Physical Watermarking for Securing Cyber Physical Systems via Packet Drop Injections

Omur Ozel Sean Weerakkody Bruno Sinopoli
Department of Electrical and Computer Engineering
Carnegie Mellon University, Pittsburgh, PA USA
{*oozel, sweerakk, brunos*}*@andrew.cmu.edu*

***Abstract*—Physical watermarking is a well known solution for detecting integrity attacks on Cyber-Physical Systems (CPSs) such as the smart grid. Here, a random control input is injected into the system in order to authenticate physical dynamics and sensors which may have been corrupted by adversaries. Packet drops may naturally occur in a CPS due to network imperfections. To our knowledge, previous work has not considered the role of packet drops in detecting integrity attacks. In this paper, we investigate the merit of injecting Bernoulli packet drops into the control inputs sent to actuators as a new physical watermarking scheme. With the classical linear quadratic objective function and an independent and identically distributed packet drop injection sequence, we study the effect of packet drops on meeting security and control objectives. Our results indicate that the packet drops could act as a potential physical watermark for attack detection in CPSs.**

I. Introduction

Cyber-Physical Systems (CPSs) such as the smart grid are complex engineering systems that could involve any combination of sensing, processing, networking and control functionalities. A major consideration in the design of CPSs is security. There have been many types of attacks considered in the context of cyber-physical security. In this paper, we focus on integrity attacks where malicious agents inject inputs at sensors and actuators. In particular, we consider replay attacks and focus on detecting them. Classical methods to detect anomalies in the system use passive detection techniques where the defender uses finely tuned algorithms to make a decision about the health of the system. However, they are ineffective against stealthy adversaries who can construct attacks that produce viable sensor measurements. In this paper, we consider an active detection method called physical watermarking.

To verify the health of a CPS, the components and dynamics of the system must be authenticated. The authentication has to be performed not only in the cyber world, but also within the framework of the physical dynamics. In the cyber space, this is enabled by cryptographic tools; however, such tools may be vulnerable to attacks. Therefore, the extra security dimension in the physical dynamics plays a crucial role in reinforcing security. Physical watermarking is a well known solution to authenticate the correct operation of a control system [1]–[7]. In physical watermarking a randomly generated input or watermark that is known to legitimate parties is injected into the physical system. It is expected that this input can be traced in the measurement of the true output. If an attacker is unaware of this physical watermark, (s)he cannot adequately emulate the system as it is not possible to consistently generate the component of the output associated with this known random input. In this paper, we investigate the role of packet drop injections as a new physical watermarking scheme.

Detecting and acting against attacks on CPSs have been active topics of research. In particular, stealthy attacks such as false data injection attacks [8], [9], zero dynamics attacks [10]–[12], and replay attacks [13] are among the many considered in the literature. Pasqualetti et al.'s work in [11] introduces a general continuous-time control system where an adversary can insert arbitrary errors to an unknown subset of sensors and actuators with particular applications to smart grid scenarios. In [12], the authors investigate the problem of robust control and estimation in the presence of an adversary that is capable of inserting arbitrary errors in sensor measurements. Bad data detection techniques, such as the largest residue test [14], have long been used for systems with a static model including the smart grid. Reference [8], on the other hand, considers false data injection attacks where the attacker is aware of the grid's configuration and can inject a stealthy input into the measurements that lies in the range space of the observation matrix to change state estimation. In reference [15], false data injection attacks are considered on the grid from the system operator's point of view.

Prior works do not consider the probable scenario where there are packet drops in the communication channels. Packet drops occur naturally in the context of smart grid communications. In particular, both command and measurement channels could be subjected to packet drops due to, e.g., imperfections at the wireless and/or wired communication networks [16], [17]. Packet drops at the command and measurement channels change the system dynamics in a specific form, see e.g. [18], [19]. In this paper, we view the packet drops as a means to create watermarked dynamics and we explore the possibility to authenticate the system via intentional packet drop injections. We allow the controller to inject independent and identically distributed packet drops at the channel to the actuators with certain probability. Such a mechanism is easy to implement using, e.g., switches and pulses and they are applicable for a wide range of applications. The motivation to inject packet

This work is supported by the Department of Energy grant DE-OE0000779 and by the National Science Foundation grant CCF 1646526. S. Weerakkody is also supported in part by a Department of Defense National Defense Science & Engineering Graduate Fellowship.

drops at the controller is to obtain authenticated dynamics that enable detecting adversarial actions with higher probability. We investigate the feasibility of replay attacks under packet drop injections. We provide extensive numerical results for attack detection performance in systems enhanced by our watermarked dynamics. In particular, as applied to smart grids, we investigate the use of a correlation based detector along with a Bernoulli pulse watermark to detect attacks on automatic power generation control.

