2019 18th European Control Conference (ECC)
Napoli, Italy, June 25-28, 2019

Sequential Detection of Deception Attacks in Networked Control Systems with Watermarking

Somayeh Salimi, Subhrakanti Dey and Anders Ahlén

Abstract— **In this paper, we investigate the role of a physical watermarking signal in quickest detection of a deception attack in a scalar linear control system where the sensor measurements can be replaced by an arbitrary stationary signal generated by an attacker. By adding a random watermarking signal to the control action, the controller designs a sequential test based on a Cumulative Sum (CUSUM) method that accumulates the log-likelihood ratio of the joint distribution of the residue and the watermarking signal (under attack) and the joint distribution of the innovations and the watermarking signal under no attack. As the average detection delay in such tests is asymptotically (as the false alarm rate goes to zero) upper bounded by a quantity inversely proportional to the Kullback-Leibler divergence(KLD) measure between the two joint distributions mentioned above, we analyze the effect of the watermarking signal variance on the above KLD. We also analyze the increase in the LQG control cost due to the watermarking signal, and show that there is a tradeoff between quick detection of attacks and the penalty in the control cost. It is shown that by considering a sequential detection test based on the joint distributions of residue/innovations and the watermarking signal, as opposed to the distributions of the residue/innovations only, we can achieve a higher KLD, thus resulting in a reduced average detection delay. We also present some new structural results involving the associated KLD and its behaviour with respect to the attacker's signal power and the watermarking signal power. These somewhat non-intuitive structural results can be used by either the attacker to choose their power to minimize the KLD, and/or by the system designer to choose its watermarking signal variance appropriately to increase the KLD. Numerical results are provided to support our claims.**

I. Introduction

Attacks on cyber-physical systems (CPS) can affect the integrity, availability and confidentiality in CPS. Examples range from deception based attacks such as *false-data-injection* [1], sensor and actuator attacks, *replay attacks*, and also *denial of service attacks* [2] on the underlying networked control system (NCS). Deception attacks refer to scenarios where integrity of control packets or measurements are compromised by altering the behaviour of sensors and actuators. In particular, false data injection attacks are introduced by injecting incorrect or misleading measurements or control inputs. Replay attacks are carried out by hijacking the sensors, recording the sensor measurements for a period of time, and then repeating such measurements to the controller while injecting a harmful exogenous signal into the system. On the other hand, denial of service attacks can be carried out by an adversary compromising the availability of resources to the CPS, e.g., by jamming the communication channel. Documented defence mechanisms can range from attack identification and detection, intrusion detection as well as physical watermarking of valid control signals. Most of these defence mechanisms have been developed to tackle specific types of attacks, whereas a generalized unified approach for attack identification and detection is developed by adopting a descriptor system modelling framework for CPS [3] and applications illustrated for power and water networks. Linear state estimation with corrupted measurements has been also studied in [4] where the maximum number of faulty sensors is characterized and a decoding algorithm for detecting corrupted measurements is presented.

The defence mechanism of relevance to this paper is the idea of *physical watermarking of control signals*. Traditionally, digital watermarking has been used extensively in audio and image processing for authentication purposes, where a specific signal is embedded in the transmitted message/document, and is later used to identify the rightful owner of the message. The idea of physical watermarking in NCS is similar, where a random signal is added to the control signal, and under normal operations, the effect of this watermarking signal should be present in the system output. However, when the system is attacked or compromised and sensor measurements are substituted by injection of false data, the expected effect of the watermarking signal will be absent or perturbed, thus leading to a statistical test which can detect the presence of an attacker. Two most recent works that deal with design and analysis of physical watermarking for NCS are [2] and [10]. In [2], the authors consider a linear state space model under a replay attack, and first design an optimal detector based on the Neyman-Pearson test. However, as the performance of this detector is hard to analyze, they design an optimal watermarking signal (added to the true control signal) which maximizes the Kullback-Leibler Divergence (KLD) measure between the densities of the residual before and after the attack, subject to a constraint on the loss of linear quadratic (LQ) control cost due to the addition of the watermarking signal. In [10], the authors proposed a model where the attacker also replaces the sensor measurements by its own simulated signal, which tries to mimic the nominal system without the knowledge of the watermarking signal in the control input. The key result in [10] develops two tests at the actuator that the attacker has to pass to remain stealthy, but this is only possible if the attacker replaces the true sensor outputs by a signal of zero average energy. Further recent results on this topic include design of statistical watermarking for

The authors are with the Division of Signals and Systems, Uppsala University, Box 534, Uppsala, SE-75121, Sweden (e-mail: subhrakanti.dey@signal.uu.se)

joint sensor and communication attack detection [13], and Gaussian watermarking with packet drops in [14]. In a separate line of enquiry, a trade-off between controller utility and detectability of an attacker is studied using known input statistics at the controller, without actually formulating the problem in the context of a true feedback control system, or investigating sequential detection of attacks [7].

In detecting attacks in CPS, it is of paramount importance that attack detection happens with minimum delay, thus favouring *quickest sequential detection* based methods. The importance of this can be easily illustrated by a simple experiment where the attacker replaces the sensor data by a stationary Gaussian signal mimicking the properties of the sensor output, and subsequently, the estimator/controller (unaware of the attack) uses this false data to design their estimation/control algorithm, resulting in system instability exponentially fast. The watermark design techniques employed in [2], [10] are not designed specifically for quickest detection of attacks, and the statistical detection tests developed in [10] are *asymptotic* in nature, thus relying on collecting a large number of system outputs in practice. In this paper, we will therefore focus on design and analysis of physical watermarking signals that minimize the average detection delay in sequential detection methods, while still keeping the system performance within a prescribed safety limit - as demanded by resilience requirements of CPS under attacks [5].

In particular, we consider a scalar networked linear control system, where the attacker launches a deception attack at a certain unknown but deterministic time point, by injecting a stationary temporally correlated Gaussian false measurement sequence (thus replacing and mimicking) the true sensor measurements, while the estimator/controller employs standard optimal linear quadratic Gaussian (LQG) control based on the received measurement sequence without knowing whether there has been an attack or not. In order to aid the detection of the attacker, which on the other hand tries to remain stealthy, the controller adds a random (independent and identically distributed Gaussian watermarking signal) to the control signal, which is only known to the controller/actuator, and not the adversary.

