Active Detection for Exposing Intelligent Attacks in Control Systems

Sean Weerakkody Omur Ozel Paul Griffioen Bruno Sinopoli

***Abstract*— In this paper, we consider approaches for detecting integrity attacks carried out by intelligent and resourceful adversaries in control systems. Passive detection techniques are often incorporated to identify malicious behavior. Here, the defender utilizes finely-tuned algorithms to process information and make a binary decision, whether the system is healthy or under attack. We demonstrate that passive detection can be ineffective against adversaries with model knowledge and access to a set of input/output channels. We then propose active detection as a tool to detect attacks. In active detection, the defender leverages degrees of freedom he has in the system to detect the adversary. Specifically, the defender will introduce a physical secret kept hidden from the adversary, which can be utilized to authenticate the dynamics. In this regard, we carefully review two approaches for active detection: physical watermarking at the control input, and a moving target approach for generating system dynamics. We examine practical considerations for implementing these technologies and discuss future research directions.**

I. Introduction

Cyber-Physical Systems (CPSs), engineered systems which apply computing, communication, and control in physical spaces, continue to evolve. CPSs involve devices with widespread sensing, networking, and processing functionalities and are used in applications that require safety, efficiency, and reliability including manufacturing, water distribution, waste management, health care, and the smart grid.

Ensuring security is a major challenge in CPSs. Our interest lies in detecting integrity attacks, where malicious agents inject inputs at sensors and actuators that drive the dynamics of the system towards undesired paths. Passive detection techniques, where the defender leverages finely-tuned algorithms to make a decision about the health of the system, can be ineffective against intelligent adversaries. As such, we consider active detection, where the defender alters the CPS to reveal otherwise stealthy attacks. We discuss two approaches for active detection and system authentication: physical watermarking at the control input, and a moving target approach for generating system dynamics.

Authentication enables the defender to verify the identity of components and the system as a whole. In CPSs, this operation has to be performed not only in the cyber realm, but also within the framework of the physical dynamics. Indeed, cryptographic tools are often vulnerable to attacks and the extra security dimension in the physical dynamics plays a crucial role in reinforcing security. Physical watermarking and the moving target allow for this extra dimension.

In watermarking a known noisy input, or watermark, is injected into the CPS. It is expected that the effect of this input can be found in the measurement of the true output, due to the dynamics. If an attacker is unaware of the watermark, (s)he cannot adequately emulate the system dynamics. Thus, the watermark acts as a cyber-physical nonce, forcing an attacker to generate outputs unique to the given inputs at a chosen time. In the moving target approach, we aim to address cases where physical watermarking can not provide sufficient security such as when the defender's inputs are compromised or when the adversary leverages model knowledge to inject stealthy attacks. In our first proposal, we introduce extraneous states correlated to the ordinary states of the system. The time varying uncertain dynamics of the extra states can reveal integrity attacks affecting the normal states. As an alternative moving target approach, we examine the formulation of the system dynamics as a hybrid control system that transitions across multiple modes. We conclude by discussing application related challenges for these methods and identifying necessary future directions.

Our research on active detection is motivated significantly by recent research. In particular, previous work investigated stealthy attack scenarios such as zero dynamics attacks [1]–[3], false data injection attacks [4], [5], covert attacks [6], and replay attacks [7]. Alternative active detection methods have also been investigated. Specifically, [8] examines a one time change in the system matrices, [9] investigates coding sensor outputs, and [10] attempts to design controllers which prevent system identification from the attacker's perspective. When compared to the moving target [11], each approach suffers from some drawbacks. In particular, a one time change in parameters [8] neglects system identification capabilities from an attacker, coding outputs does not account for physical attacks on sensors [9], and changing a controller degrades performance [10].

II. Modeling

A. System Model

We assume that our system is modeled as a discrete time linear time invariant (LTI) control system as follows:

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

Here, $x_k \in \mathbb{R}^n$ is the state vector, $u_k \in \mathbb{R}^p$ is the set of control inputs implemented, and $y_k \in \mathbb{R}^m$ is a collection of sensor measurements, all at time k. In addition, $w_k \sim$

S Weerakkody, O. Ozel, P. Griffioen, and B. Sinopoli are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA 15213. Email: {sweerakk, oozel, pgriffi1}@andrew.cmu.edu, brunos@ece.cmu.edu,

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

$\mathcal{N}(0,Q)$ is independent and identically distributed (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 assumed to be stabilizable. While the above model is relatively simple, the methods to be presented can, in many cases, be applied to realistic nonlinear systems.

A Kalman filter is used to obtain minimum mean squared error estimates $\hat{x}_{k|k} = \mathbb{E}[x_k|y_{0:k}]$ and $\hat{x}_{k+1|k} = \mathbb{E}[x_{k+1}|y_{0:k}]$. We assume that the system has been running for a long time (i.e. since $k=-\infty$) so that the Kalman filter has converged to a fixed gain linear estimator.

