Forty-Seventh Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 30 - October 2, 2009

Secure Control Against Replay Attacks

Yilin Mo, Bruno Sinopoli [*†]

Abstract

This paper analyzes the effect of replay attacks on a control system. We assume an attacker wishes to disrupt the operation of a control system in steady state. In order to inject an exogenous control input without being detected the attacker will hijack the sensors, observe and record their readings for a certain amount of time and repeat them afterwards while carrying out his attack. This is a very common and natural attack (we have seen numerous times intruders recording and replaying security videos while performing their attack undisturbed) for an attacker who does not know the dynamics of the system but is aware of the fact that the system itself is expected to be in steady state for the duration of the attack. We assume the control system to be a discrete time linear time invariant gaussian system applying an infinite horizon Linear Quadratic Gaussian (LQG) controller. We also assume that the system is equipped with a χ^2 failure detector. The main contributions of the paper, beyond the novelty of the problem formulation, consist in 1) providing conditions on the feasibility of the replay attack on the aforementioned system and 2) proposing a countermeasure that guarantees a desired probability of detection (with a fixed false alarm rate) by trading off either detection delay or LQG performance, either by decreasing control accuracy or increasing control effort.

1. Introduction

Cyber Physical Systems (CPS) refer to the embedding of widespread sensing, computation, communication and control into physical spaces [1]. Application areas are as diverse as aerospace, chemical processes, civil infrastructure, energy, manufacturing and transportation. Many of these applications are safety-critical. The availability of cheap communication technologies as the internet makes such infrastructures susceptible to cyber security threats. National security may be affected as infrastructures such as the power grid, the telecommunication networks are vital to the normal operation of our society. Any successful attack may significantly hamper the economy, the environment or may even lead to loss of human life. As a result, the role security of CPS is of primary importance to guarantee safe operation of CPS. The research community has acknowledged the importance of addressing the challenge of designing secure CPS [2] [3].

The impact of attacks on the cyber physical systems is addressed in [4]. The authors consider two possible classes of attacks on CPS: Denial of Service (DoS) and deception attacks. The DoS attack prevents the exchange of information, usually either sensor readings or control inputs between subsystems, while the deception attack affects the data integrity of packets by modifying their payloads. A robust feedback control design against DoS attack is further discussed in [5]. We feel that the deception attack can be subtler than DoS attack as it is in principle more difficult to detect and it has not adequately addressed. Hence, in this paper, we will develop a methodology to detect a particular kind of deception attack.

A significant amount of research effort has been carried out to analyze, detect and handle failures in CPS. Sinopoli et al. study the impact of random packet drops on controller and estimator performance [6] [7]. In [8], the author reviews several failure detection algorithm in dynamic systems. Results from robust control [9], a discipline that aims to design controllers that function properly under uncertain parameter or unknown disturbances, is applicable to some CPS scenarios. However, a large proportion of the literature assumes that the failure is either random or benign. On the other hand, a cunning attacker can carefully design his attack strategy and deceive both detectors and robust controllers. Hence, the applicability of failure detection algorithms is questionable in the presence of a smart attacker.

[*] The authors are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA. Email: ymo@andrew.cmu.edu, brunos@ece.cmu.edu

[†] This research is supported in part by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office. Foundation. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of ARO, CMU, or the U.S. Government or any of its agencies.

978-1-4244-5871-4/09/$26.00 ©2009 IEEE

In this paper, we study the effect of a data replay attack on control systems. We assume an attacker wishes to disrupt the operation of a control system in steady state. In order to inject an exogenous control input without being detected the attacker will hijack the sensors, observe and record their readings for a certain amount of time and repeat them afterwards while carrying out his attack. This is a very common and natural attack (we have seen numerous times intruders recording and replaying security videos while performing their attack undisturbed) for an attacker who does not know the dynamics of the system but is aware that the system itself is expected to be in steady state for the duration of the attack. We assume the control system to be a discrete time linear time invariant (LTI) Gaussian system applying an infinite horizon Linear Quadratic Gaussian (LQG) controller. We also assume that the system is equipped with a χ^2 failure detector. The main contributions of the paper, beyond the novelty of the problem formulation, consist in providing conditions on the feasibility of the replay attack on the aforementioned attack and suggesting a countermeasure that guarantees a desired probability of detection (with a fixed false alarm rate) by trading off either detection delay or LQG cost, i.e. either by decreasing control accuracy or increasing control effort.

