Detecting Integrity Attacks on SCADA Systems

Yilin Mo, *Member, IEEE*, Rohan Chabukswar, *Student Member, IEEE*,
and Bruno Sinopoli, *Member, IEEE*

***Abstract*— Ensuring security of systems based on supervisory control and data acquisition is a major challenge. The goal of this paper is to develop the model-based techniques capable of detecting integrity attacks on the sensors of a control system. In this paper, the effect of integrity attacks on the control systems is analyzed and countermeasures capable of exposing such attacks are proposed. The main contributions of this paper, beyond the novelty of the problem formulation, lies in enumerating the conditions of the feasibility of the replay attack, and suggesting countermeasures that optimize the probability of detection by conceding control performance. The methodologies are shown and the theoretical results are validated using several sets of simulations.**

***Index Terms*— Control, cyber-physical systems (CPS), secure, supervisory control and data acquisition (SCADA).**

I. Introduction

CYBER-PHYSICAL systems (CPS) are systems with tight coordination between the computational and physical elements [1]. Such systems often employ distributed networks of embedded sensors and actuators that interact with the physical environment, and are monitored and controlled by a supervisory control and data acquisition (SCADA) system. CPS are observed in multifarious applications such as smart grids, process control systems, air traffic control (ATC), medical monitoring, and so on.

A recent concern in distributed control system security is that an attacker could gain access to a set of sensing and actuation devices and modify their software or environment to launch a coordinated attack against the system infrastructure. The Stuxnet worm, specially designed to reprogram certain industrial centrifuges and make them fail in a way that was virtually undetectable [2], is an example of digital warfare [3]. This worm has brought to light serious security susceptibilities in industrial control systems. In view of the omnipresent threat of organized terrorism, a power grid failure, a local breakdown of telecommunication systems, or a disruption of ATC at a major hub, could all be executed as antecedents of a full-fledged invasion. Such threats have been predicted for a long time [4]. CPS infrastructures vital to the normal operation of a society are safety critical, and any attack on one, or a coordinated attack on two or more of them, can significantly hamper the economy and endanger human lives. Unscrupulous entities can also use such attacks to affect market pricing for making illegal profits. The secure design of CPS is thus of paramount importance.

A conventional security measure is employing encrypted communications, but cryptographic keys can be broken or stolen, or the attacker could directly attack the physical elements of the system, without hijacking communications. Such attacks are feasible when sensors and actuators are distributed in remote locations. Therefore, system knowledge and cybersecurity are essential to ensure secure operation of CPS.

A. Previous Work

The importance of security of CPS has been stressed by the research community in [5] and [6] among others. Cardenas *et al.* [7] discuss the cyber-physical impact of denial-of-service (DoS) attacks, which interrupt information flow from the sensors, actuators, and the control system, and deception attacks that compromise the integrity of data packets. DoS attacks and a feedback control design resilient to them are further discussed in [8]. The authors are of the opinion that a deception attack is more subtle, and in principle more difficult to detect, than a DoS attack. As this issue has not been adequately addressed in the literature, a methodology to detect a specific kind of deception attack is proposed in this paper.

A substantial amount of research has been carried out in analyzing, detecting, and failure-handling CPS. Sinopoli *et al.* [9], [10] studied the effect of random packet drops on controller and estimator performance. Several failure-detection schemes in dynamic systems are reviewed in [11]. Some CPS scenarios, e.g., those proposed in [12], are capable of using results from robust control, where the authors concentrate on designing the controllers for systems with unknown or uncertain parameters. While these works assume that failures are either random or benign, a shrewd attacker, such as is considered in this paper, can carefully construct an attack strategy to deceive detectors and make robust controllers fail.

Alpcan and Başar [13] applied game theoretic principles formally to intrusion detection for developing a decision and control framework. Their work considers the treatment of intrusion-detection sensors, not on the actual scheme of

Manuscript received October 29, 2012; revised May 6, 2013; accepted August 11, 2013. Manuscript received in final form September 3, 2013. Date of publication September 26, 2013; date of current version June 16, 2014. This work was supported in part by CyLab with Carnegie Mellon under Grant DAAD19-02-1-0389 from the Army Research Office, in part by the Northrop Grumman Information Technology, Inc., Cybersecurity Consortium grant NGIT2009100109, and in part by the National Science Foundation under Grant 0955111. Recommended by Associate Editor L. Xie.

Y. Mo was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA. He is now with the Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: yilinmo@caltech.edu).

R. Chabukswar and B. Sinopoli are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: rchabuks@ece.cmu.edu; brunos@ece.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCST.2013.2280899

detection that each sensor employs. Controllability and observability of linear systems has been analyzed using graph theory in [14], which provide methods for reaching consensus in the presence of malicious agents. The proposed methods are combinatorial in nature, and computationally expensive. Robust estimation using sensors in untrusted environments has been investigated in [15], and again in [16], where Lazons *et al.* propose robust localization algorithms, which concentrate on the location information of the sensors, not the sensor data itself. Pasqualetti *et al.* [17], [18] consider intentional malicious data attack, and address the problem of distributed monitoring and intrusion detection. Distributed formation control in the presence of attackers is studied in [19], where a distributed control algorithm using online adaptation is proposed. These scenarios, unlike the present work consider a noiseless process and environment.

Giani *et al.* [20] address the problem of secure and resilient power transmission and distribution, and point out several potential threats in modern power systems. A comprehensive survey of the current results in networked control systems has been carried out in [21]. Dán and Sandberg [22] analyze stealth attacks on power system state estimators, and use a static system formulation unlike this paper. Sandberg *et al.* [23] study the analysis of large scale power networks of using proposed security indices. Secure state estimation and control of systems under attack is further investigated in [24] and [25]. The security of power networks, however, focuses on static systems, contrary to the fundamental formulation of a linear time-invariant (LTI) system analyzed in this paper.

This paper builds on the previous theoretical results of the authors, [26], [27]. Mo and Sinopoli [26] proposed the original problem and attack strategy, and introduced the concept of noisy control, with some simulations on a model of a moving vehicle. The subsequent work [27] provided a way to optimize the noisy control in multiinput, multioutput systems, and introduced noisy control for a cross-correlator detector, with simulations on a chemical plant and a microgrid. In this paper, the results of the above are extended, with some new results regarding the form of the optimized control signal. The simulations have been consolidated into the chemical plant, for better comparison of the application of different techniques and their performances.

B. Outline

The goal of this paper is to develop the model-based techniques capable of detecting integrity attacks on the sensors of a control system. It is assumed that the attacker wishes to disrupt the operation of a control system in steady state, to which end the attacker hijacks the sensors, observes, and records their readings for a certain amount of time, and repeats them afterward to camouflage his attack. Such an attack is common and natural, especially if the attacker does not know the dynamics of the system, but is aware that the systems is expected to be in steady state during the attack. This deception, proposed a year before Stuxnet came to light [26], was exactly what the worm used to hide its activities—It recorded normal operations at the nuclear plant, which it then played back to the operators to feign normal operation while spinning the centrifuges beyond rated values [28].

The main contributions of this paper, beyond the novelty of the problem formulation, lies in enumerating the conditions of the feasibility of the replay attack, and suggesting countermeasures that optimize the probability of detection by conceding control performance.

This paper is organized as follows. In Section II, the problem formulation is provided by revisiting and adapting the Kalman filter, the linear quadratic Gaussian (LQG) controller, and the χ^2 failure detector. The threat model is also defined, and its effect on the control schemes of Section II is analyzed. In Section III, the class of systems incapable of detecting such attacks is identified. In Section IV, three countermeasures for detecting such attacks are provided, based on adding a zero-mean Gaussian authentication signal to the optimal control. A way to design the authentication signal to minimize the performance loss while guaranteeing a certain probability of detection is also provided. The methods validated by carrying out several simulations detailed in Section V. Section VI concludes this paper, with some directions for future work. The appendix contains some proofs that would otherwise interrupt the flow of this paper.

II. Problem Formulation

This section presents the problem formulation by deriving the Kalman filter, the LQG controller, and χ^2-detector for the case under study. The notation developed below is used for the remainder of this paper.

Consider an LTI system:

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

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, $w_k \in \mathbb{R}^n$ is the process noise at time k, and x_0 is the initial state. We assume that w_k, x_0 are independent Gaussian random variables, $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $w_k \sim \mathcal{N}(0, Q)$.

A sensor network monitors the system in (1). The observation equation can be written as follows:

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

where $y_k \in \mathbb{R}^m$ is a vector of sensor measurements and $v_k \sim \mathcal{N}(0, R)$ is the measurement noise independent of x_0 and w_k.