The following is the structure we use in the rest of the paper. In section II, we provide details of the system and attack models. In section IV, we elucidate necessary details about LQG control in a control loop with packet drops. In section V, we analyze the role of packet drop injections as a watermarking scheme under replay attacks. In section VI, we provide extensive numerical results for the real life operation and performance of physical watermarking via packet drop injections. In section VII, we conclude the paper.

As notation, we use $x_{t_1:t_2}$ to refer to the set $\{x_{t_1}, x_{t_1+1}, \cdots, x_{t_2}\}$. X^T is the transpose of X. $\{a_k\}$ is used to denote a sequence.

II. System and Attack Models

A. System Model

We model the system using discrete time linear time invariant (LTI) dynamics with packet drops at the control channel as shown in Fig. 1:

$$x_{k+1} = Ax_k + \eta_k B u_k + w_k, \tag{1}$$

$$y_k = Cx_k + v_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$ is the state vector at time k, $u_k \in \mathbb{R}^p$ is the control input at time k, and $y_k \in \mathbb{R}^m$ denotes sensor measurements taken at time k. Moreover, $\eta_k \in \{0, 1\}$ is an independent identically distributed (IID) packet drop process generated at the controller and known at the actuator and the estimator. Here, $\eta_k = 0$ indicates a packet drop and $\mathrm{P}(\eta_k = 0) = p_d$ is the packet drop probability. In the model, $w_k \sim \mathcal{N}(0, Q)$ is IID process noise and $v_k \sim \mathcal{N}(0, R)$ is IID measurement noise. We assume that (A, C) is detectable. Moreover, (A, B) and $(A, Q^{\frac{1}{2}})$ are stabilizable.

Remark 1 *In our model, the packet drops are generated by the controller and injected into the system to create a dynamical component that enables security at the physical level. Our proposal can be incorporated in the case when packet drops occur due to channel imperfections while transferring commands from the controller to the actuator.*

B. Attack Model

We consider a specific integrity attack termed the replay attack. Here, the adversary has the ability to read and modify all sensor outputs. In particular, $y_{0:k} \subset \mathcal{I}_k^{ra}$, where $\mathcal{I}_k^{ra}$ is the attackers information. Without loss of generality (WLOG), we assume the replay attack starts at time 0. Additionally, the attacker may insert his own inputs. When inserting the input $B^{ra}u_k^{ra}$ where $u_k^{ra} \in \mathbb{R}^{p'}$, the attacker has the option of modifying the defender's actuators or using his/her own. This yields the following attack dynamics for $k \geq 0$:

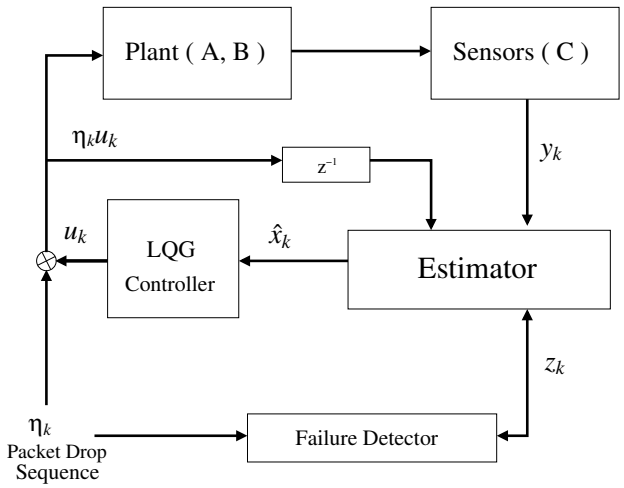


Fig. 1: System model under normal operation. When a replay attack occurs, the attacker replaces the output y_k with its time lagged version. The plant input may also be compromised.