Our main contributions are as follows: (i) We design a sequential quickest detection test at the controller, based on the cumulative sum (CUSUM) algorithm that is well known to minimize the average detection delay under a constraint on the mean time between false alarms. This sequential test is based on the log-likelihood ratio of the joint distribution of the residue (measurement prediction error) and the watermarking signal before and after the attack. (ii) Motivated by the result that the asymptotic (as the false alarm rate goes to zero) upper bound on the average detection delay is inversely proportional to the Kullback-Leibler divergence (KLD) measure between these joint distributions before and after the attack, we analyze the behaviour of the KLD measure with respect to the variance of the watermarking signal as well as the attacker signal variance, and present some structural results. These results show that for a fixed watermarking signal variance, the attacker can choose its own signal power to minimize the proposed KLD, whereas the KLD (when the attacker's signal power is fixed) is an increasing function in the watermarking signal variance if and only if the attacker signal variance is above a certain threshold. (iii) The behaviour of the increase in the LQG control cost due to the watermarking signal is also analyzed, illustrating the tradeoff between quickest detection, and the penalty in the control cost. (iv) Unlike previous works which consider KLD between the distributions of the residue signal (under attack) and the innovations (before the attack) only, we show that by considering the joint distributions of the residue/innovation and the watermarking signal, we can increase the KLD even further, thus reducing the average detection delay. Numerical results confirm our findings.

II. Problem Setup

A. System model

We consider the following architecture of a networked control system. The single-input single-output, linear time invariant system is modeled as:

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

in which $x_k \in \mathbb{R}$ is the state variable and $u_k \in \mathbb{R}$ is the control input at time k generated by the controller. $w_k \in \mathbb{R} \sim \mathcal{N}(0, Q)$ is the process noise at time k which is assumed to be an independent and identically distributed (*i.i.d*) random process. A sensor reports its (*scalar*) observations to the controller in the following form at time k:

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

in which $v_k \sim \mathcal{N}(0, R)$ is i.i.d. measurement noise that is independent of the process noise w_k. Note that although we consider a scalar state-space system, we still use uppercase letters for the system parameters A, B, C, Q, R, with a slight abuse of notation. We assume that the system has started at time $t = -\infty$ and currently is in steady-state condition, as stabilizability and detectability are guaranteed for a scalar system. Then the optimal state estimate equations, based on Kalman filtering, are given as

$$\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k \tag{3}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K\gamma_k \tag{4}$$

where $\hat{x}_{k+1|k} = E[x_{k+1}|\mathcal{Y}_k]$, and $\hat{x}_{k|k} = E[x_k|\mathcal{Y}_k]$ are the Kalman predicted and filtered state estimate, respectively based on received measurements up to time k, given by $\mathcal{Y}_k$. Also, $K = \frac{CP}{C^2P+R}$ is the steady-state Kalman gain and where P is the steady-state minimum mean-squared error (MMSE) estimation error variance $E(x_k - \hat{x}_{k|k-1})^2$ obtained from the solution to the algebraic Riccati equation

$$P = A^2P + Q - A^2C^2P^2(C^2P + R)^{-1}. \tag{5}$$

In (4) the innovation sequence γ_k is the defined as

$$\gamma_k \triangleq y_k - C\hat{x}_{k|k-1}. \tag{6}$$

We assume that the sensor is connected to the estimator/controller via a link that is susceptible to malicious

attacks. In the system equation (1), the control signal u_k is sent by the controller (which is assumed to be co-located with the actuator), to the sensor as a linear function of the filtered state estimate, such that $u_k = f(\hat{x}^k_{-\infty})$ minimizes the infinite-horizon LQG cost:

$$J = \lim_{T\to\infty} E\frac{1}{2T+1}[\sum_{k=-T}^{T}(Wx_k^2 + Uu_k^2)] \tag{7}$$

where W and U are positive weights. The LQG control policy results in a fixed-gain linear control signal as

$$u_k = L\hat{x}_{k|k},\ L = \frac{-ABS}{B^2S+U} \tag{8}$$

where S is the solution obtained from the algebraic Riccati equation

$$S = A^2S + W - A^2B^2S^2(B^2S+U)^{-1}. \tag{9}$$

B. Attack model

We assume that the adversary can launch an attack against the integrity of the sensor measurements such that the estimator/controller, instead of receiving the true measurement, y_k sent by the honest sensor, receives z_k, which is injected by the attacker. Furthermore, we assume that the attacker knows the system parameters A, B, C, Q and R and also the control policy, i.e., L but not necessarily the true sensor measurements y_k. On the other hand, we assume that the control signal is not tampered with by the adversary.

The goal of the attacker is to change the performance of the control system by sending fake observations, z_k, that replaces the true ones and while doing so remain undetected. In the following section, we consider an attack model where the attacker replaces the true measurement y_m by a fake measurement z_m for all $m \geq k$. This is a kind of spoofing attack which can be accomplished by the adversary, even without having access to y_k itself, by jamming or overpowering the true sensor signal if sent over wireless. However, if the sensor signal is not sent over wireless the adversary might be able to hijack it in another way and replacing the y_k with z_k in a so called man-in-the-middle attack. Most protocols used today would not be able to detect such an attack. Nevertheless, the objective of the attacker is to remain stealthy for a sufficiently long period of time over which the attack takes place, to cause maximum damage to the control system.

In this paper, we will assume that the attacker does not need to know the true sensor measurements but can simply alter them by injecting (as we will assume for the rest of this paper) the sequence $\{z_k\}$, which is stationary with statistics

$$E(z_k^2) = \sigma_z^2,\ E(z_k, z_{k-k'}) = \rho^{k'}\sigma_z^2 \tag{10}$$

in which $|\rho| < 1$. Depending on whether the attacker physically compromises the sensor node or simply replaces the sensor measurements by injecting a strong interfering signal, it may also need to know the encryption algorithm used by the networked control system. However, it is common to assume that the adversary has full knowledge of all system parameters and protocols, as is often done in cryptography according to the notion of "security through obscurity" known as Kerckhoffs's principle, or also according to *Shannon's maxim*, which essentially assumes that "the enemy knows the system." The attacker's knowledge of the system is a sensible assumption since then the adversary can cause maximum damage, a situation that is essential to detect as fast as possible.

C. Attack stealthiness

To determine whether an attack is present in the control system or not we shall rely on a hypothesis testing procedure based on the following two hypotheses:
H_0: No attack (the controller receives the true sequence y_k),
H_1: Attack (the controller receives the false sequence z_k).

