaussian watermark is provably effective. Next, we provide efficiently solvable optimization problems to design parameters of the Gaussian input and the Bernoulli drop process. Simulation results illustrate scenarios where packet drops improve detection performance relative to a purely Gaussian watermark.

II. System Model

We consider a discrete time LTI control system as follows

$$x_{k+1} = Ax_k + Bu_{k,c} + w_k, \quad y_k = Cx_k + v_k. \tag{1}$$

$x_k \in \mathbb{R}^n$ is the state vector at time k. A set of m sensor measurements $y_k \in \mathbb{R}^m$ is delivered to a supervisory control and data acquisition (SCADA) system at time k in order to perform remote estimation and compute an intended control input $u_k \in \mathbb{R}^p$. A set of p control inputs $u_{k,c} \in \mathbb{R}^p$ actuate the system. We differentiate $u_{k,c}$, the control input applied to the system, versus u_k, the input computed by a SCADA operator. We assume $w_k \sim \mathcal{N}(0, Q)$ is IID process noise and $v_k \sim \mathcal{N}(0, R)$ is IID measurement noise (independent of $\{w_k\}$), where $Q \succ 0, R \succ 0$. A Kalman filter performs state estimation as follows.

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_{k,c}, \ \hat{x}_{k|k} = \hat{x}_{k|k-1} + Kz_k, \tag{2}$$
$$K = PC^T(CPC^T + R)^{-1}, \ z_k = y_k - C\hat{x}_{k|k-1}, \tag{3}$$
$$P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T. \tag{4}$$

The defender minimizes a cost function J:

$$J = \lim_{N\to\infty} \frac{1}{2N+1}\mathbb{E}\left[\sum_{k=-N}^{N} x_k^T W x_k + u_{k,c}^T U u_{k,c}\right], \tag{5}$$

where $W \succ 0$ and $U \succ 0$. We assume (A, B) and $(A, Q^{\frac{1}{2}})$ are controllable and (A, C) and $(A, W^{\frac{1}{2}})$ are observable.

A. Control in Uncertain Networks

As shown in Fig. 1, the control input u_k may be dropped as it is sent from the SCADA system to the plant. Here,

$$u_{k,c} = \eta_k u_k, \tag{6}$$

where $\eta_k \in \{0, 1\}$ is a Bernoulli random variable. The control input u_k may be dropped due to network imperfections. In this case, we assume the operator receives an acknowledgement (ACK), which specifies if u_k was delivered. Alternatively, the input u_k may be intentionally dropped as a means to watermark the system, enabling the detection of integrity attacks that fail to preserve the effect of the drop process. This strategy was initially investigated in [16]. We consider both IID and Markovian drop processes.

1) IID Bernoulli Process: First, we assume $\{\eta_k\}$ is an IID Bernoulli process where $P(\eta_k = 1) = 1 - p_d$. LQG control with IID Bernoulli packet losses was studied in [18]. Consider the information set $\mathcal{F}_k \triangleq \{y_{-\infty:k}, \eta_{-\infty:k-1}, u_{-\infty:k-1}\}$. We suppose p_d is chosen (or given) so that the system (1) can have finite cost J. The

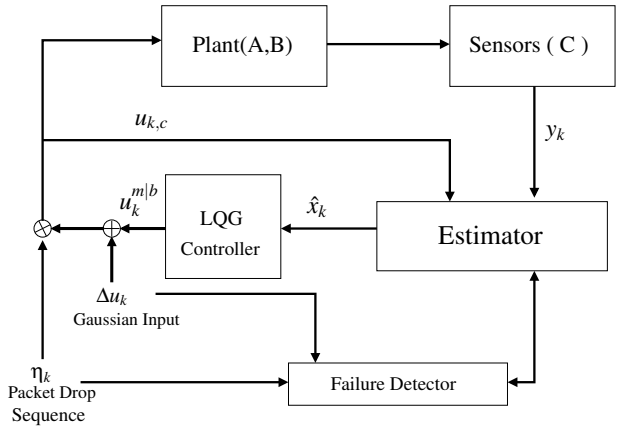


Fig. 1. System model

optimal control strategy at time k given $\mathcal{F}_k$ is as follows [16], [18]:

$$u_k^b = L_k\hat{x}_{k|k}, \ L_k = -(B^T S_{k+1}B + U)^{-1}B^T S_{k+1}A,$$
$$S_k = A^T S_{k+1}A + W + (1 - p_d)A^T S_{k+1}BL_k. \tag{7}$$

As we expect the system has been running for a long time, both L_k and S_k have converged to fixed point values so that

$$u_k^b = L_{(b)}\hat{x}_{k|k}, \ L_{(b)} = -(B^T S_{(b)}B + U)^{-1}B^T S_{(b)}A,$$
$$S_{(b)} = A^T S_{(b)}A + W + (1 - p_d)A^T S_{(b)}BL_{(b)}. \tag{8}$$

$J = J_{(b)}$ for this strategy where $J_{(b)}$ is

$$J_{(b)} = \text{tr}\left(S_{(b)}Q + (A^T S_{(b)}A + W - S_{(b)})(P - KCP)\right). \tag{9}$$

2) Markovian Bernoulli Process: In this setup, we assume there are Markovian packet losses [19] at the input where

$$\begin{bmatrix} P(\eta_{k+1} = 0|\eta_k = 0) & P(\eta_{k+1} = 1|\eta_k = 0) \\ P(\eta_{k+1} = 0|\eta_k = 1) & P(\eta_{k+1} = 1|\eta_k = 1) \end{bmatrix} = \begin{bmatrix} \bar{\alpha} & \alpha \\ \beta & \bar{\beta} \end{bmatrix} \tag{10}$$

and $\bar{\alpha} \triangleq 1 - \alpha$, $\bar{\beta} \triangleq 1 - \beta$. Here, we assume $0 < \alpha \leq 1$, $0 < \beta \leq 1$ so that η_k is irreducible. Moreover, we assume η_k is stationary, which can be obtained by letting its initial distribution be $P(\eta_{-\infty} = 0) = \frac{\beta}{\alpha+\beta}$. Finally, we assume that α and β are selected (or given) so that the system (1) can have finite cost J. The optimal strategy at time k given $\mathcal{F}_k$ is

$$u_k^m = L_{(m)}\hat{x}_{k|k}, \ L_{(m)} = -(B^T R_{(m)}B + U)^{-1}B^T R_{(m)}A,$$
$$R_m = A^T(\beta S_{(m)} + \bar{\beta}R_{(m)})A + W + \bar{\beta}A^T R_{(m)}BL_{(m)},$$
$$S_m = A^T(\bar{\alpha}S_{(m)} + \alpha R_{(m)})A + W + \alpha A^T R_{(m)}BL_{(m)},$$

where $L_{(m)}, R_{(m)}, S_{(m)}$ are parameters which converged to their steady state values. The resulting cost of control is

$$J = J_{(m)} = \frac{\text{tr}(\beta S_{(m)}Q + \alpha R_{(m)}Q)}{\alpha + \beta} \tag{11}$$
$$+ \frac{\text{tr}((A^T(\bar{\alpha}S_{(m)} + \alpha R_{(m)})A + W - S_{(m)})(P - KCP))}{\alpha + \beta}.$$