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k,\ \hat{x}_{k|k} = \hat{x}_{k|k-1} + Kz_k, \tag{2}$$

$$K = PC^T(CPC^T+R)^{-1},\ z_k = y_k - C\hat{x}_{k|k-1}, \tag{3}$$

$$P = APA^T + Q - APC^T(CPC^T+R)^{-1}CPA^T. \tag{4}$$

B. Attack Model

We consider an adversary who is able to perform integrity attacks on a subset of actuators and sensors. Without loss of generality (WLOG) an attack begins at time 0.

$$x_{k+1} = Ax_k + Bu_k + B^a u_k^a + w_k, \tag{5}$$

$$y_k = Cx_k + D^a d_k^a + v_k. \tag{6}$$

When inserting $B^a u_k^a$, $u_k^a \in \mathbb{R}^{p'}$, the attacker has the option of modifying the defender's actuators or inserting his/her own. WLOG, B^a has full column rank. Additionally, assuming an adversary modifies a set of sensors $\{s_1, s_2, \cdots, s_{m'}\}$, we can define $D^a \in \mathbb{R}^{m\times m'}$ entrywise as $D^a_{uv} = \mathbf{1}_{u=s_i, v=i}$. When performing an integrity attack, we assume the adversary's goal is to adversely affect the system without being detected. Such a policy allows an attacker to affect a system for long periods of time without defender interference.

Remark 1: One may consider authenticated encryption to detect integrity attacks; however, this technique fails to detect physical attacks. Physical attacks can violate security while bypassing countermeasures from cyber security. For instance, the secrecy, integrity, and availability of measurements from a temperature sensor can be violated by adding an additional unencrypted sensor, locally heating the sensor, and placing a metal cover over the sensor, respectively.

We assume the defender knows the system model $\mathcal{M} = \{A,B,C,Q,R,\hat{x}_{0|-1}\}$ 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 $d^a_{0:k}$. We assume the defender performs passive detection to detect adversaries. In passive detection, the defender constructs algorithms which leverage his/her information $\mathcal{I}_k$ to make a decision about the system: if it is operating normally $\mathcal{H}_0$ or under attack $\mathcal{H}_1$. In a threshold based detector, this can be formulated as $g(\mathcal{I}_k) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta_k$. As an example, the defender could construct residue detectors based on the parameter z_k, which characterize the difference between observed and expected behavior. Under normal operation, z_k has IID Gaussian distribution $\mathcal{N}(0, CPC^T+R)$. One example is a χ^2 detector

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

Here, the probability of detection β_k and the probability of false alarm α_k are given by

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

While passive detection is often effective against benign faults, powerful adversaries can damage a CPS if passive detection is used alone. We next examine such scenarios.

III. Stealthy Attack Scenarios

We now examine stealthy attacks against passive detectors.

A. Replay Attack

In [7], [12]–[14], a replay attack is considered. Here, the adversary has the ability to read and modify all sensor outputs. The adversary performs the following:

1) (s)he records a long sequence of outputs $y_{0:T'}$.
2) Starting at time T, (s)he replaces y_k with y_{k-T}. Thus, $D^a d_k^a = y_{k-T} - Cx_k - v_k$ for $T \le k \le T+T'$.
3) The attacker adds some harmful input $B^a u_k^a$.

Under certain control policies, replay attacks are asymptotically stealthy so that $\lim_{k\to\infty} \beta_k - \alpha_k = 0$, see Theorem 3 [13]. For instance, if the defender uses state feedback, $u_k = L\hat{x}_{k|k}$, a replay attack is stealthy if $\mathcal{A} \triangleq (A+BL)(I-KC)$ is Schur stable. Additionally, if $u_k = h(y_k)$ for some function h, a replay attack is also stealthy. As an example, the Stuxnet malware [15] in part used a replay attack to hide attacks on centrifuges in uranium enrichment plants in Iran.

B. Model Aware Adversaries

Knowledge of the system model can be a powerful tool for adversaries, allowing them to construct stealthy attack sequences. Dynamics may be well known from first principles or available to malicious insiders. In this section, we will briefly revisit zero dynamics attacks and stealthy false data injection attacks. In a zero dynamics attack [8], the nonzero attack inputs $\{u_k^a\}, \{d_k^a\}$ satisfy the following equation

$$x'_{k+1} = Ax'_k + B^a u_k^a,\quad 0 = Cx'_k + D^a d_k^a, \tag{8}$$

for some $x'_0 \in \mathbb{R}^n$. A nontrivial zero dynamics attack exists as long as A, B^a, C, D^a is not strongly observable [16]. Here, x'_0 must belong to the weakly unobservable subspace $\mathcal{V}(A,B^a,C,D^a)$. $\mathcal{V} \subset R^n$ can be described as the largest subspace [16] for which there are maps F_1 and F_2 satisfying

$$(A+B^aF_1)\mathcal{V} \subseteq \mathcal{V},\quad (C+D^aF_2)\mathcal{V} = 0. \tag{9}$$