Let p_k^F represent the false alarm probability, i.e., deciding H_1 when H_0 is true and let p_k^D represent the detection probability, i.e., deciding H_1 when H_1 is true, at time k. Furthermore, define $\widetilde{\gamma}_k$ to be the innovation signal $z_k - C\hat{x}^F_{k|k-1}$, where $\hat{x}^F_{k|k-1}$ is the inaccurate Kalman predictor designed in the presence of an attack based on the received sequence $\{z_k\}$. Let $\widetilde{\gamma}_1^k$ and γ_1^k represent the sequences $\{\widetilde{\gamma}_j\}_{j=1}^k$ $\{\gamma_j\}_{j=1}^k$, respectively. The goal is to design a detector which, with high probability can detect an attack while keeping the false alarm probability as small as possible. It is common to design a hypothesis testing procedure that decides in favour of H_0 or H_1 based on testing the innovation sequence $\widetilde{\gamma}_1^k$ (under attack) and the true innovation sequence γ_1^k.

In detection theory, the performance of the detector can be characterized by the trade-off between p_k^F and p_k^D. Following [6], [8], we introduce the following definition of a stealthy attack:

Definition 1: For $\epsilon > 0$ and $0 < \delta < 1$, an attack is ϵ-stealthy if for any detector that satisfies $0 < 1 - p_k^D \leq \delta$, it holds that

$$\limsup_{k\to\infty} -\frac{1}{k}\log(p_k^F) \leq \epsilon \tag{11}$$

It was shown in [8] that condition (11) is equivalent to

$$\limsup_{k\to\infty} \frac{1}{k} D(f_{\widetilde{\gamma}} \| f_\gamma) \leq \epsilon \tag{12}$$

when the hypothesis H_0 for no attack assumes the innovation sequence γ_1^k, and the residues $\widetilde{\gamma}_1^k$ for H_1. Here, $D(f_{\widetilde{\gamma}} \| f_\gamma)$ is the Kullback-Leibler Divergence (KLD) between the sequences $\widetilde{\gamma}_1^k$ and γ_1^k defined as:

$$D(f_{\widetilde{\gamma}} \| f_\gamma) = \int_{-\infty}^{\infty} f_{\widetilde{\gamma}}(\gamma_1^k) \log \frac{f_{\widetilde{\gamma}}(\gamma_1^k)}{f_\gamma(\gamma_1^k)} d\gamma_1^k\ . \tag{13}$$

where $f_{\widetilde{\gamma}}, f_\gamma$ are the (stationary) distributions of the sequences $\{\widetilde{\gamma}_k\}$ and $\{\gamma_k\}$, respectively. Clearly, the objective for the control system designer is to detect the attacker, and hence increase the value of the quantity $D = \limsup_{k\to\infty} \frac{1}{k} D(f_{\widetilde{\gamma}_k} \| f_{\gamma_k})$, an expression for which was provided in [8], while the attacker tries to minimize the KLD as much as possible (i.e, make ϵ as small as possible). This leads us to the next section, where we employ a physical watermarking mechanism to increase an appropriate KLD

measure based on the joint distributions of the innovations/residues and the random watermarking signal (rather than the residue signals only), thus making it difficult for the attacker to remain undetected through a sequential detection test designed accordingly. Note that we do not discuss how the adversary designs ϵ in response to the sequential detection test employed by the control system designer in this paper, a topic which will be further investigated in a game theoretic setting in future work.

D. Defence mechanism based on physical watermarking

As explained above, the attacker can choose an intelligent policy to inject false observations and tries to remain undetected. This however relies on the the fact that the control system is influenced by process and measurement noises, which produce uncertainty in favour of the attacker.

To protect the system against these active attacks, a key idea is to add a random watermarking signal, known only by the controller (and not to the attacker, although the attacker may know the statistics of the watermarking signal), to the control sequence u_k. In particular, the controller adds the watermarking sequence e_k to the control signal, i.e.,

$$u_k = L\hat{x}_{k|k} + e_k \tag{14}$$

where e_k is assumed to be an i.i.d. zero-mean Gaussian sequence with variance σ_e^2. The idea of adding such a physical watermarking signal was proposed in [2] in the context of detecting *replay attacks*, and further extended and analyzed in the context of dynamic watermarking in [10]. In general, the signal e_k can be a stationary Gauss-Markov process as shown in [2], although for the purpose of this paper, we assume it to be i.i.d.

By adding this sequence the controller is provided with a tool to check if the received signal from the sensors bear any correlation with the watermarking sequence or not. If the attacker injects a false observation z_k, which is naturally independent of the watermarking signal, then this can be detected by the controller, even though the attacker may know the statistics of the watermarking signal.

In [2], a Neyman-Pearson test based failure detector using the residue vector (which is either γ_k or $\widetilde{\gamma}_k$) was suggested for detecting an attack, whereas in [10], two asymptotic tests were proposed to detect an attack. Both of these schemes do not address the problem of *quickest detection* of the attack, which is of utmost importance, and this motivates us to consider a non-Bayesian sequential detection method under the assumption that the attack takes place at a fixed but unknown point of time. In particular, the cumulative-sum (CUSUM) method which minimizes the average detection delay subject to a constraint on the mean time between false alarms, also known as Lorden's method [11]. However, instead of comparing directly the distribution of residues γ_k and $\widetilde{\gamma}_k$, we propose a detection mechanism as follows. The controller, upon receiving the observation y_k (which is not known to be the true y_k or the false z_k) calculates γ_k (or $\widetilde{\gamma}_k$)and computes

$$S_k = \max(0, S_{k-1} + \log \frac{f_{\widetilde{\gamma}_k, e_{k-1}}(\widetilde{\gamma}_k, e_{k-1})}{f_{\gamma_k, e_{k-1}}(\widetilde{\gamma}_k, e_{k-1})}) \ . \tag{15}$$

where $f_{\widetilde{\gamma}_k, e_{k-1}}$ and $f_{\gamma_k, e_{k-1}}$ denote the joint distribution between the residue signal and the watermarking signal. The controller then decides on "attack" or "no attack" based on the following policy:

$$\begin{aligned} &\text{The system is under attack if} \quad S_k > \alpha, \\ &\text{The system is not under attack if} \quad S_k < \alpha \end{aligned} \tag{16}$$

where $\alpha \triangleq |\log p^F|$.

The above policy can be justified in the way that if the received observation by the controller is the true one, then

$$\begin{aligned} \gamma_k &= y_k - C\hat{x}_{k|k-1} \\ &= Cx_k + v_k - C(A+BL)\hat{x}_{k-1|k-1} - CBe_{k-1} \\ &= CA(x_{k-1} - \hat{x}_{k-1|k-1}) + Cw_{k-1} + v_k \ . \end{aligned} \tag{17}$$

meaning that γ_k is uncorrelated with the watermarking signal e_{k-1}. On the contrary, if the received observation by the controller is the false z_k, then

$$\begin{aligned} \tilde{\gamma}_k &= z_k - C\hat{x}^F_{k|k-1} \\ &= z_k - C(A+BL)\hat{x}^F_{k-1|k-1} - CBe_{k-1}. \end{aligned} \tag{18}$$

Thus, it is evident that the false innovations $\tilde{\gamma}_k$ is correlated with watermarking signal e_{k-1} and we can conclude that the control system is under attack.