Remark 1: The prior strategies are optimal when the defender only has knowledge of the observed drop sequence

$\eta_{-\infty:k-1}$. However, if the drop sequence is intentionally introduced using a pseudo random number generator (PRNG), the defender knows future values of η_k. The design of a controller that uses this information is left for future work.

B. Joint Bernoulli-Gaussian Physical Watermarking

To account for adversarial behavior, we consider additive Gaussian physical watermarks Δu_k. First introduced in [5] to detect replay attacks, an additive Gaussian watermark can be leveraged to verify the freshness of outputs. We aim to intelligently combine the Gaussian watermarks considered in [8] and [9] with a Bernoulli drop process at the input. Such a design accomplishes two goals: 1) to expand the analysis of physical watermarking to a more realistic network setting with packet drops and 2) to potentially improve performance by considering a more general joint Bernoulli-Gaussian watermark.

We consider two main joint designs.

Watermark 1: IID Gaussian Input + Markovian Drops

$$u_{k,c} = \eta_k(u_k^m + \Delta u_k). \tag{12}$$

$\{\eta_k\}$ is a Markovian Bernoulli process and $\Delta u_k \sim \mathcal{N}(0, \mathcal{Q})$ is an IID Gaussian watermark [5]. We assume Δu_k is independent of other stochastic processes in the system.

Watermark 2: Stationary Gaussian Input + IID Drops

$$u_{k,c} = \eta_k(u_k^b + \Delta u_k). \tag{13}$$

In this case, $\{\eta_k\}$ is an IID Bernoulli process. The Gaussian input Δu_k is assumed to be a stationary process generated by a hidden Markov model (HMM) as considered in [9].

$$\zeta_{k+1} = A_\omega \zeta_k + \psi_k, \quad \Delta u_k = C_h \zeta_k. \tag{14}$$

ζ_k is the hidden state of the HMM, A_ω has spectral radius $\rho(A_\omega) \leq \bar{\rho} \leq 1$, and $\psi_k \sim \mathcal{N}(0, \Psi)$ is IID Gaussian noise. For stationarity, $\text{Cov}(\zeta_0) = A_\omega \text{Cov}(\zeta_0) A_\omega^T + \Psi$. Δu_k is independent of other stochastic processes in the system.

Remark 2: Here, $\bar{\rho}$, the maximum allowable spectral radius, is a design parameter for the defender. We observe a larger $\bar{\rho}$ improves expected detection performance. However, a larger $\bar{\rho}$ means a larger correlation between watermarks and this could facilitate the prediction of future watermarks if the attacker guesses an initial Gaussian input Δu_k.

III. Attack Model

In this section we describe a model of our adversary in terms of knowledge, capabilities, and potential strategies.

A. Attacker Capabilities

Without loss of generality, we assume an attack begins at time $k = 0$. We make the following assumptions.

1) The attacker can modify all measurements y_k, $k \geq 0$. The falsified outputs at time k are denoted by y_k^v.
2) The attacker inserts an input $B^a u_k^a$ into the system.
3) The attacker is unable to read the true control inputs $u_{k,c}$. As a result, he is unaware of the drop sequence $\{\eta_k\}$ and the Gaussian watermark $\{\Delta u_k\}$.

The system under attack is given by

$$x_{k+1} = Ax_k + Bu_{k,c} + B^a u_k^a + w_k, \tag{15}$$

$$\hat{x}_{k+1|k+1} = (I - KC)(A\hat{x}_{k|k} + Bu_{k,c}) + Ky_{k+1}^v. \tag{16}$$

Remark 3: Attackers can inject $B^a u_k^a$ by appropriating the defender's actuators or inserting their own. The attacker could possibly modify inputs without being able to read them if the inputs are encrypted. Alternatively, the attacker can cause damage even if $B^a u_k^a = 0$. For example, the attacker can destabilize the plant if A is open loop unstable.

B. Attack Strategy

The attacker generates y_k^v through a virtual system:

$$x_{k+1}^v = Ax_k^v + \eta_k^v B(L_{m|b}\hat{x}_{k|k}^v + \Delta u_k^v) + w_k^v, \tag{17}$$

$$\begin{aligned}\hat{x}_{k+1|k+1}^v = {} & (I - KC)(A + \eta_k^v B L_{m|b})\hat{x}_{k|k}^v + Ky_{k+1}^v \\ & + \eta_k^v (I - KC) B \Delta u_k^v,\end{aligned} \tag{18}$$

$$y_k^v = Cx_k^v + v_k^v. \tag{19}$$

In the case of Watermark 1, $L_{m|b} = L_{(m)}$, η_k^v follows a Markovian process (10) with parameters α and β and $\Delta u_k^v \sim \mathcal{N}(0, \mathcal{Q})$ is an IID Gaussian process. In the case of Watermark 2, $L_{m|b} = L_{(b)}$, η_k^v is an IID Bernoulli process with drop probability p_d and Δu_k^v is a stationary Gaussian process which satisfies (14). Additionally, $v_k^v \sim \mathcal{N}(0, R)$ and $w_k^v \sim \mathcal{N}(0, Q)$ are IID processes. Finally, we assume the stochastic processes $\{\eta_k^v, \Delta u_k^v, w_k^v, v_k^v\}$ are independent of the real system's stochastic parameters $\{\eta_k, \Delta u_k, w_k, v_k\}$.