A feasible attack sequence is to select $x'_0 \in \mathcal{V}$ and have

$$x'_{k+1} = (A+B^aF_1)x'_k,\quad u_k^a = F_1x'_k,\quad d_k^a = F_2x'_k. \tag{10}$$

A zero dynamics attack is stealthy when the defender has no knowledge of x_0 [8]. Even when the defender has uncertain knowledge, the attack leads to vanishing bias on the residues z_k. In particular let z_k^n be the residue under normal operation

and assume the system is under attack. The residue's bias is given by $\Delta z_k \triangleq z_k - z_k^n$. We have the following result, with proof omitted due to spacing.

Theorem 2: Under a zero dynamics attack,

$$\Delta z_k = -C(A - AKC)^k x_0'. \tag{11}$$

Because (A, C) is detectable, we know that $(A - AKC)$ is stable and thus Δz_k converges to 0. It can be easily seen that in a χ^2 detector, $\lim_{k\to\infty} \beta_k - \alpha_k = 0$. More generally, smaller magnitudes of Δz_k result in poor detection performance. In fact we have the following result from [17].

Theorem 3: Let $0 \leq \delta \leq 1$ and assume $\limsup_{N\to\infty} \frac{1}{2N} \sum_{k=0}^{N-1} \Delta z_k^T (CPC^T + R)^{-1} \Delta z_k \geq \epsilon$. Then, there exists a detector such that

$$\beta_k \geq 1 - \delta, \ \forall k, \ \limsup_{k\to\infty} -\frac{1}{k} \log(\alpha_k) \geq \epsilon. \tag{12}$$

The converse also holds if y_k is ergodic.

Motivated by the relationship of residue biases to detection performance, Mo et al. in [18] and [19] examine stealthy false data injection attacks and integrity attacks where $\Delta z_k^T (CPC^T + R)^{-1} \Delta z_k$ is bounded. The authors investigate the perturbations an adversary can introduce to the state while remaining stealthy.

IV. Active Detection: Watermarking

Previously, we observed that some stealthy attacks can not be detected passively. In this section, we consider active detection using physical watermarking to expose adversaries.

A. Motivation from Cyber Security: Nonces

Let us consider the Needham Schroeder protocol [20], which establishes a session key between 2 users, Alice $\underline{A}$ and Bob $\underline{B}$, by leveraging access to a trusted third party, server $\underline{S}$. In this protocol, Alice shares a session key K_{AB} with Bob by sending $\{K_{AB}, \underline{A}\}_{K_{BS}}$ where K_{BS} is Bob's shared key with $\underline{S}$ and $\{\}_{K^*}$ denotes encryption with key K^*. This message is vulnerable to a replay attack. For instance, suppose Eve $\underline{E}$ recovers an old session key K_{AB}^*. She can replay the message $\{K_{AB}^*, \underline{A}\}_{K_{BS}}$ to Bob. Bob now believes he shares key K_{AB}^* with Alice, when he truly shares a key with Eve. This lets Eve engage in a man in the middle attack.

To counter this attack, Alice receives a nonce or random number, N_B, from Bob encrypted with K_{BS}. After communicating with $\underline{S}$, Alice sends $\{K_{AB}, \underline{A}, N_B\}_{K_{BS}}$ to Bob. The random nonce serves as a challenge to Alice. By including the encrypted nonce in her response to Bob, Alice proves that the message is fresh, and has not been replayed.

B. Physical Watermarking

Motivated by the use of nonces in cyber security, we propose watermarking to detect replay attacks in control systems. A physical watermark, Δu_k, is a secret noisy control input inserted on top of the optimal control input u_k^* to authenticate the system. In particular, u_k is given by

$$u_k = u_k^* + \Delta u_k. \tag{13}$$

Here, the adversary can not read the defender's control input u_k and does not know real time watermarks. Physical watermarking was first introduced in [7] as an IID additive input $\Delta u_k \sim \mathcal{N}(0, \mathcal{W})$. Extensions have been examined in [12]–[14], [21]–[27]. We consider a stationary Gaussian watermark generated by a hidden Markov model [14] below:

$$\begin{aligned} &\zeta_{k+1} = A_h \zeta_k + \psi_k, \quad \psi_k \sim \mathcal{N}(0, \Psi), \\ &\Delta u_k = C_h \zeta_k, \quad \text{Cov}(\zeta_{-\infty}) = A_h \text{Cov}(\zeta_{-\infty}) A_h^T + \Psi, \end{aligned} \tag{14}$$

where A_h is Schur stable with spectral radius less than or equal to a user defined constant ρ. As the watermarks are 0 mean with bounded covariance, the closed-loop system remains stable. The watermarks act as a cyber-physical nonce. Under normal conditions, the watermark will be embedded in the sensor outputs due to the system dynamics, a valid response to this challenge. However, under replay attack, the measurements contain physical responses to an earlier sequence of watermarks. Unable to detect recent watermarks in the sensor outputs, the defender can not verify freshness.

