Physical Authentication of Control Systems

Designing watermarked control inputs to detect counterfeit sensor outputs

Yilin Mo, Sean Weerakkody, Bruno Sinopoli

Cyber-Physical Systems (CPS) refer to the embedding of widespread sensing, networking, computation, and control into physical spaces with the goal of making them safer, more efficient, and reliable. Driven by the miniaturization and integration of sensing, communication, and computation in cost effective devices, CPSs are bound to transform several industries such as aerospace, transportation, built environments, energy, health-care, and manufacturing, to name a few. This great opportunity, unfortunately, is matched by even greater challenges. Taming the complexity of design and analysis of these systems poses a fundamental problem as a new paradigm is needed to bridge various scientific domains, which, through the years, have developed significantly different formalisms and methodologies. In addition, while the use of dedicated communication networks has so far sheltered systems from the outside world, use of off-the-shelf networking and computing, combined with unattended operation of a plethora of devices, provides several opportunities for malicious entities to inject attacks on CPSs. A wide variety of motivations exists for launching an attack on CPSs, ranging from economic reasons such as drawing a financial gain, all the way to terrorism, for instance, threatening an entire population by manipulating life-critical resources. Any attack on safety-critical CPSs may significantly hamper the economy and lead to the loss of human lives. While the threat of attacks on CPSs tends to be underplayed at times, the Stuxnet worm provided a clear sample of the future to come. This malware, targeting a uranium enriching facility in Iran, managed to reach the Supervisory Control

and Data Acquisition (SCADA) system controlling the centrifuges used in the enrichment process. Stuxnet modified the control system, increasing pressure in the centrifuges in a first version of the worm and spinning centrifuges in an erratic fashion in a second version. As a result, Stuxnet caused significant damage to the plant [1]. For details, see "The Stuxnet Attack".

This article proposes a control theoretic method, called physical watermarking, to authenticate the correct operation of a control system. Authentication in cyber-physical systems requires that the system operator verify the identity of a component not only in the cyber world, but also within the framework of the physical dynamics of the system. While tools exist in cryptography to perform authentication, historical data show that attackers often manage to break these security mechanisms. Moreover, such tools are ineffective against physical attacks on the system or insiders who are usually authenticated users. There arises the need to expand and complement the existing set of tools to improve the resilience of CPSs. The presence of a dynamical system, while presenting new challenges, offers new opportunities to improve detection and resilience. The existence of accurate mathematical models describing the underlying physical phenomena enables prediction of future behavior and more importantly unforeseen deviations from it. The ability to recognize irregularities in the dynamics of a system enables a principled approach to intrusion detection and resilient design. The concept of physical watermarking emerges in this context. Its utility lies in its ability to allow physical authentication of cyber-physical components. By injecting a known noisy input to a physical system, it is expected that the effect of this input can be found in the measurement of the true output, due to the system dynamics. As such, if an attacker is unaware of this physical watermark, he cannot adequately emulate the system because he is unable to consistently generate the component of the output associated with this known noisy input. Consequently, the watermark acts as a cyber-physical nonce, forcing an attacker to generate outputs unique to the given inputs at a chosen time. The next section further examines existing work and techniques related to CPS security.

Related Work

In the context of system theory, fault detection and isolation methods, used to detect anomalies in the system by measuring the discrepancy between the measured behavior and behavior predicted by the

system model, have been extensively studied over the past decades. In [2], the author provides a survey of detection schemes for the class of stochastic linear discrete time systems. First, he considers failure sensitive filters where standard estimation techniques such as the Kalman filter are altered so that state estimates are sensitive to particular failures, for instance the failure of a sensor or actuator. He next considers voting schemes, which leverage the potential presence of redundant hardware to detect failures. The author also discusses multiple hypothesis testing where the designer postulates several failure modes, for instance potential failure directions of the state, and performs statistical tests on received measurements to determine the presence of any of the specified contingencies. Finally, the author considers detection schemes based on the analysis of innovations, which are the differences between observed and expected data, for instance the χ^2 test discussed later in the article. He characterizes these schemes in terms of complexity, flexibility in the types of failures modes that can be detected, as well as performance both in detecting the presence of a fault as well as identifying the source and magnitude of a fault. While the given approaches can detect random failures, these methods are less effective against malicious attacks on control systems. This conclusion stems from the observation that while random faults can be seen as a special class of attacks, the set of intelligent attacks is much richer. Therefore, fault detection algorithms may fail under the attack of an intelligent adversary.

For example, traditional bad data detection techniques such as the largest residue test [3] have been widely used for systems with a static model, such as the power grid. Here a linear model with Gaussian noise is often used to directly relate the state, bus angles, to the outputs, power flows or power injections. However, in [4], the authors show that an attacker who is aware of the grid's configuration can inject a stealthy input into the measurements to change state estimation on the power grid. This attack, known as a false data injection attack, is implemented by inserting errors, which lie in the range space of the observation matrix, into sensor measurements. The authors separately consider the case where an adversary has access to only a given subset of sensors and the case where an adversary only has enough resources to modify a certain number of sensors. Within this context, the authors also consider the scenario where an adversary wants to insert a bias along a specific direction into the state estimate and the scenario where the attacker is indifferent to the direction of the error inserted into the state estimate. For each scenario, the authors

provide algebraic conditions to verify the existence of stealthy attack vectors, which result in no change to the residue, and methods to compute a feasible attack vector. Furthermore, the authors similarly introduce and analyze generalized false data injection attacks where small additional errors in the residue are tolerated in order to achieve more powerful attacks. Alternatively, [5] considers false data injection attacks on the grid from the system operator's point of view. The authors consider an adversary who wishes to manipulate a specific sensor measurement. The authors then propose two security indices for each sensor to quantify the least effort needed for an adversary to achieve a feasible attack while avoiding triggering the bad data detector. The first index characterizes the total number of sensors that must be modified to insert a bias into a single given sensor measurement. Meanwhile, the second index roughly characterizes the energy required to bias a single sensor measurement. These indices allow a system operator to identify potentially sparse attacks and allocate additional resources such as redundant sensors or encryption schemes as needed.

On the other hand, dynamical system models raise additional challenges for an adversary. Here, to remain undetected, an adversary must choose attack vectors that are consistent not only with the static observation model, but also with state dynamics at all times. In [6], the authors consider a general continuous-time control system. The adversary here can insert arbitrary errors to a unknown subset of sensors and actuators. The authors consider the notions of attack detectability and identifiability. In particular, the authors provide algebraic conditions that indicate if the defender can detect and identify all attacks for a given set of vulnerable sensors and actuators as well as graph theoretic conditions to characterize undetectable attacks. Furthermore, the authors propose centralized and distributed failure sensitive filters to perform attack detection as well as a centralized filter to perform attack identification. The proposed filters provide perfect detection and identification when feasible. The results given are applied to a dynamic model of the power grid. In [7], the authors consider the problem of robust control and estimation in the presence of adversaries. The authors first consider an adversary who can insert arbitrary errors in sensor measurements and show that perfect estimation is infeasible when the adversary manipulates half of the sensor measurements. Moreover, the authors show that changing system dynamics through state feedback can allow the defender to perform perfect estimation when less than half the sensor measurements are altered. The limiting assumption here is that there exists a local controller with complete access to the state. Next, the authors consider a scenario

where an adversary can arbitrarily manipulate up to q sensors and actuators and derive algebraic conditions under which the defender can perform perfect state estimation and identify the attack vectors. In their main result, the authors consider the problem of stabilizing a plant under output feedback and establish a principle of separation of estimation and control while under sensor attacks. Finally, an efficient practical decoder to perform this estimation is given in the case of just sensor attacks as well as sensor and actuator attacks.

The feasibility and detection of cyber physical attacks has also been considered in distributed systems. For example, [8] considers the setting of a wireless control network, which consists of a set of nodes with a sparse underlying communication network. A subset of nodes communicates directly with sensors, a subset of nodes communicates directly with actuators, and stabilizing output feedback is performed using distributed linear iterations. Additionally, an unknown subset of nodes is malicious and can inject arbitrary errors into its components of the state. The authors consider the design of an intrusion detection system, which can recover sensor outputs for data logging purposes and can identify malicious nodes. Furthermore, the authors derive both algebraic and graphical conditions to determine the feasibility of perfect estimation and detection. Moreover, the authors propose a systematic procedure to estimate sensor outputs and identify malicious nodes. Additionally, [9] and [10] consider the feasibility and detection of attacks in distributed systems. Specifically, [9] considers a system where nodes attempt to compute functions of their initial states by means of distributed linear iterations. The authors assume there exists a subset of malicious nodes, which broadcast incorrect values of their states to their neighbors. Here, the authors prove that a given node calculating an arbitrary function can tolerate up to f faulty agents if and only if there exists at least $2f + 1$ vertex disjoint paths to any non-neighboring node. Moreover, a combinatorial procedure to determine the entire initial state is provided. Alternatively, [10] considers the special case of consensus algorithms where a set of agents all attempt to compute the same function of their initial states through distributed linear iterations. The authors consider a scenario where there exist malicious agents who collude and may know the structure of the network and a scenario where there exists faulty agents who do not collude. Here, algebraic and graphical conditions are given for when faulty and malicious agents can be detected and identified. The authors then characterize the effect unidentifiable inputs have on the consensus value and propose three

failure sensitive filters to detect and identify malicious or faulty nodes.

In scenarios [6, 7, 8, 9, 10], the attacker can either arbitrarily perturb the system along certain directions without being detected by any filter or cannot induce any perturbation, without incurring detection. However, in these contributions, the system model is assumed to be noiseless, which greatly favors the failure detector, since the evolution of the system is deterministic and any deviation from the predetermined trajectory can be detected. A more realistic scenario needs to account for a noisy environment. In this case, it is harder to detect malicious behavior since the adversary may inject an attack that renders the compromised system statistically indistinguishable from the healthy system.

In [11], the authors consider attacks on control systems in a noisy environment. The adversary in this system is aware of the plant model, noise statistics, and the controller and state estimator. The attacker can also manipulate a subset of sensors. The authors derive sufficient and necessary conditions for the feasibility of a dynamic false data injection attack where an attacker can cause unbounded errors in the state estimate without substantially increasing the probability of detection by a residue detector. Additionally, the authors derive an algorithm to perform such an attack. This method involves rendering unstable modes in the system unobservable. To improve resilience to such an attack, the authors suggest using redundant sensors to measure unstable modes. Similarly, [12] considers false data injection attacks in a noisy wireless sensor network without control inputs, where again the adversary has access to the system model as well as a subset of sensors. Here, the authors bound the reachable region of the estimator biases that an adversary can inject without substantially increasing his probability of detection. The reachable region is characterized by formulating the attack as a constrained optimal control problem.