The previous attack strategy can be generated (approximately) by a replay attack where the attacker records a long sequence of outputs $y_{-T':-T'+T}$ and, starting at time 0, replaces y_k with $y_k^v = y_{k-T'}$ for $0 \leq k \leq T$. Attackers who do not have precise knowledge of the model may engage in replay attacks, which only require access to the outputs [5], [8], [9]. Alternatively, this attack strategy can be constructed by an adversary who is familiar with the model, for instance a malicious insider. In this case, the attacker simulates a virtual copy of the system dynamics to fool a bad data detector. It was previously shown [9] that if $p_d = 0$ and there is no Gaussian watermark, the given strategies are asymptotically stealthy when $\mathcal{A} \triangleq (A + BL_{(b)})(I - KC)$ is Schur stable.

A model aware attacker could also potentially pursue an additive attack, for instance a false data injection attack [20] or a zero dynamics attack [21], [22]. In these attacks, the adversary injects an additive bias into the system which preserves the watermark and allows the attacker to remain stealthy. However, there are scenarios where additive attacks on sensor measurements are not feasible. As an example, suppose the defender uses public key cryptography, where a public key is used to encrypt the measurements while a private key is used to decrypt the associated cipher text. An attacker could send his own virtual measurements encrypted with the public key. However, such an attack could not leverage information in the true measurement as that would require access to the defender's private key to learn y_k. In this case, additive attacks constructed by replacing a true output

packet with a virtual packet would be infeasible. By assumption, an additive networked-based attack on the defender's control input is also impossible because the adversary is unable to read the defender's input.

We argue that alternative attack strategies which manipulate all sensors y_k in a setting with public key cryptography also fail due to the fact that the resulting attack sequence $\{y_k^v\}$ is independent of the watermarks $\{\Delta u_k, \eta_k\}$. Specifically, an attacker who is unable to read the inputs or outputs will have no information about the watermarks. As a result, the outputs he can construct will fail to fool the correlation detector, which we propose in the next section.

IV. A Correlation Detector

We consider a correlation detector, proposed in [8]. The defender computes a virtual output y'_k, which explicitly characterizes the effect of watermarks on y_k.

$$x'_{k+1} = Ax'_k + \eta_k B(L_{m|b}\hat{x}'_{k|k} + \Delta u_k), \quad y'_k = Cx'_k, \tag{20}$$

$$\hat{x}'_{k+1|k+1} = (I - KC)(A + \eta_k BL_{m|b})\hat{x}'_{k|k} + Ky'_{k+1} + \eta_k(I - KC)B\Delta u_k, \tag{21}$$

where with some abuse of notation $x'_{-\infty} = 0, \hat{x}'_{-\infty|-\infty} = 0$. We can simplify (20) and (21) to obtain

$$x'_{k+1} = (A + \eta_k BL_{m|b})x'_k + \eta_k B\Delta u_k, \quad y'_k = Cx'_k. \tag{22}$$

This virtual process created by the defender is driven entirely by the sequence of Bernoulli-Gaussian watermarks $\{\Delta u_k, \eta_k\}$. Thus, if we were to multiply the true outputs y_k with the defender's virtual outputs y'_k we would expect a positive correlation. However, if an attacker introduces measurements y_k^v, which are driven by an independent sequence of watermarks, the expected correlation drops to 0. This motivates consideration of the detection statistic $y_k^T y'_k$, where a large statistic is indicative of normal behavior while a small statistic indicates malicious behavior. Observe due to the random real time selection of watermarks, $\|y'_k\|_2$ may be close to 0, impacting detector performance since the correlation will likely also approach 0 even under normal operation. As a result, we propose an event triggered detector:

$$\begin{aligned} &\textbf{If } \|y'_k\|_2^2 \geq \mu \qquad \textit{Perform Detection} \\ &\kappa = \kappa + 1, \quad t_\kappa = k \\ &\sum_{j=\kappa-\mathcal{W}+1}^{\kappa} g_j \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \tau, \quad g_\kappa = y_{t_\kappa}^T y'_{t_\kappa}. \end{aligned} \tag{23}$$

The null hypothesis $\mathcal{H}_0$ is that the system is operating without malicious behavior while the alternative hypothesis $\mathcal{H}_1$ is that the system is under attack. $\mathcal{W}$ is the size of the detector's window. A detection event is triggered if $\|y'_k\|_2^2$ is greater than some user defined threshold μ, preventing false alarms from being raised when y'_k is small, while sacrificing time to detection. This tradeoff can be addressed by tuning μ. Note that κ corresponds to the time index of the event triggered correlation detector and increases at instants when a new detection statistic is computed. Identifying attacks on an individual sensor i can be done by focusing on the correlation between individual measurements. An appropriate statistic g_κ^i would be $y_{t_\kappa}^i {y_{t_\kappa}^i}'$ where $y_{t_\kappa}^i$ is the ith entry of y_{t_κ}.

Remark 4: A detector with an adaptive threshold could address issues of small y'_k. However, such a detector is more prone to misses, mistaking an attack for noise. Incorporation and analysis of such a detector is left for future work.

Remark 5: An adversary that can not read $\{u_k\}, \{y_k\}$ can not take advantage of instances when detection does not occur, because such instances are entirely dependent on the realization of previous watermarks. An attacker who is forced to act independently of the real time watermarking sequence cannot determine if a detection has been triggered.

We now verify that the expected correlation is 0, if the outputs y_k^v are generated independently of the watermarks.

Theorem 6: If y_k^v and $\{\Delta u_k, \eta_k\}$ are independent, then

$$\mathbb{E}\left[y_k^{vT} y'_k \middle| \ \|y'_k\|_2^2 \geq \mu \right] = 0.$$

Proof: Observe that y'_k can be written as a linear function of the Gaussian watermarks Δu_k so that

$$y'_k = \sum_{j=-\infty}^{k-1} G_j(\eta_{j:k-1})\Delta u_j, \tag{24}$$

where G_j is some linear gain, determined by the sequence of Bernoulli drops $\eta_{j:k-1}$. Thus, we have

$$\begin{aligned} \mathbb{E}[y_k^{vT} y'_k] &= \mathbb{E}\left[y_k^{vT} \sum_{j=-\infty}^{k-1} G_j(\eta_{j:k-1})\Delta u_j \middle| \ \|y'_k\|_2^2 \geq \mu \right] \\ &= \sum_{j=-\infty}^{k-1} \mathbb{E}\left[y_k^v\right]^T \mathbb{E}\left[G_j(\eta_{j:k-1})\Delta u_j \middle| \ \|y'_k\|_2^2 \geq \mu \right] = 0. \end{aligned}$$

■

The proposed detector can often differentiate between faulty and malicious scenarios. During a fault, we expect to see the effect of the embedded watermarks in the output and it could be measured through correlation. Alternatively, residue based detectors such as a χ^2 detector ($g_\kappa = -z_{t_\kappa}^T(CPC^T + R)^{-1}z_{t_\kappa}$), which measures the difference between measured and expected behavior, will likely raise an alarm during faulty behavior and malicious behavior. Both detectors can be used in tandem. A residue based detector can raise alarms in the case of faulty or malicious behavior, while a correlation detector can distinguish these events. In this article, we focus on the correlation detector.