To perform detection, the defender can exploit the difference in distributions of the residue z_k under attack and normal operation. For instance, if $u_k^* = L\hat{x}_{k|k}$, then

$$z_k \sim \mathcal{N}(0, CPC^T + R), \quad \mathcal{H}_0 : \text{no attack}, \tag{15}$$

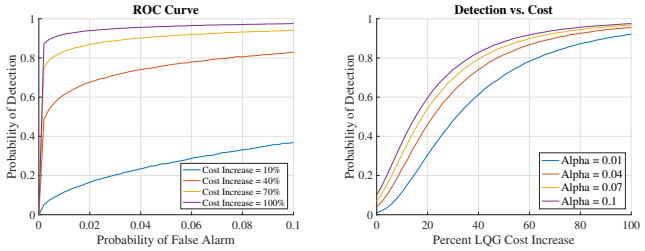
$$z_k \sim \mathcal{N}(\mu_k, CPC^T + R + \Sigma), \quad \mathcal{H}_1 : \text{attack}, \tag{16}$$

where $\mu_k \triangleq -C \sum_{i=-\infty}^{k-1} \mathcal{A}^{k-1-i} B \Delta u_i$ and Σ is a linear increasing (in the semidefinite sense) function of the watermark covariance in the IID case, and a linear increasing function of the autocovariances in the stationary case, see [14]. A χ^2 or Neyman Pearson detector can be used. Alternatively, a correlator detector [13] can be considered. Here, the defender computes an output y_k^* by running a simulated version of the system with the same chosen watermarks, and then calculates $g(\mathcal{I}_k) = -y_k^T y_k^*$. Under replay attack, y_k and y_k^* have no correlation because they are generated by different independent watermarks. However, under normal operation, a positive correlation between y_k and y_k^* exists. The correlation detector can distinguish between faulty scenarios and malicious attacks [12]. Additional asymptotic detectors (with finite statistical approximations) that guarantee zero average distortion power in sensors are proposed in [24], [25].

When deviating from an optimal strategy, performance loss must be weighed. We can use the degrees of freedom in our watermark, the (auto)covariance functions, to balance trade-offs between detection and control. While increasing the size of the covariance improves security, watermarks with larger magnitudes decay control. Semidefinite programs have been proposed in both [13] and [14] to address this trade-off. For instance, the authors in [13] maximize the expected bias inserted into the χ^2 detector subject to an upper bound on additional linear quadratic Gaussian (LQG) costs due to watermarking. Watermarking can be effective in other scenarios. Under certain attack strategies, watermarking can detect adversaries who know the system model [22] and attackers who attempt to learn the model [26], [27].

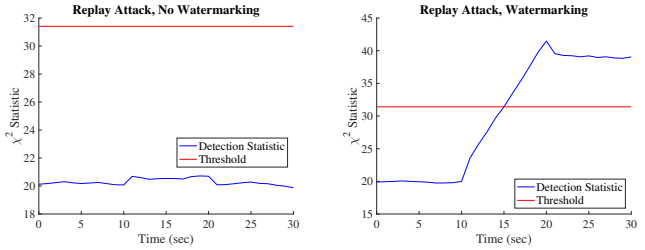
We test physical watermarking on the quadruple tank process, a four state system [28]. The goal is to control the water level of two of four tanks using two pumps. Two sensors measure water heights. We use an LQG controller with weights following suggestions in [29]. Q and R are created by generating a matrix from a uniform distribution, multiplying it by its transpose, and dividing by 100. A stationary Gaussian watermark [14] ($\rho = 0.5$) is inserted. Experiments were averaged over 1000 trials.

In Fig. 1 we examine security and performance trade-offs through relationships between the probability of false alarm, the probability of detection, and the percent increase in LQG cost due to watermarking. A χ^2 detector with window size 10 is implemented for detection. In Fig. 1a, we plot ROC curves at different additional LQG costs. In Fig. 1b, we plot the probability of detection as a function of the additional LQG cost for different fixed false alarm probabilities. In Fig. 2, we plot our χ^2 detection statistic (window size 10) during a replay attack for a system without watermarking (Fig. 2a) and a system with watermarking (Fig. 2b). Replay attacks commence at time 10 sec. The probability of false alarm in Fig. 2 is 0.05 and the LQG cost is increased by 30%.