It is worth mentioning that one might be tempted to conduct a sequential test based on the log-likelihood ratio $\log \frac{f_{\widetilde{\gamma}_k}(\widetilde{\gamma}_k)}{f_{\gamma_k}(\gamma_k)}$ (i.e, based on the log-likelihood ratio of the distributions of the residue under attack and innovations (no attack)), instead of S_k defined in (15), which is based on the joint distributions of the residue/innovations and the watermarking signal. In the following sections, we will illustrate how our suggested test quantity S_k can reduce the average detection delay as opposed to using the log-likelihood ratio based on the residue/innovations only, as mentioned above.

III. Main results

To analyze our suggested detection approach further we will use the Average Detection Delay (ADD) as a measure to quantify performance. It is well known that [9], [11] when the observations before and after the change are i.i.d, it is shown that, as the mean time between false alarms goes to infinity (or false alarm rate p^F goes to zero) the ADD is asymptotically upper bounded by $\frac{|\log p^F|}{I_1}$ where I_1 corresponds to the KLD between the distributions after and before the change. Although originally derived for i.i.d. sequences, these asymptotic upper bound results have been extended to the case of dependent but stationary sequences in [12], which allows us to write the following asymptotic upper bound on the ADD for the proposed sequential test based on (15), (16):

$$\frac{|\log p^F|}{\lim_{k\to\infty} D(f_{\widetilde{\gamma}_k, e_{k-1}} \| f_{\gamma_k, e_{k-1}})} \ . \tag{19}$$

Clearly, for a fixed p^F, the upper bound on the ADD is inversely proportional to the KLD between the joint distributions before and after the attack. In the following theorem, we obtain an expression for $D(f_{\tilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}})$ corresponding to our proposed detection approach.

Theorem 1: Consider the joint distributions between the watermarking signal and the true and false innovations, respectively, i.e., $f_{\gamma_k,e_{k-1}}$ and $f_{\tilde{\gamma}_k,e_{k-1}}$. The KLD between these joint *stationary* distributions, $\lim_{k\to\infty} D(f_{\tilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}})$, is then given by:

$$\lim_{k\to\infty} D(f_{\tilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}}) = \frac{1}{2}\log(\frac{1}{1-\lambda^2}) + \frac{1}{2}(\frac{\sigma^2_{\tilde{\gamma}}}{\sigma^2_{\gamma}} - 1 - \log\frac{\sigma^2_{\tilde{\gamma}}}{\sigma^2_{\gamma}}) \quad (20)$$

in which $\lambda = \frac{-BC\sigma_e}{\sigma_{\tilde{\gamma}}}$, and

$$\sigma^2_{\tilde{\gamma}} = \Bigg[(1-\frac{\rho CK(A+BL)}{1-\rho\mathcal{A}})^2 + \frac{(1-\rho^2)C^2K^2(A+BL)^2}{(1-\mathcal{A}^2)(1-\rho\mathcal{A})^2}\Bigg]\sigma^2_z + \frac{B^2C^2}{1-\mathcal{A}^2}\sigma^2_e \quad (21)$$

$$\sigma^2_{\gamma} = C^2P + R \quad (22)$$

where $\mathcal{A} \triangleq (1-CK)(A+BL)$ (note also that $\mathcal{A} < 1$ from stabilizability and detectability which is automatic for the scalar case) and P is calculated according to (5). Finally, $|\lambda| < 1$, as shown in the proof.

Proof: See Section Appendix A. ∎

In Theorem 1, $\lim_{k\to\infty} D(f_{\tilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}})$ is the KL divergence between the joint stationary distributions between the innovations and the watermarking signal, i.e., $f_{\gamma_k,e_{k-1}}(\gamma_k,e_{k-1})$ and $f_{\tilde{\gamma}_k,e_{k-1}}(\tilde{\gamma}_k,e_{k-1})$, as $k\to\infty$, for the healthy and attacked systems, respectively. With a fixed false alarm probability, we need to make these distributions as distinguishable as possible to avoid unnecessary detection delays., or increase the KLD as much as possible, whereas the attacker tries to minimize the KLD.

Properties of KLD as functions of σ^2_z and σ^2_e

Here we briefly present two results that illustrate the behaviour of the KLD expression in (20) with respect to σ^2_z, and σ^2_e, respectively. We assume that ρ is fixed in these discussions. First, we introduce the following simplifying notations. One can rewrite (32) in the proof of the above theorem as $\sigma^2_{\tilde{\gamma}} = M_1\sigma^2_z + M_2\sigma^2_e$, where $M_1 > 0, M_2 > 0$ are given by

$$M_1 = [(1-\frac{\rho CK(A+BL)}{1-\rho\mathcal{A}})^2 + \frac{(1-\rho^2)C^2K^2(A+BL)^2}{(1-\mathcal{A}^2)(1-\rho\mathcal{A})^2}],$$

$$M_2 = \frac{B^2C^2}{1-\mathcal{A}^2}.$$

Define also $M_3 = \frac{\lambda^2}{1-\mathcal{A}^2} = \frac{B^2C^2}{1-\mathcal{A}^2}\frac{\sigma^2_e}{\sigma^2_{\gamma}}$. Then we can show the following result:

Theorem 2: For a fixed σ^2_e, ρ, the KLD given by (20) is convex in σ^2_z. Furthermore, if $\frac{\sigma^2_e}{\sigma^2_{\gamma}} < \frac{(1/\mathcal{A}^2-1)}{B^2C^2}$, then it is first decreasing and then increasing in σ^2_z, attaining a minimum at $\sigma^2_{z\,opt} = \frac{\sigma^2_{\gamma}}{M_1}\left(M_3 + 1 - M_3/(1-\mathcal{A}^2)\right)$. Otherwise, if $\frac{\sigma^2_e}{\sigma^2_{\gamma}} \geq \frac{(1/\mathcal{A}^2-1)}{B^2C^2}$, then the KLD is monotonically increasing in σ^2_z.