In [13, 14], the authors analyze the replay attack model inspired by the Stuxnet example. Here, an adversary can read and modify all sensor signals in the system, which is assumed to be in steady state. The attacker, rather than causing physical damage to the system by perturbing sensor measurements along specific directions as is done in a false data injection attack discussed above, can insert a harmful input into the system. To evade detection, the attacker replays previous sensor measurements to the operator. These outputs are statistically identical to the true outputs in steady state. Furthermore, unlike the false data injection attack, the adversary requires no knowledge of the system model to generate stealthy outputs.

Because the adversary hijacks all sensors, resilient control as done by [7] cannot stabilize the system and as such the main focus in countering a replay attack is detection. The authors of [13, 14] create a physical authentication scheme, where a random "watermark" signal is added to the optimal control signal. A digital watermark, traditionally seen in audio and image processing, embeds information in a carrier signal, which is later used to verify authenticity or integrity of the owner. One application of digital watermarking is in source tracking of illegally copied movies where a watermark is used to determine the owner of the original signal. Similarly, in [13, 14] if the system is operating normally, then the effect of the chosen watermark signal is present in the sensor measurements. However, if the system is malfunctioning or under attack, the effect of the watermark signal chosen by the system operators cannot be detected. Conceptually, this approach is similar to a challenge-response authentication scheme in information security, where the watermark signal and the sensor measurements can be seen as the "challenge" and "response" respectively. Table 1 summarizes previous work mentioned in this article related to control system security. However, note that Table 1 does not exhaustively summarize all previous work in control system security and apologies are extended for any omissions.

This article further investigates the problem of designing the optimal watermark signal in the class of stationary Gaussian processes to maximize a relaxed version of the expected Kullback-Leibler divergence between the distributions of the compromised and healthy residue vector, while satisfying a constraint on the control performance. This approach can be seen as a generalization of [13, 14], where only independent and identically distributed (i.i.d.) Gaussian processes are considered.

Observe that the fundamental requirements of physical security in CPS are applicable to the notion of security in general control systems. As such, the watermarking scheme described in this article is applied to the class of discrete linear time invariant state-space models. The optimization problem, when carried out in the frequency domain, can be separated into two steps where the optimal direction of the signal for each frequency is first computed and then all possible frequencies are considered to find the optimal watermark signal.

The rest of the article is organized as follows: First, the system description is given. Here the linear-quadratic-Gaussian controller is revisited and adapted, the concept of a watermarked input is described, and

properties of the failure detector are defined. In the next section, a replay attack model is provided and the class of systems incapable of detecting such attacks are identified. In the following section, a systematic method to compute the watermarking signal is given. Specifically, a Neyman-Pearson detector is derived, and optimal statistical properties of the watermark input are obtained subject to some upper bound on the total cost of control. Due to the impossibility of computing the detection probability in closed form, only a relaxed version of the original optimization problem is solved. While the solution provided is near optimal with respect to the originally formulated optimization problem, the rest of the article still refers to the solution as the optimal watermark for easier reading. An algorithm to generate this input is also given. Afterwards a numerical example is provided to compare probability of detection with the probability of false alarm as well as the cost of control. The final section concludes the article, with some directions for future work. The appendix contains some proofs that would otherwise interrupt the flow of the article.

System Description

The physical watermarking strategy is given for a class of general control systems. The control system is modeled as a linear, time invariant (LTI) system, the state dynamics of which are given by

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the vector of state variables at time k, $u_k \in \mathbb{R}^p$ is the control input, and $w_k \in \mathbb{R}^n$ is the process noise at time k. w_k is assumed to be an i.i.d. Gaussian process with $w_k \sim \mathcal{N}(0, Q)$. Since the control system usually operates for an extended period of time, it is assumed that the system starts at time $-\infty$.

A sensor suite monitors the system described in (1). At each step, all the sensor readings are collected by a base station. The observation equation can be written as

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

where $y_k \in \mathbb{R}^m$ is a vector of measurements from the sensors and $v_k \sim \mathcal{N}(0, R)$ is i.i.d. measurement noise independent of w_k. It is assumed that (A, B) is stabilizable and (A, C) is detectable.

It is assumed that the system operator wants to minimize the infinite-horizon linear-quadratic-Gaussian

(LQG) cost

$$J = \lim_{T\to\infty} E\frac{1}{2T+1}\left[\sum_{k=-T}^{T}\left(x_k^T W x_k + u_k^T U u_k\right)\right], \tag{3}$$

where W, U are positive definite matrices. Since the separation principle holds in this case, the optimal solution of (3) is a combination of the Kalman filter and LQG controller [15]. The Kalman filter provides the optimal state estimate $\hat{x}_k$. Since the system is assumed to start at $-\infty$, the Kalman filter converges to a fixed gain linear estimator

$$\hat{x}_{k+1|k} = A\hat{x}_k + Bu_k,\ \hat{x}_k = \hat{x}_{k|k-1} + Kz_k. \tag{4}$$

where $z_k \triangleq y_k - C\hat{x}_{k|k-1}$ is the residue vector and the Kalman gain K is given by

$$K \triangleq PC^T\left(CPC^T + R\right)^{-1}, \tag{5}$$

where P is the solution of the Riccati equation

$$P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T. \tag{6}$$

The estimation error at time k is defined to be $e_k = x_k - \hat{x}_k$.

The LQG controller is a fixed gain linear controller based on the optimal state estimate $\hat{x}_k$. Specifically,

$$u_k^* = L\hat{x}_k, \tag{7}$$

where u_k^* is the optimal control input. The control gain matrix L is defined to be

$$L \triangleq -\left(B^T S B + U\right)^{-1} B^T S A, \tag{8}$$

where S satisfies the Riccati equation

$$S = A^T S A + W - A^T S B\left(B^T S B + U\right)^{-1} B^T S A. \tag{9}$$

Consider the case where, instead of directly applying the optimal LQG control u_k^* to the physical system, a physical watermarking scheme is used, in which the true control input u_k is given by

$$u_k = u_k^* + \zeta_k, \tag{10}$$

where u_k^* is the optimal LQG control and ζ_k is the watermark signal. The watermark signal $\{\zeta_k\}$ is assumed to be a $p-$dimensional stationary zero-mean Gaussian process independent from the noise processes $\{w_k\}$, $\{v_k\}$.

Define the autocovariance function $\Gamma : \mathbb{Z} \to \mathbb{R}^{p\times p}$ to be

$$\Gamma(d) \triangleq \mathrm{Cov}(\zeta_0, \zeta_d) = \mathbb{E}\zeta_0\zeta_d^T. \tag{11}$$

In this article, the watermark is assumed to be generated by a Hidden-Markov Model (HMM)

$$\xi_{k+1} = A_h\xi_k + \psi_k, \quad \zeta_k = C_h\xi_k, \tag{12}$$

where $\psi_k \in \mathbb{R}^{n_h}$, $k \in \mathbb{Z}$ is a sequence of i.i.d. zero-mean Gaussian random variables with covariance Ψ, and $\xi_k \in \mathbb{R}^{n_h}$ is the hidden state. To make $\{\zeta_k\}$ a stationary process, the covariance of ξ_0 is assumed to be the solution of the following Lyapunov equation

$$\mathrm{Cov}(\xi_0) = A_h\,\mathrm{Cov}(\xi_0)A_h^T + \Psi.$$

All the matrices are of proper dimensions.

Remark 1. *It is worth noticing that $\{\zeta_k\}$ is completely described by its finite dimensional distribution and hence the autocovariance function Γ. However, the watermarking signal is restricted to be generated from an HMM since any autocovariance function Γ can be approximated by an HMM, given that the dimension n_h of the hidden state is large enough. On the other hand, the HMM is easy to implement if n_h is small, which is the case for the optimal watermarking signal, as is illustrated later by Theorem 6.*

To ensure the freshness of the watermark signal, A_h is assumed to be strictly stable, which implies that the correlation between the current watermark signal ζ_k and the future watermark signal $\zeta_{k'}$ decays to 0 exponentially when $k' - k \to \infty$. The spectral radius of A_h is denoted as $\rho(A_h) < 1$. In this article, it is assumed that the watermark signal is chosen from a Hidden-Markov Model with $\rho(A_h) \leq \rho$, where $\rho < 1$ is a design parameter. A value of ρ close to 1 gives the system operator more freedom to design the watermark signal, while a value of ρ close to 0 improves the freshness of the watermark signal by reducing the correlation of ζ_k at different time steps. To simplify notations, define the feasible set $\mathcal{G}(\rho)$ as

$$\mathcal{G}(\rho) = \{\Gamma : \Gamma \text{ is generated by an HMM (12) with } \rho(A_h) \leq \rho\}. \tag{13}$$

Remark 2. *Since it is assumed that (A, B) is stabilizable and (A, C) is detectable, the closed-loop system is stable regardless of the watermark signal. Furthermore, by the separation principle, the Kalman filter is*

the optimal filter regardless of the watermark signal ζ_k. However, the addition of ζ_k incurs an LQG control performance loss and the control input u_k is not optimal. The necessity of adding the watermark signal ζ_k is illustrated later in Theorem 1. Conceptually, if the system is under normal operation, then the effect of the watermark signal ζ_k can be found in the sensor measurements y_k. The presence of the watermark is possibly lost when the system is malfunctioning or under attack, which can be detected by the failure detector.

If no watermark signal is present, that is if $\zeta_k = 0$, then the optimal objective function J^* given by the Kalman filter and LQG controller is

$$J^* = \operatorname{tr}(SQ) + \operatorname{tr}\left[\left(A^T S A + W - S\right)(P - KCP)\right]. \tag{14}$$

A failure detector is used to detect abnormality of the system. In this article, the failure detector is assumed to trigger an alarm at time k if and only if the condition,

$$g(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots) \geq \eta, \tag{15}$$

is met where $g(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots)$ is a continuous real valued function of $z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots$ and η is the threshold, which is a design parameter of the system. Under normal operation, denote the probability of false alarm to be α, defined as

$$\alpha \triangleq P(g(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots) \geq \eta). \tag{16}$$

False alarms usually occur with low probability for practical systems. When the system is operating normally, z_k is a stationary process and hence α is a constant.

Remark 3. *A widely used failure detector is the χ^2 detector ([16], [17]), which satisfies*

$$g(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots) = z_k^T (CPC^T + R)^{-1} z_k.$$

Fig 1 shows the system diagram described in this section.

Attack Model

In this section, a model for a replay attack motivated by Stuxnet is given. To cause physical damage, a first version of Stuxnet implements control logic to increase pressure in the centrifuge while a second version

of the worm varies rotor speeds. To prevent detection in the first scenario, Stuxnet replayed previous sensor outputs, recorded prior to inserting harmful control inputs, to the SCADA system [1]. Since the system was in steady state, outputs from the past, collected in steady state, were statistically identical to outputs under normal operation, and as such were not detected. Motivated by Stuxnet, the following replay attack model is considered in this article.

Attacker's Knowledge and Resources

The adversary is first described through its knowledge and available resources.

1. The attacker has knowledge of all real time sensor measurements. In particular, he knows the true sensor outputs y_k for all k.