(a) Detection vs False Alarm (b) Detection vs Cost

Fig. 1. Probability of Detection as a Function of % Increase in LQG Cost and Probability of False Alarm.



(a) No Watermarking (b) Watermarking

Fig. 2. χ^2 Detection Statistic vs Time.

V. Active Detection: Moving Target

Watermarking detection unfortunately fails against model aware zero dynamics attacks and false data injection attacks where the bias on residues Δz_k are independent of the defender's inputs. Accordingly, we investigate a new approach.

A. Motivation from Cyber Security: MACs

In cyber security, message authentication codes or MACs can verify the integrity of a message. Suppose Alice sends message $\underline{m}$ to Bob. Eve can modify the contents of the message before forwarding it to Bob, leaving Bob unable to verify the integrity of $\underline{m}$. To prevent this, Alice appends a MAC $f_{K_{AB}}(\underline{m})$ to the message $\underline{m}$. Here, K_{AB} is a shared secret key between Alice and Bob and $f_{K_{AB}}(\underline{m})$ is a keyed pseudo-random function of the message. Upon receiving $\underline{m}$, Bob computes $f_{K_{AB}}(\underline{m})$. Under normal operation, the MACs Bob receives and computes are identical. Alternatively, if Eve modifies $\underline{m}$ to $\underline{m}'$, she will likely be unable to compute a valid $f_{K_{AB}}(\underline{m}')$ unless she knows K_{AB} or replays a previous message. A replay attack can be prevented by including a timestamp in $\underline{m}$, while K_{AB} is only shared between Alice and Bob. Thus, the MAC assures message integrity.

B. Moving Target

We now introduce the moving target approach, first examined in [11]. Here, an authenticating subsystem with time varying dynamics is introduced on top of the original system:

$$\begin{bmatrix}\tilde{x}_{k+1}\\ x_{k+1}\end{bmatrix} = \begin{bmatrix}A_{1,k} & A_{2,k}\\ 0 & A\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ x_k\end{bmatrix} + \begin{bmatrix}B_k\\ B\end{bmatrix} u_k + \begin{bmatrix}\tilde{w}_k\\ w_k\end{bmatrix}. \quad (17)$$

Moreover, we introduce additional sensors $\tilde{y}_k \in \mathbb{R}^{\tilde{m}}$ to measure the extraneous states,

$$\begin{bmatrix}\tilde{y}_k\\ y_k\end{bmatrix} = \begin{bmatrix}C_k & 0\\ 0 & C\end{bmatrix}\begin{bmatrix}\tilde{x}_k\\ x_k\end{bmatrix} + \begin{bmatrix}\tilde{v}_k\\ v_k\end{bmatrix}. \quad (18)$$

We assume that the process and sensor noise is IID with $\begin{bmatrix}\tilde{w}_k\\ w_k\end{bmatrix} \sim \mathcal{N}(0, \mathcal{Q})$, $\begin{bmatrix}\tilde{v}_k\\ v_k\end{bmatrix} \sim \mathcal{N}(0, \mathcal{R})$. The system matrices are assumed to be IID variables which are independent of the sensor and process noise with distribution

$$A_{1,k}, A_{2,k}, B_k, C_{k+1} \sim f(A_1, A_2, B, C). \quad (19)$$

The sequence of time varying system matrices is determined by a pseudo-random number generator (PRNG), the seed of which is known by the defender, but hidden from the adversary. We consider an attacker who knows (A, B, C, Q, R) and has the ability to read and modify all inputs and outputs $u_{0:k}, y_{0:k}, \tilde{y}_{0:k}$. $\tilde{y}_k$ acts as a cyber-physical MAC, preventing the adversary from constructing a stealthy integrity attack.

In particular, suppose the message $\underline{m}$ corresponds to outputs y_k while the MAC is $\tilde{y}_k$. The MAC $\tilde{y}_k$ is correlated to the message y_k through the state x_{k-1} and the input u_{k-1}. The key is the seed which determines the sequence of system matrices. The defender uses knowledge of y_k and the sequence of system matrices to estimate $\tilde{y}_k$. Under normal operation, $\tilde{y}_k$ and its estimate $\hat{\tilde{y}}_k$ closely agree, as seen by a residue based detector, and as a result the MAC is verified.