$$x_{k+1} = Ax_k + Bu_k + B^{ra}u_k^{ra} + w_k, \tag{3}$$

$$y_k = Cx_k + D^{ra}d_k^{ra} + v_k. \tag{4}$$

The adversary is modeled to act as follows:

- Record a sufficiently long sequence of outputs $y_{0:T'}$.
- Starting at time T, replace y_k with y_{k-T}. Thus $D^{ra}d_k^{ra} = y_{k-T} - Cx_k - v_k$ for $T \leq k \leq T + T'$.
- The attacker adds some harmful input $B^{ra}u_k^{ra}$.

III. Attack Detection in a Physically Watermarked System

Starting from [1], the use of physical watermarking to detect replay attacks in control systems has been considered in the literature, e.g., in [2]–[7]. In these works, a physical watermark Δu_k is a secret random control input which we insert on top of the optimal control input u_k^* to authenticate the system:

$$u_k = u_k^* + \Delta u_k. \tag{5}$$

Here, the adversary can not read the defender's control input u_k and does not know realizations of the watermark sequence. Let us consider an IID Gaussian watermark[1] with $\Delta u_k \sim \mathcal{N}(0, \mathcal{Q})$. Note that the IID property allows the watermark to act like a secret nonce that the operator can use to verify normal operation. Under normal operation, the defender expects to see his/her watermark embedded in the sensor outputs through the dynamics of the system. Under replay attack, the sensor measurements instead contain responses to a sequence of independently selected watermarks which are tracked by the operator. This causes an alert when the freshness of the watermarks cannot be verified.

We assume the defender constructs algorithms which leverage his/her information $\mathcal{I}_k$ to make a decision, whether the

[1]Please see [4] for the treatment of a general stationary Gaussian watermark.

system operates normally $\mathcal{H}_0$ or under attack $\mathcal{H}_1$. In a threshold based detector, this can be formulated as

$$g_k(\mathcal{I}_k) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \tau_k. \tag{6}$$

We assume the defender knows the system model $\{A, B, C, Q, R, \hat{x}_{0|-1}\}$ as well as the input and output histories given by $u_{-\infty:k}$ and $y_{-\infty:k}$. However, the defender is in general unaware of the parameters of the attack model including B^a, D^a, $u^a_{0:k-1}$, and $d^a_{0:k}$.

A particular set of detectors that the defender could utilize are the residue detectors based on the parameter $z_k \triangleq y_k - CA\hat{x}_{k-1} - \eta_{k-1}CBu_{k-1}$, which is the difference between observed and expected behavior. A well known example is a χ^2 detector:

$$g_k(\mathcal{I}_k) = \sum_{t=k-WS+1}^{k} z_t^T (CPC^T + R)^{-1} z_t \tag{7}$$

where WS denotes window size. Here, given the threshold τ_k, the probability of detection β_k and the probability of false alarm α_k are:

$$\beta_k = \Pr\left(g_k\left(\mathcal{I}_k\right) > \tau_k | \mathcal{H}_1\right), \quad \alpha_k = \Pr\left(g_k\left(\mathcal{I}_k\right) > \tau_k | \mathcal{H}_0\right).$$

Another set of detectors that are used in the context of replay attacks are correlation detectors. We examine the application of a correlation detector in a microgrid example in Section VI.

IV. LQG Control with Packet Drops

Let us assume that the following information set $\mathcal{I}_k = \{y_{-\infty,k}, \eta_{-\infty,k-1}\}$ is available to the estimator at time k. Along with model knowledge, this information is leveraged to obtain an estimate $\hat{x}_k$ and generate an input u_k. We consider LQG cost optimization:

$$J = \min \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{2N+1}\sum_{k=-N}^{N}\left(x_k^T W x_k + \eta_k u_k^T U u_k\right)\right] \tag{8}$$