2. The attacker can violate the integrity of all sensor measurements. Specifically, he can modify the true sensor signals y_k to arbitrary sensor signals y_k^v.

   Remark 4. *The attack on the sensors can be carried out by breaking the cryptography algorithm. Another way to perform an attack, which is potentially much harder to defend, is to use physical attacks. Physical attacks can violate the basic properties of secrecy, integrity and availability without the need to attack the cyber part of the system. Consider for example a temperature sensor. Secrecy, integrity and availability of its sensing data can be affected by placing a sensor nearby, affecting the local temperature around the sensor, and enclosing the sensor with a metal cover respectively. In addition, the insider threat is critical in large infrastructures, as these systems usually involve many employees. These kinds of attacks may be easy to carry out when sensors are spatially distributed in remote locations.*

3. The attacker has access to a set of external actuators with control matrix $B^a \in \mathbb{R}^{n \times p_a}$ and can thus insert an external input $B^a u_k^a$ where $u_k^a \in \mathbb{R}^{p_a}$ is the control input. Moreover, assuming that u_k^a is intelligently chosen, the set of actuators B^a allows the adversary to achieve a malicious objective, for instance causing physical damage to the plant.

   Remark 5. *The attacker could inject the external control input by controlling a subset of actuators of*

the system and/or deploying its own actuators. For example, to change the temperature distribution in a building, the attacker could take control of the HVAC (heating, ventilation, and air conditioning) system, deploy heaters of its own, or even commit arson.

4. The attacker does not need to have full knowledge of the system parameters, namely the A, B, C, Q, R, K, L matrices and the Γ function. However, the attacker has enough knowledge of the system model to design an input $u_k^a \in \mathbb{R}^{p_a}$, which may achieve its malicious objective such as physically damaging the plant.

Attack Strategy

Given the adversary's knowledge and resources, the following attack strategy is considered.

1. The attacker records a sequence of sensor measurements from time $-T$ to time -1, where T is a large enough number to ensure that the attacker can replay the sequence later for an extended period of time.

2. Starting at time 0 to time $T-1$, the attacker modifies the sensor signals to y_k^v, which is the same as the measurements recorded by the attacker at time $k-T$. In other words,

$$y_k^v = y_{k-T},\ 0 \leq k \leq T-1.$$

Remark 6. *For simplicity, the time that the replay starts is denoted as time* 0*. In reality, the attacker can freely choose the starting time, which is unknown to the system operator.*

3. Starting at time 0, the attacker injects an external control input $B^a u_k^a$, where $u_k^a \in \mathbb{R}^{p_a}$ is the control input and $B^a \in \mathbb{R}^{n \times p_a}$ denotes its direction.

Remark 7. *When the system is under attack, the controller cannot perform closed loop control since the true sensory information is not available. Therefore, control performance of the system cannot be guaranteed during the attack. In fact, the attacker can inject a bias on the state of the physical system along its controllable subspace, which is the column space of* $[B^a, AB^a, \ldots, A^{n-1}B^a]$*. The only way to counter this attack is to detect its presence.*

System Model Under Attack

To simplify notations, time-shifted variables,

$$\hat{x}^v_{k|k-1} \triangleq \hat{x}_{k-T|k-T-1},\ z^v_k = z_{k-T},\ \zeta^v_k = \zeta_{k-T},\ 0 \le k \le T-1, \tag{17}$$

are defined. During the replay ($0 \le k \le T-1$), the system dynamics changes to

$$x_{k+1} = Ax_k + Bu_k + B^a u^a_k + w_k,\ y_k = Cx_k + v_k, \tag{18}$$

$$\hat{x}_{k+1|k} = A\hat{x}_k + Bu_k,\ \hat{x}_k = \hat{x}_{k|k-1} + K\left(y^v_k - C\hat{x}_{k|k-1}\right), \tag{19}$$

$$u_k = L\hat{x}_k + \zeta_k,\ z_k = y^v_k - C\hat{x}_{k|k-1}. \tag{20}$$

Notice that the fake measurement y^v_k is used instead of y_k for calculating the state estimate and residue. In addition, the probability of detection at time k is defined to be β_k given as

$$\beta_k \triangleq P(g(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots) \ge \eta), \quad 0 \le k \le T-1. \tag{21}$$

Fig 2 shows the diagram of the control system under attack.

The following theorem characterizes the feasibility of the replay attack in the absence of the watermark signal ζ_k, which illustrates the necessity of the physical watermark.

Theorem 1. *Suppose $\zeta_k = 0$ for all k. If $\mathcal{A} \triangleq (A+BL)(I-KC)$ is stable, $\rho((A+BL)(I-KC)) < 1$, then the detection rate β_k of all detectors g converges to the false alarm rate α during the attack, that is,*

$$\lim_{k\to\infty} \beta_k = \alpha. \tag{22}$$

On the other hand, if $\mathcal{A}$ is strictly unstable and g satisfies

$$\lim_{||z||\to\infty} g(z, 0, 0, \ldots) = \infty, \tag{23}$$

for some norm $||\cdot||$, then the detection rate β_k converges to 1, that is,

$$\lim_{k\to\infty} \beta_k = 1. \tag{24}$$

Proof. Part of the proof is reported in [13]. However, for the sake of completeness, the whole proof is included here. Manipulating (17)-(20) yields

$$\hat{x}_{k+1|k} = \mathcal{A}\hat{x}_{k|k-1} + (A+BL)Ky^v_k + B\zeta_k \tag{25}$$

$$\hat{x}^v_{k+1|k} = \mathcal{A}\hat{x}^v_{k|k-1} + (A+BL)Ky^v_k + B\zeta^v_k \tag{26}$$

$$z_{k+1} = z^v_{k+1} - C\mathcal{A}^{k+1}\left(\hat{x}_{0|-1} - \hat{x}^v_{0|-1}\right) - C\sum_{i=0}^{k}\mathcal{A}^{k-i}B\left(\zeta_i - \zeta^v_i\right). \tag{27}$$

If $\mathcal{A}$ is stable and $\zeta_k = \zeta^v_k = 0$, then the residue z_k of the system under the replay attack converges to the residue z^v_k of the virtual system, which is essentially z_{k-T}. Hence,

$$\lim_{k\to\infty}\beta_k = \lim_{k\to\infty}P(g(z_k,0,0,\ldots)\geq\eta) = P(g(z^v_k,0,0,\ldots)\geq\eta) = P(g(z_{k-T},0,0,\ldots)\geq\eta) = \alpha.$$

On the other hand, if $\mathcal{A}$ is strictly unstable, the second term on the RHS (right hand side) of (27) goes to infinity almost surely. Hence, if $g(z,0,0,\ldots)\to\infty$ when $||z||\to\infty$,

$$\lim_{k\to\infty}\beta_k = \lim_{k\to\infty}P(g(z_k,0,0,\ldots)\geq\eta) = 1,$$

which concludes the proof. □

Remark 8. *Notice that the stability of the "healthy" system depends only on the $A+BL$ and $A-KCA$ matrices, not on $\mathcal{A}$. Hence, it is entirely possible that the closed-loop system is stable while $\mathcal{A}$ is unstable. As seen from (25) and (26), the stability of $\mathcal{A}$ implies that the open-loop cyber system, consisting of the controller and estimator, is stable. In the one dimensional case, the stability of $\mathcal{A}$ is easy to analyze since $\mathcal{A} = (A+BL)(A-KCA)A^{-1}$. Thus, due to the stability of $A+BL$ and $A-KCA$, $\mathcal{A}$ is stable if A is unstable. Such analysis does not hold for higher dimensional systems since the product of stable matrices may not be stable.*

Remark 9. *Additionally, observe that Theorem 1 considers the alarm rate β_k when k goes to infinity while in the attack model it is assumed that the replay is performed from time 0 to time $T-1$. However, since T is assumed to be large and β_k typically converges quickly, as is illustrated by the numerical examples, the asymptotic performance of β_k serves as an indicator of the detection performance of the system.*

Based on Theorem 1, if $\mathcal{A}$ is strictly unstable, then the attacker can be detected efficiently as the detection rate β_k converges to 1. However, if $\mathcal{A}$ is stable, then the attacker can perform the replay attack for an extended period of time given that the false alarm rate α is insignificant, which implies that the system is not resilient to this type of attack. In that case, one possible countermeasure is to redesign the

estimation and control gain matrices K and L so that the closed-loop system is stable, while enforcing $\mathcal{A}$ strictly unstable. However, this approach is not always desirable, since the control and estimation gain matrices are usually designed to satisfy certain safety and performance constraints and hence cannot be changed arbitrarily. In these scenarios, instead of redesigning K and L, the watermark signal can be used to enable intrusion detection.

Watermark Design and Detection

This section is devoted to developing a design methodology for the watermark signal and the anomaly detector. To begin, the following assumption is made on the control system.

Assumption 1. *$\mathcal{A}$ is stable. That is, $\rho((A+BL)(I-KC)) < 1$.*

Throughout this section, it is assumed that $\mathcal{A}$ is stable, since otherwise the watermark signal would be unnecessary as a consequence of Theorem 1. To simplify notations, define the symmetric part of a matrix X as

$$\mathrm{sym}(X) \triangleq \frac{X + X^T}{2}. \tag{28}$$

LQG Performance Loss

The addition of noisy watermarks on top of optimal LQG inputs naturally degrades the performance of the system as described by the LQG cost. The following theorem provides the LQG control performance loss incurred by the watermark signal.

Theorem 2. *The LQG performance of the system described by* (1), (2), (4) *and* (10) *is given by*

$$J = J^* + \Delta J, \tag{29}$$

where J^* is the optimal LQG cost without the watermark signal and

$$\Delta J = \mathrm{tr}\left\{U\Gamma(0) + 2U\,\mathrm{sym}\left[L\sum_{d=0}^{\infty}(A+BL)^d B\Gamma(1+d)\right]\right\} + \mathrm{tr}\left[(W + L^T U L)\Theta_1\right], \tag{30}$$

where

$$\Theta_1 \triangleq 2\sum_{d=0}^{\infty}\mathrm{sym}\left[(A+BL)^d\mathcal{L}_1(\Gamma(d))\right] - \mathcal{L}_1(\Gamma(0)),$$

and $\mathcal{L}_1 : \mathbb{C}^{p\times p} \to \mathbb{C}^{n\times n}$ *is a linear operator defined as*

$$\mathcal{L}_1(X) = \sum_{i=0}^{\infty} (A+BL)^i BXB^T((A+BL)^i)^T = (A+BL)\mathcal{L}_1(X)(A+BL)^T + BXB^T.$$

Remark 10. *While the expression for* ΔJ *is complicated, it is linear with respect to the autocovariance functions* $\Gamma(d)$*. This linearity enables the optimization in the frequency domain, as is illustrated in Theorem 6.*

Optimal Detector

This subsection derives the asymptotically optimal detector. As seen from Fig 1, the detector has real time knowledge of the residue z_k, obtained from the estimator, as well as real time knowledge of the trajectory of the watermark, $\{\zeta_k\}$. Define the covariance of the residue z_k of the healthy system to be

$$\mathcal{P} \triangleq CPC^T + R. \tag{31}$$

For the "healthy" system, z_k is Gaussian distributed with mean 0 and covariance $\mathcal{P}$.