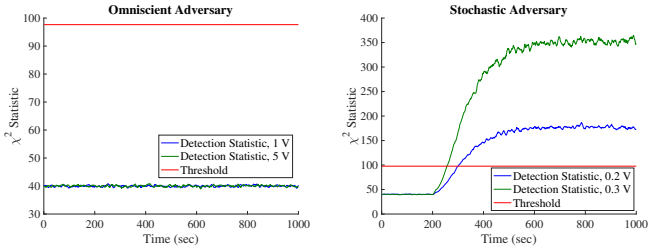
However, suppose an adversary performs integrity attacks using knowledge of (A, B, C, Q, R). The attacker could generate convincing outputs y_k, while biasing the states x_k through a false data injection or zero dynamics attack. At the same time, (s)he will also bias the states $\tilde{x}_k$ and thus the MAC outputs $\tilde{y}_k$ if the time varying matrices are properly chosen. Having no knowledge of the seed, the adversary can not know the time varying matrices. Moreover, the time varying dynamics act as a moving target, hindering system

identification. As a result, the attacker can not generate a convincing cyber-physical MAC output $\tilde{y}_k$. Bounds characterizing how well an adversary can construct outputs y_k and $\tilde{y}_k$ to fool residue detectors are obtained in [11, Theorem 3].

We apply the moving target to the quadruple tank process. 4 extra states and 2 extra outputs are added. The time varying matrices $A_{1,k}, A_{2,k}, B_k, C_{k+1}$ are somewhat sparse (50% of entries nonzero). The non-zero elements follow a multivariate Gaussian distribution with means generated from $U(-0.5, 0.5)$. The covariances of the nonzero parameters are created by generating a matrix from a uniform distribution, multiplying it by its transpose, and dividing by 100.

We consider an adversary who, starting at time 200 sec, adds a constant input (in Volts) to the optimal LQG input and avoids detection by trying to subtract his own influence [11] from the measurements. First, in Figs. 3a, 4a, we assume the attacker knows the time varying system matrices. Secondly, we assume the attacker does not know the realization of $A_{1,k}, A_{2,k}, B_k, C_{k+1}$, but instead performs his attack by sampling the matrices from (19), (Figs 3b, 4b). We plot the χ^2 detector statistic (window 10, $\alpha = 10^{-7}$) in Fig. 3 and system performance in Fig. 4, both averaged over 1000 trials.

Given full knowledge of the system matrices, the attacker can significantly affect water levels while remaining perfectly stealthy. However, with stochastic knowledge of the system matrices, the attack is easily revealed, even for small system perturbations and small α. In practice, the attack can be improved by using the measurements $\tilde{y}_k$ to perform system identification. We expect improvements to be marginal since the system changes at each time step. As a result, characterizing the effectiveness of an attacker who performs machine learning in a scenario where the moving target changes at a lower frequency is an immediate goal. The theory of estimation with stochastically varying parameters [30] might also be useful for developing stronger attack strategies.



(a) Full knowledge (b) Stochastic knowledge

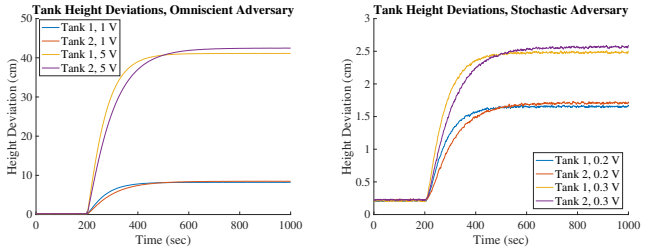
Fig. 3. χ^2 Detection Statistic vs Time for an attacker who subtracts his influence with a) full knowledge of the moving target matrices b) stochastic knowledge of moving target matrices as given by $f(A_1, A_2, B, C)$.

C. Alternative Formulation of Moving Target

Consider the scenario where the plant is modeled as a hybrid system which transitions across multiple modes:

$$x_{k+1} = A_k x_k + B u_k + w_k, \quad y_k = C_k x_k + v_k. \tag{20}$$

We assume that (A_k, C_k) belongs to some finite set $\Gamma \triangleq \{(A(1), C(1)), \cdots, (A(l), C(l))\}$. This alternative moving



(a) Full knowledge (b) Stochastic knowledge

Fig. 4. Mean absolute deviation of heights (cm) of 2 tanks from desired equilibrium.

target formulation was considered in [31]. Here, the system periodically and randomly transitions across modes in Γ, for instance according to an aperiodic and irreducible Markov Chain. The authors consider how this moving target could be used to identify sensor attacks from an adversary whose attack inputs do not leverage knowledge of the prior inputs/outputs. In this case, the goal is not only to detect sensor attacks, but to determine which sensors have been corrupted.

Prior results in deterministic scenarios state that a system must remain observable after removing any $2q$ sensors to identify q sensor attacks [3]. Weerakkody et al. argue [31] that a moving target approach increases the number of sensors that can be identified. Design recommendations are made which guarantee the perfect identifiability of all sensor attacks in deterministic systems. Additionally, in stochastic systems, the authors show they can construct robust fusion based estimators and detectors, which identify sensors that are subject to destabilizing integrity attacks.