V. The First Watermark Design

We consider the design of a watermark consisting of an IID Gaussian input and Markovian drops. This requires the evaluation of a detection and performance trade-off. We wish to maximize the correlation of y_k and y'_k to distinguish the system under attack from normal operation. However, we also need to ensure the system meets an adequate level of performance as characterized by the cost $\bar{J}$, starting at $k = 0$.

$$\bar{J} = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}\left[\sum_{k=0}^{N-1} x_k^T W x_k + u_{k,c}^T U u_{k,c} \right]. \tag{25}$$

As such, we design the parameters $\alpha, \beta, \mathcal{Q}$ by solving the following optimization problem

$$\begin{aligned} \underset{\alpha,\beta,\mathcal{Q}}{\text{maximize}} \quad & \lim_{k\to\infty} \mathbb{E}[y_k^T y_k' | \mathcal{H}_0] \\ \text{subject to} \quad & \bar{J} \leq \delta,\ 0 < \alpha, \beta \leq 1. \end{aligned} \tag{26}$$

To begin with, we use [19, Theorem 3] to analytically compute the cost $\bar{J}$ as follows.

Theorem 7: Suppose α and β are chosen so that the system has finite cost $J_{(m)}$ in the absence of a Gaussian watermark. The LQG cost $\bar{J}$ of the control system (1) with IID Gaussian and Markovian watermark (12) is:

$$\bar{J} = J_{(m)}(\alpha, \beta) + \frac{\alpha}{\alpha+\beta} \text{tr}\left((B^T R_{(m)} B + U)\mathcal{Q}\right). \tag{27}$$

Proof: Consider the cost to go in a finite horizon, $V_k(x_k) \triangleq \sum_{j=k}^{N} \mathbb{E}\left[x_j^T W x_j + u_{j,c}^T U u_{j,c} | \mathcal{F}_k\right]$, and let $u_{N,c} = 0$. Similar to, [19], it can be shown that

$$V_k(x_k) = \begin{cases} \mathbb{E}[x_k^T S_k x_k | \mathcal{F}_k] + c_k & (\eta_{k-1} = 0) \\ \mathbb{E}[x_k^T R_k x_k | \mathcal{F}_k] + d_k & (\eta_{k-1} = 1) \end{cases}, \tag{28}$$

where $c_N = d_N = 0,\ R_N, S_N = W, \bar{P} = P - KCP$ and

$$\begin{aligned} F &= (A + BL_{(m)}), \\ R_k &= W + \beta A^T S_{k+1} A + \bar{\beta} F^T R_{k+1} F + \bar{\beta} L_{(m)}^T U L_{(m)}, \\ S_k &= W + \bar{\alpha} A^T S_{k+1} A + \alpha F^T R_{k+1} F + \alpha L_{(m)}^T U L_{(m)}, \\ c_k &= -\alpha \text{tr}((F^T R_{k+1} F - A^T R_{k+1} A + L_{(m)}^T U L_{(m)})(\bar{P})) \\ &\quad + \alpha[\text{tr}(R_{k+1} Q) + d_{k+1} + \text{tr}((B^T R_{k+1} B + U)\mathcal{Q})] \\ &\quad + \bar{\alpha}[\text{tr}(S_{k+1} Q) + c_{k+1}], \end{aligned} \tag{29}$$

$$\begin{aligned} d_k &= -\bar{\beta} \text{tr}((F^T R_{k+1} F - A^T R_{k+1} A + L_{(m)}^T U L_{(m)})(\bar{P})) \\ &\quad + \bar{\beta}[\text{tr}(R_{k+1} Q) + d_{k+1} + \text{tr}((B^T R_{k+1} B + U)\mathcal{Q})] \\ &\quad + \beta[\text{tr}(S_{k+1} Q) + c_{k+1}]. \end{aligned} \tag{30}$$

Let $\bar{J}_N = \mathbb{E}\left[\sum_{k=0}^{N} x_k^T W x_k + u_{k,c}^T U u_{k,c}\right] = \mathbb{E}[V_0(x_0)]$. We find that

$$\begin{aligned} \bar{J}_N &= P(\eta_{-1} = 0)\left(\mathbb{E}[x_0^T S_0 x_0 | \eta_{-1} = 0] + c_0\right) \\ &\quad + P(\eta_{-1} = 1)\left(\mathbb{E}[x_0^T R_0 x_0 | \eta_{-1} = 1] + d_0\right). \end{aligned}$$

Leveraging the fact that $\{\eta_k\}$ is stationary with $P(\eta_k = 0) = \frac{\beta}{\alpha+\beta}$ as well as (29) and (30), we obtain

$$\begin{aligned} \bar{J}_N = \frac{1}{\alpha+\beta} \sum_{k=0}^{N-1} \Bigg(& -\alpha \text{tr}((F^T R_{k+1} F - A^T R_{k+1} A \\ & + L_{(m)}^T U L_{(m)})(\bar{P})) + \text{tr}((\beta S_{k+1} + \alpha R_{k+1}) Q) \\ & + \alpha \text{tr}((B^T R_{k+1} B + U)\mathcal{Q})\Bigg) \\ & + \frac{\beta \mathbb{E}[x_0^T S_0 x_0 | \eta_{-1}^0] + \alpha \mathbb{E}[x_0^T R_0 x_0 | \eta_{-1}^1]}{\alpha+\beta}, \end{aligned}$$

where η_{-1}^j refers to the condition $\eta_{-1} = j$. It can be shown (in a similar manner to the proof of Theorem 8) that the last term is bounded. Note $\bar{J} = \lim_{N\to\infty} \frac{1}{N} \bar{J}_{N-1}$. Moreover, from [19][Theorem 3, Lemma 4], $\{S_k\}, \{R_k\}$ converge to $S_{(m)}, R_{(m)}$, respectively. This proves the desired result. ■

We now compute the expected correlation without attacks.