Proof: The proof follows simply by investigating the first and second derivatives of the KLD with respect to σ^2_z, while keeping σ^2_e, ρ fixed, and is omitted for space restrictions. ∎

In a similar fashion, by investigating the first and second derivatives of the KLD with respect to σ^2_e, while keeping σ^2_z, ρ fixed, one can show that the KLD is a convex function of σ^2_e, and is monotonically increasing for all values of $\sigma^2_e > 0$, if and only if $\sigma^2_z > \frac{\mathcal{A}^2}{M_1}\sigma^2_{\gamma}$. The proof involves tedious calculations, and hence is not included due to space restrictions. A similar behaviour of the KLD can be observed as a function of ρ, if σ^2_z and σ^2_e are kept fixed.

Remark 1: The above observations indicate that the attacker (defender) can choose their signal power (watermarking signal variance) based on their knowledge of each other's parameters, and while the attacker can choose its signal power to minimize the KLD, increasing the watermarking variance does not necessarily increase the KLD for the defender unless the attack signal variance is above a certain threshold. We illustrate this non-intuitive behaviour in the simulations section with a numerical example.

However, increasing the watermarking variance to increase the KLD (decrease the ADD) comes at a cost: by increasing the watermarking signal power we also diverge from the optimal LQG cost given by (7), (8). Hence, there is a tradeoff between reducing the average detection delay and the system performance in terms of the increase of the LQG cost. We elaborate on this issue in the following theorem, by considering the difference in the LQG cost for the healthy system (no attack), as under attack an unstable open loop system can be easily destabilized. See [2] for a similar treatment.

Theorem 3: Consider the LQG cost (7) with weighting factors W and U and let ΔLQG represent the acceptable increase in the LQG cost from the optimal LQG cost for the system under no attack. Then, the watermarking signal variance is related to the increase in LQG cost as follows:

$$\sigma^2_e = \frac{\Delta LQG}{U + \frac{B^2(W+L^2U)}{1-(A+BL)^2}} \quad (23)$$

Proof: See Section Appendix B. ∎

In practice, the system designer will choose a certain σ^2_e to tolerate a maximum ΔLQG as given by the above theorem, and since the attacker knows σ^2_e and other system parameters as assumed, it can design its signal variance σ^2_z to minimize the KLD. For the system designer, the knowledge of the attacker parameters ρ, σ^2_z can help increase the KLD by choosing an appropriate σ^2_e, by allowing a larger ΔLQG if necessary. This interaction between the defender and the attacker can be formulated in a dynamic game scenario, which will be investigated in future work.

Remark 2: As stated in the previous section, instead of the distribution of the innovations in our detection policy, we considered the joint distribution between the innova-

tions and the watermarking signal. The benefit of using the joint distribution instead of using the innovations only can be immediately observed in the expression of the $D(f_{\widetilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}})$ in (20). By considering the innovations only, the corresponding $D(f_{\widetilde{\gamma}_k}\|f_{\gamma_k})$ expression equals the second term in (20) i.e., $\frac{1}{2}(\frac{\sigma^2_{\widetilde{\gamma}}}{\sigma^2_{\gamma}}-1-\log\frac{\sigma^2_{\widetilde{\gamma}}}{\sigma^2_{\gamma}})$. By using the joint distributions we also obtain the first term in (20), which is positive, leading to a larger KLD, thus making it more difficult for the attacker to remain stealthy.

IV. Numerical results

In this section we will investigate the tradeoff between ADD and ΔLQG. This is shown in Fig. 1 for two different values of $p^F = .001$ and $p^F = .01$. The blue dashed line represents the simulated ADD according to our detection policy in (16) while the solid pink line shows the ADD upper bound according to our result in Theorem 1. The red dash-dot line shows the ADD upper bound for the detection policy based on the KLD between the distributions of the residue signal (under attack) and the true innovations (no attack). This is referred to in the graphs as "ADD bound based on innovations only". It should be noted that in previous works such as [10], some numerical results based on a sequential detection are presented, although the actual detection policy is not clearly stated. The results in Fig. 1 are depicted for the values $A = .7, B = C = R = Q = W = 1, U = .4$, $\sigma_z^2 = 4$, $\rho = .5$ and the real ADD blue dashed line is calculated based on the average over 100 random realizations of the sequential detection algorithm. Comparison between the solid pink line and the red dash-dot line in Fig. 1 demonstrates the reduction in the ADD upper bounds due to using our proposed sequential detection test compared to a sequential detection based on innovations/residues only.

In Fig. 2 we investigate the behaviour of the KLD (as given by (20)) as a function of σ_z^2 (while keeping ρ, σ_e^2 fixed), and as a function of σ_e^2 (while keeping ρ, σ_z^2 fixed). In these simulations, $A = 1.1$, $B = C = Q = R = W = U = 1$, $\rho = 0.7$, and σ_e^2 is chosen by using $\Delta LQG = 0.7$. The first decreasing and then increasing behaviour of the KLD is clearly visible, as in this case it can be checked that $\frac{\sigma_e^2}{\sigma_\gamma^2} < \frac{(1/\mathcal{A}^2-1)}{B^2C^2}$, and the theoretically calculated value of $\sigma^2_{z\,opt} = 4.0253$. In the bottom figure, we plot the KLD as a function of σ_e^2 while keeping ρ, σ_z^2 fixed. The parameter values are the same as before, except here $\sigma_z^2 = 0.02$, which is less than the threshold given by $\frac{\mathcal{A}^2}{M_1}$. The first decreasing and then increasing behaviour of the KLD as a function of σ_e^2 is clearly visible, implying that if the attacker signal variance is lower than a required threshold, increasing the watermarking variance may not always increase the KLD. In fact, it is obvious from the figure that if the currently used $\sigma_e^2 < 2$, then it is beneficial to actually decrease σ_e^2 if only a moderate increase in KLD is required, as this also lowers the increase in LQG cost if there is no attack.

Fig. 3 illustrates the ADD for an unstable system with $A > 1$. The ADD is shown in two different values for ρ for $A = 1.2$ while $B = C = R = Q = W = 1, U = .4, \sigma_z^2 = 4.0, p^F = .01$. It is seen that in the unstable case with $A > 1$, there exists a higher gap between the ADD bounds for the sequential tests based on the joint distribution as proposed in this paper and the one based on innovations/residues only.