By (27), for the system under the replay attack

$$z_{k+1} = -C\mathcal{A}^{k+1}\left(\hat{x}_{0|-1} - \hat{x}^v_{0|-1}\right) - C\sum_{i=0}^{k}\mathcal{A}^{k-i}B\zeta_i + C\sum_{i=0}^{k}\mathcal{A}^{k-i}B\zeta^v_i + z^v_{k+1}. \tag{32}$$

The first term on the RHS of (32) converges to 0 since $\mathcal{A}$ is stable. The second term is a function of the watermark signal, which is generated and thereby known by the control system and the failure detector. The third and fourth terms are independent from each other since z_k is the residue vector of the Kalman filter. Further define

$$\mu_k \triangleq -C\sum_{i=-\infty}^{k}\mathcal{A}^{k-i}B\zeta_i, \tag{33}$$

and

$$\Sigma \triangleq \lim_{k\to\infty}\mathrm{Cov}\left[C\sum_{i=0}^{k}\mathcal{A}^{k-i}B\zeta^v_i\right] = \mathrm{Cov}\left[C\sum_{i=0}^{\infty}\mathcal{A}^{i}B\zeta_{-i}\right]. \tag{34}$$

Expanding the RHS of (34),

$$\Sigma = 2\sum_{d=0}^{\infty} C\,\mathrm{sym}\left[\mathcal{A}^d\mathcal{L}_2(\Gamma(d))\right]C^T - C\mathcal{L}_2(\Gamma(0))C^T, \tag{35}$$

where $\mathcal{L}_2 : \mathbb{C}^{p\times p} \to \mathbb{C}^{n\times n}$ is a linear operator on the space of $p\times p$ matrices, which is defined as

$$\mathcal{L}_2(X) \triangleq \sum_{i=0}^{\infty}\mathcal{A}^i BXB^T(\mathcal{A}^i)^T = \mathcal{A}\mathcal{L}_2(X)\mathcal{A}^T + BXB^T.$$

Therefore, z_k converges to a Gaussian with mean μ_{k-1} and covariance $\mathcal{P}+\Sigma$. As a result, the null hypothesis is

$$H_0: \text{ the residue } z_k \text{ follows a Gaussian distribution } \mathcal{N}_0(0, \mathcal{P}).$$

The alternative hypothesis is

$$H_1: \text{ the residue } z_k \text{ follows a Gaussian distribution } \mathcal{N}_1(\mu_{k-1}, \mathcal{P}+\Sigma).$$

By the Neyman-Pearson lemma [18], the optimal detector is given by the Neyman-Pearson detector as discussed in Theorem 3.

Theorem 3. *The optimal Neyman-Pearson detector rejects H_0 in favor of H_1 if*

$$g_{NP}(z_k, \zeta_{k-1}, \zeta_{k-2}, \ldots) = z_k^T \mathcal{P}^{-1} z_k - (z_k - \mu_{k-1})^T (\mathcal{P}+\Sigma)^{-1}(z_k - \mu_{k-1}) \geq \eta. \tag{36}$$

Otherwise, hypothesis H_0 is accepted.

To characterize the performance of the detector, ideally the asymptotic detection rate $\lim_{k\to\infty} \beta_k$ or expected time to detection is considered. However, the detection rate and expected time to detection involve integrating a Gaussian distribution, which usually does not have an analytical solution. In this article, the Kullback-Leibler (KL) divergence, which measures the "distance" between the two distributions, is used to characterize the detection performance. This choice rests on the observation that as the KL divergence between two distributions increases, the distributions become, roughly speaking, easier to distinguish. For details, see "The Kullback-Liebler Divergence." The KL divergence of the two Gaussian distributions in H_0 and H_1 is given by the next theorem

Theorem 4. *The expected KL divergence of distribution $\mathcal{N}_1$ and $\mathcal{N}_0$ is*

$$\mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) = \operatorname{tr}(\Sigma \mathcal{P}^{-1}) - \frac{1}{2} \log \det(I + \Sigma \mathcal{P}^{-1}). \tag{37}$$

Furthermore, the expected KL divergence satisfies the inequality

$$\frac{1}{2} \operatorname{tr}(\Sigma \mathcal{P}^{-1}) \leq \mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) \leq \operatorname{tr}(\Sigma \mathcal{P}^{-1}) - \frac{1}{2} \log \left[1 + \operatorname{tr}(\Sigma \mathcal{P}^{-1})\right], \tag{38}$$

where the upper bound is tight if C is of rank 1.

It is worth noticing that the expected KL divergence is a convex function of Σ. However, both the upper and lower bound of the expected KL divergence are monotonically increasing with respect to $\mathrm{tr}(\Sigma\mathcal{P}^{-1})$, which is linear in Σ.

Optimal Watermark Signal

This subsection derives the optimal watermark signal. Ideally, the following optimization problem should be solved.

$$\begin{aligned} \underset{\Gamma(d)\in\mathcal{G}(\rho)}{\text{maximize}} \quad & \mathbb{E}\, D_{KL}(\mathcal{N}_1\|\mathcal{N}_0) \\ \text{subject to} \quad & \Delta J \leq \delta, \end{aligned} \tag{39}$$

where $\delta > 0$ is a design parameter.

However, it is computationally hard to solve this maximization problem since the expected KL divergence is not a concave function of $\Gamma(d)$. Hence, the ensuing optimization problem is solved.

$$\begin{aligned} \underset{\Gamma(d)\in\mathcal{G}(\rho)}{\text{maximize}} \quad & \mathrm{tr}(\Sigma\mathcal{P}^{-1}) \\ \text{subject to} \quad & \Delta J \leq \delta, \end{aligned} \tag{40}$$

Notice that the expected KL divergence is relaxed to $\mathrm{tr}(\Sigma\mathcal{P}^{-1})$, using the upper and lower bound derived in Theorem 4. Furthermore, if C is of rank 1, then by Theorem 4, optimizing $\mathrm{tr}(\Sigma\mathcal{P}^{-1})$ is equivalent to optimizing the expected KL divergence. For general cases, the optimality gap can be quantified using the upper and lower bound.

Although Σ and ΔJ are linear functionals of Γ, convex optimization techniques cannot be directly applied to solve (40), since Γ is in an infinite dimensional space. As a result, (40) is transformed into the frequency domain. Before continuing on, the following definition is needed.

Definition 1. *ν is a positive Hermitian measure of size $p \times p$ on the interval $(-0.5, 0.5]$ if for a Borel set $S_B \subseteq (-0.5, 0.5]$, $\nu(S_B)$ is a positive semidefinite Hermitian matrix with size $p \times p$.*

The following theorem establishes the existence of a frequency domain representation for $\Gamma(d)$.

Theorem 5 (Bochner's Theorem [19, 20])**.** $\Gamma(d)$ *is the autocovariance function of a stationary Gaussian process* $\{\zeta_k\}$ *if and only if there exists a unique positive Hermitian measure* ν *of size* $p \times p$*, such that*

$$\Gamma(d) = \int_{-1/2}^{1/2} \exp(2\pi j d\omega)\, \mathrm{d}\, \nu(\omega). \tag{41}$$

$\mathrm{d}\, \nu(\omega)$ can be interpreted as the discrete-time Fourier transform of the function $\Gamma(d)$. In fact, if $\nu(\omega)$ is absolutely continuous with respect to the Lebesgue measure, then

$$\mathrm{d}\, \nu(\omega) = f(\omega)\, \mathrm{d}\, \omega,$$

and

$$\Gamma(d) = \int_{-1/2}^{1/2} \exp(2\pi j d\omega) f(\omega)\, \mathrm{d}\, \omega,$$

where f is a mapping from $(-0.5, 0.5]$ to the set of positive semidefinite Hermitian matrices. f is exactly the "entrywise" Fourier transform of $\Gamma(d)$.

By the fact that $\Gamma(d)$ is real, the Hermitian measure ν satisfies the following property, which can be applied to the Fourier transform of the real valued signals.

Proposition 1. $\Gamma(d)$ *is real if and only if for all Borel-measureable sets* $S_B \subseteq (-0.5, 0.5]$*,*

$$\nu(S_B) = \overline{\nu(-S_B)}. \tag{42}$$

By (42), (41) can be simplified as

$$\Gamma(d) = 2\, \mathfrak{Re} \left(\int_{0}^{1/2} \exp(2\pi j d\omega)\, \mathrm{d}\, \nu(\omega) \right). \tag{43}$$

Theorem 6. *The optimal solution (not necessarily unique) of* (40) *is*

$$\Gamma_*(d) = 2\rho^{|d|}\, \mathfrak{Re}\left[\exp(2\pi j d\omega_*) H_* \right], \tag{44}$$

where ω_* *and* H_* *are the solution of the ensuing optimization problem.*

$$\begin{aligned} \underset{\omega, H}{\textit{maximize}} \quad & \operatorname{tr}\left[\mathcal{F}_2(\omega, H) C^T \mathcal{P}^{-1} C \right] \\ \textit{subject to} \quad & \mathcal{F}_1(\omega, H) \leq \delta,\ 0 \leq \omega \leq 0.5, \\ & H \textit{ Hermitian and Positive Semidefinite,} \end{aligned} \tag{45}$$

where the function $\mathcal{F}_1$ is defined as

$$\mathcal{F}_1(\omega, H) \triangleq \operatorname{tr}\left[U\Theta_2\right] + \operatorname{tr}\left[(W + L^T U L)\Theta_3\right], \tag{46}$$

$$\Theta_2 \triangleq 2\,\mathfrak{Re}\left\{2\,\mathrm{sym}\left(s\rho L[I - s\rho(A+BL)]^{-1}BH\right) + H\right\},$$

$$\Theta_3 \triangleq 2\,\mathfrak{Re}\left\{2\,\mathrm{sym}\left[(I - s\rho(A+BL))^{-1}\mathcal{L}_1(H)\right] - \mathcal{L}_1(H)\right\},$$

and $s \triangleq \exp(2\pi j\omega)$.

The function $\mathcal{F}_2$ is defined as

$$\mathcal{F}_2(\omega, H) \triangleq 2\,\mathfrak{Re}\left\{2\,\mathrm{sym}\left[(I - s\rho\mathcal{A})^{-1}\mathcal{L}_2(H)\right] - \mathcal{L}_2(H)\right\}. \tag{47}$$

Furthermore, one optimal (not necessarily unique) H_* of Problem (45) is of the form

$$H_* = hh^H, \tag{48}$$

where $h \in \mathbb{C}^p$. The corresponding HMM is given by

$$\xi_{k+1} = \rho \begin{bmatrix} \cos 2\pi\omega_* & -\sin 2\pi\omega_* \\ \sin 2\pi\omega_* & \cos 2\pi\omega_* \end{bmatrix} \xi_k + \psi_k, \quad \zeta_k = \begin{bmatrix} \sqrt{2}h_r & \sqrt{2}h_i \end{bmatrix} \xi_k, \tag{49}$$

where h_r, $h_i \in \mathbb{R}^p$ are the real and imaginary part of h respectively and $\Psi = \mathrm{Cov}(\psi_k) = (1-\rho^2)I$.

