Integrity Attacks on Cyber-Physical Systems[*]

Yilin Mo
Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213
ymo@andrew.cmu.edu

Bruno Sinopoli
Department of Electrical and Computer Engineering
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, 15213
bruons@ece.cmu.edu

ABSTRACT

In this paper we consider the integrity attack on Cyber-Physical System(CPS), which is modeled as a discrete linear time-invariant system equipped with a Kalman filter, LQG controller and χ^2 failure detector. An attacker wishes to disturb the system by injecting external control inputs and fake sensor measurements. In order to perform the attack without being detected, the adversary will need to carefully design its actions to fool the failure detector as abnormal sensor measurements will result in an alarm. The adversary's strategy is formulated as a constrained control problem. In this paper, we characterize the reachable set of the system state and estimation error under the attack, which provides a quantitative measure of the resilience of the system. To this end, we will provide an ellipsoidal algorithm to compute the outer approximation of the reachable set. We also prove a necessary condition under which the reachable set is unbounded, indicating that the attacker can successfully destabilize the system.

Categories and Subject Descriptors

B.2.3 [**Reliability, Testing, and Fault-Tolerance**]: Diagnostics

General Terms

Security

Keywords

Cyber-Physical Systems, Reachability analysis, Security

1. INTRODUCTION

The concept of Cyber-Physical System (CPS) refers to the embedding of sensing, communication, control and computation into the physical spaces. Today, CPSs can be found in areas as diverse as aerospace, automotive, chemical process control, civil infrastructure, energy, health-care, manufacturing and transportation, where secure operation is usually one of the main concerns. Any successful attack on these safety-critical CPS may significantly hamper the economy (for example, power outrage) or even lead to the loss of human lives (for example, the malfunctioning of automotive). The current CPSs are usually running in isolated networks, which protects the system by limiting the access points of the attacker. However, the next generation of "smarter" CPSs, such as Smart Grids, Vehicular ad-hoc network (VANET) and sensor networks, will make extensive use of widespread networking, which create lots of entry points for the attacker. Stuxnet, the first-ever malware that targets and subverts CPS systems, was first discovered on June 2010, which raises great concerns on CPS security.

The impact of attacks on CPS is discussed in [2]. The authors consider two possible classes of attacks on CPS: Denial of Service (DoS) and deception (or integrity) attacks. The DoS attack prevents the exchange of information, usually either sensor readings or control inputs between subsystems, while a integrity attack affects the data integrity of packets by modifying their payloads. A robust feedback control design against DoS attacks has been discussed in [1]. We feel that integrity attacks can be a subtler attack than DoS as they are in principle more difficult to detect. In this paper we want to analyze the impact of integrity attacks on CPS.

A significant amount of research has been carried out on detect, analyze and handle integrity attacks in CPS. In [5], the authors consider replay attack, which is a special kind of integrity attacks. The authors provide algebraic conditions for the feasibility of replay attack and proposed a detection technique to counter it. In [4], the authors discuss general integrity attacks in wireless sensor networks, where they propose an ellipsoidal approximation method to compute all possible biases the attacker could introduce to the state estimator. However, they are only concerned with state estimation. Therefore, it is not clear what is the impact of integrity attacks on the control performance.

For distributed control systems, Pasqualetti et al. [6] and Sundaram et al. [8] show how to detect and identify malicious behavior in consensus algorithm and wireless control networks respectively, based on the theory of structured linear systems. However, their models are noiseless, which

[*]This research is supported in part by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office Foundation and grant NGIT2009100109 from Northrop Grumman Information Technology Inc Cybersecurity Consortium. 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.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HiCoNS'12, April 17–18, 2012, Beijing, China.
Copyright 2012 ACM 978-1-4503-1263-9/12/04 ...$10.00.

greatly favor the intrusion detection system, since the evolution of the system is deterministic and any deviation from the predetermined trajectory will be detected. In a noisy environment, it is much harder to detect the malicious behavior of the attacker since it may be indistinguishable from the noise.

The effect of integrity attacks on power systems has also been extensively studied. Liu et al. [3] illustrate how an adversary can inject a stealthy input into the measurements to change the state estimation, without being detected by the bad data detector. In [7], the authors consider how to find a sparse stealthy input, which enables the adversary to launch an attack with minimum number of compromised sensors. Xie et al. [9] discuss how an adversary could use such kind of stealthy attacks to gain financial benefit in the electricity market. The main drawbacks of the above approaches is they only consider static systems and estimators. As a result, the applicability of such results to dynamic systems is questionable.

In this paper we model the CPS as a discrete linear time-invariant system equipped with a Kalman filter, LQG controller and χ^2 failure detector. We assume that an attacker wishes to disturb the system by injecting external control inputs and fake sensor measurements. In order to perform the attack without being detected, the adversary also need to carefully design its actions to fool the failure detector. We formulate the adversary's strategy as a constrained control problem and characterize the reachable set of the system state and estimation error under stealthy constraint, which provides a quantitative measure of the resilience of the system. We provide an ellipsoidal algorithm to compute the outer approximation of the reachable set. We also prove a necessary condition under which the reachable set is unbounded, indicating that the attacker can successfully destabilize the system.