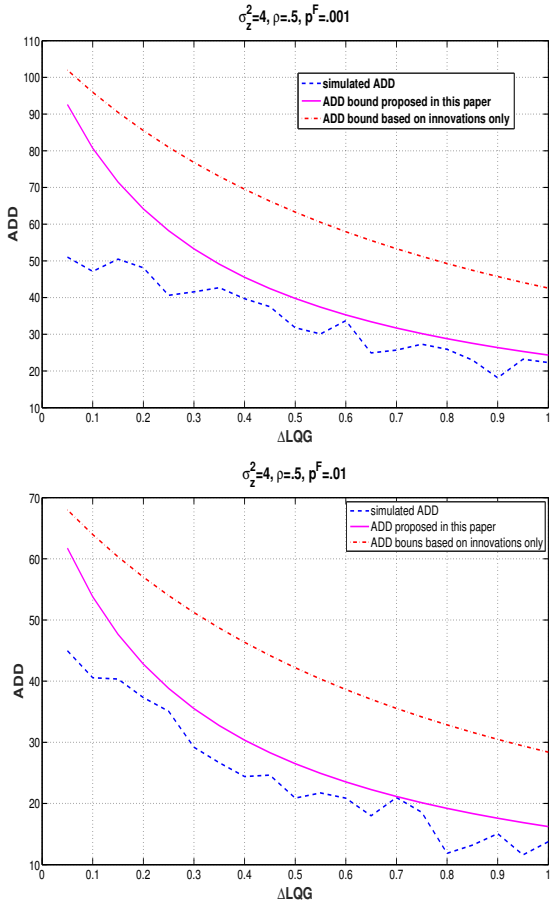

Fig. 1. ADD in terms of ΔLQG for $p^F = .001$ and $p^F = .01$

V. Conclusions and Future Work

In this paper, we investigated how a suitably designed sequential detection test can detect deception attacks in a scalar networked control system with an average detection delay that can be reduced by introducing a physical watermarking signal with a suitable variance. The tradeoff between quick detection and penalty in the control cost as a result of using the watermarking signal is also investigated. Future works will extend these results to multi-variable (vector state and measurements) systems with more general attack strategies, and propose a dynamic game between the adversary and the control system designer regarding the attacker's effort to remain stealthy, and the system designer's effort to detect the attack with minimum delay, using such physical watermarking schemes.

Appendix

A. Proof of Theorem 1

To calculate $D(f_{\widetilde{\gamma}_k,e_{k-1}}\|f_{\gamma_k,e_{k-1}})$, we need to obtain the joint distributions $f_{\gamma_k,e_{k-1}}(\gamma,e)$ and $f_{\widetilde{\gamma}_k,e_{k-1}}(\gamma,e)$ in steady

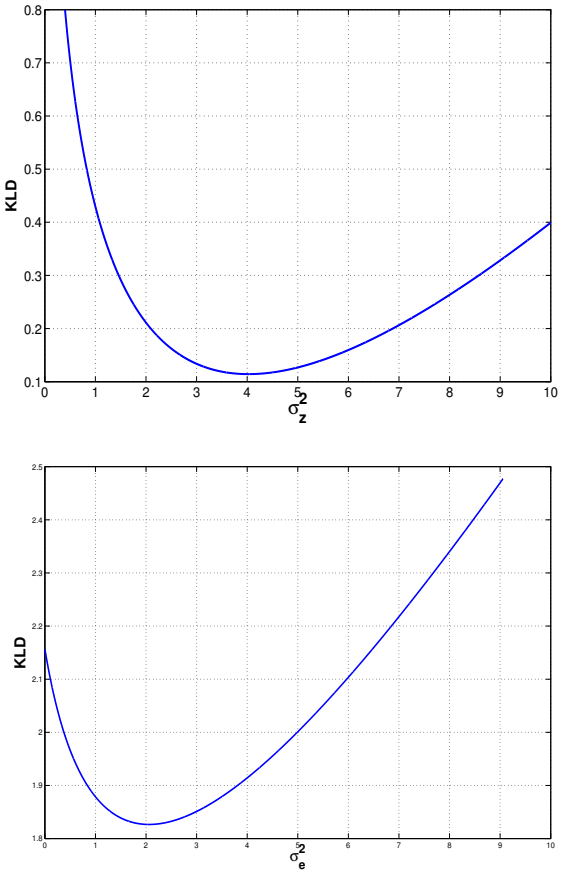

Fig. 2. Behaviour of KLD as a function of σ_z^2 and σ_e^2.

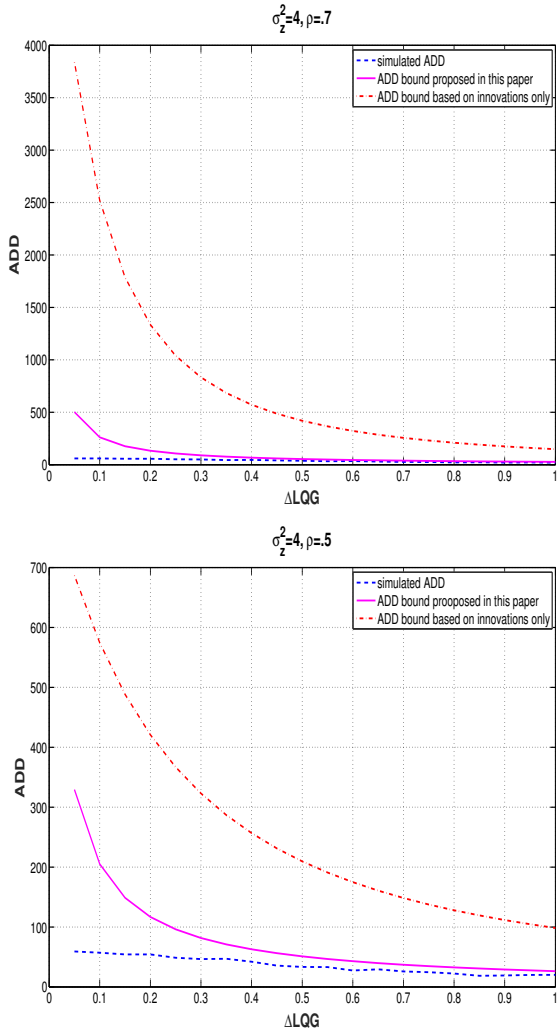

Fig. 3. ADD in terms of ΔLQG for different values of σ_z^2 and ρ with $A = 1.2$

state. As it was shown in (17), in the healthy system, γ_k and e_{k-1} are uncorrelated for i.i.d. watermarking sequence. Hence the joint distribution $f_{\gamma_k e_{k-1}}(\gamma, e)$ appears as:

$$f_{\gamma_k e_{k-1}}(\gamma, e) = \frac{1}{2\pi\sigma_\gamma\sigma_e} \exp \frac{-1}{2}\left(\frac{\gamma^2}{\sigma_\gamma^2} + \frac{e^2}{\sigma_e^2}\right) \tag{24}$$