It is assumed that the system operator wants to minimize the following infinite-horizon LQG cost:

$$J = \min \lim_{T\to\infty} E\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T U u_k\right)\right] \tag{3}$$

where W, U are positive semidefinite matrices and u_k is measurable with respect to $y_0, y_1, \ldots, y_k$, i.e., u_k is a function of the previous observations. The separation principle holds in this case and the optimal solution of (3) is a combination of Kalman filter and LQG controller. The Kalman filter provides the optimal state estimate $\hat{x}_{k|k}$:

$$\hat{x}_{0|-1} = \bar{x}_0, \quad P_{0|-1} = \Sigma \tag{4}$$

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k, \quad P_{k+1|k} = AP_kA^T + Q \\ K_k &= P_{k|k-1}C^T\left(CP_{k|k-1}C^T + R\right)^{-1} \\ \hat{x}_k &= \hat{x}_{k|k-1} + K_k\left(y_k - C\hat{x}_{k|k-1}\right) \\ P_k &= P_{k|k-1} - K_kCP_{k|k-1}. \end{aligned} \tag{5}$$

Although the Kalman filter uses a time-varying gain K_k, this gain will converge if the system is detectable. In practice, the Kalman gain usually converges in a few steps. Hence, P and K can be defined as follows:

$$P \triangleq \lim_{k\to\infty} P_{k|k-1}, \; K \triangleq PC^T\left(CPC^T + R\right)^{-1}. \tag{6}$$

As the control systems usually run for a long time, the system can be assumed to be at steady state. The initial condition $\Sigma = P$ reduces the Kalman filter to a fixed gain estimator:

$$\begin{aligned}\hat{x}_{0|-1} &= \bar{x}_0, \; \hat{x}_{k+1|k} = A\hat{x}_k + Bu_k \\ \hat{x}_k &= \hat{x}_{k|k-1} + K\left(y_k - C\hat{x}_{k|k-1}\right). \end{aligned} \tag{7}$$

The LQG controller is a fixed gain linear controller based on the optimal state estimation $\hat{x}_k$:

$$u_k = u_k^* = -\left(B^TSB + U\right)^{-1}B^TSA\hat{x}_k \tag{8}$$

where u_k^* is the optimal control input and S satisfies the Riccati equation

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

Let $L \triangleq -\left(B^TSB + U\right)^{-1}B^TSA$, then $u_k^* = L\hat{x}_k$. The optimal value of objective function in this case is

$$J = \text{trace}\left(SQ\right) + \text{trace}\left[\left(A^TSA + W - S\right)\left(P - KCP\right)\right]. \tag{10}$$

A. χ^2 Failure Detector

The χ^2 detector [29], [30] is widely employed in control systems, and uses characteristics of Kalman filter residues:

Theorem 1: For the LTI system defined in (1) with Kalman filter and LQG controller, the Kalman filter resides $y_i - C\hat{x}_{i|i-1}$ of are Gaussian independent identically distributed (i.i.d.) with zero mean and covariance $\mathscr{P} = CPC^T + R$.

Proof: The proof is given in [29]. ∎

Let

$$g_k \triangleq \sum_{i=k-\mathscr{T}+1}^{k} \left(y_i - C\hat{x}_{i|i-1}\right)^T \mathscr{P}^{-1}\left(y_i - C\hat{x}_{i|i-1}\right) \tag{11}$$

where $\mathscr{T}$ is the window size. With Theorem 1, it is known that when the system is operating normally, g_k has a χ^2 distribution with $m\mathscr{T}$ degrees of freedom,[1] implying lower probability of a larger g_k. The χ^2 detector at time k is:

$$g_k \underset{H_1}{\overset{H_0}{\lessgtr}} v \tag{12}$$

where η is the threshold, chosen for a specific false alarm probability. H_1 denotes a triggered alarm.

[1]The concept of degrees of freedom is a component of the definition of the χ^2 distribution. Please refer to Scharf and C. See [31] for more details.

III. Feasibility of Attack

In this section, it is assumed that a malicious third party wants to break the control system described in Section II. An attack model similar to the replay attack in computer security is defined and the feasibility of such kind of attacks on the control system is analyzed. The analysis is then generalized to other classes of control systems.

The attacker is assumed to have the capability to

1) inject an external control input u_k^a into the system.
2) (conservatively) read all the sensor readings and modify them arbitrarily. The readings modified by the attacker are denoted by y_k'.

Given these capabilities, the attacker is assumed to implement an attack strategy, which can be divided into two stages.

1) The attacker records a sufficient number of y_ks without giving any input to the system.
2) The attacker gives a sequence of desired control input while replaying the previous recorded y_ks.

Remark 1: The attack on the sensors can be done by breaking the cryptography algorithm. Another way to perform an attack, which is thought to be much harder to defend, is to use physical attacks. For example, the readings of a temperature sensor can be manipulated if the attacker puts a heater near the sensor.

Remark 2: When the system is under attack, the controller cannot perform closed-loop control, as the sensory information is not available. Therefore, control performance of the system cannot be guaranteed during replay attack. The only way to counter such an attack is to detect it happening.

Remark 3: In the attacking stage, the goal of the attacker is to make the fake readings y_k's look like normal y_ks. Replaying the previous y_ks is just the easiest way to achieve this goal. There are other methods, such as machine learning or system identification, to generate a fake sequence of readings. To provide a unified framework, y_k's can be thought as the output of the following virtual system under normal operation:

$$x_{k+1}' = Ax_k' + Bu_k' + w_k', \quad y_k' = Cx_k' + v_k' \tag{13}$$

$$\hat{x}_{k+1|k}' = A\hat{x}_k' + Bu_k' \tag{14}$$

$$\hat{x}_{k+1}' = \hat{x}_{k+1|k}' + K\left(y_{k+1}' - C\hat{x}_{k+1|k}'\right) \tag{15}$$

$$u_k' = L\hat{x}_k' \tag{16}$$

with initial conditions x_0' and $\hat{x}_{0|-1}'$. For the replay attack, suppose that the attacker records the sequence y_ks from time t onward. The virtual system, then, is just a time shifted version of the real system, with $x_k' = x_{t+k}$, $\hat{x}_{k|k}' = \hat{x}_{t+k|t+k}$. w_k' and v_k' will still be independent of each other and of w_k and v_k, since the original process and sensor noises are white Gaussian.

Remark 4: While the attacker can only record the readings for a finite time before the attack, in general this recording will be long enough to cause damage to the system. In addition, if the attacker does find the recording to be too short, the recorded measurements can be looped to form a longer replay, because the system is in steady state. Thus, for the sake of simplicity, we assume the length of recording to be infinite.

Theorem 2: Consider the system and detector of Section II, and an attacker running the virtual system given by (13).

Let α_k and β_k be, respectively, the false alarm and detection rates of the system at time k. If $\mathscr{A} \triangleq (A+BL)(I-KC)$ is stable

$$\lim_{k\to\infty} \beta_k = \alpha_k. \tag{17}$$

Conversely, if $\mathscr{A}$ is unstable

$$\lim_{k\to\infty} \beta_k = 1. \tag{18}$$

Proof: Suppose the system is under attack, the estimation of the Kalman filter $\hat{x}_{k|k-1}$ can be rewritten as:

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k \\ &= (A+BL)\,\hat{x}_k \\ &= (A+BL)\left[\hat{x}_{k|k-1} + K\left(y'_k - C\hat{x}_{k|k-1}\right)\right] \\ &= (A+BL)(I-KC)\,\hat{x}_{k|k-1} + (A+BL)\,Ky'_k. \end{aligned}\tag{19}$$

For the virtual system, the same equation holds true for $\hat{x}'_{k|k-1}$

$$\hat{x}'_{k+1|k} = (A+BL)(I-KC)\,\hat{x}'_{k|k-1} + (A+BL)\,Ky'_k. \tag{20}$$

Thus[2]

$$\hat{x}_{k|k-1} - \hat{x}'_{k|k-1} = \mathscr{A}^k\left(\hat{x}_{0|-1} - \hat{x}'_{0|-1}\right). \tag{21}$$

Let $\hat{x}_{0|-1} - \hat{x}'_{0|-1} \triangleq \zeta$. Now, the residue can be written as:

$$y'_k - C\hat{x}_{k|k-1} = \left(y'_k - C\hat{x}'_{k|k-1}\right) - C\mathscr{A}^k\zeta \tag{22}$$

and

$$\begin{aligned} g_k = \sum_{i=k-\mathscr{T}+1}^{k} \Big[&\left(y'_i - C\hat{x}'_{k|k-1}\right)^T \mathscr{P}^{-1}\left(y'_i - C\hat{x}'_{k|k-1}\right) \\ &+ 2\left(y'_i - C\hat{x}'_{k|k-1}\right)^T \mathscr{P}^{-1} C\mathscr{A}^i\zeta \\ &+ \zeta^T\left(\mathscr{A}^i\right)^T C^T \mathscr{P}^{-1} C\mathscr{A}^i\zeta \Big]. \end{aligned}\tag{23}$$

By the definition of virtual system, it is known that $y'_k - C\hat{x}'_{k|k-1}$ follows the exact distribution as $y_k - C\hat{x}_{k|k-1}$. Hence, if $\mathscr{A}$ is stable, the second and third terms in (23) will converge to 0. Thus, $y'_k - C\hat{x}_{k|k-1}$ will converge to the same distribution as $y_k - C\hat{x}_{k|k-1}$, and the detection rate (β) given by χ^2 detector will converge to the false alarm rate (α).

If, on the other hand, $\mathscr{A}$ is unstable, the attacker cannot replay y'_k for long, because g_k will soon become unbounded, implying $\beta_k \to 1$. In this case, the system is resilient to the replay attack, as the detector will be able to detect the attack. ∎

Remark 5: During the transient period when the attack starts, the value of g_k in the above formulation will jump to a high value. It is, however, not very difficult for a sophisticated attacker to reduce this jump in values, even remove it completely, by designing the start of the attack more carefully than in the above formulation. For example, an attacker could ramp up the introduction of false measurements with time, or he could wait till the initial part of the recording is close to the current measurements. Reliance on the transient jump in g_k is not a wise move.