The rest of the paper is organized as follows: in Section 2, we introduce the model of CPS 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 integrity attacks and formulate it as a constrained control design problem. In Section 4 we discuss how to derive the upper bound for the reachable region. We also prove a necessary condition under which the reachable region is unbounded. An illustrative example is provided in Section 5. Finally Section 6 concludes the paper.

2. SYSTEM DESCRIPTION

In this section we model the CPS as a linear control system, which is equipped with a Kalman filter, a LQG controller and a χ^2 failure detector.

2.1 Physical System

We assume that the physical system has Linear Time Invariant (LTI) dynamics, which take the following form:

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

where $x_k \in \mathbb{R}^n$ is the vector of physical 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. w_k, x_0 are independent Gaussian random variables, and $x_0 \sim \mathcal{N}(0,\ \Sigma)$, $w_k \sim \mathcal{N}(0,\ Q)$.

2.2 Kalman filter and LQG controller

A sensor network is deployed to monitor the system described in (1). At each step all the sensor readings are collected and sent to a centralized estimator. The observation equation can be written as

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

where $y_k = [y_{k,1}, \ldots, y_{k,m}]^T \in \mathbb{R}^m$ is a vector of measurements from the sensors, and $y_{k,i}$ is the measurement made by sensor i at time k. $v_k \sim \mathcal{N}(0,\ R)$ is the measurement noise independent of x_0 and w_k.

A Kalman filter is used to compute state estimation $\hat{x}_k$ from observations y_ks:

$$\begin{aligned}
\hat{x}_{0|-1} &= 0,\ P_{0|-1} = \Sigma, \\
\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k,\ P_{k+1|k} = AP_kA^T + Q, \\
K_k &= P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}, \\
\hat{x}_k &= \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}), \\
P_k &= P_{k|k-1} - K_kCP_{k|k-1}.
\end{aligned}$$

Although the Kalman filter uses a time varying gain K_k, it is well known that this gain will converge if the system is detectable. In practice the Kalman gain usually converges in a few steps. Thus, we can safely assume the Kalman filter to be already in steady state. Let us define

$$P \triangleq \lim_{k\to\infty} P_{k|k-1},\ K \triangleq PC^T(CPC^T + R)^{-1}. \tag{3}$$

The update equations of Kalman filter are as follows:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + K\left[y_{k+1} - C(A\hat{x}_k + Bu_k)\right], \tag{4}$$

For future analysis, let us define the residue z_{k+1} at time $k+1$ to be

$$z_{k+1} \triangleq y_{k+1} - C(A\hat{x}_k + Bu_k). \tag{5}$$

(4) can be simplified as

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + Kz_{k+1}. \tag{6}$$

The estimation error e_k at time k is defined as

$$e_k \triangleq x_k - \hat{x}_k. \tag{7}$$

An LQG controller is used to stabilize the system by minimizing the following objective function[1]:

$$J = \lim_{T\to\infty} \min_{u_0,\ldots,u_T} E\frac{1}{T}\left[\sum_{k=0}^{T-1}(x_k^TWx_k + u_k^TUu_k)\right], \tag{8}$$

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 optimal controller of the above minimization problem is a fixed gain controller, which takes the following form:

$$u_k = -(B^TSB + U)^{-1}B^TSA\hat{x}_k, \tag{9}$$

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

$$S = A^TSA + W - A^TSB(B^TSB + U)^{-1}B^TSA. \tag{10}$$

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

[1] We assume an infinite horizon LQG controller is implemented.

It is easy to see that x_k, e_k are the states of the CPS[2]. Let us define

$$\tilde{x}_k \triangleq \begin{bmatrix} x_k \\ e_k \end{bmatrix} \in \mathbb{R}^{2n} \tag{11}$$

Hence, we can write the dynamics of CPS as

$$\begin{aligned} \tilde{x}_{k+1} = & \begin{bmatrix} A+BL & -BL \\ 0 & A-KCA \end{bmatrix} \tilde{x}_k \\ & + \begin{bmatrix} I & 0 \\ I-KC & -K \end{bmatrix} \begin{bmatrix} w_k \\ v_k \end{bmatrix} \end{aligned}$$

It is trivial to prove that the CPS is stable if and only if both matrices $A-KCA$ and $A+BL$ are stable. In the rest of the paper, we will only consider stable CPS. Further, we assume that the CPS is already in steady state, which means that $\{x_k, y_k, \hat{x}_k\}$ are stationary random processes.

Remark 1. *At the first glance, it seems that our choice of estimator, controller is quite limited. However, the analysis in this paper can be easily generalized to any fixed gain linear estimator and controller.*

2.3 Failure Detector

Failure detectors are often used in CPS to detect abnormal operations. We assume that a χ^2 failure detector is deployed, which computes the following quantity

$$g_k = z_k^T \mathcal{P}^{-1} z_k, \tag{12}$$

where $\mathcal{P}$ is the covariance matrix of the residue z_k. Since z_k is Gaussian distributed, g_k is χ^2 distributed with m degrees of freedom. As a result, g_k cannot be far away from 0. The χ^2 failure detector will compare g_k with a certain threshold. If g_k is greater than the threshold, then an alarm will be triggered. We assume that the probability of false alarm for the χ^2 detector is P_f.