The rest of the paper is organized as follows: In Section 2, we provide the problem formulation by revisiting and adapting Kalman filter, LQG controller and χ^2 failure detector to our scenario. In Section 3, we define the threat model of replay attack and analyze its effect on the control schemes discussed in Section 2. In Section 4 we discuss one possible countermeasure, the efficiency of which is illustrated by several numerical examples in Section 5. Finally Section 6 concludes the paper. The appendix contains several proofs, some of which had to be removed due to space constraints.

2. Problem Formulation

In this section we will formulate the problem by deriving the Kalman filter, the LQG controller and χ^2 detector for our case. We will use the notation below for the remainder of the paper.

Consider the following linear, time invariant (LTI) system whose state dynamics are given by

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

where $x_k \in \mathbb{R}^n$ is the vector of state variables at time k, $w_k \in \mathbb{R}^n$ is the process noise at time k and x_0 is the initial state. We assume 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 is monitoring the system described in (1). At each step all the sensor readings are sent to a base station. The observation equation can be written as

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

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

2.1. Kalman Filter

It is well known that for the system of equations (1), (2) the Kalman filter is the optimal estimator as it provides the minimum variance unbiased estimate of the state x_k given the previous observations $y_0, \ldots, y_k$. The Kalman filter is recursive and it takes the following form:

$$\begin{aligned}
&\hat{x}_{0|-1} = \bar{x}_0,\, P_{0|-1} = \Sigma, \\
&\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k,\, P_{k+1|k} = AP_{k|k}A^T + Q, \\
&\quad K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}, \\
&\quad \hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}),\, P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1}.
\end{aligned} \tag{3}$$

Although the Kalman filter uses a time varying gain K_k, it is known that this gain will converge if the system is detectable. In practice the Kalman gain usually converges in a few steps. Hence, let us define

$$P \triangleq \lim_{k\to\infty} P_{k|k-1},\, K \triangleq PC^T(CPC^T + R)^{-1}. \tag{4}$$

Since control systems usually run for a long time, we can assume to be running at steady state from the beginning. Hence, we assume initial condition $\Sigma = P$. In that case, the Kalman filter is a fixed gain estimator, taking the following form

$$\hat{x}_{0|-1} = \bar{x}_0,\, \hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k,\, \hat{x}_{k|k} = \hat{x}_{k|k-1} + K(y_k - C\hat{x}_{k|k-1}).$$

2.2. Linear Quadratic Gaussian (LQG) Optimal Control

Given the state estimation $\hat{x}_{k|k}$, the LQG controller minimizes the following objective function[1]:

$$J = \min \lim_{T\to\infty} E\frac{1}{T}\left[\sum_{k=0}^{T-1}(x_k^T W x_k + u_k^T U u_k)\right], \tag{5}$$

where W, U are positive semidefinite matrices and u_k is measurable with respect to $y_0, \ldots, y_k$, i.e. u_k is a function of previous observations. It is well known that the

[1] Here we just discuss the case of infinite horizon LQG control problem.

solution of the above minimization problem will lead to a fixed gain controller, which takes the following form:

$$u_k = u_k^* = -(B^T SB + U)^{-1} B^T SA\hat{x}_{k|k}, \tag{6}$$

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

$$S = A^T SA + W - A^T SB(B^T SB + U)^{-1} B^T SA. \tag{7}$$

Let us define $L \triangleq -(B^T SB + U)^{-1} B^T SA$, then $u_k^* = Lx_{k|k}$.

The objective function given by the optimal estimator and controller is in our case is

$$J = trace(SQ) + trace[(A^T SA + W - S)(P - KCP)]. \tag{8}$$

2.3. χ^2 Failure Detector

The χ^2 detector [10] is widely used to detect anomalies in control systems. Before introducing the detector, we will characterize the probability distribution of the residue of the Kalman filter:

Theorem 1. *For the LTI system defined in* (1) *with Kalman filter and LQG controller, the residues $y_i - C\hat{x}_{i|i-1}$ of Kalman filter are i.i.d. Gaussian distributed with* 0 *mean and covariance $\mathscr{P}$, where $\mathscr{P} = CPC^T + R$.*

Proof. Due to space constraints, we cannot give the proof here. Please refer to [10] for the details. □