where U and W matrices are positive semidefinite and the optimization is performed over all inputs u_k that are measurable with respect to the information set $\mathcal{I}_k$. Note that the separation principle holds [19] and the optimal estimator and controller can be designed separately. A Kalman filter is used to obtain minimum mean squared error estimates $\hat{x}_k = \mathbb{E}[x_k|\mathcal{I}_k]$. The innovation or residual $z_k = y_k - CA\hat{x}_{k-1} - \eta_{k-1}CBu_{k-1}$ is used to recursively update the state estimate as follows:

$$\hat{x}_k = \hat{x}_{k|k=1} + Kz_k,\ \hat{x}_{k|k-1} = A\hat{x}_{k-1} + \eta_{k-1}Bu_{k-1},$$

where K is the stationary Kalman filter gain due to (A, C, Q, R):

$$K = PC^T(CPC^T + R)^{-1} \tag{9}$$

$$P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T \tag{10}$$

and $\hat{x}_{0|-1}$ is the initial apriori Kalman state estimate. The optimal control is in the following form $u_k^* = L_k\hat{x}_k$ where $L_k = -(B^TS_{k+1}B + U)^{-1}B^TS_{k+1}A$ and

$$S_k = A^TS_{k+1}A + W - (1-p_d)A^TS_{k+1}B(B^TS_{k+1}B + U)^{-1}B^TS_{k+1}A \tag{11}$$

We note that as $k \to \infty$, L_k converges to $L = -(B^TS_\infty B + U)^{-1}B^TS_\infty A$ where S_∞ satisfies the Riccati equation:

$$S_\infty = A^TS_\infty A + W - (1-p_d)A^TS_\infty B(B^TS_\infty B + U)^{-1}B^TS_\infty A. \tag{12}$$

We assume that p_d is sufficiently small so that (12) has a solution. The long term average LQG cost due to the packet drops is (c.f. [19]) given as follows:

Lemma 1 *Optimal cost J^* is*

$$J^* = tr(S_\infty Q) + tr[(A^TS_\infty A + W - S_\infty)(P - KCP)].$$

Proof: From equation (27) in [19], we have the optimal finite horizon cost, J_N^*, found as follows:

$$J_N^* = q_{-N} + \sum_{k=-N}^{N} \text{tr}(S_{k+1}Q) + \sum_{k=-N}^{N} \text{tr}(A^TS_{k+1}A + W - S_k)P_{k|k} \tag{13}$$

where q_{-N} is a bounded constant (specified in [19]) and

$$P_{k|k} = P_k - P_kC^T(CP_kC^T + R)^{-1}CP_k \tag{14}$$

Here, P_k denotes the apriori error covariance. As $N \to \infty$, $P_{k|k} \to P - KCP$ and $S_k \to S_\infty$. Thus, $\frac{1}{2N+1}J_N^* \to \text{tr}(S_\infty Q) + \text{tr}[(A^TS_\infty A + W - S_\infty)(P - KCP)]$. ■

It is worthwhile to note that J^* can be computed in closed form when packet drops occur only in the control channel. This is not possible in the general setting of [19] with sensor and control packet drops. We also note that the dependence of J^* on p_d is due to S_∞. In the sequel, we assume that the system has been running for a long time (i.e. from $k = -\infty$) so that the Kalman and state feedback gains have converged to K and L, respectively.

V. Packet Drop Injections as Physical Watermarking

We now analyze the role of packet drop injections in physical watermarking. In a scenario where the control packets are dropped by following an IID Bernoulli sequence η_k as in (1), the resulting dynamics have strong dependence on the realization of the drop sequence. This dependence offers an advantage to be used for attack detection in the same spirit as the Gaussian physical watermark. In particular, packet drops could be intentionally injected at the control stage possibly adding to existing drops due to imperfect communications. These injections enable a new type of secret nonce known to the controller and kept hidden from potential attackers and hence can be utilized for attack detection. We next consider packet drop injections in the context of replay attack detection.

Note that a replay attack may or may not be effective depending on the defender's control strategy. For example, in [3, Theorem 3], it is reported that replay attacks are asymptotically stealthy ($\lim_{k\to\infty}\beta_k - \alpha_k = 0$) in an LQG setting without drops provided that the matrix $\mathcal{A} \triangleq (A + BL)(I - KC)$ is Schur stable. On the other hand, [4] reports that if $\mathcal{A}$ has a spectral radius greater than 1, then replay attacks are asymptotically detectable with an exponentially growing detection statistic. We consider the use of packet drop injections when $\mathcal{A}$ is stable.