Remark 2. *We will show later that the choice of detector is not critical for the analysis to hold. In fact, our result is valid for any detector which computes g_k as*

$$g_k = f(\hat{x}_k, \dots, \hat{x}_{k-T}, y_k, \dots, y_{k-T}, z_k, \dots, z_{k-T}), \tag{13}$$

where f is an arbitrary continuous function and T is the window size.

3. ATTACK MODEL

In this section we want to describe the integrity attack model on the CPS. To distinguish the compromised system and healthy system, we will use x'_k, y'_k, u'_k to indicate the states, measurements and control inputs of the compromised system respectively. We assume that an adversary has the following capabilities:

1. The adversary knows the static parameters of the system, namely A, B, C, K, L.
2. The adversary compromised a subset of sensors, and can add arbitrary bias to the reading of compromised sensors. As a result, the modified reading received by the estimator takes the following form:

$$y'_k = Cx'_k + \Gamma y^a_k + v_k, \tag{14}$$

where $\Gamma = diag(\gamma_1, \dots, \gamma_m)$ is the sensor selection matrix such that $\gamma_i = 1$ if the ith sensor is compromised and $\gamma_i = 0$ otherwise. y^a_k is the bias introduced by the attacker.

3. The adversary can inject external control inputs to the system. As a result, the system equation becomes

$$x'_{k+1} = Ax'_k + Bu'_k + B^a u^a_k + w_k, \tag{15}$$

where $B^a \in \mathbb{R}^{n\times q}$ characterizes the direction of control inputs the attacker could inject to the system.

Let us define $\hat{x}'_k$, z'_k, e'_k as the state estimation, residue and estimation error of the compromised system respectively. Moreover, let us define the differences between the healthy and compromised system as

$$\begin{aligned} \Delta x_k \triangleq & \quad x'_k - x_k, \; \Delta y_k \triangleq \quad y'_k - y_k, \\ \Delta u_k \triangleq & \quad u'_k - u_k, \; \Delta \hat{x}_k \triangleq \quad \hat{x}'_k - \hat{x}_k, \\ \Delta z_k \triangleq & \quad z'_k - z_k, \Delta e_k \triangleq \quad e'_k - e_k \\ \Delta \tilde{x}_k \triangleq & \quad \begin{bmatrix} x'_k \\ e'_k \end{bmatrix} - \begin{bmatrix} x_k \\ e_k \end{bmatrix}. \end{aligned} \tag{16}$$

It can be proved that

$$\begin{aligned} \Delta \tilde{x}_{k+1} = & \begin{bmatrix} A+BL & -BL \\ 0 & A-KCA \end{bmatrix} \Delta \tilde{x}_k \\ & + \begin{bmatrix} B^a & 0 \\ B^a - KCB^a & -K\Gamma \end{bmatrix} \begin{bmatrix} u^a_k \\ y^a_{k+1} \end{bmatrix} \end{aligned} \tag{17}$$

and

$$\Delta z_{k+1} = \begin{bmatrix} 0 & CA \end{bmatrix} \Delta \tilde{x}_k + \begin{bmatrix} CB^a & \Gamma \end{bmatrix} \begin{bmatrix} u^a_k \\ y^a_{k+1} \end{bmatrix} \tag{18}$$

To simplify notations, let us define the following matrices:

$$\begin{aligned} \tilde{A} & \triangleq \begin{bmatrix} A+BL & -BL \\ 0 & A-KCA \end{bmatrix} \in \mathbb{R}^{2n\times 2n}, \\ \tilde{B} & \triangleq \begin{bmatrix} B^a & 0 \\ B^a - KCB^a & -K\Gamma \end{bmatrix} \in \mathbb{R}^{2n\times(q+m)} \\ \tilde{C} & \triangleq \begin{bmatrix} 0 & CA \end{bmatrix} \in \mathbb{R}^{m\times 2n}, \\ \tilde{D} & \triangleq \begin{bmatrix} CB^a & \Gamma \end{bmatrix} \in \mathbb{R}^{m\times(q+m)}. \end{aligned} \tag{19}$$

and the attacker's action ζ^a_k as

$$\zeta^a_k \triangleq \begin{bmatrix} u^a_k \\ y^a_{k+1} \end{bmatrix}. \tag{20}$$

Therefore,

$$\Delta \tilde{x}_{k+1} = \tilde{A} \Delta \tilde{x}_k + \tilde{B} \zeta^a_k, \tag{21}$$

and

$$\Delta z_{k+1} = \tilde{C} \Delta \tilde{x}_k + \tilde{D} \zeta^a_k. \tag{22}$$

It is clear that $\Delta x_k, \Delta z_k, \Delta e_k$ are functions of the attacker's actions $(\zeta^a_0, \zeta^a_1, \dots)$. Let us define $\zeta^a = (\zeta^a_0, \zeta^a_1, \dots)$ as the infinite sequence of the attacker's actions. As a result, we can write $\Delta x_k, \Delta z_k, \Delta e_k$ as $\Delta x_k(\zeta^a)$, $\Delta z_k(\zeta^a)$, $\Delta e_k(\zeta^a)$ respectively. We will omit the parameter ζ^a when there is no confusion.