Theorem 8: Suppose α and β are chosen so the resulting system has finite cost $J_{(m)}$ [19][Theorem 3] in the absence of a Gaussian watermark. Then, for the control system (1) with IID Gaussian and Markovian watermark (12), we have

$$\lim_{k\to\infty} \mathbb{E}[y_k^T y_k' | \mathcal{H}_0] = \frac{\text{tr}(C(\alpha X_1 + \beta X_0) C^T)}{\alpha+\beta}, \tag{31}$$

where

$$X_0 = A(\bar{\alpha} X_0 + \alpha X_1) A^T, \tag{32}$$
$$X_1 = (A + BL_{(m)})(\beta X_0 + \bar{\beta} X_1)(A + BL_{(m)})^T + B\mathcal{Q}B^T$$

Proof: We begin with the Lemma below.

Lemma 9: $\forall\ M \in \mathbb{R}^{2n\times n}$, $\lim_{k\to\infty} \mathcal{L}_0^k(M) = 0$ where,

$$\mathcal{L}_0 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{bmatrix} A(\bar{\alpha} X + \alpha Y) A^T \\ (A + BL_{(m)})(\beta X + \bar{\beta} Y)(A + BL_{(m)})^T \end{bmatrix}.$$

The proof is found in an extended version of this article [23]. The closed loop dynamics are

$$\begin{aligned} x_{k+1} &= (A + \eta_k B L_{(m)}) x_k - \eta_k B L_{(m)} e_k + w_k + \eta_k B \Delta u_k \\ e_{k+1} &= (A - KCA) e_k + (I - KC) w_k - K v_{k+1}, \end{aligned}$$

where $e_k = x_k - \hat{x}_{k|k}$. From (22), when $\eta_k = 1$. we obtain

$$\begin{aligned} &\mathbb{E}[x_{k+1}' x_{k+1}^T | \eta_k = 1] \\ &= (A + BL_{(m)}) \mathbb{E}[x_k' x_k^T | \eta_k = 1](A + BL_{(m)})^T - \\ &(A + BL_{(m)})(\mathbb{E}[x_k' e_k^T | \eta_k = 1] L_{(m)}^T B^T - \mathbb{E}[x_k' w_k^T | \eta_k = 1]) \\ &+ (A + BL_{(m)}) \mathbb{E}[x_k' \Delta u_k^T | \eta_k = 1] B^T \\ &+ B \mathbb{E}[\Delta u_k x_k^T | \eta_k = 1](A + BL_{(m)})^T \\ &+ B\left(\mathbb{E}[\Delta u_k w_k^T | \eta_k = 1] + \mathbb{E}[\Delta u_k \Delta u_k^T | \eta_k = 1] B^T\right) \\ &- B \mathbb{E}[\Delta u_k e_k^T | \eta_k = 1](BL_{(m)})^T, \end{aligned}$$

where we implicitly condition on $\mathcal{H}_0$. x_k' is independent of $\Delta u_k, w_k, e_k$ and Δu_k is independent of x_k, w_k, e_k. Thus,

$$\begin{aligned} &\mathbb{E}[x_{k+1}' x_{k+1}^T | \eta_k = 1] \\ &= (A + BL_{(m)}) \mathbb{E}[x_k' x_k^T | \eta_k = 1](A + BL_{(m)})^T + B\mathcal{Q}B^T. \end{aligned} \tag{33}$$

Next, since the Markov process is stationary and x_k, x_k' and η_k are conditionally independent given η_{k-1}, we observe

$$\begin{aligned} &\mathbb{E}[x_k' x_k^T | \eta_k = 1] \\ &= P(\eta_{k-1} = 1 | \eta_k = 1) \mathbb{E}[x_k' x_k^T | \eta_k = 1, \eta_{k-1} = 1] \\ &+ P(\eta_{k-1} = 0 | \eta_k = 1) \mathbb{E}[x_k' x_k^T | \eta_k = 1, \eta_{k-1} = 0], \\ &= \bar{\beta} \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 1] + \beta \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 0]. \end{aligned} \tag{34}$$

It can be similarly shown that

$$\mathbb{E}[x_{k+1}' x_{k+1}^T | \eta_k = 0] = A \mathbb{E}[x_k' x_k^T | \eta_k = 0] A^T. \tag{35}$$

$$\begin{aligned} &\mathbb{E}[x_k' x_k^T | \eta_k = 0] \\ &= \alpha \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 1] + \bar{\alpha} \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 0]. \end{aligned} \tag{36}$$

Letting $X_{k,j} = \mathbb{E}[x_k' x_k^T | \eta_{k-1} = j]$ we have

$$\begin{pmatrix} X_{k+1,0} \\ X_{k+1,1} \end{pmatrix} = \mathcal{L}_0 \begin{pmatrix} X_{k,0} \\ X_{k,1} \end{pmatrix} + \begin{bmatrix} 0 \\ B\mathcal{Q}B^T \end{bmatrix}. \tag{37}$$

Since $\mathcal{L}_0$ is stable, $\lim_{k\to\infty} \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 0]$ and $\lim_{k\to\infty} \mathbb{E}[x_k' x_k^T | \eta_{k-1} = 1]$ are obtained by solving a fixed

point equation which has a unique solution X_0 and X_1. (32) immediately follows from (37). Next, we find that

$$\lim_{k\to\infty} \mathbb{E}[x_k' x_k^T] = P(\eta_{k-1}=1)X_1 + P(\eta_{k-1}=0)X_0, \quad (38)$$
$$= \frac{\alpha X_1 + \beta X_0}{\alpha+\beta}.$$

Finally, we observe that

$$\mathbb{E}[y_k^T y_k'] = \text{tr}\left(\mathbb{E}[(y_k' y_k^T)]\right) = \text{tr}\left(C\mathbb{E}[x_k' x_k^T]C^T\right). \quad (39)$$

■

Thus, the watermark design problem (26) is given by

$$\begin{aligned} \underset{\alpha,\beta,\mathcal{Q}}{\text{maximize}} \quad & \frac{\text{tr}(C(\alpha X_1 + \beta X_0)C^T)}{\alpha+\beta} \\ \text{subject to} \quad & \begin{pmatrix} X_0 \\ X_1 \end{pmatrix} = \mathcal{L}_0 \begin{pmatrix} X_0 \\ X_1 \end{pmatrix} + \begin{bmatrix} 0 \\ B\mathcal{Q}B^T \end{bmatrix}, \\ & J_{(m)}(\alpha,\beta) + \text{tr}((B^T R_{(m)} B + U)\mathcal{Q}) \le \delta, \\ & 0 < \alpha, \beta \le 1. \end{aligned} \quad (40)$$

For fixed α and β, the problem is an efficiently solvable semidefinite program. However, to optimize over α and β, we have to solve multiple instances of the problem over a finite 2 dimensional space. Ideally a designer will sample the space sufficiently. Note, not all (α, β) in $(0,\ 1] \times (0,\ 1]$ are feasible as some selections of α and β lead to unbounded cost. Likewise, there may be naturally occurring drops which constrain α and β. For instance, if we add an artificial Markovian drop process on top of a naturally occurring IID drop process with drop probability p_d, we know that $\alpha \le (1-p_d), \bar{\beta} \le (1-p_d)$.