in which σ_γ^2 is given as in (22). On the other hand, when we have attack, $\tilde{\gamma}_k$ and e_{k-1} are correlated according to (18). Since $\tilde{\gamma}_k$ and e_{k-1} are zero-mean Gaussian, to obtain their joint distribution, we need to calculate $\sigma_{\tilde{\gamma}}^2$ and $\mathbf{cov}(\tilde{\gamma}_k, e_{k-1})$. Since e_{k-1} is uncorrelated with z_k and $\hat{x}^F_{k-1|k-1}$, we use (18) to obtain:

$$\mathbf{cov}(\tilde{\gamma}_k, e_{k-1}) = -CB\sigma_e^2 \tag{25}$$

Using the same equation, we have:

$$\begin{aligned}\sigma_{\tilde{\gamma}}^2 &= \sigma_z^2 + C^2(A+BL)^2\sigma_{\hat{x}}^2 \\ &\quad - 2C(A+BL)\mathbf{cov}(z_k, \hat{x}^F_{k-1|k-1}) + C^2B^2\sigma_e^2. \end{aligned}\tag{26}$$

where $\sigma_{\hat{x}}^2 = \mathrm{E}\left(\hat{x}^F_{k-1|k-1}\right)^2$.

To calculate $\mathbf{cov}(z_k, \hat{x}^F_{k-1|k-1})$ we proceed as follows. By combining (4) and (6), for the attacked system, one obtains:

$$\hat{x}^F_{k-1|k-1} = Kz_{k-1} + \mathcal{A}\hat{x}^F_{k-2|k-2} + B(1-CK)e_{k-2}. \tag{27}$$

By multiplying the above equation with $z_k, z_{k+1}, z_{k+2}, \ldots$ etc., and calculating expectation of the both sides for the stationary system, and defining $\mathrm{E}_{\hat{x}z}(-l) = \mathbf{cov}(z_k, \hat{x}^F_{k-l|k-l})$, we have for $k = 1, 2, \ldots,$

$$\mathrm{E}_{\hat{x}z}(-k) = KE_{zz}(k) + \mathcal{A}E_{\hat{x}z}(-(k+1)).$$

in which $\mathrm{E}_{zz}(k)$ is obtained according to (10).

Since $\mathcal{A} < 1$ and $\rho < 1$, $\mathrm{E}_{\hat{x}z}(-1)$ is obtained as a sum of an infinite geometric series which converges to:

$$\mathrm{E}_{\hat{x}z}(-1) = K\sigma_z^2\sum_{i=1}^{\infty}\rho^i\mathcal{A}^{i-1} = K\sigma_z^2\rho\sum_{i=0}^{\infty}(\rho\mathcal{A})^i = \frac{K\sigma_z^2\rho}{1-\rho\mathcal{A}} \tag{28}$$

To calculate $\sigma_{\hat{x}}^2$, we reuse (27) such that:

$$\begin{aligned}\mathrm{E}(\hat{x}^F_{k-1|k-1})^2 &= K^2\mathrm{E}(z_{k-1}^2) + \mathcal{A}^2\mathrm{E}(\hat{x}^F_{k-2|k-2})^2 \\ &\quad + 2K\mathcal{A}\mathrm{E}(z_{k-1}\hat{x}^F_{k-2|k-2}) + B^2(1-CK)^2\sigma_e^2\end{aligned}\tag{29}$$

which results in:

$$\sigma_{\hat{x}}^2 = K^2\sigma_z^2 + \mathcal{A}^2\sigma_{\hat{x}}^2 + 2K\mathcal{A}\mathrm{E}_{\hat{x}z}(-1) + B^2(1-CK)^2\sigma_e^2. \tag{30}$$

Combining (28) and (30) yields:

$$\sigma_{\hat{x}}^2 = \frac{K^2(1+\rho\mathcal{A})}{(1-\rho\mathcal{A})(1-\mathcal{A}^2)}\sigma_z^2 + \frac{B^2(1-CK)^2}{1-\mathcal{A}^2}\sigma_e^2. \quad (31)$$

and further manipulations using (26), $\sigma_{\tilde{\gamma}}^2$ is obtained as:

$$\sigma_{\tilde{\gamma}}^2 = [(1-\tfrac{\rho CK(A+BL)}{1-\rho\mathcal{A}})^2 + \tfrac{(1-\rho^2)C^2K^2(A+BL)^2}{(1-\mathcal{A}^2)(1-\rho\mathcal{A})^2}]\sigma_z^2 + \tfrac{B^2C^2}{1-\mathcal{A}^2}\sigma_e^2 \quad (32)$$

To obtain the joint distribution $f_{\tilde{\gamma}_k e_{k-1}}(\gamma, e)$, we form the cross-covariance matrix of $\gamma_k.e_{k-1}$ as:

$$\Sigma = \begin{pmatrix} \sigma_{\tilde{\gamma}}^2 & -CB\sigma_e^2 \\ -CB\sigma_e^2 & \sigma_e^2 \end{pmatrix} \quad (33)$$

and consequently:

$$f_{\tilde{\gamma}_k, e_{k-1}}(\gamma, e) = \frac{1}{2\pi\sigma_{\tilde{\gamma}}\sigma_e\sqrt{1-\lambda^2}} \times \exp\left\{\frac{-1}{2(1-\lambda^2)}\left(\frac{\gamma^2}{\sigma_{\tilde{\gamma}}^2} + \frac{e^2}{\sigma_e^2} - \frac{2\lambda e\gamma}{\sigma_e\sigma_{\tilde{\gamma}}}\right)\right\} \quad (34)$$

in which

$$\lambda = \frac{\mathbf{cov}(\tilde{\gamma}_k, e_{k-1})}{\sigma_e\sigma_{\tilde{\gamma}}} \overset{(a)}{=} \frac{-BC\sigma_e}{\sigma_{\tilde{\gamma}}} \quad (35)$$

where (a) is deduced from (25). Note also that $|\lambda| < 1$ as it is a correlation coefficient.