We assume that the attacker wants its attack to be stealthy. In other words, the attacker wants the failure detector to have a very small probability to detect its presence. Ideally,

[2] x_k is the states of the physical system and $\hat{x}_k$ is the states of the estimator. Since the transformation from x_k, $\hat{x}_k$ to x_k, e_k is invertible, we could use x_k, e_k as the states of CPS.

to achieve this goal, the attacker would choose its action ζ^a, such that the following condition holds for all $k = 0, 1, \ldots$:

$$P({z'_k}^T \mathcal{P}^{-1} z'_k > threshold) < \alpha, \tag{23}$$

where α is a threshold probability chosen by the attacker. However, such probability is hard to compute in general as it involves in integrating a Gaussian distribution over an ellipsoid and hence difficult to enforce. As a result, we assume that attacker would enforce the following condition instead:

$$\|\Delta z_k\| \leq \beta. \forall k = 0, 1, \ldots. \tag{24}$$

where

$$\|\Delta z_k\| \triangleq \sqrt{(\Delta z_k)^T \mathcal{P}^{-1} \Delta z_k}.$$

Remark 3. *By triangular inequality, it can be easily seen that*

$$\|z'_k\| \leq \|z_k\| + \beta.$$

Therefore, when $\beta \to 0$, then the probability of detection $P({z'_k}^T \mathcal{P}^{-1} z'_k > threshold)$ converges to the false alarm probability P_f, which is in general very small. Moreover, we would like to point out that for general detector of the form (13)*, the probability of detection always converges to the false alarm probability when $\beta \to 0$. Therefore, by carefully choosing β,* (24) *is valid for more general detection schemes. Due to linearity, we will assume that $\beta = 1$ for the rest of the paper.*

We define the attacker's action ζ^a to be feasible if (24) holds for all k and $\beta = 1$. Moreover, we define the reachable region R_k as

$$R_k \triangleq \{\tilde{x} \in \mathbb{R}^{2n} : \tilde{x} = \Delta\tilde{x}_k(\zeta^a), \text{ for some feasible } \zeta^a\}.$$

and

$$\mathcal{R} \triangleq \bigcup_{k=0}^{\infty} R_k. \tag{25}$$

Remark 4. *$\mathcal{R}$ indicates all possible biases an attacker could introduce to the system. Since $x'_k = x_k + \Delta x_k$ and $e'_k = e_k + \Delta e_k$, and x_k, e_k are stationary Gaussian process, we can immediately derive the statistics of x'_k, e'_k from $\mathcal{R}$. As a result, we will focus ourselves on characterizing the reachable set $\mathcal{R}$.*

4. MAIN RESULTS

In this section we want to characterize the shape of the reachable region $\mathcal{R}$. We will provide an outer approximation of $\mathcal{R}$ based on ellipsoidal approximation. Moreover, we provide a necessary condition for $\mathcal{R}$ to be unbounded, which indicates that the attacker could destabilize the system by introducing an arbitrary large bias.

4.1 Outer Approximation of $\mathcal{R}$

First, let us define the following sets:

$$\begin{aligned} T_0 &\triangleq \mathbb{R}^{2n}, \\ T_{i+1} &\triangleq \{\tilde{x} \in \mathbb{R}^{2n} : \exists\zeta, \text{ such that } \tilde{A}\tilde{x} + \tilde{B}\zeta \in T_i, \\ &\quad \text{and } \|\tilde{C}\tilde{x} + \tilde{D}\zeta\| \leq 1\}. \end{aligned} \tag{26}$$

The following theorem shows that T_i is a superset of the reach region $\mathcal{R}$.

Theorem 1. *The following properties hold for $\mathcal{R}$ and T_is:*

1. *For any $\tilde{x} \in \mathcal{R}$, there exists a ζ, such that*
$$\tilde{A}\tilde{x} + \tilde{B}\zeta \in \mathcal{R},$$
and
$$\|\tilde{C}\tilde{x} + \tilde{D}\zeta\| \leq 1.$$

2. *$T_{i+1} \subseteq T_i$, $\forall i$.*

3. *$\mathcal{R} \subseteq T_i$, $\forall i$.*

Proof. 1. Since $\tilde{x} \in \mathcal{R}$, we know that
$$\tilde{x} = \Delta\tilde{x}_k(\zeta^a),$$
for some k and feasible ζ^a. Now choose $\zeta = \zeta^a_k$. Hence,
$$\tilde{A}\tilde{x} + \tilde{B}\zeta = \Delta\tilde{x}_{k+1}(\zeta^a) \in \mathcal{R},$$
and
$$\|\tilde{C}\tilde{x} + \tilde{D}\zeta\| = \|\Delta z_{k+1}(\zeta^a)\| \leq 1,$$
which concludes the proof.

2. Since $T_1 \subseteq T_0 = \mathbb{R}^{2n}$, it is trivial to prove that $T_{i+1} \subseteq T_i$ for all i by induction.

3. Since $\mathcal{R} \subseteq T_0 = \mathbb{R}^{2n}$, it is trivial to prove that $\mathcal{R} \subseteq T_i$ for all i by induction.