It turns out the feasibility result derived for a special estimator, controller, and detector implementation is actually applicable to a large class of systems, with a slightly stronger condition. Suppose the state of the estimator at time k is s_k and it evolves according to

$$s_{k+1} = f(s_k, y_k). \tag{24}$$

Let the seminorm of f be defined as

$$\|f\| \triangleq \sup_{\Delta s\neq 0, y, s} \frac{\|f(s,y) - f(s+\Delta s, y)\|}{\|\Delta s\|}. \tag{25}$$

Suppose that the defender is using the following criterion to perform intrusion detection:

$$g(s_k, y_k)\, g_k \underset{H_1}{\overset{H_0}{\lessgtr}} v \tag{26}$$

where g is an arbitrary continuous function and η is a threshold value for g.

Theorem 3: If $\|f\| \leq 1$, then

$$\lim_{k\to\infty} g\left(s_k, y'_k\right) - g\left(s'_k, y'_k\right) = 0 \tag{27}$$

where s'_k is the state variable of the virtual system. The detection rate β_k at time k converges to

$$\lim_{k\to\infty} \beta_k - \alpha_k = 0 \tag{28}$$

where α_k is the false alarm rate of the virtual system at time k.

Proof: Because of space limit, only an outline of the proof is given. Initially, $\|f\| \leq 1$ will ensure that s_k converges to s'_k. By the continuity of g, $g(s_k, y'_k)$ converges to $g(s'_k, y'_k)$. The detection rate of the system and the false alarm rate of the virtual system are given by

$$\beta_k = \text{Prob}\left(g\left(s_k, y'_k\right) > \eta\right) \tag{29}$$

$$\alpha_k = \text{Prob}\left(g\left(s'_k, y'_k\right) > \eta\right). \tag{30}$$

Hence, β_k converges to α_k. ∎

Remark 6: If Theorem 3 is applied to the LTI system under consideration, the case of LQG controller, Kalman filter, and χ^2 detector then becomes just a special case, where the state s_k of the estimator at time k is $y_{k-\mathscr{T}+1}, y_{k-\mathscr{T}+2}, \ldots, y_k$ and $\hat{x}_{k-\mathscr{T}+1|k-\mathscr{T}}, \hat{x}_{k-\mathscr{T}+2|k-\mathscr{T}+1}, \ldots, \hat{x}_{k|k-1}$. The function f is given by (4) and g is given by (12). The condition for resiliency thus derived is that the largest singular value of $\mathscr{A}$ is less than one. This is a more restrictive condition than the one derived in Theorem 2.

Remark 7: For linear systems, the stability of $\mathscr{A}$ implies that the detection rate converges to the false alarm rate. If $\mathscr{A}$ is unstable, the detection rate goes to one. For the larger class of systems, $\|f\| \leq 1$ is a sufficient condition for the detection rate converging to the false alarm rate.

IV. Countermeasures Against Attacks

A. Using Unstable $\mathscr{A}$

The result of Theorem 2, is that if $\mathscr{A}$, is unstable, then g_k goes to infinity exponentially fast, triggering the detector. One possible way to counter the replay attack is to redesign the control system, i.e., using nonoptimal estimation and control gain matrices K and L, so that $\mathscr{A}$ becomes unstable while

[2]For simplicity, here the time the attack begins is considered as time 0.

maintaining the stability of the system. However, since K and L are not optimal in the LQG sense, the cost increase.

The LQG cost for using nonoptimal K and L is now characterized. It is known that

$$x_{k+1} = Ax_k + Bu_k + w_k = Ax_k + BL\hat{x}_k + w_k \tag{31}$$

and

$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k = (A + BL)\,\hat{x}_k \\ \hat{x}_{k+1} &= \hat{x}_{k+1|k} + K\left(y_{k+1} - C\hat{x}_{k+1|k}\right) \\ &= (I - KC)\,(A + BL)\,\hat{x}_k + Ky_{k+1} \\ &= \big(I - KC\big)\big(A + BL\big)\hat{x}_k + K\big(Cx_{k+1} + v_{k+1}\big) \\ &= KCAx_k + (A + BL - KCA)\,\hat{x}_k \\ &\quad + KCw_k + Kv_{k+1}. \end{aligned} \tag{32}$$

Equations (31) and (32) can be written in matrix form as follows:

$$\begin{aligned}\begin{pmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{pmatrix} &= \begin{pmatrix} A & BL \\ KCA & A + BL - KCA \end{pmatrix}\begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} \\ &\quad + \begin{pmatrix} I \\ KC \end{pmatrix} w_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v_{k+1}. \end{aligned} \tag{33}$$

Let $\hat{A}$ be defined as

$$\hat{A} \triangleq \begin{pmatrix} A & BL \\ KCA & A + BL - KCA \end{pmatrix}. \tag{34}$$

Let $\hat{R}$ be covariance matrix of final terms of (33)

$$\hat{R} \triangleq \begin{pmatrix} I \\ KC \end{pmatrix} Q \begin{pmatrix} I & C^T K^T \end{pmatrix} + \begin{pmatrix} 0 \\ K \end{pmatrix} R \begin{pmatrix} 0 & K^T \end{pmatrix}. \tag{35}$$

The LQG cost for nonoptimal K and L can now be derived, which is given by the following theorem.

Theorem 4: The LQG cost of using an arbitrary estimation and control gain K and L is

$$J = \operatorname{trace}\left(\begin{pmatrix} W & 0 \\ 0 & L^T U L \end{pmatrix}\hat{Q}\right) \tag{36}$$

where $\hat{Q}$ is the solution of the following Lyapunov equation:

$$\hat{Q} = \hat{A}\hat{Q}\hat{A}^T + \hat{R}. \tag{37}$$

Proof: Since a fixed gain controller and estimator is used

$$J = \lim_{k\to\infty} x_k^T W x_k + u_k^T U u_k \tag{38}$$

which can then be written in matrix form as

$$\begin{aligned} J &= \lim_{k\to\infty} \begin{pmatrix} x_k^T & u_k^T \end{pmatrix}\begin{pmatrix} W & 0 \\ 0 & U \end{pmatrix}\begin{pmatrix} x_k \\ u_k \end{pmatrix} \\ &= \lim_{k\to\infty} \operatorname{trace}\left(\begin{pmatrix} W & 0 \\ 0 & U \end{pmatrix}\begin{pmatrix} x_k \\ u_k \end{pmatrix}\begin{pmatrix} x_k^T & u_k^T \end{pmatrix}\right) \\ &= \lim_{k\to\infty} \operatorname{trace}\left(\begin{pmatrix} W & 0 \\ 0 & L^T U L \end{pmatrix}\operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix}\right)\right). \end{aligned} \tag{39}$$

Let

$$\hat{Q} \triangleq \lim_{k\to\infty} \operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix}\right). \tag{40}$$

By (33)

$$\operatorname{Cov}\left(\begin{pmatrix} x_{k+1} \\ u_{k+1} \end{pmatrix}\right) = \hat{A}\operatorname{Cov}\left(\begin{pmatrix} x_k \\ u_k \end{pmatrix}\right)\hat{A}^T + \hat{R}. \tag{41}$$

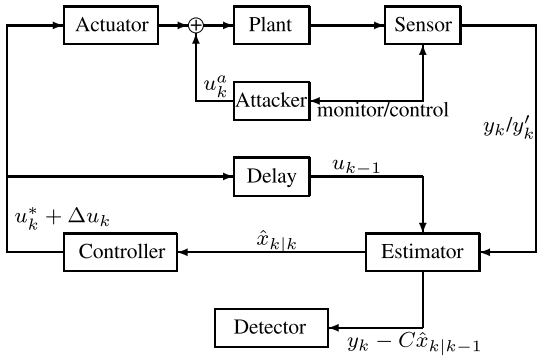

Fig. 1. System diagram.

Taking the limit on both sides,

$$\hat{Q} = \hat{A}\hat{Q}\hat{A}^T + \hat{R}.$$

Therefore, the LQG cost is given by

$$J = \operatorname{trace}\left(\begin{pmatrix} W & 0 \\ 0 & L^T U L \end{pmatrix}\hat{Q}\right).$$

■

Remark 8: There might not be enough freedom to redesign the control, which is required for this countermeasure to be implemented. The inclusion of this method is, however, not just for the sake of completeness—as g_k increases exponentially, this method therefore provides the highest asymptotic probability of detection, in the case that it is feasible.

It is, however, likely that the design constraints do not allow $\mathscr{A}$ to be unstable, due to constraints on operating costs, safety parameters, etc. In such cases, two other countermeasures are proposed to detect the replay attacks in the following section.

B. Noisy Control

The main problem of the combination of a LQG controller and a Kalman filter is that the whole control system is fairly static, which renders it vulnerable to a replay attack. To detect such a replay attack, one methodology is to redesign the control signal as

$$u_k = u_k^* + \Delta u_k \tag{42}$$

where u_k^* is the optimal LQG control signal and the sequence Δu_k is drawn from an i.i.d. Gaussian distribution with zero mean and covariance $\mathscr{Q}$, and independent of u_k^*. Fig. 1 shows the system diagram, including the attacker and the noisy control.