Remark 11. *By (44), $\Gamma_*(d)$ can be seen as a sinusoidal signal with a decay factor ρ, where ω_* and H_* can be interpreted as the optimal frequency and direction respectively. Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are linear with respect to H, when ω is fixed, (45) is a semidefinite programing problem and hence can be solved efficiently. Therefore, (45) can be solved in two steps by first calculating the optimal signal direction for every frequency $0 \le \omega \le 0.5$ and then searching over all possible frequencies ω. In practice, (45) can be solved for enough sample frequencies to obtain a near optimal watermarking signal.*

It is worth noticing that regardless of the dimensions of the physical system n or the control input p, the dimension of the hidden ξ_k is always 2, which is desirable from a computational perspective when dealing with a high-dimensional linear system.

Numerical Example

This section illustrates the utility of the watermarking scheme by analyzing detection performance on a control system, with parameters

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}. \tag{50}$$

The cost matrices in this system, W and U, are equal to the identity. The covariance matrices, Q and R, are equal to 0.8 times the identity and the identity respectively. As a result, the eigenvalues of $\mathcal{A}$ are -0.339 and -0.105. Consequently, $\mathcal{A}$ is stable, thus motivating the use of a watermark signal for detection. Two watermarking designs are analyzed. First, a stationary watermark is generated using (49) where $\rho = 0.6$. In the second case, an i.i.d. Gaussian process is considered, similar to the design presented in [13, 14]. Designing a stationary Gaussian watermark requires solving a semidefinite program for a set of frequencies sampled in $0 \leq \omega \leq 0.5$. A step size of 0.01 is chosen for this system, which requires solving 51 semidefinite programs. On a Macbook Pro with a 2.4 GHz processor, solving all 51 semidefinite programs takes 12.9 seconds using CVX [21, 22].

First, the asymptotic detection rate $\lim_{k\to\infty} \beta_k$ versus the false alarm rate α for each design is plotted in Fig 3. The additional cost ΔJ imposed by the watermark is 10 for each design, roughly 40 percent of the optimal cost $J^* = 23.1$. The relationship between the asymptotic detection rate and false alarm rate is again considered in Fig 4. Here, α is chosen to be less than 0.1, which is typical for real systems, where the cost considerations of investigating possible attacks make it undesirable to have frequent false alarms during normal operation. The stationary watermarking design offers a visible improvement in the asymptotic rate of detection over an i.i.d. design. The percent improvement in asymptotic detection rate $\lim_{k\to\infty} \beta_k$ of the stationary Gaussian design with $\rho = 0.6$ over the i.i.d. approach is explicitly examined in Fig 5 for $\alpha \leq 0.1$. It can be seen that the stationary watermark achieves its best relative performance for α in this range. In fact, a 60 percent improvement over the i.i.d. design in the asymptotic rate of detection is obtained when $\alpha \approx 0.005$ and $\rho = 0.6$.

Fig 6 and Fig 7 illustrate the tradeoff between the asymptotic detection rate $\lim_{k\to\infty} \beta_k$ and the LQG cost ΔJ for $\Delta J \leq 100$ and $\Delta J \leq 20$ respectively. For this simulation, the false alarm rate α is fixed to be

0.02. For practical systems, ΔJ needs to be carefully chosen to balance the control cost and the detection performance.

Fig 8 shows the detection rate as a function of time k where $\Delta J = 10$ for the watermarking approaches and $\alpha = 0.02$. In this scenario, detection performance in the absence of physical watermarking is also considered. For this case, a χ^2 detector is used. It is assumed that the attacker gathers measurements from $-50 \leq k \leq -1$ and replays these measurements from $0 \leq k \leq 49$. For all chosen designs, the probability of detection quickly rises to a maximum detection rate at $k = 0$ due to a mismatch between the expected and received measurements at the beginning of a replay attack. However, since $\mathcal{A}$ is stable, the detection rate quickly decreases back to false alarm rate without watermarking. Meanwhile, in the watermarking strategies β_k converges quickly. As a result, it is reasonable to design the watermark signal to optimize the asymptotic detection performance.

Finally, Fig 9 examines the relationship between the expected time of detection and the additional LQG cost ΔJ when $\alpha = 0.02$. In the absense of physical watermarking, which corresponds to $\Delta J = 0$, the expected time of detection is roughly given by $k = 34.3$. Watermarking strategies can significantly reduce the time of detection. For instance, for $\Delta J = 10$, the expected time of detection for the stationary watermark is $k = 5.82$ and the expected time of detection for the i.i.d. watermark is $k = 6.27$.

Conclusion

In this article, a replay attack against control systems was defined. Specifically, the adversary could record a sequence of sensor measurements and later deliver these previous outputs to the system operator. If the system is operating in steady state, it was shown that the replayed outputs are statistically identical to the outputs of the system under normal operation. It was noted that for some control systems, the classical estimation, control, and failure detection strategy is not resilient to a replay attack. In these systems, an authenticating watermarked input was super-imposed on the LQG optimal input, providing improved detection at the expense of control performance. The watermarked input was assumed to be stationary and Gaussian, extending previous results, which only considered the i.i.d. case. An optimal Neyman-Pearson detector was given to determine if the system is under attack. Furthermore, a methodology to select the

statistical properties associated with the watermark was provided based on the tradeoff between desired detection performance and allowable control performance loss. In addition, an algorithm to generate a watermark with these statistical properties was provided. Simulations were carried out to examine asymptotic detection performance as a function of the rate of false alarms, asymptotic detection performance as a function of cost to the system, detection performance as a function of time, and expected time of detection as a function of cost. Future work consists of applying the watermarking techniques to detect other attacks, where the adversary cannot perfectly replicate the effect of the watermark in the sensor measurements.

Appendix

Proof of Theorem 2. The "healthy" control system follows

$$\begin{bmatrix} x_{k+1} \\ e_{k+1} \end{bmatrix} = \begin{bmatrix} A+BL & -BL \\ 0 & A-KCA \end{bmatrix} \begin{bmatrix} x_k \\ e_k \end{bmatrix} + \begin{bmatrix} I & 0 \\ I-KC & -K \end{bmatrix} \begin{bmatrix} w_k \\ v_{k+1} \end{bmatrix} + \begin{bmatrix} B\zeta_k \\ 0 \end{bmatrix}, \tag{51}$$

and

$$u_k = L\hat{x}_k + \zeta_k = Lx_k - Le_k + \zeta_k. \tag{52}$$

Since the control system is closed-loop stable, $\{x_k\}$, $\{e_k\}$ and $\{u_k\}$ are all stationary Gaussian processes. Hence,

$$J = \mathbb{E}(x_1^T W x_1 + u_1^T U u_1) = \operatorname{tr}(W \operatorname{Cov}(x_1)) + \operatorname{tr}(U \operatorname{Cov}(u_1)).$$

By (51),

$$x_1 = l_1(w_0, w_{-1}, \dots, v_0, v_{-1}, \dots) + \sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i},\ e_1 = l_2(w_0, w_{-1}, \dots, v_1, v_0 \dots),$$

where l_1 and l_2 are linear functions. As a result,

$$u_1 = l_3(w_0, w_{-1}, \dots, v_1, v_0, \dots) + L\sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i} + \zeta_1,$$

where l_3 is another linear function. Since the watermark signal is independent from the process noise $\{w_k\}$ and sensor noise $\{v_k\}$,

$$\operatorname{Cov}(x_1) = \operatorname{Cov}\left(l_1(w_0, w_{-1}, \dots, v_0, v_{-1}, \dots)\right) + \operatorname{Cov}\left(\sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i}\right),$$

and

$$\operatorname{Cov}(u_1) = \operatorname{Cov}\left(l_3(w_0, w_{-1}, \dots, v_1, v_0, \dots)\right) + \operatorname{Cov}\left(L\sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i} + \zeta_1\right).$$

By the fact that when $\zeta_k = 0$, the optimal LQG cost is J^* and

$$J = J^* + \Delta J,$$

where

$$\Delta J = \operatorname{tr}\left[W \operatorname{Cov}\left(\sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i}\right)\right] + \operatorname{tr}\left[U \operatorname{Cov}\left(L\sum_{i=0}^{\infty} (A+BL)^i B\zeta_{-i} + \zeta_1\right)\right]. \tag{53}$$

Manipulating the RHS of (53) leads to (30), which finishes the proof. □

Proof of Theorem 4. By the definition of KL divergence, it is known that

$$\begin{aligned} D_{KL}(\mathcal{N}_1\|\mathcal{N}_0) =& \frac{1}{2}\operatorname{tr}\left[(\mathcal{P}+\Sigma)\mathcal{P}^{-1}\right] - \frac{m}{2} - \frac{1}{2}\log\det\left[(\mathcal{P}+\Sigma)\mathcal{P}^{-1}\right] + \frac{1}{2}{\mu_k}^T\mathcal{P}^{-1}\mu_k, \\ =& \frac{1}{2}\operatorname{tr}(\Sigma\mathcal{P}^{-1}) - \frac{1}{2}\log\det(I+\Sigma\mathcal{P}^{-1}) + \frac{1}{2}\operatorname{tr}(\mu_k{\mu_k}^T\mathcal{P}^{-1}). \end{aligned}$$

Take the expectation on both sides. It is easy to verify that $\Sigma = \mathbb{E}\,\mu_k{\mu_k}^T$, which proves (37).

Now assume that the eigenvalues of $\Sigma\mathcal{P}^{-1}$ are $\lambda_1, \cdots, \lambda_m$. As a result,

$$\operatorname{tr}(\Sigma\mathcal{P}^{-1}) = \sum_{i=1}^{m}\lambda_i,$$

and

$$\log\det(I+\Sigma\mathcal{P}^{-1}) = \sum_{i=1}^{m}\log(1+\lambda_i).$$

Since $\mathcal{P}$ is positive semidefinite, there exists a positive semidefinite matrix $\mathcal{P}^{1/2}$, where $\mathcal{P}^{1/2}\mathcal{P}^{1/2} = \mathcal{P}$. Hence, $\Sigma\mathcal{P}^{-1}$ shares the same eigenvalues as $\mathcal{P}^{-1/2}\Sigma\mathcal{P}^{-1/2}$, which implies all λ_is are real and nonnegative. As a result, by the concavity of log function, it is known that

$$\log\left[1+\operatorname{tr}(\Sigma\mathcal{P}^{-1})\right] \le \log\det(I+\Sigma\mathcal{P}^{-1}) \le m\log\left(1+\frac{\operatorname{tr}(\Sigma\mathcal{P}^{-1})}{m}\right) \le \operatorname{tr}(\Sigma\mathcal{P}^{-1}). \tag{54}$$

The first inequality holds when $\lambda_1 = \operatorname{tr}(\Sigma\mathcal{P}^{-1})$ and $\lambda_2 = \cdots = \lambda_m = 0$. The second inequality holds when $\lambda_1 = \cdots = \lambda_m = \operatorname{tr}(\Sigma\mathcal{P}^{-1})/m$. The third inequality uses the fact that $\log(1+x) \le x$. Combining (54) and (37), (38) holds.