□

Due to Theorem 1, we know that T_i is an outer approximation of $\mathcal{R}$. Moreover, T_i is monotonically decrease. Therefore, we can use $\lim_{i\to\infty} T_i$ as the outer approximation of $\mathcal{R}$. However, the exact shape of T_i is still numerically difficult to compute as i goes to infinity. Therefore, we will try to compute an ellipsoidal superset of T_i. To this end, let us suppose that T_i is outer approximated by the following ellipsoid:

$$T_i \subseteq \mathcal{E}_{2n}(\tilde{Q}_i),$$

where $\tilde{Q}_i \in \mathbb{R}^{2n\times 2n}$ is positive semidefinite, and $\mathcal{E}_l(S)$ is defined as the following ellipsoid

$$\mathcal{E}_l(S) \triangleq \{x \in \mathbb{R}^l : x^T S x \leq 1\}.$$

It is trivial to see that since $T_0 = \mathbb{R}^{2n}$, $\tilde{Q}_0 = 0$. Our goal is to find a recursive algorithm to evaluate $\tilde{Q}_i$.

Let us choose an arbitrary $\tilde{x} \in T_{i+1}$ and ζ, such that

$$\tilde{A}\tilde{x} + \tilde{B}\tilde{\zeta} \in T_i, \|\tilde{C}\tilde{x} + \tilde{D}\zeta\| \leq 1.$$

Since T_i is outer approximated by $\mathcal{E}_{2n}(\tilde{Q}_i)$, we know that

$$\begin{bmatrix} \tilde{x}^T & \zeta^T \end{bmatrix} \begin{bmatrix} \tilde{A}^T\tilde{Q}_i\tilde{A} & \tilde{A}^T\tilde{Q}_i\tilde{B} \\ \tilde{A}^T\tilde{Q}_i\tilde{B} & \tilde{B}^T\tilde{Q}_i\tilde{B} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \zeta \end{bmatrix} \leq 1.$$

Moreover,

$$\begin{bmatrix} \tilde{x}^T & \zeta^T \end{bmatrix} \begin{bmatrix} \tilde{C}^T\mathcal{P}^{-1}\tilde{C} & \tilde{C}^T\mathcal{P}^{-1}\tilde{D} \\ \tilde{C}^T\mathcal{P}^{-1}\tilde{D} & \tilde{D}^T\mathcal{P}^{-1}\tilde{D} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \zeta \end{bmatrix} \leq 1.$$

Therefore, $[\tilde{x}^T, \zeta^T]^T$ must jointly lies in the intersection of the following two $2n + p + q$ dimension ellipsoids:

$$\begin{bmatrix} \tilde{x} \\ \zeta \end{bmatrix} \in \mathcal{E}_{2n+q+m}\left(\begin{bmatrix} \tilde{A}^T\tilde{Q}_i\tilde{A} & \tilde{A}^T\tilde{Q}_i\tilde{B} \\ \tilde{B}^T\tilde{Q}_i\tilde{A} & \tilde{B}^T\tilde{Q}_i\tilde{B} \end{bmatrix}\right)$$
$$\bigcap \mathcal{E}_{2n+q+m}\left(\begin{bmatrix} \tilde{C}^T\mathcal{P}^{-1}\tilde{C} & \tilde{C}^T\mathcal{P}^{-1}\tilde{D} \\ \tilde{D}^T\mathcal{P}^{-1}\tilde{C} & \tilde{D}^T\mathcal{P}^{-1}\tilde{D} \end{bmatrix}\right).$$

One can prove that an outer approximation of the intersection is given by

$$\begin{aligned}\left[\begin{array}{c}\tilde{x}\\ \zeta\end{array}\right] &\in \mathcal{E}_{2n+q+m}\left(\left[\begin{array}{cc}\tilde{A}^T\tilde{Q}_i\tilde{A} & \tilde{A}^T\tilde{Q}_i\tilde{B}\\ \tilde{B}^T\tilde{Q}_i\tilde{A} & \tilde{B}^T\tilde{Q}_i\tilde{B}\end{array}\right]\right)\\ &\bigcap \mathcal{E}_{2n+q+m}\left(\left[\begin{array}{cc}\tilde{C}^T\mathcal{P}^{-1}\tilde{C} & \tilde{C}^T\mathcal{P}^{-1}\tilde{D}\\ \tilde{D}^T\mathcal{P}^{-1}\tilde{C} & \tilde{D}^T\mathcal{P}^{-1}\tilde{D}\end{array}\right]\right)\\ &\subseteq \mathcal{E}_{2n+q+m}\left(\frac{1}{2}\left[\begin{array}{cc}\tilde{A}^T\tilde{Q}_i\tilde{A} & \tilde{A}^T\tilde{Q}_i\tilde{B}\\ \tilde{B}^T\tilde{Q}_i\tilde{A} & \tilde{B}^T\tilde{Q}_i\tilde{B}\end{array}\right]\right.\\ &\left.+\frac{1}{2}\left[\begin{array}{cc}\tilde{C}^T\mathcal{P}^{-1}\tilde{C} & \tilde{C}^T\mathcal{P}^{-1}\tilde{D}\\ \tilde{D}^T\mathcal{P}^{-1}\tilde{C} & \tilde{D}^T\mathcal{P}^{-1}\tilde{D}\end{array}\right]\right).\end{aligned}$$

Finally, using the Schur complement, we can project a high dimensional ellipsoid in $\mathbb{R}^{2n+q+m}$ to $\mathbb{R}^{2n}$ to obtain $\tilde{Q}_{i+1}$ as follows:

$$\begin{aligned}\tilde{Q}_{i+1} = \frac{1}{2}\Big[&\tilde{A}^T\tilde{Q}_i\tilde{A} + \tilde{C}^T\mathcal{P}^{-1}\tilde{C}\\ &-(\tilde{A}^T\tilde{Q}_i\tilde{B} + \tilde{C}^T\mathcal{P}^{-1}\tilde{D})\\ &\times(\tilde{B}^T\tilde{Q}_i\tilde{B} + \tilde{D}^T\mathcal{P}^{-1}\tilde{D})^+\\ &\times(\tilde{B}^T\tilde{Q}_i\tilde{A} + \tilde{D}^T\mathcal{P}^{-1}\tilde{C})\Big],\end{aligned} \tag{27}$$

where $^+$ is the Moore-Penrose pseudoinverse.