By Theorem 1, we know that the probability to get the sequence $y_{k-\mathscr{T}+1}, \dots, y_k$ when the system is operating normally is

$$P(y_{k-\mathscr{T}+1}, \dots, y_k) = \left[\frac{1}{(2\pi)^{N/2}|\mathscr{P}|}\right]^{\mathscr{T}} exp(-\frac{1}{2}g_k), \tag{9}$$

where

$$g_k = \sum_{i=k-\mathscr{T}+1}^{k} (y_i - C\hat{x}_{i|i-1})^T \mathscr{P}^{-1} (y_i - C\hat{x}_{i|i-1}). \tag{10}$$

When this probability is low, it means that the system is likely to be subject to certain failure. In order to check the probability, we only need to compute g_k. Hence, the χ^2 detector at time k takes the following form

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

where $\mathscr{T}$ is the window size of detection. By Theorem 1, the left of the equation is χ^2 distributed with $m\mathscr{T}$ degrees of freedom[2]. Hence, it is easy to calculate the false alarm rate from χ^2 distribution. If g_k is greater than the threshold, then the detector will trigger an alarm.

3. Replay Attack against Control System

In this section, we assume that a malicious third party wants to break the control system described in Section 2. We will define an attack model similar to the replay attack in computer security and analyze the feasibility of such kind of attack on the control system. We will later generalize our analysis to other classes of control systems.

We suppose the attacker has the capability to perform the following actions:

1. It can inject a control input u_k^a into the system anytime.

2. It knows all sensor readings and can modify them. We will denote the reading modified by the attacker by y_k'.

Given these abilities, the attacker will implement the following 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 we think is much harder to defend, is to induce false sensor readings by changing the local conditions around it. Such attack may be easy to carry out when sensors are spatially distributed in remote locations.*

Remark 2. *We assume that the attacker has control over all the sensors. This could be accomplished for a smaller system consisting of few sensors. For a large system, usually the whole system can be break down to several small and weakly coupled subsystems. For example consider the temperature control problem in a building. One can think of the temperature in each room as subsystems, which will hardly affects each other. Hence, the attacker only needs to control the sensors of a small subsystem in order to perform the replay attack on the subsystem.*

[2] The degrees of freedom is from the definition of χ^2 distribution. Please refer to [11] for more details.

Remark 3. *The attack strategy is fairly simple. In principle, if the attacker has more knowledge of the system model, the controller design, it can perform a much more subtle and powerful attack. However, to identify the underlying model of the system is usually a hard problem and not all the attackers have the knowledge and power to do so. Hence, we will only focus on a simple, easy to implement attack strategy which is easy to implement.*

Remark 4. *When the system is under attack, the central computer will be unable to perform close loop control on the system since the sensory information is not available. Hence, we cannot guarantee any control performance of the system under this attack. Any counter-attack will need to be able to detect the attack.*

It is worth noticing that in the attacking stage, the goal of the attacker is to make the fake readings y'_ks look 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, to generate a fake sequence of readings. In order to provide a unified framework to analyze such kind of attack, we can think of y'_ks as the output of the following virtual system (this does not necessarily mean that the attacker runs a virtual system):

$$\begin{aligned} x'_{k+1} &= Ax'_k + Bu'_k + w'_k,\ y'_k = Cx'_k + v'_k, \\ \hat{x}'_{k+1|k} &= A\hat{x}'_{k|k} + Bu'_k,\ \hat{x}'_{k+1|k+1} = \hat{x}'_{k+1|k} + K(y'_k - \hat{x}'_{k+1|k}), \\ u'_k &= L\hat{x}'_{k|k}, \end{aligned}$$

with initial conditions x'_0 and $\hat{x}'_{0|-1}$. If the attacker actually learns the system, then the virtual system will be the system the attacker runs. For the replay attack, suppose that the attacker records the sequence y_ks from time t time. Then the virtual system is just a time shifted version of the real system, with $x'_k = x_{t+k}$, $\hat{x}'_{k|k} = \hat{x}_{t+k|t+k}$ (Note that the attacker may not know x_{t+k} and $\hat{x}_{t+k|t+k}$).