The sequence Δu_k acts as a time-stamped authentication signal. It is chosen to be zero mean so as not to introduce any bias into the system. The presence of this extra authentication signal will cause the controller to not be optimal—to decrease the vulnerability of the system to the attack, the control performance must be sacrificed. Theorem 5 characterizes the dependence of the loss of LQG performance on the strength of the authentication signal.

Theorem 5: The LQG performance after adding Δu_k is given by

$$J' = J + \underbrace{\operatorname{trace}\big[\big(U + B^T S B\big)\mathscr{Q}\big]}_{\Delta J}. \tag{43}$$

Proof: The proof is given in Appendix I. ■

Remark 9: As the LQG performance is still bounded, the system is stable.

1) χ^2 Detector: Theorem 6 shows the effectiveness of the detector using the noisy-control scheme.

Theorem 6: In the absence of an attack

$$E\left[\left(y_k - C\hat{x}_{k|k-1}\right)^T \mathscr{P}^{-1}\left(y_k - C\hat{x}_{k|k-1}\right)\right] = m. \tag{44}$$

Under attack

$$\lim_{k\to\infty} E\left[\left(y'_k - C\hat{x}_{k|k-1}\right)^T \mathscr{P}^{-1}\left(y'_k - C\hat{x}_{k|k-1}\right)\right] = m + \underbrace{2\cdot\mathrm{trace}\left(C^T\mathscr{P}^{-1}C\mathscr{U}\right)}_{\Delta g_k} \tag{45}$$

where $\mathscr{U}$ is the solution to the following Lyapunov equation:

$$\mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T. \tag{46}$$

Proof: Equation (44) can be easily proved using Theorem 1. $\hat{x}_{k+1|k}$ can be rewritten as

$$\hat{x}_{k+1|k} = \mathscr{A}\hat{x}_{k|k-1} + (A+BL)Ky'_k + B\Delta u_k. \tag{47}$$

Similarly, for the virtual system

$$\hat{x}'_{k+1|k} = \mathscr{A}\hat{x}'_{k|k-1} + (A+BL)Ky'_k + B\Delta u'_k. \tag{48}$$

Thus

$$\hat{x}_{k|k-1} - \hat{x}'_{k|k-1} = \mathscr{A}^k\left(\hat{x}_{0|-1} - \hat{x}'_{0|-1}\right) + \sum_{i=0}^{k-1}\mathscr{A}^{k-i-1}B\left(\Delta u_i - \Delta u'_i\right). \tag{49}$$

Hence

$$y'_k - C\hat{x}_{k|k-1} = y'_k - C\hat{x}'_{k|k-1} - C\mathscr{A}^k\left(\hat{x}_{0|-1} - \hat{x}'_{0|-1}\right) - C\sum_{i=0}^{k-1}\mathscr{A}^{k-i-1}B\left(\Delta u_i - \Delta u'_i\right). \tag{50}$$

The first term in (50) has the same distribution as $y_k - C\hat{x}_{k|k-1}$, and the second term converges to zero when $\mathscr{A}$ is stable. One can observe that Δu_i is independent of $y'_k - C\hat{x}'_{k|k-1}$ of the virtual system. In addition, for the virtual system, $y'_k - C\hat{x}'_{k|k-1}$ is independent of $\Delta u'_i$. Hence

$$\begin{aligned}\lim_{k\to\infty}\mathrm{Cov}\left(y'_k - C\hat{x}_{k|k-1}\right) &= \lim_{k\to\infty}\mathrm{Cov}\left(y'_k - C\hat{x}'_{k|k-1}\right) + \sum_{i=0}^{\infty}\mathrm{Cov}\left(C\mathscr{A}^iB\Delta u_i\right) \\ &\quad + \sum_{i=0}^{\infty}\mathrm{Cov}\left(C\mathscr{A}^iB\Delta u'_i\right) \\ &= \mathscr{P} + 2\sum_{i=0}^{\infty}C\mathscr{A}^iB\mathscr{Q}B^T\left(\mathscr{A}^i\right)^TC^T. \end{aligned}\tag{51}$$

By the definition of $\mathscr{U}$ from Theorem 5, the Lyapunov equation (46) can be solved to yield $\mathscr{U}$ as

$$\mathscr{U} = \sum_{i=0}^{\infty}\mathscr{A}^iB\mathscr{Q}B^T\left(\mathscr{A}^i\right)^t. \tag{52}$$

Hence

$$\lim_{k\to\infty}\mathrm{Cov}\left(y'_k - C\hat{x}_{k|k-1}\right) = \mathscr{P} + 2C\mathscr{U}C^T \tag{53}$$

and

$$\begin{aligned}\lim_{k\to\infty}E\left[\left(y'_k - C\hat{x}_{k|k-1}\right)^T\mathscr{P}^{-1}\left(y'_k - C\hat{x}_{k|k-1}\right)\right] &= \mathrm{trace}\left[\lim_{k\to\infty}\mathrm{Cov}\left(y'_k - C\hat{x}_{k|k-1}\right)\times\mathscr{P}^{-1}\right] \\ &= m + 2\cdot\mathrm{trace}\left(C^T\mathscr{P}^{-1}C\mathscr{U}\right).\end{aligned}\tag{54}$$

■

Corollary 1: In the absence of an attack, the expectation of the χ^2 detector is

$$E[g_k] = m\mathscr{T}. \tag{55}$$

Under attack, the asymptotic expectation becomes

$$\lim_{k\to\infty}E[g_k] = m\mathscr{T} + 2\cdot\mathrm{trace}\left(C^T\mathscr{P}^{-1}C\mathscr{U}\right)\mathscr{T}. \tag{56}$$

The difference in the expectations of g_k with and without attack proves that the detection rate does not converge to the false alarm rate.

In a SISO system, there is only one way to insert the random signal, and only one way to observe it. Thus, to achieve a certain detection rate, a certain performance loss would have to be accepted. In MIMO systems, the authentication signal can be inserted on one input or on many, with different strengths, independent or not. Similarly, the responsiveness of the system to the signal can be checked for one output or many. The authentication signal Δu_k can be optimized such that the detection requirements are met while minimizing the effect on controller performance. As the authentication signal has to be zero mean, the design hinges on the covariance matrix $\mathscr{Q}$. Let the optimal value of $\mathscr{Q}$, based on the design requirements, be denoted by $\mathscr{Q}^*$.

The optimization problem can be setup in two ways. Initially, the LQG performance loss (ΔJ) can be constrained to be less than some design parameters Θ, and the increase (Δg_k) in the expected value of the quadratic residues in case of an attack maximized. In this case, the optimal $\mathscr{Q}^*$ is the solution to the optimization problem

$$\begin{aligned}\max_{\mathscr{Q}} \quad & \mathrm{trace}\left(C^T\mathscr{P}^{-1}C\mathscr{U}\right) \\ \text{s.t.} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q}\succeq 0 \\ & \mathrm{trace}\left[\left(U + B^TSB\right)\mathscr{Q}\right] \le \Theta.\end{aligned}\tag{57}$$

Theorem 7: There exists an optimal $\mathscr{Q}^*$ for (57) of the following form:

$$\mathscr{Q}^* = \alpha\omega\omega^T \tag{58}$$

where $\alpha > 0$ is a scalar and ω is a vector such that $\omega^T\omega = 1$.

Proof: Suppose that $\mathscr{Q}^*$ is the optimal solution of (57) and $\mathscr{U}^*$ is the solution of

$$\mathscr{U}^* - B\mathscr{Q}^*B^T = \mathscr{A}\mathscr{U}^*\mathscr{A}^T. \tag{59}$$

Because $\mathscr{Q}^*$ is positive semidefinite, it is known that

$$\mathscr{Q}^* = \Omega \underbrace{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}}_{\Lambda} \Omega^T \tag{60}$$

where $\lambda_i \geq 0$s are the eigenvalues of $\mathscr{Q}^*$ and $\Omega = (\omega_1, \omega_2, \ldots, \omega_p)$ is an orthonormal matrix, such that $\omega_i \in \mathbb{R}^p$. Thus, $\mathscr{Q}^*$ can be written as the sum of p rank 1 matrices

$$\mathscr{Q}^* = \sum_{i=1}^{p} \lambda_i \omega_i \omega_i^T. \tag{61}$$

Let $\mathscr{Q}_i$ be defined as

$$\mathscr{Q}_i \stackrel{\Delta}{=} \alpha_i \omega_i \omega_i^T \tag{62}$$

where $\alpha_i > 0$ is chosen such that

$$\text{trace}\big[(U + B^T S B)\mathscr{Q}_i\big] = \Theta. \tag{63}$$

In addition, let $\mathscr{U}_i$ be defined as the solution of the following Lyapunov equation:

$$\mathscr{U}_i - B\mathscr{Q}_i B^T = \mathscr{A}\mathscr{U}_i\mathscr{A}^T. \tag{64}$$

It is clear that the optimal $\mathscr{Q}^*$ must satisfy

$$\text{trace}\big[(U + B^T S B)\mathscr{Q}^*\big] = \Theta. \tag{65}$$

Therefore, as

$$\mathscr{Q}^* = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathscr{Q}_i \tag{66}$$

it can be observed that

$$\begin{aligned} \Theta &= \text{trace}\big[(U + B^T S B)\mathscr{Q}^*\big] \\ &= \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \text{trace}\big[(U + B^T S B)\mathscr{Q}_i\big] \\ &= \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \Theta \end{aligned} \tag{67}$$

which proves that

$$\sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} = 1. \tag{68}$$

Furthermore, it is easy to observe that since Lyapunov equation is linear

$$\mathscr{U}^* = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \mathscr{U}_i. \tag{69}$$

Hence

$$\text{trace}\big(C^T \mathscr{P}^{-1} C \mathscr{U}^*\big) = \sum_{i=1}^{p} \frac{\lambda_i}{\alpha_i} \text{trace}\big(C^T \mathscr{P}^{-1} C \mathscr{U}_i\big). \tag{70}$$

Thus, $\mathscr{Q}^*$ is a convex combination of p feasible $\mathscr{Q}_i$s. Because $\mathscr{Q}^*$ is optimal, we know that for any $\lambda_i > 0$, the corresponding $\mathscr{Q}_i$ must also be optimal, which finishes the proof. ■

Remark 10: The fact that $\mathscr{Q}^*$ has rank 1, has a direct bearing on the computation requirement. The number of independent random noise generators required is equal to the rank of $\mathscr{Q}^*$. Naïvely, one would have to use one independent random noise generator per system input, to protect all of them. Irrespective of the number of system inputs, the rank of $\mathscr{Q}^*$ is, however, always one, which means that a single random noise generator will suffice for a system with any number of inputs.