Then $D(f_{\tilde{\gamma}_k, e_{k-1}} \| f_{\gamma_k, e_{k-1}})$ is calculated as:

$$\iint_{-\infty}^{\infty} f_{\tilde{\gamma}_k e_{k-1}}(\gamma, e) \log \frac{f_{\tilde{\gamma}_k, e_{k-1}}(\gamma, e)}{f_{\gamma_k, e_{k-1}}(\gamma, e)} de d\gamma. \quad (36)$$

Replacing the joint distributions as in (24) and (34) in (36) yields

$$\begin{aligned} D(f_{\tilde{\gamma}_k, e_{k-1}} \| f_{\gamma_k, e_{k-1}}) &= \log(\frac{\sigma_\gamma}{\sigma_{\tilde{\gamma}}\sqrt{1-\lambda^2}}) \\ &+ \frac{-\sigma_{\tilde{\gamma}}^2}{2}(\frac{1}{(1-\lambda^2)\sigma_{\tilde{\gamma}}^2} - \frac{1}{\sigma_\gamma^2}) + \frac{-1}{2}(\frac{1}{(1-\lambda^2)} - 1) \\ &+ \frac{\lambda \mathbf{cov}(\tilde{\gamma}_k, e_{k-1})}{(1-\lambda^2)\sigma_{\tilde{\gamma}}\sigma_e} \\ &= \frac{1}{2}\log(\frac{1}{1-\lambda^2}) + \frac{1}{2}(\frac{\sigma_{\tilde{\gamma}}^2}{\sigma_\gamma^2} - 1 - \log\frac{\sigma_{\tilde{\gamma}}^2}{\sigma_\gamma^2}) \end{aligned}$$

Using the same approach as in [8], it can be shown that the averaged KLD is equal to the single letter $D(f_{\tilde{\gamma}_k, e_{k-1}} \| f_{\gamma_k, e_{k-1}})$ calculated above in which the distributions are the steady state ones.

B. Proof of Theorem 2

To calculate the cost of using the watermarking for the purpose of detection, we obtain the difference of LQGs between the cases depending on whether watermarking is used or not used in the healthy system. According to (7), we use (14) to obtain:

$$\Delta LQG = J_w - J_n = W(\mathrm{E}(X_w^2) - \mathrm{E}(X_n^2)) + UL^2(\mathrm{E}(\hat{X}_w^2) - \mathrm{E}(\hat{X}_n^2)) + U\sigma_e^2 \quad (37)$$

where subscript 'w' refers to the case where we use watermarking and subscript 'n' refers to the case that we don't use watermarking. To calculate ΔLQG, we need to calculate $\mathrm{E}(X_w^2)$ and $\mathrm{E}(\hat{X}_w^2)$. Based on the fact that γ_k is uncorrelated with $\hat{x}_{k|k-1}$, we use (4) in steady-state condition to obtain:

$$\mathrm{E}(\hat{X}_w^2) = \frac{1}{1-(A+BL)^2}(B^2\sigma_e^2 + K^2(C^2P+R)) \quad (38)$$

To calculate $\mathrm{E}(X_w^2)$, (6) is obtained as:

$$\begin{aligned} \mathrm{E}(Y_w^2) &= \sigma_\gamma^2 + C^2((A+BL)^2\mathrm{E}(\hat{X}_w^2) + B^2\sigma_e^2) \\ &= (C^2P+R)(1 + \tfrac{C^2K^2(A+BL)^2}{1-(A+BL)^2}) + \tfrac{B^2C^2}{1-(A+BL)^2}\sigma_e^2 \end{aligned} \quad (39)$$

where we used (38) to conclude (39).

Then (2) is used to calculate $\mathrm{E}(X_w^2)$ as

$$\begin{aligned} \mathrm{E}(X_w^2) &= \tfrac{\mathrm{E}(Y_w^2) - R}{C^2} \\ &= P + \tfrac{K^2(A+BL)^2(C^2P+R)}{1-(A+BL)^2} + \tfrac{B^2}{1-(A+BL)^2}\sigma_e^2 \end{aligned} \quad (40)$$

Combining (38) and (40) with (37) gives us:

$$\Delta LQG = \left(U + \frac{B^2(W+L^2U)}{1-(A+BL)^2}\right)\sigma_e^2 \quad (41)$$

REFERENCES

[1] A. Teixeira, K. C. Sou, H. Sandberg and K.H. Johansson, "Secure Control Systems: A quantitative Risk-Management Approach," *IEEE Control Syst. Mag.*, pp. 24-45, Feb. 2015.

[2] Y. Mo, S. Weerakkody and B. Sinopoli, "Physical Authentication of Control Systems," *IEEE Control Syst. Mag.*, pp. 93-109, Feb. 2015.

[3] F. Pasqualetti, F. Dörfler and F. Bullo, "Control-Theoretic Methods for Cyberphysical Security," *IEEE Control Syst. Mag.*, pp. 110-127, Feb. 2015.

[4] H. Fawzi, P. Tabuada and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. on Auto. Contr.*, vol. 59, no. 6, pp. 1454-1467, June 2014.

[5] A.A. Cardenas, S. Amin and S. Sastry, "Secure Control: Towards Survivable Cyber-Physical Systems," *Proc. 28th International Conference on Distributed Computing Systems Workshops*, pp. 495-500, Beijing, China, June 2008.

[6] C. Bai, F. Pasqualetti, V. Gupta, "Security in Stochastic Control Systems: Fundamental Limitations and Performance Bounds," *American Control Conference*, Chicago, IL, USA, pp. 195–200, 2015.

[7] P. Pradhan and P. Venkitasubramaniam, "Stealthy Attacks in Dynamical Systems: Tradeoffs Between Utility and Detectability With Application in Anonymous Systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 779-792, April 2017.

[8] Enoch Kung, Subhrakanti Dey, and Ling Shi, "The Performance and Limitations of ϵ-Stealthy Attacks on Higher Order Systems," *IEEE Transactions ON Automatic Control*, vol. 62, no. 2, pp. 941–947, Feb. 2017.

[9] A. G. Tartakovsky, V. V. Veeravalli, "Asymptotically Optimal Quickest Change Detection in Distributed Sensor Systems," *Sequential Analysis*, vol. 27, pp. 441–475, 2008.

[10] B. Satchidanandan and P. R. Kumar, "Dynamic Watermarking: Active Defense of Networked Cyber-Physical Systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219-240, Feb. 2017.

[11] H.V. Poor and O. Hadjiliadis, *Quickest Detection*, Cambridge University Press, Cambridge, UK, 2009.

[12] S. Pergamenchtchikov and A. G. Tartakovsky, "Asymptotically optimal pointwise and minimax quickest change-point detection for dependent data," *Statistical Inference for Stochastic Processes (Springer)*, DOI 10.1007/s11203-016-9149-x, published online, Oct. 2016.

[13] P. Hespanhol, M. Porter, R. Vasudevan and A. Aswani, "Statistical Watermarking for Networked Control Systems," *Proc. ACC 2018*, pp. 5467-5472, Milwaukee, USA, June 2018.

[14] S. Weerakkody, O. Ozel and B. Sinopoli, "A Bernoulli-Gaussian Physical Watermark for Detecting Integrity Attacks in Control Systems," *Proc. 55th Annual Allerton Conference*, pp. 966-973, Illinois, USA, Oct. 2017.