Furthermore, if C is of rank 1, then by (35),

$$\operatorname{rank}(\Sigma\mathcal{P}^{-1}) \le \operatorname{rank}(\Sigma) \le 1.$$

As a result, the first inequality of (54) is tight, which implies that the upper bound in (38) is tight. □

Proof of Theorem 6. The proof is divided into steps.

Step 1 Consider another function $\tilde{\Gamma}$ and prove that this function is indeed an autocovariance function.

Define function $\tilde{\Gamma} : \mathbb{Z} \to \mathbb{R}^{p\times p}$ as

$$\tilde{\Gamma}(d) \triangleq \rho^{-|d|}\Gamma(d). \tag{55}$$

It can be shown that $\tilde{\Gamma}$ is the autocovariance function of the stationary Gaussian distribution,

$$\tilde{\xi}_{k+1} = (A_h/\rho)\,\tilde{\xi}_k + \tilde{\psi}_k,\ \tilde{\zeta}_k = C_h\tilde{\xi}_k,$$

where $\mathrm{Cov}(\tilde{\xi}_0)$ is the solution of the Lyapunov equation,

$$\mathrm{Cov}(\tilde{\xi}_0) = A_h \,\mathrm{Cov}(\tilde{\xi}_0) A_h^T + \Psi,$$

and $\{\tilde{\psi}_k\}$ is an i.i.d. zero-mean Gaussian process with covariance equals to $\mathrm{Cov}(\tilde{\xi}_0) - A_h \,\mathrm{Cov}(\tilde{\xi}_0) A_h^T/\rho^2$.

Step 2 Rewrite the objective function and the contraint of Problem (40) in terms of the Fourier transform $\tilde{\nu}$ of $\tilde{\Gamma}$.

Consider a partition of $[0, 1/2]$ into disjoint intervals $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_q$, where

$$\mathcal{I}_i \bigcap \mathcal{I}_j = \emptyset, \bigcup_{i=1}^{q} \mathcal{I}_i = [0, \frac{1}{2}].$$

Define σ as the maximum length of interval $\mathcal{I}_i$s. By Riemann-Stieltjes integral and Theorem 5, $\tilde{\Gamma}(d)$ can be written as

$$\tilde{\Gamma}(d) = \lim_{\sigma \to 0} 2\,\mathfrak{Re}\left[\sum_{i=1}^{q} \exp(2\pi j d\omega_i)\tilde{\nu}(\mathcal{I}_i)\right],$$

where $\omega_i \in \mathcal{I}_i$. By (35) and (55),

$$\begin{aligned}
\Sigma &= \lim_{\sigma \to 0} C \sum_{i=1}^{q} 2\,\mathfrak{Re}\left\{2\sum_{d=0}^{\infty} \mathrm{sym}\left[\exp(2\pi j d\omega_i)\,(\rho\mathcal{A})^d\,\mathcal{L}_2(\tilde{\nu}(\mathcal{I}_i))\right] - \mathcal{L}_2(\tilde{\nu}(\mathcal{I}_i))\right\} C^T, \\
&= \lim_{\sigma \to 0} C \sum_{i=1}^{q} 2\,\mathfrak{Re}\left\{2\,\mathrm{sym}\left[(I - \exp(2\pi j\omega_i)\rho\mathcal{A})^{-1}\mathcal{L}_2(\tilde{\nu}(\mathcal{I}_i))\right] - \mathcal{L}_2(\tilde{\nu}(\mathcal{I}_i))\right\} C^T, \\
&= \lim_{\sigma \to 0} C \sum_{i=1}^{q} \mathcal{F}_2(\omega_i, \tilde{\nu}(I_i)) C^T.
\end{aligned}$$

Notice that the order of summation and limit changes, which is feasible as $\mathcal{A}$ is stable. As a result,

$$\mathrm{tr}(\Sigma\mathcal{P}^{-1}) = \lim_{\sigma \to 0} \sum_{i=1}^{q} \mathrm{tr}\left[\mathcal{F}_2(\omega_i, \tilde{\nu}(I_i)) C^T \mathcal{P}^{-1} C\right]. \tag{56}$$

Similarly,

$$\Delta J = \lim_{\sigma \to 0} \sum_{i=1}^{q} \mathcal{F}_1(\omega_i, \tilde{\nu}(I_i)). \tag{57}$$

Step 3 Prove that the upper bound for Problem (40) is the optimal value of the objective function of Problem (45).

Since ΔJ and Σ are always nonnegative, for all $\omega \in [0, 1/2]$ and H positive semidefinite,

$$\mathcal{F}_1(\omega, H) \geq 0,\ \mathcal{F}_2(\omega, H) \geq 0. \tag{58}$$

Suppose that the optimal solution of (45) is ω_*, H_* and the optimal value of the objective function is φ. Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are linear with respect to H, it can be shown that

$$\mathcal{F}_1(\omega_*, H_*) = \delta.$$

Hence, for all $\tilde{\nu}(I_i)$ and $\omega_i \in [0, 1/2]$

$$\text{tr}\left[\mathcal{F}_2(\omega_i, \tilde{\nu}(I_i))C^T\mathcal{P}^{-1}C\right] \leq \frac{\varphi}{\delta}\mathcal{F}_1(\omega_i, \tilde{\nu}(I_i)). \tag{59}$$

By (56)-(59), for all watermark signals $\{\zeta_k\}$ with $\Delta J \leq \delta$,

$$\text{tr}(\Sigma\mathcal{P}^{-1}) \leq \varphi.$$

Step 4 Prove that the upper bound is tight.

Consider the point mass measure $\tilde{\nu}_*$,

$$\tilde{\nu}_*(S_B) = H_* \mathbb{I}_{\{\omega_* \in S_B\}} + \overline{H_*}\, \mathbb{I}_{\{-\omega_* \in S_B\}},$$

where $\mathbb{I}$ is the indicator function. It can be shown that $\Gamma_*(d)$ is generated by $\tilde{\nu}_*$. Furthermore, by (56) and (57), the corresponding $\Delta J = \delta$ and $\text{tr}(\Sigma\mathcal{P}^{-1}) = \varphi$. Hence, $\Gamma_*(d)$ achieves the upper bound of Problem (40). Now it only remains to prove that $\Gamma_*(d)$ can be generated by an HMM with $\rho(A_h) \leq \rho$. Notice that the boundary of the cone of positive semidefinite Hermitian matrices is of the form hh^H. Furthermore, since $\mathcal{F}_1$ and $\mathcal{F}_2$ are linear with respect to H, for fixed ω, the optimization problem (45) attains its maximum on the boundary of the cone (through it is possible that an interior point is also optimal), which proves (48). As a result,

$$H_* = (h_r + jh_i)(h_r^T - jh_i^T) = h_r h_r^T + h_i h_i^T - j(h_r h_i^T - h_i h_r^T).$$

It can be shown that the watermark signal $\{\zeta_k\}$ generated by the HMM (49) follows (44), which proves that (44) is the optimal autocovariance function for Problem (40).

□

References

[1] R. Langner, "To kill a centrifuge: A technical analysis of what stuxnet's creators tried to achieve," Langner Communications, Tech. Rep., November 2013. [Online]. Available: www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf

[2] A. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, pp. 601–611, Nov 1976.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] S. Sundaram, M. Pajic, C. Hadjicostis, R. Mangharam, and G. J. Pappas, "The wireless control network: monitoring for malicious behavior," in *IEEE Conference on Decision and Control (CDC)*, Atlanta, Georgia, Dec 2010, pp. 5979–5984.

[9] S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[10] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, Jan 2012.

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

[12] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *Proc. 49th IEEE Conf. Decision and Control (CDC)*, Atlanta, Georgia, 2010, pp. 5967–5972.

[13] 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.

[14] 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.

[15] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, 1986.

[16] 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, Sep. 1971.

[17] P. E. Greenwood and M. S. Nikulin, *A guide to chi-squared testing*. John Wiley & Sons, Apr. 1996.

[18] L. L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation And Time Series Analysis*. Addison-Wesley Pub. Co., 1991.

[19] T. Chonavel and J. Ormrod, *Statistical Signal Processing: Modelling and Estimation*, ser. Advanced Textbooks in Control and Signal Processing. SPRINGER VERLAG GMBH, 2002.

[20] P. Delsarte, Y. Genin, and Y. Kamp, "Orthogonal polynomial matrices on the unit circle," *IEEE Transactions on Circuits and Systems*, vol. 25, no. 3, pp. 149–160, 1978.

[21] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," www.cvxr.com/cvx, Sep. 2013.

[22] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, vol. 371, pp. 95–110.

[S1] D. P. Fidler, "Was stuxnet an act of war? Decoding a cyberattack," *IEEE Security & Privacy*, vol. 9, no. 4, pp. 56–59, 2011.

[S2] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *37th Annual Conference on IEEE Industrial Electronics Society.* Melbourne, Australia: IEEE, 2011, pp. 4490–4494.

[S3] T. M. Chen, "Stuxnet, the real start of cyber warfare? [editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[S4] R. Langner, "Stuxnet: dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[S5] P. Syverson, "A taxonomy of replay attacks [cryptographic protocols]," in *Computer Security Foundations Workshop VII*, Franconia, New Hampshire, 1994, pp. 187–191.

[S6] S. Kullback and R. Leiber, "On information and sufficiency," in *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[S7] S. Eguchi and J. Copas,"Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma," in *Journal of Multivariate Analysis*, vol. 97, no. 9, pp. 2034–2040.

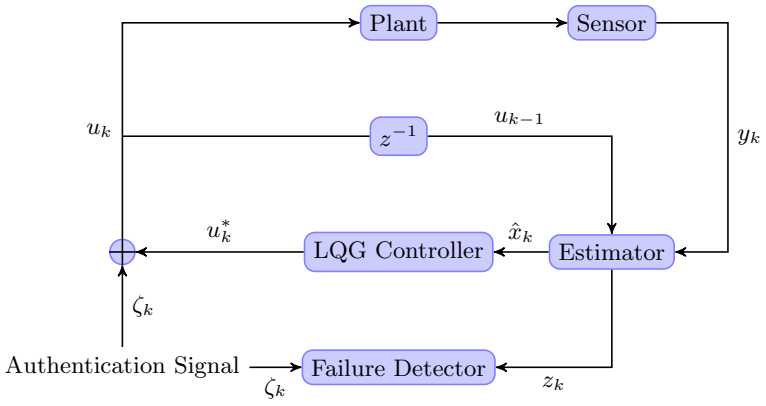

Figure 1: Diagram of the control system under normal operation. In this system, no adversary is present and as a result, the watermark input is present in the sensor outputs. By confirming the presence of a watermark in the sensor measurements, the failure detector can verify that the system is not under a replay attack.