Remark 11: Ideally, if there is a design constraint on the LQG cost, one would try to optimize the detection rate. It, however, can be shown that under attack g_k follows a generalized χ^2 distribution, and no analytical form for the detection rate can be accrued. Thus, only the maximization of the expectation in the case of an attack is attempted, with the intuition that the detection rate in such a case will be close to the maximum possible.

Remark 12: It can be observed from Theorems 5 and 6 that the increase (ΔJ) in LQG cost and increase (Δg_k) in the expectation of the quadratic residues are linear functions of the noise covariance matrix $\mathscr{Q}$. Thus, the optimization problem is a semidefinite programming problem, and hence can be solved efficiently. Furthermore, it can be observed that if the constraints are changed from Θ to $\alpha\Theta$, the optimal $\mathscr{Q}^*$ will be changed to $\alpha\mathscr{Q}$.

Another way of optimizing is to constrain the increase (Δg_k) in the expected values of the quadratic residues to be above a fixed value Γ, thereby guaranteeing a certain rate of detection, and the performance loss (ΔJ) can be minimized. The optimal $\mathscr{Q}^*$ is now the solution to the optimization problem

$$\begin{aligned} \min_{\mathscr{Q}} \quad & \text{trace}\big[(U + B^T S B)\mathscr{Q}\big] \\ \text{s.t.} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q} \succeq 0 \\ & \text{trace}\big(C^T \mathscr{P}^{-1} C \mathscr{U}\big) \geq \Gamma. \end{aligned} \tag{71}$$

Remark 13: The solutions of the two optimization problems given in 57 and 71 will be scalar multiples of each other, thus solving either optimization problem guarantees same performance. An intuitive way to observe this, is that $\mathscr{Q}^*$ measures the sensitivity of the system output to the different inputs, thus making it a system property.

The results of Remarks 12 and 13 can be applied to decouple the design of the signal into two steps. Because there is a linear relationship between the performance loss or increase in residues to the amplitude of the signal, the form of the $\mathscr{Q}^*$ can first be ascertained. The norm of $\mathscr{Q}^*$ can then be designed in the second step, considering either the detector performance or the controller performance. These design steps are further shown in Section V-D.

2) Cross Correlator: Implementing the χ^2 detector requires the implementation of a Kalman estimator. In some systems, a Kalman estimator, however, might not be feasible, because of noise characteristics or system observability. The noisy-control countermeasure, however, can still be applied, to virtually any controller and detector, as long as a virtual system can be implemented. We add a signal $\Delta u_k \sim \mathcal{N}(0, \sigma^2)$. The effect of the control input on the virtual system can be calculated and the outputs are compared. The system of the previous

section, with a Kalman estimator and an LQG control, can be used as an example to show this countermeasure. The system evolution equation is

$$\begin{pmatrix} x_{k+1} \\ \hat{x}_{k+1} \end{pmatrix} = \underbrace{\begin{pmatrix} A & BL \\ KCA & A+BL-KCA \end{pmatrix}}_{\hat{A}} \begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} + \underbrace{\begin{pmatrix} B \\ B \end{pmatrix}}_{\hat{B}} \Delta u_k + \begin{pmatrix} I \\ KC \end{pmatrix} w_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v_{k+1} \tag{72}$$

and the measurement equation is

$$y_k = \underbrace{\begin{pmatrix} C & 0 \end{pmatrix}}_{\hat{C}} \begin{pmatrix} x_k \\ \hat{x}_k \end{pmatrix} + v_k. \tag{73}$$

Note that $\hat{A}$ is the same, as defined in (34). For the virtual system, the system evolution equation is

$$\begin{pmatrix} x'_{k+1} \\ \hat{x}'_{k+1} \end{pmatrix} = \hat{A} \begin{pmatrix} x'_k \\ \hat{x}'_k \end{pmatrix} + \hat{B} \Delta u'_k + \begin{pmatrix} I \\ KC \end{pmatrix} w'_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v'_{k+1} \tag{74}$$

and the measurement equation is

$$y'_k = \hat{C} \begin{pmatrix} x'_k \\ \hat{x}'_k \end{pmatrix} + v'_k. \tag{75}$$

It is assumed that $x_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $x'_0 \sim \mathcal{N}(\bar{x}_0, \Sigma)$, $\Delta u \sim \mathcal{N}(0, \mathscr{Q})$, $w_k \sim \mathcal{N}(0, Q)$, $w'_k \sim \mathcal{N}(0, Q)$, $v_k \sim \mathcal{N}(0, R)$, and $v'_k \sim \mathcal{N}(0, R)$ are all independent of each other. Let the detector run another virtual system, which is connected directly to the controller and cannot be attacked by the attacker

$$\begin{pmatrix} x''_{k+1} \\ \hat{x}''_{k+1} \end{pmatrix} = \hat{A} \begin{pmatrix} x''_k \\ \hat{x}''_k \end{pmatrix} + \hat{B} \Delta u_k + \begin{pmatrix} I \\ KC \end{pmatrix} w''_k + \begin{pmatrix} 0 \\ K \end{pmatrix} v''_{k+1} \tag{76}$$

and the measurement equation is

$$y''_k = \hat{C} \begin{pmatrix} x''_k \\ \hat{x}''_k \end{pmatrix} + v''_k. \tag{77}$$

Consider the detector variable $g_k = y'^T y'' = \text{trace}\left(y' y''^T\right)$. It can be proved that in the absence of a replay attack

$$E\left[y' y''^T\right] = \hat{C} \mathscr{R} \hat{C}^T \tag{78}$$

where $\mathscr{R}$ is the solution of the following Lyapunov equation:

$$\hat{A} \mathscr{R} \hat{A}^T + \hat{B} \mathscr{Q} \hat{B}^T = \mathscr{R}. \tag{79}$$

If the attacker replays the outputs y or if he is running another virtual system, the $\Delta u'$ generated by the attacker will be independent of the Δu used in the controller's virtual system. In case of either form of attack, $\mathscr{R}$ becomes zero, causing $E[y' y''^T]$ to drop to zero as well. We can thus detect the absence of the authentication signal in the output and hence, the attack.

Similar to the χ^2 detector, in the case of MIMO systems, the covariance matrix $\mathscr{Q}$ can be optimized, such that the detection requirements are met while minimizing the effect on controller performance. Just like the previous case, the optimization problem can be setup in two ways. First, the LQG performance loss (ΔJ) can be constrained to be less than some design parameters Θ, and the increase (Δg_k) in the expected value of the correlator output in case of an attack maximized. In this case, the optimal $\mathscr{Q}^*$ is the solution to the optimization problem

$$\begin{aligned} \max_{\mathscr{Q}} \quad & \text{trace}\left(\hat{C} \mathscr{R} \hat{C}^T\right) \\ \text{s.t.} \quad & \hat{A} \mathscr{R} \hat{A}^T + \hat{B} \mathscr{Q} \hat{B}^T = \mathscr{R} \\ & \mathscr{Q} \succeq 0 \\ & \text{trace}\left[\left(U + B^T S B\right) \mathscr{Q}\right] \leq \Theta. \end{aligned} \tag{80}$$

Second, the increase (Δg_k) in the expected values of the quadratic residues can be constrained to be above a fixed value Γ, thereby guaranteeing a certain rate of detection, and the performance loss (ΔJ) can be minimized. The optimal $\mathscr{Q}^*$ is now the solution to the optimization problem

$$\begin{aligned} \min_{\mathscr{Q}} \quad & \text{trace}\left[\left(U + B^T S B\right) \mathscr{Q}\right] \\ \text{s.t.} \quad & \hat{A} \mathscr{R} \hat{A}^T + \hat{B} \mathscr{Q} \hat{B}^T = \mathscr{R} \\ & \mathscr{Q} \succeq 0 \\ & \text{trace}\left(\hat{C} \mathscr{R} \hat{C}^T\right) \geq \Gamma. \end{aligned} \tag{81}$$

Theorem 8: There exists and optimal $\mathcal{Q}^*$ for (80) of the following form:

$$\mathcal{Q}^* = \alpha \omega \omega^T \tag{82}$$

where $\alpha > 0$ is a scalar and ω is a vector with $\omega^T \omega = 1$.