Remark 10: The optimal design of Watermark 1 requires solving multiple instances of a convex optimization problem with parameters varying over a bounded 2 dimensional space. This will also be true for Watermark 2. A formulation that considers a stationary Gaussian input with a Markovian drop process is nontrivial. Even if analysis can be performed, optimal design will likely require searching over 3 dimensions. This more complicated case is left for future work.

VI. The Second Watermark Design

We now investigate a watermark consisting of stationary Gaussian noise generated by a HMM (14) and an IID Bernoulli drop process at the control input with drop probability equal to p_d. Again, we design a watermark to address a performance and security trade-off. We wish to solve:

$$\begin{aligned} \underset{p_d, A_\omega, C_h, \Psi}{\text{maximize}} \quad & \lim_{k\to\infty} \mathbb{E}[y_k^T y_k' | \mathcal{H}_0] \\ \text{subject to} \quad & \bar{J} \le \delta, \quad \rho(A_\omega) \le \bar{\rho}, \\ & 0 \le p_d \le 1. \end{aligned} \quad (41)$$

Rather than optimizing over the parameters of the HMM, we instead optimize over the autocovariance functions $\Gamma(d) \triangleq \mathbb{E}[\Delta u_k \Delta u_{k+d}^T]$. For tractable analysis we replace the constraint $\rho(A_\omega) \le \bar{\rho}$ with the following related assumption.
Assumption 1: Let $\Gamma(d)$ be an autocovariance function for a Gaussian process generated by an HMM (A_ω, C_h, Ψ). Then $(A_\omega, C_h, \Psi, \bar{\rho})$ is feasible only if $\tilde{\Gamma}(d) \triangleq \bar{\rho}^{-|d|}\Gamma(d)$ is a autocovariance function of a stationary Gaussian process.

$\tilde{\Gamma}(d)$ can be potentially realized by an alternate HMM

$$\tilde{\zeta}_{k+1} = (A_\omega/\bar{\rho})\tilde{\zeta}_k + \tilde{\psi}_k, \quad \Delta\tilde{u}_k = C_h\tilde{\zeta}_k, \quad (42)$$
$$\text{Cov}(\tilde{\zeta}_0) = A_\omega \text{Cov}(\tilde{\zeta}_0) A_\omega^T + \Psi, \quad (43)$$
$$\tilde{\psi}_k \sim \mathcal{N}(0, \text{Cov}(\tilde{\zeta}_0) - A_\omega \text{Cov}(\tilde{\zeta}_0) A_\omega^T/\bar{\rho}^2). \quad (44)$$

Note, that if $\rho(A_\omega) > \bar{\rho}$, (42) can not be a stationary process. This HMM can be realized if and only if $\text{Cov}(\tilde{\zeta}_0) - A_\omega \text{Cov}(\tilde{\zeta}_0) A_\omega^T/\bar{\rho}^2$ is positive semidefinite. Intuitively, if $\rho(A_\omega)$ is marginally less than $\bar{\rho}$, there is a larger chance that $\text{Cov}(\tilde{\zeta}_0) - A_\omega \text{Cov}(\tilde{\zeta}_0) A_\omega^T/\bar{\rho}^2$ is positive semidefinite.

Remark 11: When $\bar{\rho} = 1$, Assumption 1, introduces no relaxation. In fact, the resulting formulation optimizes all stationary Gaussian processes in general. However, in the case $\bar{\rho} = 1$, we will prove that the resulting Gaussian process $\{\Delta u_k\}$ is entirely deterministic except for the initial watermark. A lower parameter $\bar{\rho}$ reduces average performance, but prevents an attacker who learns or guesses the current hidden state from adequately predicting future watermarks.
We arrive at a relaxed formulation to (41) below.

Theorem 12: Consider the control system (1) with IID Bernoulli and stationary Gaussian watermark (14). Suppose p_d is chosen so that the system has finite cost $J_{(b)}$ [19][Theorem 3] in the absence of a Gaussian watermark. An equivalent formulation to (41) after replacing the constraint $\rho(A_\omega) \le \bar{\rho}$ with Assumption 1 is given by

$$\begin{aligned} \underset{\omega, H, p_d}{\text{maximize}} \quad & \text{tr}(CF_2(\omega, H, p_d)C^T) \\ \text{subject to} \quad & J_{(b)}(p_d) + F_1(\omega, H, p_d) \le \delta, \\ & 0 \le p_d \le 1, \quad 0 \le \omega \le 0.5, \\ & H \in \mathbb{C}^{p\times p}, \quad H \succeq 0. \end{aligned} \quad (45)$$

where

$$\begin{aligned} F_2(\omega, H, p_d) &= 2\text{Re}\left(2\text{sym}\left[L_1(M_2 H B^T)\right] + L_1(BHB^T)\right) \\ F_1(\omega, H, p_d) &= \text{tr}(U\Theta) + \text{tr}((W + \bar{p}_d L_{(b)}^T U L_{(b)})F_2), \\ \Theta(\omega, H, p_d) &= 2\text{Re}\left(2\text{sym}\left[\bar{p}_d M_1 H\right] + \bar{p}_d H\right), \\ M_2 &= \bar{p}_d \bar{\rho} s(A + BL_{(b)})\left[I - s\bar{\rho}(A + \bar{p}_d B L_{(b)})\right]^{-1} B, \\ M_1 &= \bar{p}_d \bar{\rho} s L_{(b)}\left[I - s\bar{\rho}(A + \bar{p}_d B L_{(b)})\right]^{-1} B, \\ L_1(X) &= \bar{p}_d\left((A + BL_{(b)})L_1(X)(A + BL_{(b)})^T + X\right) \\ &\quad + p_d A L_1(X) A^T, \\ \text{sym}(X) &= \frac{X + X^T}{2}, \quad s = \exp(2\pi j\omega), \quad \bar{p}_d = 1 - p_d. \end{aligned}$$