In particular, we consider residue detector performance under replay attack. Let us denote z_k and z_k^{ra} as the residues under normal and replay attack operations, respectively with packet drop dynamics. We start by noting that

$$\begin{aligned} z_k =& z_k^{ra} - C\mathcal{A}_k(\eta_0^{k-1})\hat{x}_{0|-1} + C\mathcal{A}_k(\tilde{\eta}_0^{k-1})\hat{x}_{0|-1}^{ra} \\ & - C\sum_{i=1}^{k}\Big(\mathcal{A}_{k-i}(\eta_i^{k-1})(A+\eta_{i-1}BL) \\ & \qquad - \mathcal{A}_{k-i}(\tilde{\eta}_i^{k-1})(A+\tilde{\eta}_{i-1}BL)\Big)Ky_{i-1}^{ra}, \quad (15) \end{aligned}$$

where $\hat{x}_{0|-1}$ and $\hat{x}_{0|-1}^{ra}$ are the initial apriori Kalman state estimates under normal and replay attack operations, respectively, and y_k^{ra} is the replayed output. Moreover, for any $\ell_1 \leq \ell_2$:

$$\mathcal{A}_{\ell_2-\ell_1}(\eta_{\ell_1+1}^{\ell_2}) = \Pi_{j=\ell_1+1}^{\ell_2}(A+\eta_j BL)(I-KC) \quad (16)$$

where $\mathcal{A}_0 = I$ and $\eta_{\ell_1+1}^{\ell_2}$ denotes the sequence $(\eta_{\ell_1+1}\dots\eta_{\ell_2})$. In (15), $\{\eta_k\}$ and $\{\tilde{\eta}_k\}$ are two binary drop sequences independent from each other and i.i.d. across k. We note that even when $\mathcal{A}_k(\eta_0^{k-1})$ vanishes, the additive term $\nu_k \triangleq C\sum_{i=1}^{k}(\mathcal{A}_{k-i}(\eta_i^{k-1})(A+\eta_{i-1}BL) - \mathcal{A}_{k-i}(\tilde{\eta}_i^{k-1})(A+\tilde{\eta}_{i-1}BL))Ky_{i-1}^{ra}$ renders the residue z_k different than the residue z_k^{ra} under attack. For example, we can show that if $||(A+BL)||^{1-p_d}||A||^{p_d}||(I-KC)|| < 1$ where $||.||$ denotes the matrix norm, then $\mathcal{A}_k(\eta_0^{k-1})$ vanishes in probability. However, the additive term ν_k does not vanish and creates a difference in the distributions of z_k and z_k^{ra}. This additive term has a similar effect to that of the additive watermark in [4] and can be leveraged to detect replay attacks. As an example, one can characterize explicit or approximate distributions of the additive term and analyze detection performance. Also note that when $p_d = 0$ or $p_d = 1$ or (possibly) the packet drop sequence is periodic, the effect of the additive term is lost since $\mathcal{A}_{\ell_2-\ell_1}(\eta_{\ell_1+1}^{\ell_2})$ is equivalent to $\mathcal{A}_{\ell_2-\ell_1}(\tilde{\eta}_{\ell_1+1}^{\ell_2})$. In these cases, the asymptotic stealthiness condition in, e.g., [3, Theorem 3], could be adapted to the current setting. In the next section, we provide real life examples and extensive numerical results to determine the effects of packet drop injection watermarking on both detection performance and overall cost.

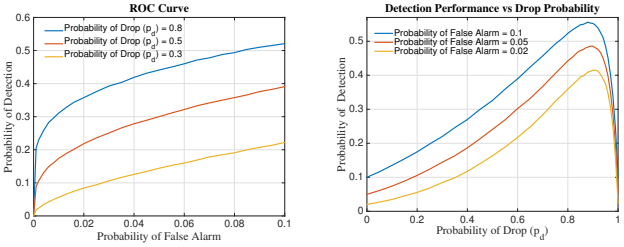
VI. Numerical Results

In this section we evaluate the performance of physical watermarking via packet drop injections on two systems. We first consider replay attacks in the quadruple tank process [20]. Then, we examine a microgrid example [2].