Proof: The proof is very similar to that of Theorem 7, hence is omitted. ■

Remark 14: Like the χ^2 detector, only the maximization of the expectation is attempted. The optimization problems are linear and generate optimal $\mathscr{Q}^*$s, which are multiples of each other.

V. Simulation

In this section, some simulation results pertaining to the detection of replay attacks on one system using different countermeasures is given. For the system, a simplified version of the Tennessee Eastman control challenge oroblem [32] is used. Ricker [33] derived an LTI dynamic model of the plant in its base state, and a corresponding robust controller. The system is given as a transfer function of four outputs and inputs[3]

$$\mathbf{y} = \begin{pmatrix} F_4 \\ P \\ y_{A3} \\ V_L \end{pmatrix} = \mathbf{G}\mathbf{u} = \begin{pmatrix} g_{11} & 0 & 0 & g_{14} \\ g_{21} & 0 & g_{23} & 0 \\ 0 & g_{32} & 0 & 0 \\ 0 & 0 & 0 & g_{44} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}. \tag{83}$$

The attacker is considered to know the readings of all[4] the sensors, with the ability to hijack and modify them, but not the dynamics of the system. The only known fact is that the system is expected to be in a steady state for the duration of the attack. Of the 30 min for which the system is simulated, the attacker records the sensor readings for the first 15 min, and replays them to the controller for the next 15 min. The attack consists

[3]For detailed values of the transfer functions, see [33].

[4]The requirement of control over all sensors can be weakened if the system can be decomposed into several weakly coupled subsystems, compromising sensors for one subsystem may be sufficient.

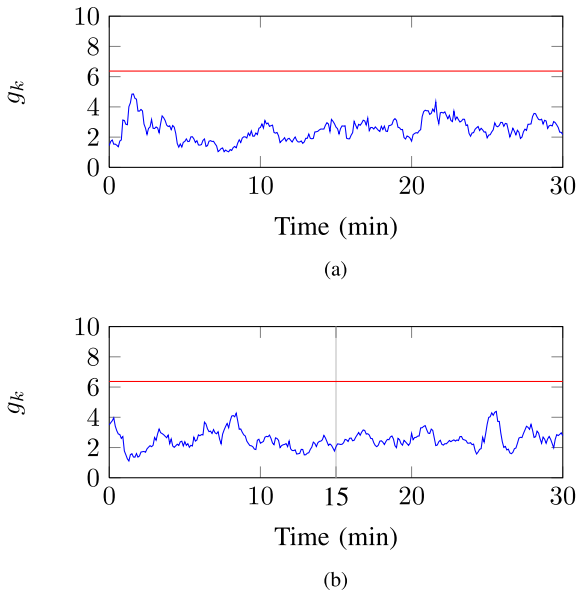

Fig. 2. g_k as a function of time during normal operation and a replay attack. This shows that the detector (with threshold at 99% shown) fails to detect the fall in g_k due to an attack. (a) Normal operation. (b) Replay attack.

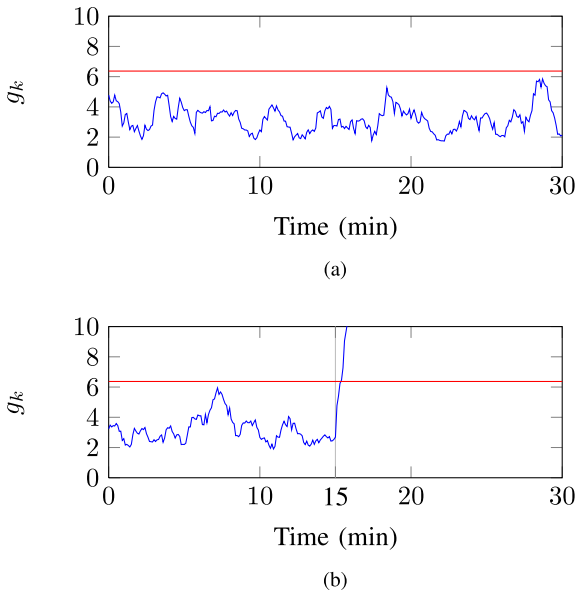

Fig. 3. g_k as a function of time during normal operation, and a replay attack, using a controller with unstable $\mathscr{A}$. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack. (a) Normal operation. (b) Replay attack.

for varying the control inputs of the plant, to try and evolve it into a potentially dangerous state. As no information from the system is conveyed to the controller, the system becomes open loop, without guarantees on the control performance. The only way to obtain the system back into the controlled state is to detect and mitigate the attack.

A. Feasibility of Attack

For the chemical plant, a W and U were chosen such that $\mathscr{A}$ is stable. A χ^2 detector with a window size of 10 samples (1 min) is used. Fig. 2(a) shows the value of g_k for a χ^2 detector, for the duration of 30 min, when no attack is present. Fig. 2(b) shows the value of g_k when an attack occurs after the first 15 min. It can be observed that there is no appreciable statistical difference in g_k when an attack is present, making detection impossible.

Thus, executing the attack without being detected is feasible.

B. Unstable $\mathscr{A}$

It is assumed that the design parameters are flexible enough to allow $\mathscr{A}$ to be unstable. K and L are generated randomly such that they form a good estimator-controller pair, such that $\mathscr{A}$ is unstable. A χ^2 detector with a window size of 10 samples (1 min) is used. Fig. 3 shows the value of g_k in normal operation and when an attack occurs after the first 15 min. It can be observed that the instability in $\mathscr{A}$ causes a change in g_k when an attack is present, which can be detected.

C. χ^2 Detector, Nonoptimal

For this simulation, the estimator and controller are reverted to the original case of Section V-A. The countermeasure of noisy control is now used for the system. A χ^2 detector with a window size of 10 samples (1 min) is implemented. In this case, the authentication signal is not optimized. The expected increase in LQG cost is 10% of the optimal LQG cost. In this case, Fig. 4(a) shows the value of g_k for a χ^2 detector, for the duration of 30 min, when no attack is present. Fig. 4(b) shows the value of g_k when an attack occurs after the first 15 min. It can be observed that there is some differences in the statistical distribution of g_k with and without an attack.

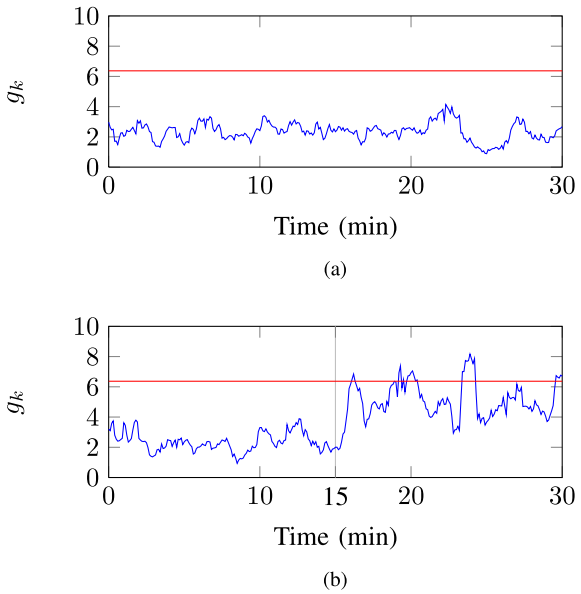

Fig. 4. g_k as a function of time during normal operation and a replay attack. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack. (a) Normal operation. (b) Replay attack.

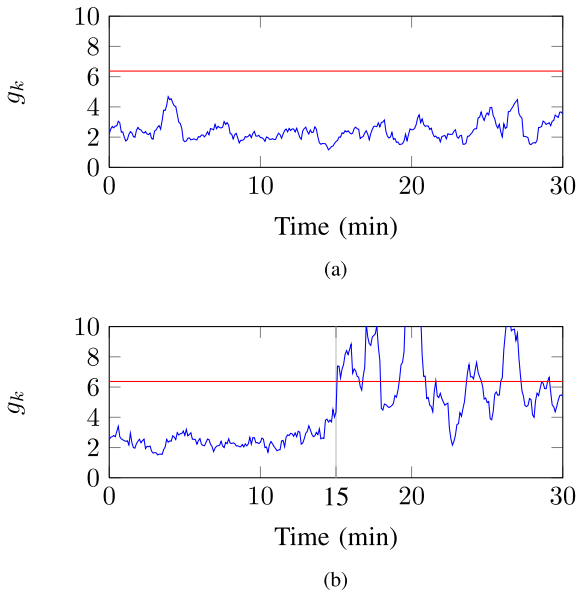

Fig. 5. g_k as a function of time during normal operation and a replay attack. This shows that the detector (with threshold at 99% shown) is able to detect the fall in g_k due to an attack. (a) Normal operation. (b) Replay attack.

D. χ^2 Detector, Optimal

This simulation is similar to the one in Section V-C, except that the authentication signal is optimized such that the expected increase in LQG cost is 10% of the optimal LQG cost. In this case, Fig. 5(a) shows the value of g_k for a χ^2 detector, for the duration of 30 min, when no attack is present. Fig. 5(b) shows the value of g_k when an attack occurs after the first 15 min. It can be observed there is a significant difference in the statistical distribution of g_k with and without an attack. The results of this simulation over that of Section V-C show the importance of optimizing the form of $\mathscr{Q}$.