There is also an optimal solution (H_*, ω_*, p_{d*}) such that $H_* = hh^H$ where h^H denotes the conjugate transpose of $h \in \mathbb{C}^p$. Letting Re and Im be the real and imaginary parts of a matrix/vector, respectively, an optimal A_ω, C_h, Ψ is

$$A_\omega = \bar{\rho}\begin{bmatrix} \cos(2\pi\omega_*) & -\sin(2\pi\omega_*) \\ \sin(2\pi\omega_*) & \cos(2\pi\omega_*) \end{bmatrix},$$
$$C_h = \sqrt{2}\begin{bmatrix} \text{Re}(h) & \text{Im}(h) \end{bmatrix}, \quad \Psi = (1-\bar{\rho}^2)I. \quad (46)$$

The proof is similar in nature to the proof of Theorem 6 in [9]. A sketch is found in [23]. For fixed p_d and ω,

the proposed problem is an efficiently solvable semidefinite program. To approximate a global maximum, we solve the problem repeatedly over the space $0 \leq \omega \leq 0.5$ and $0 \leq p_d \leq 1$. For sufficiently large p_d, the cost $\bar{J}$ becomes infinite in open loop unstable systems [18], limiting the feasible space. We can account for natural packet drops in the system as before. For instance, if the input is dropped naturally with probability p'_d, we have $p'_d \leq p_d \leq 1$.

Remark 13: An optimal watermark for a given $p_d \neq p_{d*}$ may have better detection performance than the globally optimal watermark. Future work aims to use objective functions that better highlight the relative performance of watermarks.

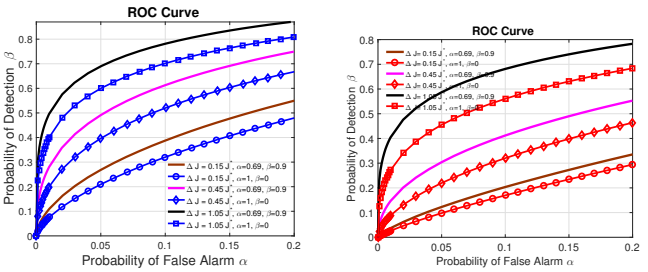
Remark 14: While packet drops at the sensor measurements are not modeled in this paper, our framework could be extended to address this behavior without significantly changing the formulations of the proposed optimization problems. The main effect of packet drops at the sensor side is a time varying Kalman gain. The objective function and increase in cost $\bar{J}$ due to the Gaussian portion of the watermark are not affected by time variations in the Kalman gain in both watermarking settings. Both $J_{(m)}$ and $J_{(b)}$ can be empirically evaluated for fixed (α, β) and p_d, respectively, to account for packet drops at the sensor measurements.

VII. Simulations

In this section, we illustrate the performance of the proposed watermarking designs through extensive numerical results. We tested our watermark designs in various randomly generated systems and, unless otherwise stated, averaged results over 1500 trials. Replay attacks are considered.

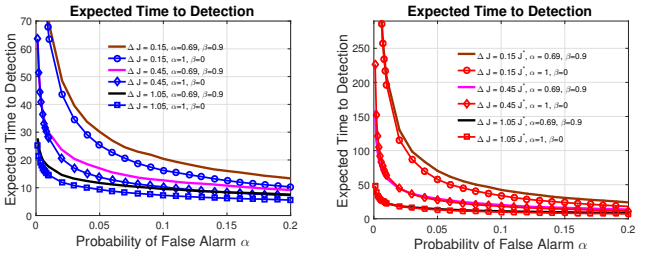
In Fig. 2, we utilize Watermark 1, which has a Markovian drop process defined by parameters (α, β) and an IID Gaussian watermark. The watermark is tested on a randomly generated open loop stable system with 5 states, 4 inputs, and 2 outputs. We plot the receiver operating characteristic (ROC) curve for both the proposed correlation detector and a χ^2 detector. The χ^2 detector serves as a benchmark, having been previously used for attack detection [5], [8], [10], [16] in watermarked systems. The threshold μ is chosen to be a constant multiple of $\lim_{k\to\infty} E[y_k^T y'_k]$. The ROC curves are collected at multiple different costs $\Delta J = 1.05J^*$, $\Delta J = 0.45J^*$ and $\Delta J = 0.15J^*$. Here, ΔJ represent the increase in the cost $\bar{J}$ relative to optimal cost J^* without drops or a Gaussian watermark. We compare a system with drops $(\alpha = 0.69, \beta = 0.9)$ to a system without drops $(\alpha = 1, \beta = 0)$. The proposed detector outperforms the χ^2 detector in all cases and packet drops improve the ROC curve for both detectors. The improvement appears to be higher for moderately valued ΔJ before saturating. In Fig. 3, we plot the expected time to detection for both detectors in a system with Watermark 1. The packet drop process introduces an additional delay in the time to detection though this additional time is less significant as ΔJ is increased.

In Fig. 4, we introduce Watermark 2, which has IID drops (with probability of drop p_d) and a stationary Gaussian watermark. The watermark is added to a randomly generated open loop stable system with 6 states, 5 inputs, and 5 outputs.



(a) Correlation Detector (b) χ^2 Detector

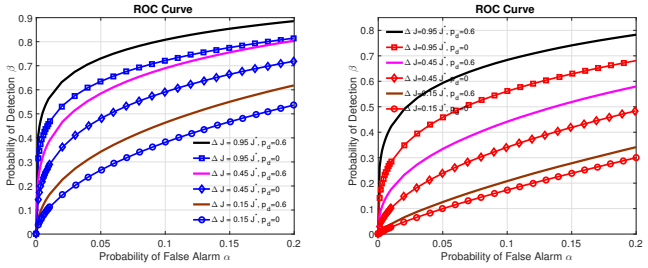
Fig. 2. Detection probability versus false alarm rate for χ^2 and correlation detectors for a system using Watermark 1.



(a) Correlation Detector (b) χ^2 Detector

Fig. 3. Expected time to detection for χ^2 and correlation detectors for a system using Watermark 1.