VI. Considerations and Future Work

In this section, we discuss next steps towards implementing physical watermarking and the moving target.

A. Physical Watermarking

In physical watermarking, previous work [13], [14], [22], addressed trade-offs between detection and control performance by solving semidefinite programs to obtain watermark (auto)covariances. These semidefinite programs attempt to maximize a statistic related to detection, subject to a constraint on LQG cost. In practice, however, LQG controllers may not be used. Ideally, system designers will formulate convex constraints and objectives, which adequately summarize necessary system goals for different controllers. Alternatively, designers can formulate LQG constraints which best capture performance requirements and perform simulations to verify proper behavior. We must also account for a system's nonlinearity. A watermark should never be so large that it takes a system out of the region of operation.

Designers must consider the composition of watermarks with other security primitives. To implement watermarking effectively, a nonempty subset of the actuators must deliver control inputs which are kept secret from the adversary [22]. For instance, knowledge of the input can be used to completely break security if a model aware adversary uses

this information to generate virtual outputs by simulation. The secrecy of control inputs can be preserved for instance through encryption. Actuators typically have more computational resources than sensors, allowing them to implement cryptographical primitives. Prior work assumes that watermarks are generated by true random number generators. In practice, we are restricted to PRNGs. Thus, the selection of cryptographically secure PRNGs must be done carefully.

As an example, suppose an IID Gaussian watermark is generated by: 1) A linear congruential generator which evolves according to $X_{n+1} = (aX_n + c) \bmod M, 0 \leq a, c, \leq M$. Here, $\frac{X_n}{M}$ approximates a uniform random variable, 2) The Box-Muller transform which converts a uniform random variable to a Gaussian random variable, 3) A linear transformation to obtain a watermark of the appropriate covariance. A malicious insider with model knowledge, and some prior input and output history, can learn the secret key of the PRNG (a, c, M) as follows. 1) Run a linear filter using the input/output history to estimate prior watermarks. The convergence to the true watermarks is exponential. 2) Take the inverse of the linear and Box-Muller transformations to obtain the sequence of uniform random variables. 3) Intelligently search over space of M (as low as 2^{32} possible values), and algebraically solve for a and c. If the key is not changed, the attacker can predict all future watermarks.

We recommend the examination of alternative watermarking designs. In principle, there is no need to restrict watermarks to be Gaussian or stationary. For example, results in [23] suggest that Gaussian watermarks could be optimal against Gaussian attackers. See also [27] for non-stationary and non-Gaussian watermark designs. In our current work, we are evaluating the effectiveness of a watermark obtained by dropping the control input randomly according to IID Bernoulli and Markovian strategies in combination with a Gaussian additive input. From an application perspective, implementations are necessary to validate watermark designs. Ko et al. [32] have examined watermarking in vehicular systems while Rubio-Hernan et al. [27] have experimented on a SCADA testbed. We are investigating watermark implementations with quadrotors and plan testing on a smart grid testbed.

B. Moving Target

Developing a mechanism for generating time varying dynamics is important in the moving target. It will be ideal to leverage the dynamics that already exist in a system, (e.g. the heat released by a reaction or the frictional signatures of mechanical components). Alternatively, we can introduce physical hardware to generate the authenticating subsystem. When devoting resources to develop new hardware, economic costs must be balanced with potential benefits of security against powerful attacks.

We envision that a moving target device could contain simple circuits which take a portion of the state as an input and emit $\tilde{y}_k$ as the output. We can affix such a device to system components (e.g an electric generator or reactor). The time varying behavior can be implemented by using adaptive components in the circuit, for instance variable capacitors and resistors. Such a device would likely also contain a PRNG whose output would determine real time resistances and capacitances and whose seed is known to the defender. In the alternative moving target, we assume that there exist sufficiently many degrees of freedom in the plant, allowing the defender to periodically transfer across modes of operation as determined by a PRNG.

The integration of the moving target with other security primitives has to be investigated. The root of the trust is the seed of our PRNG. Thus, it is desirable to use a cryptographically secure PRNG. Additionally, a secure key sharing protocol [33] must be used to ensure both the plant and operator agree on the secret seed so that the defender can perform proper estimation and detection.

We should consider trade-offs between detection and control performance when implementing the moving target. In [11], it is assumed that the defender does not care about controlling $\tilde{x}_k$. If this assumption is true, then we can utilize optimal control inputs u_k^* derived from $y_{0:k}$ and avoid a loss in performance. Such a design also prevents a control input from leaking information about the time varying system dynamics. However, this assumption should be rigorously verified in practice. The alternative moving target, as a consequence of unplanned transitions, likely does not obtain peak control performance. Theoretical and practical studies must be performed to understand the trade-offs.