In the next set of simulations, $\mathscr{Q}$ is scaled by 0.2, 0.4, 0.6, 0.8, and 1, which corresponds to setting Θ to 2%, 4%, 6%, 8%, and 10%, respectively. A sample set of 500 simulations was carried out to calculate the receiver operating characteristic (ROC) curves for each signal strength. These curves are shown in Fig. 6. In this case, probability of detection 1 min after the onset of the attack has been considered. It is easy to observe that the performance of the detector improves with increase in $\|\mathscr{Q}^*\|$, so an appropriate signal strength can be designed considering the tradeoff between the required ROC curves and allowed performance loss.

E. Cross-Correlator Detector, Optimal

In this simulation, we use a cross-correlator detector with a window size of 30 samples (3 min) and the authentication signal is optimized such that the expected increase in LQG cost is 20% of the optimal LQG cost. The expected value of the correlator output g_k is 30.996. Fig. 7(a) shows the correlator output, for the duration of 30 min, when no attack is present. Fig. 7(b) shows the correlator output when an attack occurs after the first 15 min. It can be observed that g_k drops significantly when an attack is in progress.

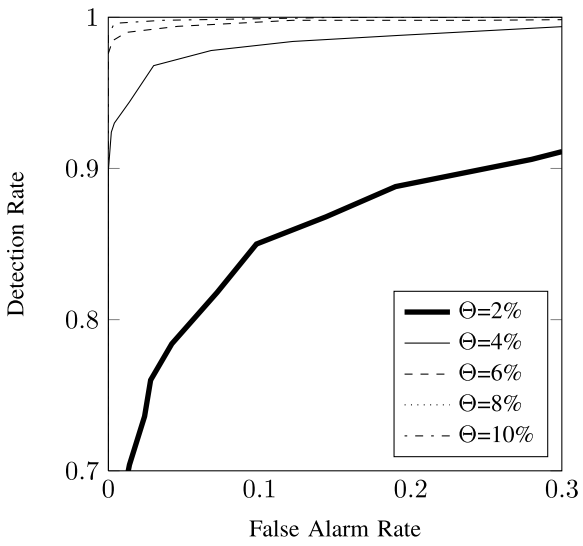

Fig. 6. ROC curves for detector, when Θ is 2% (dark solid line), 4% (thin solid line), 6% (dashed line), 8% (dotted line), and 10% (dashed-dotted line). Detection up to 1 s after attack is considered.

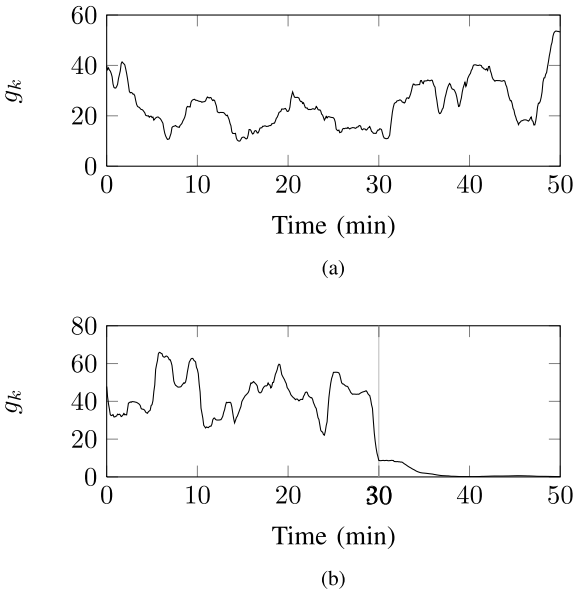

Fig. 7. g_k as a function of time during normal operation and a replay attack. This shows that the detector is able to detect the fall in g_k due to an attack. (a) Normal operation. (b) Replay attack.

VI. Conclusion

In this paper, a replay attack model on CPS was defined and the performance of the control system under the attack was analyzed. It was noted that for some control systems, the classical estimation-control-failure detection strategy is not resilient to a replay attack. For such a system, a technique using a noisy control authentication signal was provided to improve detection at the expense of control performance. The relationships between the performance loss, detection rate, and the strength of the authentication signal were characterized. A methodology for optimizing the signal was also provided, based on the tradeoff between the desired detection performance and allowable control performance loss. Three different sets of simulations were carried out to verify the theoretical results and show the optimization of the control signal.

In a real-world scenario, several engineering considerations could be employed to improve the proposed designs. For example, the authentication signal can be introduced into the system at random intervals instead of continuously, thus only affecting the performance for some time. Future work will concentrate on extending these techniques to more sophisticated attack models and to distributed control systems.

Appendix I
Proof of Theorem 5

Because of space constraints, only the basic outlines of the proof are given below. Let the sigma-algebra generated by $y_k, y_{k-1}, \ldots, y_0, \Delta u_{k-1}, \Delta u_{k-2} \ldots, \Delta u_0$ be defined as $\mathscr{F}_k$. For the proof of Theorem 5, Lemmas 1–3 are required.

Lemma 1: The following equations hold for the Kalman filter:

$$\hat{x}_{k|k} = E\left[x_k \middle| \mathscr{F}_k\right] \tag{84}$$

and

$$P_{k|k} = E\left[e_{k|k} e_{k|k}^T \middle| \mathscr{F}_k\right] \tag{85}$$

where $e_{k|k} = x_k - \hat{x}_{k|k}$.

Lemma 2: The following equation holds:

$$E\left[x_k^T \mathscr{S} x_k \middle| \mathscr{F}_k\right] = \operatorname{trace}\left(\mathscr{S} P_{k|k} + \hat{x}_{k|k}^T \mathscr{S} \hat{x}_{k|k}\right) \tag{86}$$

where $\mathscr{S}$ is any positive semidefinite matrix.

Now define

$$J_N \triangleq \min E\left[\sum_{i=0}^{N-1}\left(x_i^T W x_i + u_i^T U u_i\right)\right]. \tag{87}$$

By the definition of J', we know that

$$J' = \lim_{N\to\infty} \frac{J_N}{N}. \tag{88}$$

For fixed N

$$V_k(x_k) \triangleq \min E\left[\sum_{k=i}^{N-1}\left(x_i^T W x_i + u_i^T U u_i\right) \middle| \mathscr{F}_k\right] \tag{89}$$

and $V_N(x_N) = 0$. By definition, it is known that $E[V_0] = J_N$. In addition, from dynamic programming, V_k satisfies the following backward recursive equation:

$$V_k(x_k) = \min_{u_k^*} E\left[x_k^T W x_k + u_k^T U u_k + V_{k+1}(x_{k+1}) \middle| \mathscr{F}_k\right]. \tag{90}$$

Let

$$S_{k-1} \triangleq A^T S_k A + W - A^T S_k B\left(B^T S_K B + U\right)^{-1} B^T S_k A \tag{91}$$

$$\begin{aligned} c_{k-1} \triangleq\ & c_k + \operatorname{trace}\left[\left(W + A^T S_k A - S_{k-1}\right) P_{k-1|k-1}\right] \\ & + \operatorname{trace}(S_k Q) + \operatorname{trace}\left[\left(B^T S_k B + U\right)\mathscr{Q}\right] \end{aligned} \tag{92}$$

with $S_N = 0$ and $c_N = 0$.

Lemma 3: $V_k(x_k)$ is given by

$$V_k(x_k) = E\left[x_k^T S_k x_k \middle| \mathscr{F}_k\right] + c_k,\ k = N, N-1, \ldots, 0. \tag{93}$$

Proof: Equation (93) will be proved using backward induction. The induction hypothesis with $V_N = 0$ trivially satisfies (93). Now, suppose that V_{k+1} satisfies (93). Then, by (90)

$$\begin{aligned} V_k(x_k) &= \min E\left[x_k^T W x_k + u_k^T U u_k + V_{k+1}(x_{k+1}) \middle| \mathscr{F}_k\right] \\ &= \min E\left[x_k^T W x_k + \left(u_k^* + \Delta u_k\right)^T U\left(u_k^* + \Delta u_k\right)\right. \\ &\qquad \left. + x_{k+1}^T S_{k+1} x_{k+1} + c_{k+1} \middle| \mathscr{F}_k\right]. \end{aligned} \tag{94}$$

As it is known that u_k^* is measurable to $\mathscr{F}_k$ and Δu_k is independent of $\mathscr{F}_k$

$$E\left[\left(u_k^* + \Delta u_k\right)^T U\left(u_k^* + \Delta u_k\right)\right] = u_k^{*T} U u_k^* + \operatorname{trace}(U\mathscr{Q}). \tag{95}$$

x_{k+1} can be rewritten as

$$x_{k+1} = A x_k + B u_k^* + B\Delta u_k + w_k. \tag{96}$$

As Δu_k and w_k are independent of $Ax_k + Bu_k^*$

$$\begin{aligned} E\left(x_{k+1}^T S_{k+1} x_{k+1} \middle| \mathscr{F}_k\right) &= E\left(x_k^T A^T S_{k+1} A x_k \middle| \mathscr{F}_k\right) \\ &\quad + 2u_k^{*T} B^T S_{k+1} A\hat{x}_{k|k} + u_k^{*T} B^T S_{k+1} B u_k^* \\ &\quad + \operatorname{trace}\left(S_{k+1} Q\right) + \operatorname{trace}\left(B^T S_{k+1} B\mathscr{Q}\right). \end{aligned} \tag{97}$$