Suppose the system is under attack and the defender is using the χ^2 detector to perform intrusion detection. We will rewrite the estimation of the Kalman filter $\hat{x}_{k|k-1}$ in the following recursive way:

$$\begin{aligned} \hat{x}_{k+1|k} &= A\hat{x}_{k|k} + Bu_k = (A+BL)\hat{x}_{k|k} \\ &= (A+BL)[\hat{x}_{k|k-1} + K(y'_k - C\hat{x}_{k|k-1})] \\ &= (A+BL)(I-KC)\hat{x}_{k|k-1} + (A+BL)Ky'_k. \end{aligned} \tag{12}$$

For the virtual system, it is easy to see that the same equation holds true for $\hat{x}'_{k|k-1}$:

$$\hat{x}'_{k+1|k} = (A+BL)(I-KC)\hat{x}'_{k|k-1} + (A+BL)Ky'_k. \tag{13}$$

Define $\mathscr{A} \triangleq (A+BL)(I-KC)$, then[3]

$$\hat{x}_{k|k-1} - \hat{x}'_{k|k-1} = \mathscr{A}^k(\hat{x}_{0|-1} - \hat{x}'_{0|-1}). \tag{14}$$

Define $\hat{x}_{0|-1} - \hat{x}'_{0|-1} \triangleq \zeta$. Now write the residue as

$$y'_k - C\hat{x}_{k|k-1} = (y'_k - C\hat{x}'_{k|k-1}) + C\mathscr{A}^k\zeta, \tag{15}$$

and

$$\begin{aligned} g_k = \sum_{i=k-\mathscr{T}+1}^{k} \Big[&(y'_i - C\hat{x}'_{k|k-1})^T \mathscr{P}^{-1} (y'_i - C\hat{x}'_{k|k-1}) \\ &+ 2(y'_i - C\hat{x}'_{k|k-1})^T \mathscr{P}^{-1} C\mathscr{A}^i\zeta + \zeta^T(\mathscr{A}^i)^T C^T \mathscr{P}^{-1} C\mathscr{A}^i\zeta \Big]. \end{aligned} \tag{16}$$

By the definition of the virtual system, we know that $y'_k - C\hat{x}'_{k|k-1}$ follows exactly the same distribution as $y_k - C\hat{x}_{k|k-1}$. Hence, if $\mathscr{A}$ is stable, the second term and the third term in (16) will converge to 0. As a result, $y'_k - C\hat{x}_{k|k-1}$ will converges to the same distribution as $y_k - C\hat{x}_{k|k-1}$, and the detection rate given by χ^2 detector will be the same as false alarm rate. In other words, the detector is useless.

On the other hand, if $\mathscr{A}$ is unstable, the attacker cannot replay y'_k for long since g_k will soon become unbounded. In this case, the system is resilient to the replay attack, as the detector will be able to detect the attack. It turns out the feasibility result derived for a special estimator, controller, and detector implementations is actually applicable to virtually any system. In fact we can generalize the technique used here to analyze more general controller, estimator and detectors. Suppose the state of the estimator at time k is s_k and it evolves according to

$$s_{k+1} = f(s_k, y_k). \tag{17}$$

Define the norm of f to be

$$\|f\| \triangleq \sup_{\Delta s \neq 0, y, s} \frac{\|f(s,y) - f(s+\Delta s, y)\|}{\|\Delta s\|}. \tag{18}$$

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

$$g(s_k, y_k) \lesseqgtr threshold, \tag{19}$$

where g is an arbitrary continuous function.

Theorem 2. *If* $\|f\| \leq 1$, *then*

$$\lim_{k\to\infty} g(s_k, y'_k) = g(s'_k, y'_k), \tag{20}$$

[3] For simplicity, here we consider the time the attack begins as time 0.

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

$$\lim_{k\to\infty} \beta_k = \alpha_k, \tag{21}$$

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

Proof. Due to space limit, we will just give an outline of the proof. First, $\|f\| \leq 1$ 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

$$\begin{aligned} \beta_k &= Prob(g(s_k, y'_k) > threshold), \\ \alpha_k &= Prob(g(s'_k, y'_k) > threshold). \end{aligned} \tag{22}$$

Hence β_k converges to α_k. □

The LQG controller, Kalman filter and χ^2 detector becomes just a special case, where the state s_k of the estimator at time k is $y_{k-\mathscr{T}+1}, \ldots, y_k$ and $\hat{x}_{k-\mathscr{T}+1|k-\mathscr{T}}, \ldots, \hat{x}_{k|k-1}$. The f function is given by (3) and g is given by (11).

Remark 5. *The convergence of detection rate under the replay attack to the false alarm rate indicates that the information given by the detector will asymptotically go to* 0*. In the other word, the detector becomes useless and the system is not resilient to replay attack.*

4. Detection of Replay Attack

As discussed in the previous section, there exist control systems that are not resilient to the replay attack. In this section, we want to design a detection strategy against replay attacks. Throughout this section we will always assume that $\mathscr{A}$ is stable.