Further theoretical developments are also needed. In the original moving target, it is important to examine the properties of the time varying matrices $A_{1,k}, A_{2,k}, B_k, C_{k+1}$. Some matrix parameters may not be time varying and are possibly 0. For instance C_k can be constant and B_k can be 0, without eliminating the advantages of the moving target. Likewise, system transitions need not be at every time step, but can instead occur after a realistic period (accounting for system inertia) that still hampers an attacker's system identification. Moreover, in practice, time varying parameters in a moving target will be sampled from a discrete (though possibly quite rich) distribution. Nonlinear designs of a moving target should also be investigated.

A number of necessary conditions for the design of system matrices can be obtained. For instance, the original states x_k should be observable from the sensors $\tilde{y}_k$ Additionally, given partial channel access for the attacker, the zero dynamics in the nominal system should not remain as zero dynamics when the authenticating subsystem is added.

While ample design suggestions have been provided for the alternative moving target [31], the attack model is limiting. In particular, it is assumed that the attacker does not leverage channel information. In a stochastic scenario, it would be worthwhile to examine what guarantees can be provided when measurements and control inputs are public and used by the adversary.

Ample testing is also required in the moving target. We wish to perform simulations to investigate the bounds characterizing how well an adversary can construct outputs y_k and $\tilde{y}_k$ to fool residue based detectors (Theorem 3, [11]).

We also plan to determine effective attack strategies which utilize system identification. Once we identify an application and build suitable hardware for generating the moving target dynamics, a testing phase could commence.

The development of a unified approach to detection is a clear goal. A defender is not restricted to use only one of these strategies, especially since the proposed approaches vary in terms of their effectiveness for detecting specific attacks. In this respect, it is important to identify an adversarial model which characterizes the resources and knowledge an operator anticipates an attacker could procure. From here, the defender can choose defenses to expose feasible attack strategies. For instance, if the system model is unknown and the inputs are private, a defender may determine that physical watermarking is effective on its own. Alternatively, against stronger adversaries, the moving target and physical watermarking can be strategically combined to detect attacks while limiting costs and performance loss to the system.

Another direction to investigate is a game theoretic formulation of the attack detection problem [21] where an attacker wishes to maximize performance loss while ensuring stealthiness, while the defender wishes to limit the impact of a stealthy attacker through active strategies. Analyzing potential equilibria can illustrate the effectiveness of the proposed approaches against strategic, intelligent adversaries.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

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

[6] R. Smith, "A decoupled feedback structure for covertly appropriating network control systems," in *IFAC World Congress*, Milan, Italy, 2011, pp. 90–95.

[7] 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.

[8] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, 2012, pp. 1806–1813.

[9] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *53rd IEEE Conference on Decision and Control (CDC)*, Los Angeles, California, 2014, pp. 5776–5781.

[10] Y. Yuan and Y. Mo, "Security in cyber-physical systems: Controller design against known-plaintext attack," in *54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5814–5819.

[11] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *54th IEEE Conference on Decision and Control*. IEEE, 2015, pp. 5820–5826. [Online]. Available: https://arxiv.org/abs/1706.08182

[12] 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.

[13] 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.

[14] 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.

[15] 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

[16] H. Trentelman, A. A. Stoorvogel, and M. Hautus, *Control theory for linear systems*. Springer Science & Business Media, 2012.

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

[18] Y. Mo and B. Sinopoli, "Integrity attacks on cyber-physical systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 47–54.

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

[20] R. M. Needham and M. D. Schroeder, "Using encryption for authentication in large networks of computers," *Communications of the ACM*, vol. 21, no. 12, pp. 993–999, 1978.

[21] 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.

[22] 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.

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

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

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

[26] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "Event-triggered watermarking control to handle cyber-physical integrity attacks," in *Nordic Conference on Secure IT Systems*. Springer, 2016, pp. 3–19.

[27] ——, "On the use of watermark-based schemes to detect cyber-physical attacks," *EURASIP Journal on Information Security*, vol. 2017, no. 1, 2017.

[28] 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.

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

[30] T. F. Cooley and E. C. Prescott, "Estimation in the presence of stochastic parameter variation," *Econometrica*, vol. 44, no. 1, pp. 167–184, January 1976.

[31] S. Weerakkody and B. Sinopoli, "A moving target approach for identifying malicious sensors in control systems," in *54th Annual Allerton Conference on Communication, Control, and Computing*, 2016, pp. 1149–1156. [Online]. Available: https://arxiv.org/abs/1609.09043

[32] W.-H. Ko, B. Satchidanandan, and P. Kumar, "Theory and implementation of dynamic watermarking for cybersecurity of advanced transportation systems," in *2016 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2016, pp. 416–420.

[33] L. Eschenauer and V. D. Gligor, "A key-management scheme for distributed sensor networks," in *Proceedings of the 9th ACM conference on Computer and communications security*. ACM, 2002, pp. 41–47.