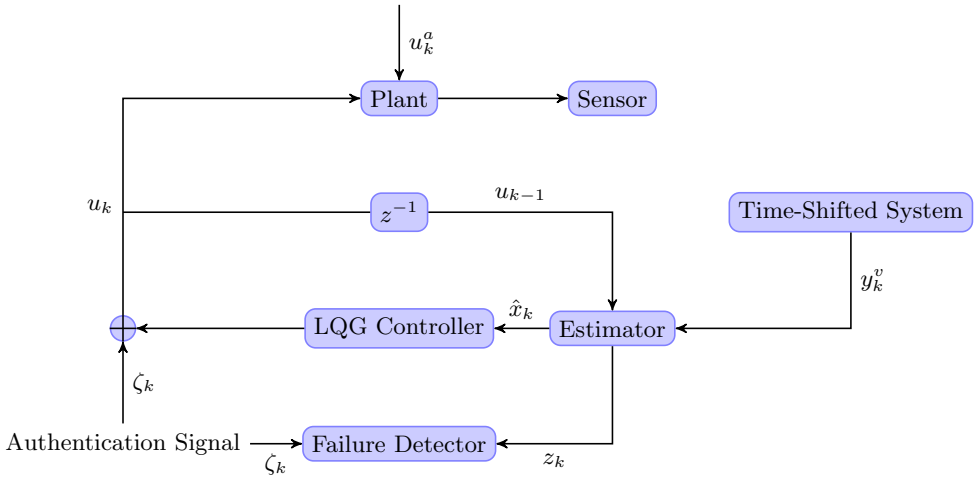

Figure 2: Diagram of the control system under attack. Here, the adversary performs a replay attack, providing replayed outputs y_k^v to the system operator while injecting a potentially damaging input to the system. When under attack, the watermark is asymptotically independent of the replayed sensor measurements. The failure detector leverages this independence to determine if there has been a replay attack.

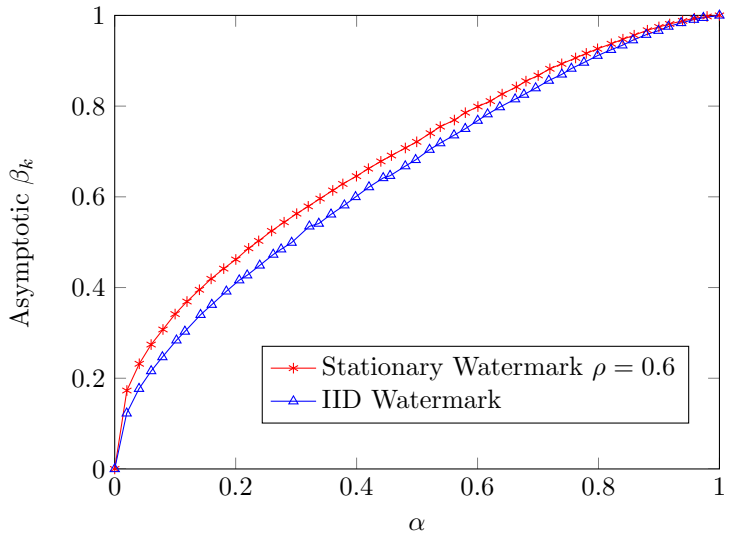

Figure 3: $\lim_{k\to\infty} \beta_k$ as a function of α for a stationary watermark with $\rho = 0.6$, and an independent and identically distributed (i.i.d.) watermark with $\Delta J = 10$. This figure shows that the use of a stationary gaussian watermark with $\rho = 0.6$ visibly increases asymptotic detection performance as the rate of false alarm varies, providing improvements over the i.i.d. approach where ρ is equivalently 0. With respect to the optimization problem (40), increasing ρ corresponds to increasing the set of feasible autocovariance functions for the stationary watermark.

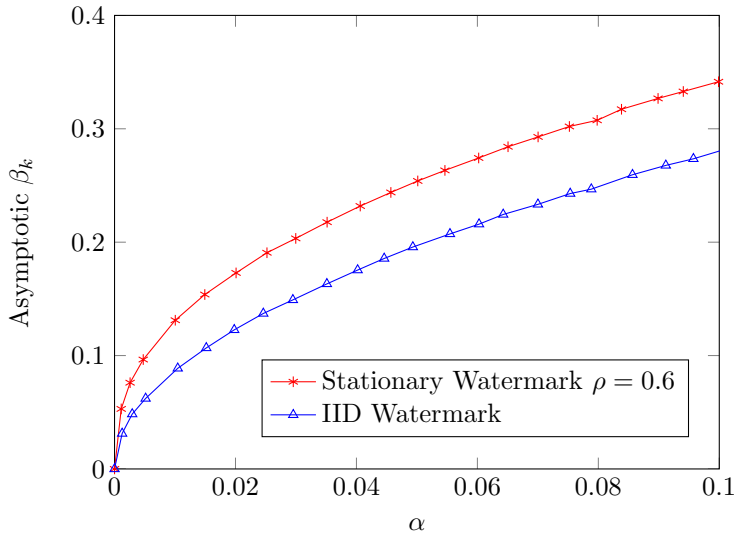

Figure 4: $\lim_{k\to\infty} \beta_k$ as a function of α for $\alpha \leq 0.1$ and $\Delta J = 10$. It is desirable to implement detectors with infrequent false alarms in real life systems and consequently detection performance in this region of operation is essential. The stationary watermarking scheme with $\rho = 0.6$ obtains its best relative performance in comparison to independent and identically distributed (i.i.d.) watermarking schemes when the probability of false alarm approaches 0.

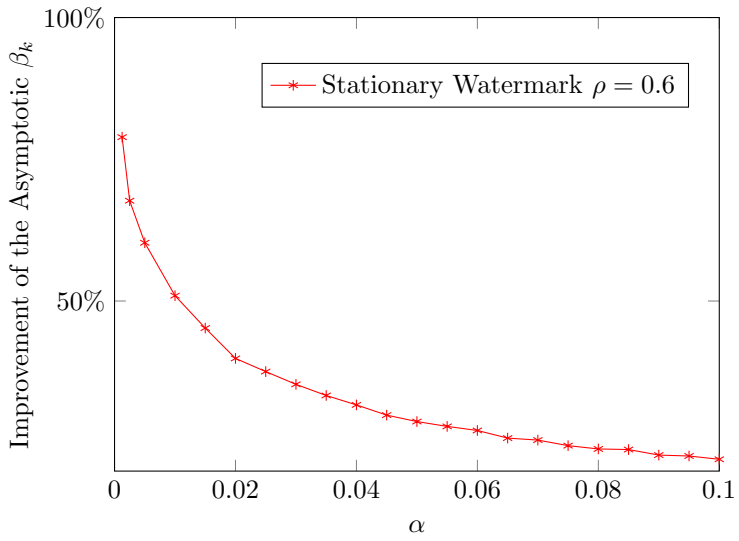

Figure 5: Percentage improvement in $\lim_{k\to\infty} \beta_k$ over the independent and identically distributed (i.i.d.) design versus α for a stationary watermarking scheme with $\rho = 0.6$ and $\Delta J = 10$. This figure shows that the stationary watermarking scheme with $\rho = 0.6$ can offer up to a 70 percent improvement in asymptotic detection performance when the probability of false alarm is near 0

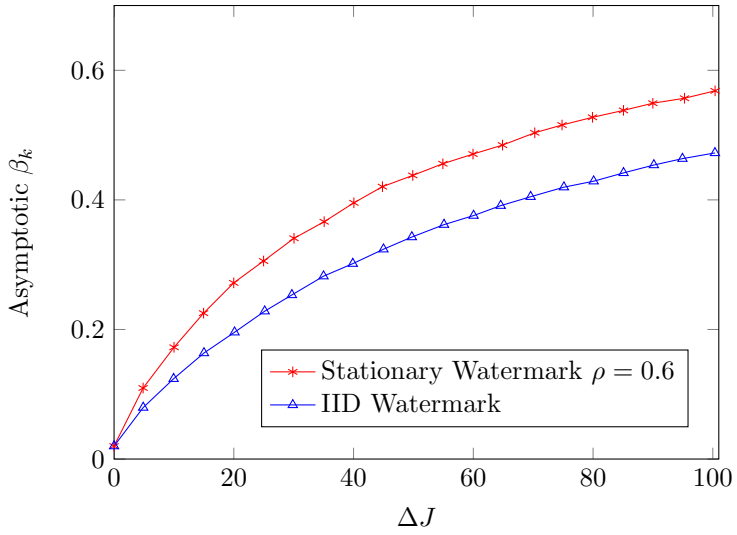

Figure 6: $\lim_{k\to\infty} \beta_k$ versus ΔJ for a stationary watermark with $\rho = 0.6$ and an independent and identically distributed (i.i.d.) watermark. Here, $\alpha = 0.02$. This figure shows that as more control effort is expended, the rate of detection increases. In particular, additional linear-quadratic-Gaussian (LQG) cost corresponds to increasing the magnitude of the watermark's autocovariances. Through the dynamics of the system, watermarks with larger autocovariances increase descrepances between the replayed sensor outputs and the expected sensor outputs, thus resulting in a higher probability of detection.

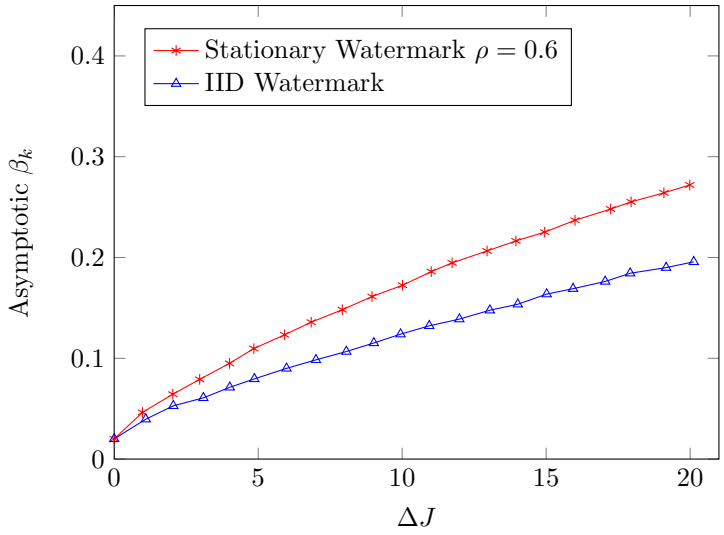

Figure 7: $\lim_{k\to\infty} \beta_k$ versus ΔJ for a stationary watermarks with $\rho = 0.6$ and an independent and identically distributed (i.i.d.) watermark. Here, $\alpha = 0.02$ and $\Delta J \leq 20$, which is roughly 86% of the optimal linear-quadratic-Gaussian (LQG) cost. It is desirable to implement a watermark with limited additional LQG cost in real systems to maintain performance in the control system. This figure illustrates the tradeoff between control and detection performance within this desired region of operation.