The main problem of LQG controller and Kalman filter is that they use a fixed gain, or a gain that converges really fast. Hence, the whole control system is static in some sense. In order to detect replay attack, we redesign the controller as

$$u_k = u_k^* + \Delta u_k, \tag{23}$$

where u_k^* is the optimal LQG control signal and Δu_ks are drawn from an i.i.d. Gaussian distribution with zero mean and covariance $\mathscr{Q}$, and Δu_ks are chosen to be also independent of u_k^*. Figure 1 shows the diagram of the whole system.

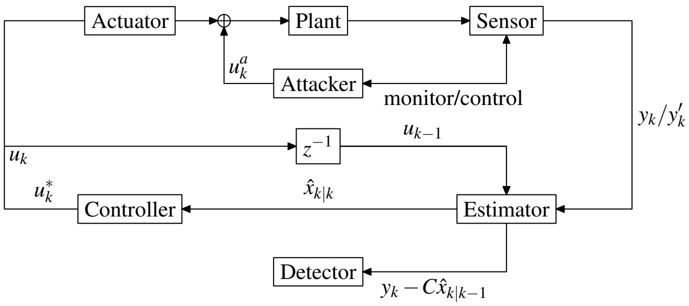

Figure 1. System Diagram

We add Δu_k as an authentication signal. We choose it to be zero mean because we do not wish to introduce any bias to x_k. It is clear that without the attack, the controller is not optimal in the LQG sense anymore, which means that in order to detect the attack, we need to sacrifice control performance. The following theorem characterizes the loss of LQG performance when we inject Δu_k into the system:

Theorem 3. *The LQG performance after adding Δu_k is given by*

$$J' = J + trace[(U + B^T SB)\mathscr{Q}]. \tag{24}$$

Proof. See the appendix. □

We now wish to consider the χ^2 detector after adding the random control signal. The following theorem shows the effectiveness of the detector under the modified control scheme.

Theorem 4. *In the absence of an attack,*

$$E[(y_k - C\hat{x}_{k|k-1})^T \mathscr{P}^{-1}(y_k - C\hat{x}_{k|k-1})] = m. \tag{25}$$

Under attack

$$\begin{aligned} &\lim_{k\to\infty} E[(y'_k - C\hat{x}_{k|k-1})^T \mathscr{P}^{-1}(y'_k - C\hat{x}_{k|k-1})] \\ &= m + 2trace(C^T \mathscr{P}^{-1} C\mathscr{U}), \end{aligned} \tag{26}$$

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

$$\mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T. \tag{27}$$

Proof. The first equation is trivial to prove using Theorem 1. Rewrite $\hat{x}_{k+1|k}$ as

$$\hat{x}_{k+1|k} = \mathscr{A}\hat{x}_{k|k-1} + (A + BL)Ky'_k + B\Delta u_k. \tag{28}$$

For the virtual system

$$\hat{x}'_{k+1|k} = \mathscr{A}\hat{x}'_{k|k-1} + (A + BL)Ky'_k + B\Delta u'_k. \tag{29}$$

Hence,

$$\hat{x}_{k|k-1} - \hat{x}'_{k|k-1} = \mathscr{A}^k(\hat{x}_{0|-1} - \hat{x}'_{0|-1}) + \sum_{i=0}^{k-1} \mathscr{A}^{k-i-1} B(\Delta u_i - \Delta u'_i). \tag{30}$$

As a result,

$$\begin{aligned} y'_k - C\hat{x}_{k|k-1} = y'_k - C\hat{x}'_{k|k-1} + C\mathscr{A}^k(\hat{x}_{0|-1} - \hat{x}'_{0|-1}) \\ + C\sum_{i=0}^{k-1}\mathscr{A}^{k-i-1}B(\Delta u_i - \Delta u'_i). \end{aligned} \tag{31}$$

The first term has exactly the same distribution as $y_k - C\hat{x}_{k|k-1}$. The second term will converge to 0 when $\mathscr{A}$ is stable. Also Δu_i is independent of the virtual system and for the virtual system, $y'_k - C\hat{x}'_{k|k-1}$ is independent of $\Delta u'_i$. Hence

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

By the definition of $\mathscr{U}$, it is easy to see that

$$\mathscr{U} = \sum_{i=0}^{\infty}\mathscr{A}^i B\mathscr{Q}B^T(\mathscr{A}^i)^T.$$