A. Quadruple Tank Process

In the quadruple tank process, the desired system goal is to control the water level of two tanks by leveraging two input pumps. Two sensors are used to measure the water heights of two tanks. The chosen sample period is 1 second. We use an LQG controller with weighting matrices determined using suggestions made in [21]. When examining the quadruple tank process, the optimal state feedback matrix L is dependent on the probability of drop p_d. A χ^2 detector with window size 10 is implemented to perform detection.

In Fig. 2 we examine security and performance trade-offs through relationships between the probability of false alarm, the probability of detection, and the packet drop rate. Results were averaged over 1500 trials where each trial consists of a run with 1000 time steps. In Fig. 2a, we plot several ROC curves examining the probability of detection as a function of the probability of false alarm for different packet drop rates. In Fig. 2b, we plot the probability of detection as a function of the drop rate for different false alarm probabilities ranging from 0.02 to 0.1. Note that detection performance peaks before the drop rate equals one. This can be understood in the extreme case where $p_d = 1$. Here, the system is operating in an open loop without control. Thus, when using a stable estimator, a replay attack will always be asymptotically stealthy.

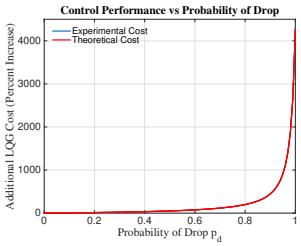


(a) Detection vs False Alarm (b) Detection vs Drop Rate

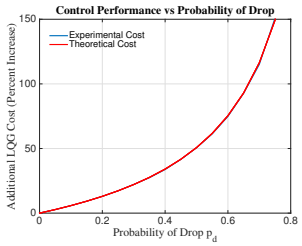
Fig. 2: Probability of Detection as a Function of Probability of Drop and Probability of False Alarm

In Fig. 3, we further characterize the tradeoff between security and control performance by mapping the probability of drop to the increased LQG cost (as a percentage of the optimal LQG cost when $p_d = 0$). In Fig. 3a, we observe the relationship between control performance and drop probability over the domain of p_d. In Fig. 3b, we examine this relationship over a smaller domain where the cost increase is restricted to be less than 150% of the optimal cost. Both the empirical cost, obtained by averaging results over 4,500 trials, and the theoretical cost are shown. We observe that they closely agree.

In Fig. 4, we plot our χ^2 detection statistic (with window size 10) averaged over 10,000 trials during a replay attack as a function of time for a system without packet drop injections (Fig. 4a) and a system with packet drop injections (Fig. 4b). Replay attacks commence at time 20. The probability of false alarm in Fig. 4 is fixed to be 0.1 and $p_d = 0.7$. The noticeable temporary bumps in detection performance seen in both Fig. 4a and Fig. 4b are likely due to initial state mismatches between



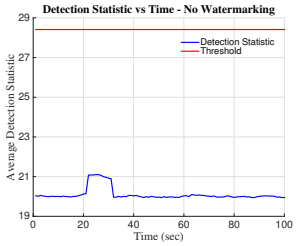
(a) LQG Cost vs Drop Rate: Full



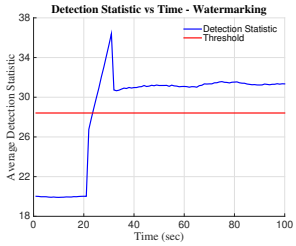
(b) LQG Cost vs Drop Rate: Partial

Fig. 3: Percent Increase in LQG Cost as a Function of Drop Probability

the true and replayed systems. An intelligent attacker can choose to delay the start of a replay attack until the true and replayed states closely match.