4.2 Stability analysis

In this subsection, we want to characterize the boundedness of $\mathcal{R}$. An unbounded $\mathcal{R}$ indicates that the attacker could destabilize the system by introducing an arbitrary large bias. The following theorem provides a necessary condition for $\mathcal{R}$ to be unbounded.

Theorem 2. *The reachable region $\mathcal{R}$ is unbounded only if there exist a vector $v \in \mathbb{R}^n$ and a matrix $L^a \in \mathbb{R}^{q\times n}$, such that*

1. *v is an eigenvector of $A + B^aL^a$, the corresponding eigenvalue of which is λ.*

2. *Cv belongs to the column space of Γ or $\lambda = 0$.*

Remark 5. *It is worth noticing that the shape of $\mathcal{R}$ depends on the choice of estimation and controller gain K, L in general. However, the necessary condition is independent of K, L. In other words, if the necessary condition holds, then the system has an inherent vulnerability which cannot be fixed by simply redesigning the estimator and controller.*

The rest of the subsection is devoted to proving Theorem 2. First we want to show that Δx_k is bounded if and only if Δe_k is bounded. From definition,

$$\Delta x_k = \Delta e_k + \Delta\hat{x}_k,$$

and

$$\Delta\hat{x}_{k+1} = (A + BL)\Delta\hat{x}_k + K\Delta z_{k+1}.$$

Since we assume that the system is closed-loop stable, which implies that $A + BL$ is stable, and $\|\Delta z_k\| \le 1$, $\Delta\hat{x}_k$ must be bounded. Thus, the boundedness of Δx_k is equivalent to the boundedness of Δe_k. Therefore, we only need to focus on Δe_k and thus we could simplify (21),(22) as

$$\Delta e_{k+1} = (A-KCA)\Delta e_k+(B^a-KCB^a)u_k^a-K\Gamma y_{k+1}^a, \tag{28}$$

$$\Delta z_{k+1} = CA\Delta e_k + CB^au_k^a + \Gamma y_{k+1}^a. \tag{29}$$

To further simplify notations, we define the following matrices:

$$\begin{aligned}&\mathcal{A} \triangleq A - KCA \in \mathbb{R}^{n\times n},\\ &\mathcal{B} \triangleq \left[\begin{array}{cc}B^a - KCB^a & -K\Gamma\end{array}\right] \in \mathbb{R}^{n\times(q+m)}\\ &\mathcal{C} \triangleq CA \in \mathbb{R}^{m\times n},\\ &\mathcal{D} \triangleq \tilde{D} \in \mathbb{R}^{m\times(q+m)}.\end{aligned} \tag{30}$$

Thus, we can write (28) and (29) as

$$\begin{aligned}\Delta e_{k+1} &= \mathcal{A}\Delta e_k + \mathcal{B}\zeta_k,\\ \Delta z_{k+1} &= \mathcal{C}\Delta e_k + \mathcal{D}\zeta_k,\end{aligned}$$

Lemma 1. *There exists feasible attacker's action ζ^a such that $\Delta e_k(\zeta^a)$ is unbounded only if there exist a vector $v \in \mathbb{R}^n$ and a matrix $\mathcal{L} \in \mathbb{R}^{(q+m)\times n}$, such that*

1. *v is an eigenvector of $\mathcal{A} + \mathcal{B}\mathcal{L}$,*

2. *$(\mathcal{C} + \mathcal{D}\mathcal{L})v = 0$*

Proof. The proof is quite long and is hence reported in the appendix for the sake of legibility. □

Now we are ready to prove the Theorem 2.

Proof. By Lemma 1, we know that $\mathcal{R}$ is unbounded only when there exist a vector $v \in \mathbb{R}$. and a matrix $\mathcal{L} \in \mathbb{R}^{(q+m)\times n}$, such that

1. v is an eigenvector of $\mathcal{A} + \mathcal{B}\mathcal{L}$,

2. $(\mathcal{C} + \mathcal{D}\mathcal{L})v = 0$

Now let us write $\mathcal{L}$ as

$$\mathcal{L} = \left[\begin{array}{c}L^a\\ M\end{array}\right]$$

where $L^a \in \mathbb{R}^{q\times n}$ and $M \in \mathbb{R}^{m\times n}$. Since v satisfies $(\mathcal{C} + \mathcal{D}\mathcal{L})v = 0$, we have

$$(CA + CB^aL^a + \Gamma M)v = 0. \tag{31}$$

Now by the fact that v is an eigenvector of $\mathcal{A}+\mathcal{B}\mathcal{L}$, we have

$$(A - KCA + B^aL^a - KCB^aL^a - K\Gamma M)v = \lambda v, \tag{32}$$

where λ is the corresponding eigenvalue. Combining (31) and (32), we have

$$\begin{aligned}(A + B^aL^a)v - K(CA + CB^aL^a + \Gamma M)v &= (A + B^aL^a)v\\ &= \lambda v.\end{aligned} \tag{33}$$

Therefore, v is also an eigenvector of $A + B^aL^a$. Now by (31) and (33), we have

$$\lambda Cv = -\Gamma Mv.$$

The right-hand side of equation belongs to the column space of Γ. Therefore, either $\lambda = 0$ or Cv belongs to the column space of Γ, which concludes the proof. □

5. ILLUSTRATIVE EXAMPLE

In this section, we will provide a numerical example to illustrate the effects of integrity attack on CPS.