Hence, $\lim_{k\to\infty} Cov(y'_k - C\hat{x}_{k|k}) = \mathscr{P} + 2C\mathscr{U}C^T$ and

$$\begin{aligned} &\lim_{k\to\infty} E[(y'_k - C\hat{x}_{k|k-1})^T\mathscr{P}^{-1}(y'_k - C\hat{x}_{k|k-1})] \\ &= trace\left[\lim_{k\to\infty} Cov(y'_k - C\hat{x}_{k|k}) \times \mathscr{P}^{-1}\right] \\ &= m + 2trace(C^T\mathscr{P}^{-1}C\mathscr{U}). \end{aligned} \tag{32}$$

□

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

$$E(g_k) = m\mathscr{T}. \tag{33}$$

Under attack, the asymptotic expectation becomes

$$\lim_{k\to\infty} E(g_k) = m\mathscr{T} + 2trace(C^T\mathscr{P}^{-1}C\mathscr{U})\mathscr{T}. \tag{34}$$

The difference in the expectation of g_k illustrates that the detection rate will not converges to the false alarm rate, which will also be shown in the next section. Another thing worth noticing is that to design $\mathscr{Q}$, one possible criterion is to minimize $J' - J = trace[(U + B^T SB)\mathscr{Q}]$ while maximizing $trace(C^T\mathscr{P}^{-1}C\mathscr{U})$.

5. Simulation Result

In this section we provide some simulation results on the detection of replay attack. Consider the control system described in Section 2 is controlling the temperature inside one room. Let T_k be the temperature of the room at time k and T^* to be the desired temperature. Define the state as $x_k = T_k - T^*$. Suppose that

$$x_{k+1} = x_k + u_k + w_k, \tag{35}$$

where u_k is the input from air conditioning unit and w_k is the process noise. Suppose that just one sensor is measuring the temperature, which is

$$y_k = x_k + v_k, \tag{36}$$

where v_k is the measurement noise. We choose $R = 0.1, Q = W = U = 1$. One can compute that $P = 1.092, K = 0.9161, L = -0.6180$. Hence $\mathscr{A} = 0.0321$ and the system is vulnerable to replay attack. The LQG cost is $J = 1.7076, J' = J + 2.618\mathscr{Q}$.

We will first fix the window size $\mathscr{T} = 5$ and show the detection rate for different $\mathscr{Q}$s. We assume that the attacker records the y_ks from time 1 to time 10 and then replays it from time 11 to time 20. We also fixed the false alarm rate to be 5% at each step.

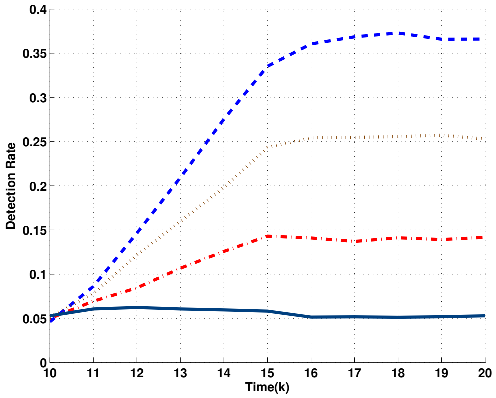

Figure 2. Detection rate at each time step for $\mathscr{Q} = 0.6$ (blue dashed line), $\mathscr{Q} = 0.4$ (brown dot line), $\mathscr{Q} = 0.2$ (red dash-dot line) and $\mathscr{Q} = 0$ (black solid line).

Figure 2 shows the detection rate at each time step for different $\mathscr{Q}$s. Each detection rate is the average of 10,000 experiments. Note that the attack starts at time 11. Hence, each line starts at the false alarm rate 5% at time 10. One can see that without additional input signal, the detection rate will soon converge to 5%, which proves that the detector is inefficient for replay

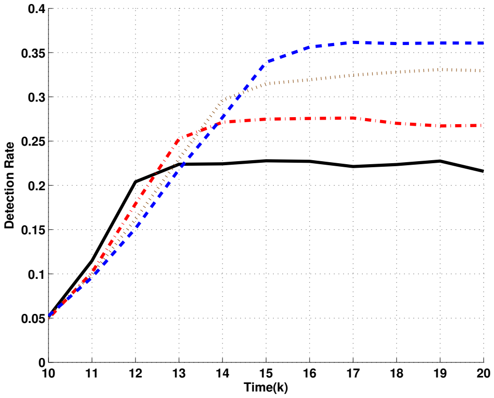

Figure 3. Detection rate at each time step for $\mathscr{T} = 5$ (blue dashed line), $\mathscr{T} = 4$ (brown dot line), $\mathscr{T} = 3$ (red dash-dot line) and $\mathscr{T} = 2$ (black solid line).