(a) No Watermarking



(b) Watermarking, $p_d = 0.7$

Fig. 4: χ^2 Detection Statistic vs Time

B. Microgrid

We now investigate a microgrid example borrowed from [2], using an alternative watermarking design. Here, there are 5 loads and frequency control by a mechanical speed governor is used to address small imbalances (roughly 1 percent) between load and demand. The frequency should be kept close to constant near 60 Hz. If the demand in a system far exceeds the generation, resulting in a measured drop in frequency, loads are shed to account for the imbalance. We use the linear generator model found in [22, p. 386, Fig. 11.8], see also [2]. ΔP_c, a control input which moves a steam valve in the generator, is used for watermarking. Additionally, we use $\Delta\omega$ to denote a change in angular frequency.

In the attack model, the attacker has the ability to manipulate the system's frequency sensors. The goal is to make the operator believe the frequency in the system is dropping. The defender in response sheds loads one at a time to address perceived imbalances. The attacker, once a third load is shed, relinquishes control on the frequency sensor and this way the attacker forces the operator to supply power to only two loads.

As a response, we assume the defender inserts a watermark at ΔP_c. As opposed to the packet drop injection watermark considered in this paper, we evaluate a similar zero-mean Bernoulli pulse watermark. In particular, we have

$$\Delta P_c(k) = \eta_k M(-1)^k. \tag{17}$$

where M is the magnitude of the pulse and η_k is an IID Bernoulli random variable where $\mathrm{P}(\eta_k = 0) = p_d$. Observe that a χ^2 detector is ineffective against the proposed attack because it will send an alarm in both the case that an attacker modifies a frequency sensor as well as the case that a real drop in frequency has occurred. As a result, we consider the correlation based detectors used in [2]. Here, a virtual model of the system with input ΔP_c is simulated by the defender. The response $\Delta\hat{\omega}_k$ is multiplied by the true frequency $\Delta\omega$ to obtain a correlation detector statistic g_k. Under normal operation,

$$\mathbb{E}[\Delta\hat{\omega}_k \Delta\omega_k] = \mathbb{E}[g_k] = \sigma'^2 > 0. \tag{18}$$

Under a replay attack $\mathbb{E}[\Delta\hat{\omega}_k \Delta\omega_k] = 0$. Note that unlike (6), a higher detection statistic indicates normal operation.

We simulate the microgrid over 70 seconds. Control inputs are modified every 0.1 seconds. The amplitude M controls the variance of the watermark, $\mathbb{E}[\Delta P_c^2(k)]$. A correlation detector with window of length 10 seconds is used. We first assume the sensor is not under attack, but the frequency in the system is dropping. The average frequency profile considered is given by Fig. 5. Secondly, an attacker replays the same profile (with noise independent of the watermark) from time 10 sec to 57.5 sec to force the defender to incorrectly shed loads.

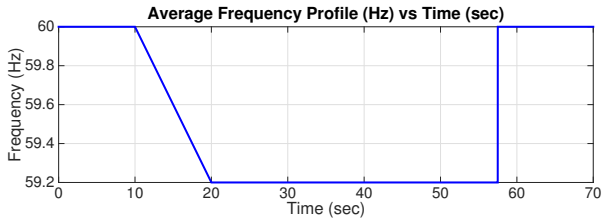


Fig. 5: Average Frequency Profile during Fault and Attack

In Fig. 6, we plot several ROC curves averaged over 1500 trials. The probability of detection is computed over the region where g_k has reached a steady state (20 to 57.5 seconds). Three different watermark variances $\mathbb{E}[\Delta P_c^2(k)]$ and p_d's are evaluated where we observe that increasing $\mathbb{E}[\Delta P_c^2(k)]$ improves detection performance.

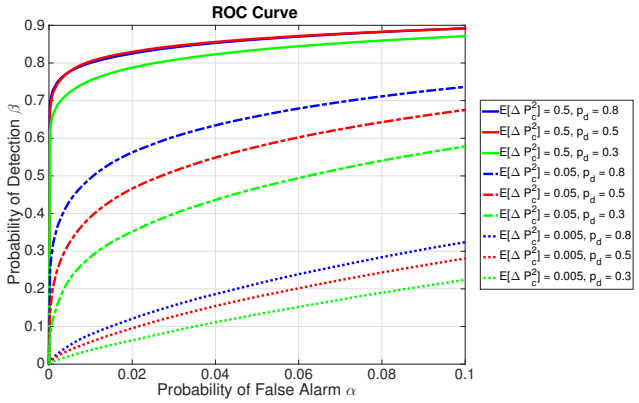


Fig. 6: Probability of Detection vs. Probability of False Alarm

In Fig. 7, we observe the detection statistics used by the correlation detector under system fault and replay attack scenarios as a function of time, averaged over 1500 trials. In this setting, the variance of the watermark is set to 0.5. Since the replayed profile is independent of the pulse watermark under a replay attack the correlation drops to 0. Detection delays occur due to the chosen 10 second detector window.