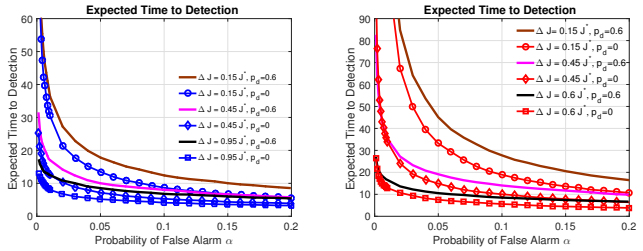
We plot ROC curves generated by both the correlation detector and χ^2 detector for a system with drops $(p_d = 0.6)$ and a system without drops $(p_d = 0)$, at various costs of control $\Delta J = 0.95J^*$, $\Delta J = 0.45J^*$ and $\Delta J = 0.15J^*$. Time to detection plots are provided in Fig. 5. The results and patterns observed here are similar to the results seen in the system with Watermark 1.



(a) Correlation Detector (b) χ^2 Detector

Fig. 4. Detection probability versus false alarm rate for χ^2 and correlation detectors for a system using Watermark 2.

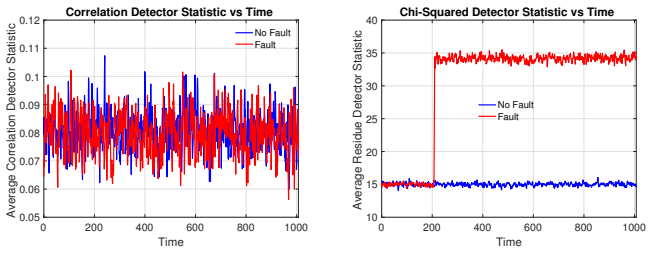
In Figs. 6 and 7, we plot χ^2 detector and correlation detector statistics (averaged over 500 trials) during a fault in the system. The fault introduced (at time 210) is a constant additive bias added to a subset of sensors (i.e. due to disturbances/sensor drift). While the χ^2 detector raises an alarm, the correlation detector does not since the watermark is preserved in the system. This motivates the use of both the correlation and χ^2 detector to distinguish faults from attacks. If both detectors raise an alarm, indicating the watermark is absent in the outputs, we consider a likely attack scenario.



(a) Correlation Detector (b) χ^2 Detector

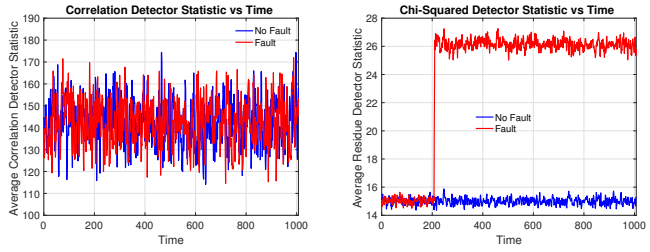
Fig. 5. Expected time to detection for χ^2 and correlation detectors for a system using Watermark 2.

If only the χ^2 detector raises an alarm, we expect that the watermark is preserved while the dynamics are inconsistent with modeling. As such, we anticipate a fault.



(a) Correlation Detector (b) χ^2 Detector

Fig. 6. Average correlation detector and χ^2 detector statistics under a fault at the sensor output for a system using Watermark 1.



(a) Correlation Detector (b) χ^2 Detector

Fig. 7. Average correlation detector and χ^2 detector statistics under a fault at the sensor output for a system using Watermark 2.

VIII. Conclusion and Future Work

In this paper, we showed how to incorporate Bernoulli packet drops at the control input in the design of physical watermarks. We argued that packet drops can be beneficial for detection and consequently considered the design of a joint Bernoulli-Gaussian watermark to detect integrity attacks. We proposed two main watermark designs in conjunction with a correlation detector and provided efficiently solvable optimization problems to address the trade-off between detection and control performances. In future work we aim to generalize our watermarking approach to allow us to drop either the entire control input or the Gaussian portion of the watermark. We also hope to conduct testing in real systems.

References

[1] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Secure Control: Towards Survivable Cyber-Physical Systems," in *Distributed Computing Systems Workshops, 2008. ICDCS '08. 28th International Conference on DOI - 10.1109/ICDCS.Workshops.2008.40*. IEEE, 2008, pp. 495–500.

[2] T. M. Chen, "Stuxnet, the real start of cyber warfare? [editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[3] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," in *Critical Infrastructure Protection*. Springer US, 2008, pp. 73–82.

[4] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 ukraine blackout: Implications for false data injection attacks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3317–3318, 2017.

[5] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference on Communication, Control, and Computing*, Sept 2009, pp. 911–918.

[6] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *American Control Conference (ACC), 2015*. IEEE, 2015, pp. 195–200.

[7] S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta, "Information flow for security in control systems," in *55th IEEE Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 5065–5072.

[8] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[9] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93 – 109, 2015.

[10] F. Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2013, pp. 1854–1859.

[11] B. Satchidanandan and P. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2017.

[12] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general LTI systems," *arXiv preprint arXiv:1703.07760*, 2017.

[13] M. Hosseini, T. Tanaka, and V. Gupta, "Designing optimal watermark signal for a stealthy attacker," in *2016 European Control Conference (ECC)*. IEEE, 2016, pp. 2258–2262.

[14] 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, 2017.

[15] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *53rd IEEE Conference on Decision and Control (CDC)*, Los Angeles, California, 2014, pp. 3757–3764.

[16] O. Ozel, S. Weerakkody, and B. Sinopoli, "Physical watermarking for securing cyber-physical systems via packet drop injections," in *To appear, 8th IEEE International Conference on Smart Grid Communications*, 2017.

[17] R. Chabukswar, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," in *18th IFAC World Congress*, Milan, Italy, Aug 2011, pp. 11 239–11 244.

[18] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of control and estimation over lossy networks," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[19] Y. Mo, E. Garone, and B. Sinopoli, "LQG control with Markovian packet loss," in *Control Conference (ECC), 2013 European*. IEEE, 2013, pp. 2380–2385.

[20] Y. Mo and B. Sinopoli, "False data injection attacks in cyber physical systems," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, April 2010.

[21] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[22] A. Teixeira, D. Perez, H. Sandberg, and K. H. Johannson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*, Beijing, China, 2012, pp. 55–64.

[23] S. Weerakkody, O. Ozel, and B. Sinopoli, "A Bernoulli-Gaussian physical watermark for detecting integrity attacks in control systems," *arXiv preprint arXiv:1710.01105*, 2017.