By (95) and (97)

$$\begin{aligned} V_k(x_k) = \min_{u_k^*}\big[& u_k^{*T}\left(U + B^T S_{k+1} B\right) u_k^* \\ & + 2u_k^{*T} B^T S_{k+1} A\hat{x}_{k|k}\big] + \operatorname{trace}\left(S_{k+1} Q\right) \\ & + E\left[x_k^T\left(W + A^T S_{k+1} A\right) x_k \middle| \mathscr{F}_k\right] \\ & + E\left[c_{k+1} \middle| \mathscr{F}_k\right] + \operatorname{trace}\left[\left(B^T S B + U\right)\mathscr{Q}\right]. \end{aligned} \tag{98}$$

Hence, the optimal u_k^* is given by

$$u_k^* = -\left(U + B^T S_{k+1} B\right)^{-1} B^T S_{k+1} A\hat{x}_{k|k} \tag{99}$$

and

$$\begin{aligned} V_k(x_k) &= \hat{x}_{k|k}^T A^T S_{k+1} B\left(B^T S_{k+1} B + U\right)^{-1} B^T S_{k+1} A\hat{x}_{k|k} \\ &\quad + E\left[x_k^T\left(W + A^T S_{k+1} A\right) x_k \middle| \mathscr{F}_k\right] + c_{k+1} \\ &\quad + \operatorname{trace}\left(S_{k+1} Q\right) + \operatorname{trace}\left[\left(B^T S B + U\right)\mathscr{Q}\right] \\ &= E\left(x_k^T S_k x_k \middle| \mathscr{F}_k\right) + \operatorname{trace}\left[\left(B^T S B + U\right)\mathscr{Q}\right] \\ &\quad + \operatorname{trace}\left[\left(W + A^T S_{k+1} A\hat{x}_{k|k}\right) P_{k|k}\right] + c_{k+1} \\ &\quad + \operatorname{trace}\left(S_{k+1} Q\right) \\ &= E\left(x_k^T S_k x_k \middle| \mathscr{F}_k\right) + c_k \end{aligned} \tag{100}$$

which completes the induction step and the proof. ■

Proof of Theorem 5 follows.

Proof: Since

$$\begin{aligned} J_n &= E(V_0) \\ &= E\left(x_0^T S_0 x_0\right) + \operatorname{trace}\left[\sum_{k=0}^{N-1}\left(B^T S_{k+1} B + U\right)\mathscr{Q}\right] \\ &\quad + \operatorname{trace}\left[\sum_{k=0}^{N-1}\left(W + A^T S_{k+1} A - S_k\right) P_{k|k}\right] \\ &\quad + \operatorname{trace}\left[\sum_{k=0}^{N-1} S_{k+1} Q\right] \end{aligned} \tag{101}$$

$$\tag{102}$$

$$J' = \frac{J_N}{N} \tag{103}$$
$$= \operatorname{trace}\big[(W + A^T S A - S)(P - KCP)\big] + \operatorname{trace}(SQ) + \operatorname{trace}\big[(B^T S B + U)\mathcal{Q}\big]$$
$$= J + \operatorname{trace}\big[(B^T S B + U)\mathcal{Q}\big]. \tag{104}$$

■

REFERENCES

[1] E. A. Lee, "Cyber physical systems: Design challenges," in *Proc. 11th IEEE ISORC*, May 2008, pp. 363–369.

[2] J. Markoff, "A silent attack, but not a subtle one," *New York Times*, vol. 160, no. 55176, pp. 1–6, Sep. 2010.

[3] D. E. Sanger, "Obama order sped up wave of cyberattacks against Iran," *New York Times*, vol. 161, no. 55789, Jun. 2012.

[4] J. Carlin. (1997, May). *A Farewell to Arms* [Online]. Available: http://www.wired.com/wired/archive/5.05/netizen.html

[5] E. J. Byres and J. Lowe, "The myths and facts behind cyber security risks for industrial control systems," in *Proc. VDE Congr.*, vol. 116. Oct. 2004, pp. 1–6.

[6] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Research challenges for the security of control systems," in *Proc. 3rd Conf. Hot Topics Sec.*, Mar. 2008, pp. 1–6.

[7] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Proc. 28th ICDCS*, Jun. 2008, pp. 495–500.

[8] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Proc. 12th Int. Conf. Hybrid Syst. Comput. Control*, 2009, pp. 31–45.

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

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

[11] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, no. 6, pp. 601–611, Nov. 1975.

[12] R. F. Stengel and L. R. Ray, "Stochastic robustness of linear time-invariant control systems," *IEEE Trans. Autom. Control*, vol. 36, no. 1, pp. 82–87, Jan. 1991.

[13] T. Alpcan and T. Başar, "A game theoretic approach to decision and analysis in network intrusion detection," in *Proc. 42nd IEEE Conf. Decision Control*, vol. 3. Dec. 2003, pp. 2595–2600.

[14] S. Sundaram and C. N. Hadjicostis, "Structural controllability and observability of linear systems over finite fields with applications to multi-agent systems," *IEEE Trans. Autom. Control*, vol. 58, no. 1, pp. 60–73, Jan. 2013.

[15] L. Lazos and R. Poovendran, "SeRLoc: Robust localization for wireless sensor networks," *ACM Trans. Sensor Netw.*, vol. 1, no. 1, pp. 73–100, Aug. 2005.

[16] L. Lazos, R. Poovendran, and S. Čapkun, "ROPE: Robust position estimation in wireless sensor networks," in *Proc. 4th Int. Symp. Inf. Process. Sensor Netw.*, 2005, pp. 1–8.

[17] F. Pasqualetti, A. Bicchi, and F. Bullo, "Distributed intrusion detection for secure consensus computations," in *Proc. 46th IEEE Conf. Decision Control*, Dec. 2007, pp. 5594–5599.

[18] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *Proc. IEEE Conf. Decision Control*, Dec. 2012, pp. 1–8.

[19] M. Zhu and S. Martínez, "Attack-resilient distributed formation control via online adaptation," in *Proc. 50th IEEE CDC-ECC*, Dec. 2011, pp. 6624–6629.

[20] A. Giani, S. S. Sastry, K. H. Johansson, and H. Sandberg, "The VIKING project: An initiative on resilient control of power networks," in *Proc. 2nd ISRCS*, Aug. 2009, pp. 31–35.

[21] J. P. Hespanha, P. Naghshtabrizi, and Y. Xu, "A survey of recent results in networked control systems," *Proc. IEEE*, vol. 95, no. 1, pp. 138–162, Jan. 2007.

[22] G. Dán and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Proc. 1st IEEE Int. Conf. Smart Grid Commun.*, Oct. 2010, pp. 214–219.

[23] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *Proc. 1st Workshop Secure Control Syst., Cyber Phys. Syst.*, Apr. 2010, pp. 1–6.

[24] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2011, pp. 337–344.

[25] H. Fawzi, P. Tabuada, and S. Diggavi, *Secure Estimation and Control for Cyber-Physical Systems Under Adversarial Attacks*. Ithaca, NY, USA: Cornell Univ. Press, May 2012.

[26] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun., Control, Comput.*, Oct. 2009, pp. 911–918.

[27] R. Chabukswar, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," in *Proc. 18th World Congr. IFAC*, Mar. 2011, pp. 1–6.

[28] W. J. Broad, J. Markoff, and D. E. Sanger, "Israeli test on worm called crucial in Iran nuclear delay," *New York Times*, vol. 160, no. 55287, p. 1, Jan. 2011.

[29] 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.

[30] P. E. Greenwood and M. S. Nikulin, *A Guide to Chi-Squared Testing*. New York, NY, USA: Wiley, Apr. 1996.

[31] L. L. Scharf and C. Demeure, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*. Reading, MA, USA: Addison-Wesley, 1991.

[32] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Comput. Chem. Eng.*, vol. 17, no. 3, pp. 245–255, Jan. 1993.

[33] N. L. Ricker, "Model predictive control of a continuous, nonlinear, two-phase reactor," *J. Process Control*, vol. 3, no. 2, pp. 109–123, Sep. 1995.

Yilin Mo (M'13) received the Bachelor of Engineering degree from the Department of Automation, Tsinghua University, Beijing, China, in 2007, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2012.

He is a Post-Doctoral Researcher with the Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, CA, USA. His current research interests include secure control systems and networked control systems with applications in sensor networks.

Rohan Chabukswar (S'10) received the Bachelor of Technology degree in engineering physics from the Indian Institute of Technology Bombay, Mumbai, India, in 2008, and the Master of Science degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, where he is currently pursuing the Ph.D. degree.

His current research interests include cyber-physical systems security and secure control systems with applications to smart grids.

Bruno Sinopoli (M'05) received the Dr.Eng. degree from the University of Padova, Padova, Italy, in 1998, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Berkeley, Berkeley, CA, USA, in 2003 and 2005, respectively.

He joined the faculty at Carnegie Mellon University, Pittsburgh, PA, USA, where he is an Associate Professor with the Department of Electrical and Computer Engineering with courtesy appointments in mechanical engineering and in the Robotics Institute. His current research interests include networked embedded control systems, distributed estimation and control with applications to wireless sensor-actuator networks and cyber-physical systems security.

Dr. Sinopoli received 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 National Science Foundation Career Award in 2010.