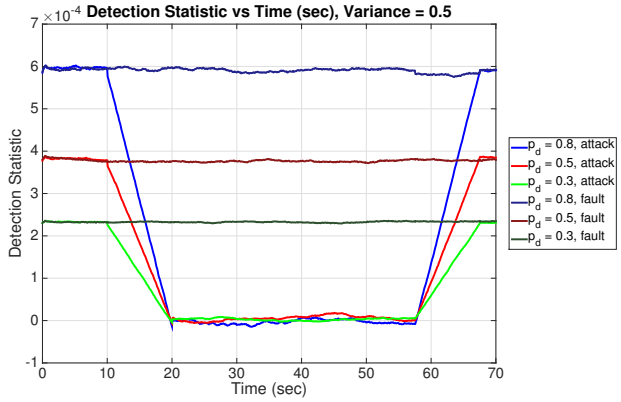


Fig. 7: Detection Statistic During Fault and Attack

As an additional measure of the watermark's affect on system performance, we consider the mean absolute deviation of the measured frequency from the average frequency profile with watermarking. Note that in the absence of watermarking, the mean deviation is 0.0252 Hz for the simulation setting.

TABLE I: Mean Abs. Deviation from Avg. Freq. Profile (Hz)

$\mathbb{E}[\Delta P_c^2]$	$p_d = 0.3$	$p_d = 0.5$	$p_d = 0.8$
0.005	0.0254	0.0256	0.0258
0.05	0.0275	0.0288	0.0307
0.5	0.0429	0.0513	0.0604

VII. Conclusion

In this paper, we proposed a new physical watermarking scheme for securing the smart grid and CPSs in general by utilizing packet drop injections. In this scheme, the noisy control input needed to authenticate the physical dynamics of the system is obtained by dropping the control packets randomly with certain probability. With the classical linear quadratic objective function, we considered the effect of packet drops on meeting security and control objectives. We analyzed the trade-off between attack detection and control in this setting. We provided extensive numerical results for the attack detection performance of specific detectors under watermarked dynamics due to packet drops, including a correlation based detector in a microgrid system. Our results indicate that IID Bernoulli packet drops could act as a potential physical watermark for attack detection in cyber-physical systems. Current work leaves several future directions to pursue. We will explore possible ways to combine packet drop injections with additive Gaussian watermarking to obtain hybrid schemes. We will also expand upon the coexistence of network packet drops and intentional packet drops. Additionally, we will work on packet drop models that involve memory and intermittency.

References

[1] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annual Allerton Conf. Communication, Control, and Computing*, Allerton, Illinois, 2009, pp. 911–918.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] B. Satchidanandan and P. Kumar, "Secure control of networked cyber-physical systems," in *55th IEEE Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 283–289.

[7] 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.

[8] Y. Liu, M. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security*, vol. 14, no. 1, pp. 13:1–13:33, 2011.

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

[10] 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.

[11] 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.

[12] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[13] 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.

[14] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*. CRC Press, 2004.

[15] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, 2010.

[16] V. Gungor, B. Lu, and G. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 10, pp. 3557 – 3564, October 2010.

[17] L. Zheng, N. Lu, and L. Cai, "Reliable wireless communication networks for demand response control," *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 133 – 140, March 2013.

[18] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1453–1464, 2004.

[19] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of Control and Estimation Over Lossy Networks," *Proc. IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[20] K. H. Johansson, "The quadruple-tank process: A multivariable laboratory process with an adjustable zero," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 3, pp. 456–465, 2000.

[21] M. Grebeck, "A comparison of controllers for the quadruple tank system," *Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, Tech. Rep*, 1998.

[22] A. R. Bergen, *Power systems analysis*. Pearson Education India, 2009.