Consider a vehicle which is moving along the x-axis. The state space includes position x and velocity v of the vehicle. As a result, the discrete-time system dynamics are as follows:

$$\begin{aligned} v_{k+1} &= v_k + w_{k,1} + u_k, \\ x_{k+1} &= x_k + v_k + w_{k,2}, \end{aligned} \tag{34}$$

which can be written in the matrix form as

$$X_{k+1} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} X_k + w_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u_k, \tag{35}$$

where

$$X_k = \begin{bmatrix} v \\ x \end{bmatrix}, \; w_k = \begin{bmatrix} w_{k,1} \\ w_{k,2} \end{bmatrix}. \tag{36}$$

Suppose two sensors are measuring velocity and position respectively. Hence

$$y_k = X_k + v_k. \tag{37}$$

We assume that the covariance of the noise is $Q = R = I$. The steady state Kalman gain in this case is

$$K = \begin{bmatrix} 0.5939 & 0.0793 \\ 0.0793 & 0.6944 \end{bmatrix}. \tag{38}$$

Moreover, we assume that the LQG cost $W = I$ and $U = 1$. Therefore, the control gain is given by

$$L = \begin{bmatrix} -1.2439 & -0.4221 \end{bmatrix} \tag{39}$$

We consider two cases, where either the velocity sensor or the position sensor is compromised, i.e. $\Gamma = [1, 0]'$ or $\Gamma = [0, 1]'$. We assume that the attacker does not inject external control input, i.e. $B^a = [0, 0]'$ for both cases.

Figure 1 shows the outer approximation of $\mathcal{R}$ when the velocity sensor is compromised. From the simulation we can conclude that the reachable region $\mathcal{R}$ is bounded. Therefore the attacker cannot destabilize the system by simply compromising velocity sensor.

Figure 2 shows the outer approximation of $\mathcal{R}$ when the position sensor is compromised. It can be seen that the outer approximation is unbounded (The ellipse degenerates into two straight lines), which implies that the attacker can arbitrarily manipulate the position of the vehicle. In fact, one can check that $v = [0, 1]'$ and $L^a = 0$ satisfies the necessary conditions listed in Theorem 2.

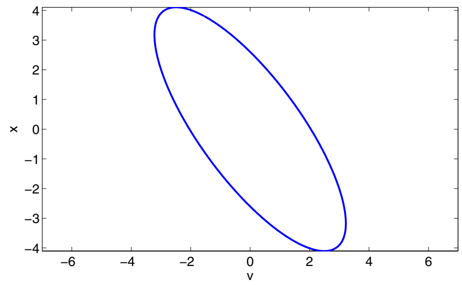

Figure 1: Outer Approximation When Velocity Sensor is Compromised

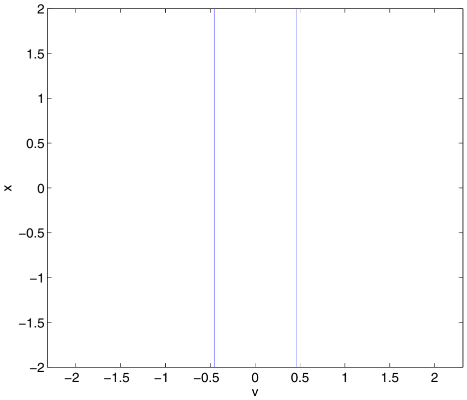

Figure 2: Outer Approximation When Position Sensor is Compromised

6. CONCLUSION

In this paper we consider the integrity attack on Cyber-Physical System(CPS). We formalize the adversary's strategy as a constrained control problem and characterize the shape and boundedness of the reachable set of the system state and estimation error under the attack, which provides a quantitative measure of the resilience of the system.

7. REFERENCES

[1] 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*, pages 31–45. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, April 2009.

[2] A. A. Cárdenas, S. Amin, and S. Sastry. Secure control: Towards survivable cyber-physical systems. In *Distributed Computing Systems Workshops, 2008. ICDCS '08. 28th International Conference on*, pages 495–500, June 2008.

[3] Y. Liu, M. Reiter, and P. Ning. False data injection attacks against state estimation in electric power grids. In *Proceedings of the 16th ACM conference on Computer and communications security*, 2009.

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

[5] Y. Mo and B. Sinopoli. Secure control against replay attacks. In *Proc. 47th Annual Allerton Conf. Communication, Control, and Computing Allerton 2009*, pages 911–918, 2009.

[6] F. Pasqualetti, A. Bicchi, and F. Bullo. Consensus computation in unreliable networks: A system theoretic approach. *IEEE Transactions on Automatic Control*, Feb. 2010. To appear.