attack. With $\mathscr{Q} = 0.6$, the loss of LQG performance is $2.618 \times 0.6/1.7076 = 91\%$ with respect to the optimal LQG cost. As a result of the high control performance lost, one can get more than 35% detection rate at each step.

Next we would like to fix $\mathscr{Q} = 0.6$ and compare the detection rate of different window size $\mathscr{T}$. We still assume the attack starts at time 11 and the false alarm rate is 5%. Fig 3 shows the detection rate for different window size. It is worth noticing that choosing a small window size will make the detector response faster to replay attack. However, the asymptotic detection rate will be lower than that of larger window size. On the other hand, by the law of large numbers, the asymptotic detection rate will converges to 1 as $\mathscr{T}$ increases. However the detector will respond very slowly to the replay attack. For more details on the choice of window size, please refer to [8].

6. Conclusions

In this paper we defined a replay attack model on cyber physical system and analyzed the performance of the control system under the attack. We discovered that for some control systems, the classical estimation, control, failure detection strategy are not resilience to the replay attack. For such kind of system, we provide a technique that can improve detection rate in the expense of control performance.

7. Appendix: Proof of Theorem 3

To simplify notation, let us first define the sigma algebra generated by $y_k, \dots, y_0, \Delta u_{k-1}, \dots, \Delta u_0$ to be $\mathscr{F}_k$. Due to space limit, we will just list the outlines of the proof. Before proving Theorem 3, we need the following lemmas:

Lemma 1. *The following equations about Kalman filter are true:*

$$\hat{x}_{k|k} = E(x_k|\mathscr{F}_k), P_{k|k} = E(e_{k|k}e_{k|k}^T|\mathscr{F}_k),$$

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

Lemma 2. *The following equations are true*

$$E(x_k^T \mathscr{S} x_k|\mathscr{F}_k) = trace(\mathscr{S} P_{k|k}) + (\hat{x}_{k|k})^T \mathscr{S} \hat{x}_{k|k}, \quad (37)$$

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

Now define

$$J_N \triangleq \min E\left[\sum_{i=0}^{N-1}(x_i^T W x_i + u_i^T U u_i)\right]. \quad (38)$$

By the definition of J', we know that $J' = \lim_{N\to\infty} J_N/N$.

Now fix N, let us define

$$V_k(x_k) \triangleq \min E\left[\sum_{k=i}^{N-1}(x_i^T W x_i + u_i^T U u_i)|\mathscr{F}_k\right], \quad (39)$$

and $V_N(x_N) = 0$. By definition, we know that $E(V_0) = J_N$. Also from dynamic programming, we know that V_k satisfies the following backward recursive equation:

$$V_k(x_k) = \min_{u_k^*} E\left[x_k^T W x_k + u_k^T U u_k + V_{k+1}(x_{k+1})|\mathscr{F}_k\right]. \quad (40)$$

Let us define

$$S_{k-1} \triangleq A^T S_k A + W - A^T S_k B(B^T S_k B + U)^{-1} B^T S_k A,$$
$$c_{k-1} \triangleq c_k + trace[(W + A^T S_k A - S_{k-1})P_{k-1|k-1}] + trace(S_k Q) + trace[(B^T S_k B + U)\mathscr{Q}],$$

with $S_N = 0$, $c_N = 0$.

Lemma 3. *$V_k(x_k)$ is given by*

$$V_k(x_k) = E[x_k^T S_k x_k|\mathscr{F}_k] + c_k, k = N, \dots, 0. \quad (41)$$

Proof. We will use backward induction to prove (41). First it is trivial to see that $V_N = 0$ satisfies (41). Now suppose that V_{k+1} satisfies (41), then by (40)

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

First we know that u_k^* is measurable to $\mathscr{F}_k$ and Δu_k is independent of $\mathscr{F}_k$, hence

$$E[(u_k^*+\Delta u_k)^T U(u_k^*+\Delta u_k)|\mathscr{F}_k] = (u_k^*)^T U u_k^* + trace(U\mathscr{Q}). \tag{42}$$