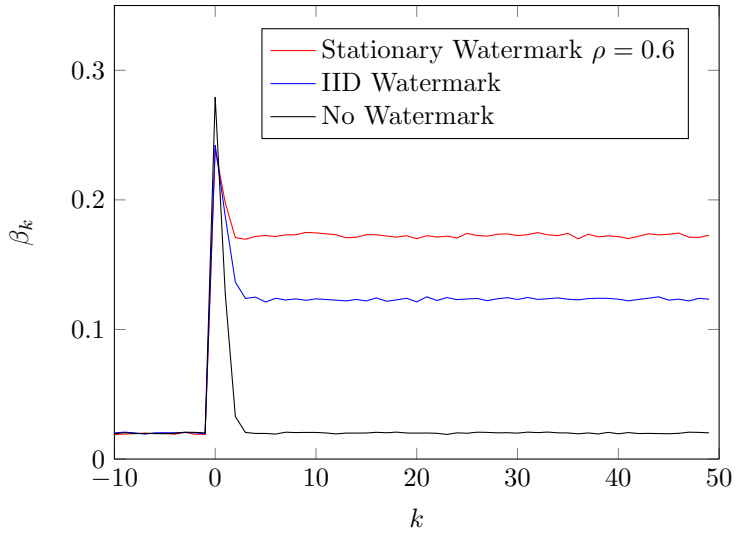

Figure 8: β_k versus time k for a stationary watermark with $\rho = 0.6$, an independent and identically distributed (i.i.d.) watermark, and no watermark. For watermarking schemes, $\Delta J = 10$, and $\alpha = 0.02$ for all schemes. When the attack begins at time $k = 0$, the detection rate quickly increases due to a mismatch in the expected and received measurements before converging quickly to the asymptotic detection rate. Without watermarking the asymptotic detection rate is the false alarm rate. Since the rate of detection quickly converges quickly to its asymptotic value for each design, it is reasonable to design a watermark to optimize asymptotic detection performance.

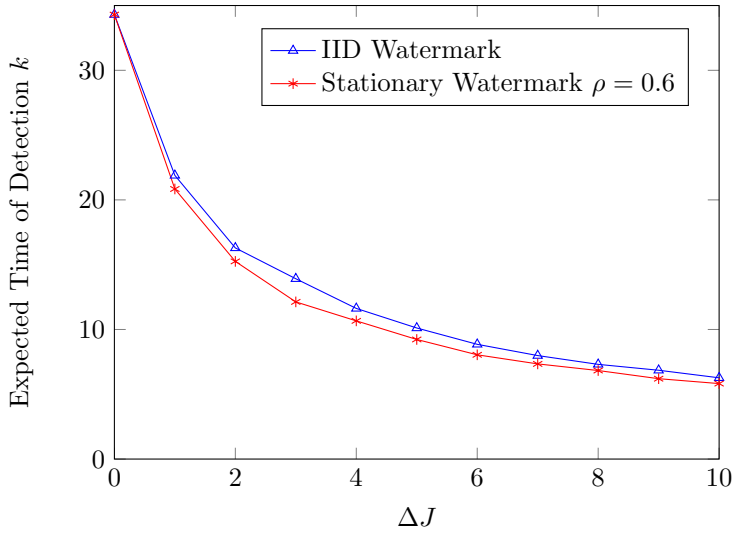

Figure 9: Expected time of detection versus ΔJ for a stationary watermark with $\rho = 0.6$, and an independent and identically distributed (i.i.d.) watermark. The probability of false alarm is fixed to be $\alpha = 0.02$. In the absense of physical watermarking, which corresponds to $\Delta J = 0$, the expected time of detection is $k = 34.3$. However, physical watermarks can significantly reduce the expected time of detection. For instance, when $\Delta J = 10$, the the expected time of detection for the i.i.d. watermark is $k = 6.27$ and the expected time of detection for the stationary watermark with $\rho = 0.6$ is $k = 5.82$. Moreover, for the given range of ΔJ, the stationary watermark with $\rho = 0.6$ allows the detector to identify replay attacks on average earlier than in the case of an i.i.d watermark.

Reference	Type of System	Noise	Attack Models	Defense Mechanisms
[2]	control system	noisy	faults	filters, voting schemes, hypothesis testing, residue detectors
[4]	static power grid	noisy	false data injection (sensor attack)	residue detector
[5]	static power grid	noisy	false data injection (sensor attack)	residue detector
[6]	control system	none	arbitrary sensor and actuator attacks	detection and identification filters
[7]	control system	none	arbitrary sensor and actuator attacks	optimization decoder
[8]	wireless control network	none	malicious nodes with arbitrary state attacks	intrusion detector, output estimation
[9]	distributed network	none	malicious nodes with arbitrary state attacks	combinatorial estimator
[10]	consensus network	none	malicious or faulty nodes	detection and identification filters
[11]	control system	noisy	dynamic false data injection (sensor attack)	residue detector
[12]	sensor network	noisy	dynamic false data injection (sensor attack)	residue detector
[13, 14]	control system	noisy	replay attack	physical watermarking, χ^2, correlation detectors

Table 1: A summary of related works in control system security. The works are characterized by the system models considered, adversary attack models, and defense mechanisms to detect and identify adversaries. Observe that this table does not exhaustively summarize all contributions in control system security and apologies are extended for any omissions

Sidebar 1: The Stuxnet Attack

Stuxnet was a complex malware that infected uranium enrichment facilities in Iran, reportedly causing damage to approximately 1,000 centrifuges at these plants [S1]. From a cyber perspective, Stuxnet was sophisticated, exploiting four previously unknown or zero-day Microsoft Windows vulnerabilities. Additionally, Stuxnet used the first known programmable logic controller (PLC) rootkit, malicious stealthy code injected into controllers, which can hide its own existence [S2]. Moreover, to reduce the chance of being detected, Stuxnet was signed by two stolen certificates from chip manufacturers Realtek and JMicron [S3]. The attack itself was targeted with the worm designed for local distribution, mainly through USB sticks or local networks. While Stuxnet eventually reached 100,000 computer systems, only specific controllers were targeted by the malware [S4]. In particular, Stuxnet targeted PLCs manufactured by Siemens and installed malicious code only after checking that model numbers, configuration details, and program code matched its target. In the first version of Stuxnet, gas pressure of uranium hexafluoride in the centrifuge is increased, which can cause stress to rotors. In a second version, the worm varies rotational speeds. When operating near certain critical speeds or harmonics, this attack can cause the rotor to vibrate or even break [1].

Another question to consider is how Stuxnet evaded detection from verification systems, which measure the system state. According to [1], rotor speed is not a controlled variable and is unlikely to be measured. Furthermore, a monitoring application would have seen rotor speed values prior to an attack since legitimate control logic reading the frequency converter was suspended during the second version of Stuxnet. In the first version, legitimate controller code continued to run but was isolated from the true dynamics of the system. Stuxnet intecepted all input and output signals. To cause damage, Stuxnet implemented control logic causing pressure in centrifuges to increase. To prevent detection, Stuxnet also delivered previously recorded measurements to the SCADA system, performing a replay attack. Replay attacks are commonly considered in information security where adversaries masquerade as a trusted party by repeating a valid data transmission, which they have intercepted [S5]. In information security, replay attacks are thwarted by ensuring a message is current through session tokens or timestamps or by authenticating the sender. Since the Stuxnet attack is linked to a physical process, as opposed to residing strictly in the cyber realm, this article proposes a system theoretic method to verify the "freshness" of sensor outputs, physical watermarking.

Sidebar 2: The Kullback-Liebler Divergence

The Kullback-Liebler divergence, first described in [S6], is a measure of the difference between two distributions $P_1(z)$ and $P_0(z)$. For continuous probability density functions P_1 and P_0, the KL divergence is given as

$$D(P_1 \| P_0) = \int_z P_1(z) \log \left(\frac{P_1(z)}{P_0(z)} \right) dz. \tag{S3}$$

It can be shown that $D(P_1 \| P_0) \geq 0$. Moreover, equality holds if and only if $P_1(z) = P_0(z)$ for almost surely all z. Thus, if the distribution $P_1(z)$ is close to $P_0(z)$, the KL divergence likely approaches 0.

The KL divergence between distributions P_1 and P_0 can be related to the Neyman-Pearson detector associated with a binary hypothesis test. Here, consider $P_1(z)$ to be the distribution of the observations z under the alternative hypothesis H_1 and $P_0(z)$ to be the distribution of the observations z under the null hypothesis H_0. The optimal Neyman-Pearson detector is a threshold detector on the log likelihood $l(z) = \log \left(\frac{P_1(z)}{P_0(z)} \right)$ (S2), where if $l(z)$ is greater than a constant c the alternative hypothesis is chosen. Observe that the KL divergence $D(P_1 \| P_0)$ satisfies

$$D(P_1 \| P_0) = \mathbb{E}[l(z) | H_1]. \tag{S4}$$

Thus, maximizing the KL divergence over a subset of possible distributions P_1 potentially increases the probability of an observation z such that $l(z) > c$, when the alternative hypothesis is true. As a result, the probability of detection also increases. For additional discussion of the relationship between the KL divergence and Neyman Pearson lemma, see [S7].

Author Biographies

Yilin Mo is a postdoctoral researcher in the Department of Control and Dynamical Systems at the California Institute of Technology. He received the Ph.D. degree In Electrical and Computer Engineering from Carnegie Mellon University in 2012 and the Bachelor of Engineering degree from Department of Automation, Tsinghua University in 2007. His research interests include secure control systems and networked control systems, with applications in sensor networks.

Sean Weerakkody received the B.S. degree in Electrical Engineering and Mathematics from the University of Maryland, College Park, in 2012. He is currently working toward the Ph.D. degree at the Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh PA. His research interests include security in cyber-physical systems and distributed estimation with applications to sensor networks.

Bruno Sinopoli received the Dr. Eng. degree from the University of Padova in 1998 and the M.S. and Ph.D. in Electrical Engineering from the University of California at Berkeley, in 2003 and 2005 respectively. After a postdoctoral position at Stanford University, he joined the faculty at Carnegie Mellon University where he is an associate professor in the Department of Electrical and Computer Engineering with courtesy appointments in Mechanical Engineering and in the Robotics Institute. He was awarded the 2006 Eli Jury Award for outstanding research achievement in the areas of systems, communications, control and signal processing at U.C. Berkeley, the 2010 George Tallman Ladd Research Award from Carnegie Mellon University and the NSF Career award in 2010. His research interests include networked embedded control systems, distributed estimation and control with applications to wireless sensor-actuator networks and cyber-physical systems security.