Then let us write x_{k+1} as

$$x_{k+1} = Ax_k + Bu_k^* + B\Delta u_k + w_k.$$

By the fact that Δu_k, w_k are independent of $Ax_k + Bu_k^*$, one can finally get

$$\begin{aligned} &E(x_{k+1}^T S_{k+1} x_{k+1}|\mathscr{F}_k) = E(x_k^T A^T S_{k+1} A x_k|\mathscr{F}_k) + \\ &2(u_k^*)^T B^T S_{k+1} A\hat{x}_{k|k} + (u_k^*)^T B^T S_{k+1} B(u_k^*) \\ &+ trace(S_{k+1}Q) + trace(B^T S_{k+1} B\mathscr{Q}). \end{aligned} \tag{43}$$

By (42) and (43), we know that

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

Hence, the optimal u_k^* is

$$u_k^* = -(U + B^T S_{k+1} B)^{-1} B^T S_{k+1} A\hat{x}_{k|k}, \tag{44}$$

and $V_k(x_k)$ is

$$\begin{aligned} V_k(x_k) &= (\hat{x}_{k|k})^T A^T S_{k+1} B(B^T S_{k+1} B + U)^{-1} B^T S_{k+1} A\hat{x}_{k|k} \\ &+ E[x_k^T (W + A^T S_{k+1} A) x_k|\mathscr{F}_k] + trace(S_{k+1}Q) \\ &+ c_{k+1} + trace[(B^T S_{k+1} B + U)\mathscr{Q}] \\ &= E(x_k^T S_k x_k|\mathscr{F}_k) + trace[(W + A^T S_{k+1} A - S_k) P_{k|k}] \\ &+ c_{k+1} + trace(S_{k+1}Q) + trace[(B^T S_{k+1} B + U)\mathscr{Q}] \\ &= E(x_k^T S_k x_k|\mathscr{F}_k) + c_k, \end{aligned} \tag{45}$$

which completes the proof[4]. □

Now we are ready to prove Theorem 3.

Proof of Theorem 3. Since

$$\begin{aligned} J_N = EV_0 = E(x_0^T S_0 x_0) + trace[\sum_{k=0}^{N-1} (W + A^T S_{k+1} A - S_k) P_{k|k}] \\ + trace(\sum_{k=0}^{N-1} S_{k+1} Q) + trace[\sum_{k=0}^{N-1} (B^T S_{k+1} B + U)\mathscr{Q}], \end{aligned}$$

we know that

$$\begin{aligned} J' = \lim_{N\to\infty} J_N/N = trace[(W + A^T SA - S)(P - KCP)] + trace(SQ) \\ + trace[(B^T SB + U)\mathscr{Q}] = J + trace[(B^T SB + U)\mathscr{Q}]. \end{aligned} \tag{46}$$

□

[4] We use Lemma 2 in the second equality.

References

[1] E. A. Lee, “Cyber physical systems: Design challenges,” EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-8, Jan 2008. [Online]. Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-8.html

[2] E. Byres and J. Lowe, “The myths and facts behind cyber security risks for industrial control systems.” VDE Congress, 2004.

[3] A. A. Cárdenas, S. Amin, and S. Sastry, “Research challenges for the security of control systems,” in *HOTSEC’08: Proceedings of the 3rd conference on Hot topics in security*. Berkeley, CA, USA: USENIX Association, 2008, pp. 1–6.

[4] ——, “Secure control: Towards survivable cyber-physical systems,” in *Distributed Computing Systems Workshops, 2008. ICDCS ’08. 28th International Conference on*, June 2008, pp. 495–500.

[5] S. Amin, A. Cardenas, and S. S. Sastry, “Safe and secure networked control systems under denial-of-service attacks.” in *Hybrid Systems: Computation and Control*. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, April 2009, pp. 31–45. [Online]. Available: http://chess.eecs.berkeley.edu/pubs/597.html

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

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

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

[9] R. Stengel and L. Ryan, “Stochastic robustness of linear time-invariant control systems,” *Automatic Control, IEEE Transactions on*, vol. 36, no. 1, pp. 82–87, Jan 1991.

[10] R. Mehra and J. Peschon, “An innovations approach to fault detection and diagnosis in dynamic systems,” *Automatica*, vol. 7, pp. 637–660, 1971.

[11] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. New York: Addison-Wesley Publishing Co., 1990.