[7] H. Sandberg, A. Teixeira, and K. H. Johansson. On security indices for state estimators in power networks. In *First Workshop on Secure Control Systems*, 2010.

[8] S. Sundaram, M. Pajic, C. Hadjicostis, R. Mangharam,

and G. J. Pappas. The wireless control network: monitoring for malicious behavior. In *IEEE Conference on Decision and Contro*, Atlanta, GA, Dec 2010.

[9] L. Xie, Y. Mo, and B. Sinopoli. False data injection attacks in electricity markets. In *IEEE Int'l Conf. on Smart Grid Communications (SmartGridComm)*, pages 226–231, Oct. 2010.

8. APPENDIX

Before proving Lemma 1, we want to define the following sets: $S_0 \triangleq \mathbb{R}^n$ and

$$S_{i+1} \triangleq \{e \in \mathbb{R}^n : \exists \zeta, \mathcal{A}e + \mathcal{B}\zeta \in S_i, \\ \|\mathcal{C}e + \mathcal{D}\zeta)\|_\infty \leq 1\}.$$

Remark 6. *Due to the equivalence of norms in $\mathbb{R}^n$ and linearity of the system, we could use infinity norm instead of $\|\cdot\|_{\mathcal{P}^{-1}}$ without affecting the stability result.*

Let us also define

$$\mathcal{S} \triangleq \bigcap_{i=0}^{\infty} S_i.$$

The following lemma characterizes some important properties of S_i and $\mathcal{S}$:

Lemma 2. *The following statements of $\mathcal{S}$ and S_i hold:*

1. *$S_{i+1} \subseteq S_i$.*
2. *S_i takes the following form:*

$$S_i = \{e \in \mathbb{R}^n : \mathcal{A}_i e \leq b_i\}, \qquad (40)$$

where $\mathcal{A}_i$ is a matrix and b_i is a vector of proper dimensions and the comparison is entry-wise.

3. *$\mathcal{S}$ is convex and closed.*
4. *If $e \in \mathcal{S}$, then $\alpha e \in \mathcal{S}$, where $\alpha \in [-1, 1]$.*
5. *Suppose that subspaces $\mathcal{V}, \mathcal{V}' \subseteq \mathcal{S}$, then the direct sum $\mathcal{V} \oplus \mathcal{V}' \subseteq \mathcal{S}$.*
6. *If $\mathcal{S}$ is unbounded, then $\mathcal{S}$ contains a subspace $\mathcal{V} \neq \{0\}$.*
7. *For any $e \in \mathcal{S}$, there exists an ζ, such that*

$$\mathcal{A}e + \mathcal{B}\zeta \in \mathcal{S}, \|\mathcal{C}e + \mathcal{D}\zeta\|_\infty \leq 1. \qquad (41)$$

Proof. 1. Since $S_0 = \mathbb{R}^n$, $S_1 \subseteq S_0$. The statement can be easily proved by induction.

2. Since we use infinite norm in the definition of S_i, this can also be shown by induction.
3. It is trivial to see that S_i is convex and closed for each i. Hence, their intersection $\mathcal{S}$ is also convex and closed.
4. Since S_i is symmetric, $\mathcal{S}$ is also symmetric. Using the convexity of $\mathcal{S}$, we can finish the proof.
5. This is a direct consequence of the convexity of $\mathcal{S}$.
6. Suppose that $e_1, e_2, \ldots$ is an unbounded sequence in $\mathcal{S}$. Without loss of generality, we assume that

$$\lim_{i\to\infty} e_i/\|e_i\|_\infty = e^0, \lim_{i\to\infty} \|e_i\|_\infty = \infty.$$

Now pick an arbitrary $\alpha \in \mathbb{R}$. There exists N, such that for all $i \geq N$, $\alpha/\|e_i\|_\infty \in [-1, 1]$, which implies that

$$\frac{\alpha}{\|e_i\|_\infty} e_i \in \mathcal{S}.$$

Take the limit on the left side and use the fact that $\mathcal{S}$ is closed, we have $\alpha e^0 \in \mathcal{S}$ for all α. Hence, $\mathcal{S}$ contains *span*(e^0).

7. From the definition of S_i, if $e \in \mathcal{S}$, there exists ζ_i for each i such that

$$\mathcal{A}e + \mathcal{B}\zeta_i \in S_i, \|\mathcal{C}e + \mathcal{D}\zeta_i\|_\infty \leq 1. \qquad (42)$$

Without loss of generality, let us pick such ζ_is with minimal infinite norm. Suppose that ζ_i converges to ζ^*. For each i, we know that if $j \geq i$, then

$$\mathcal{A}e + \mathcal{B}\zeta_j \in S_j \subseteq S_i.$$

Take the limit on the left-hand side and use the fact that S_i is closed, we have

$$\mathcal{A}e + \mathcal{B}\zeta^* = \lim_{j\to\infty} \mathcal{A}e + \mathcal{B}\zeta_j \in S_i, \forall i.$$

Therefore

$$\mathcal{A}e + \mathcal{B}\zeta^* \in \bigcap_{i=0}^{\infty} S_i = \mathcal{S}.$$

Moreover,

$$\|\mathcal{C}e + \mathcal{D}\zeta^*\|_\infty = \lim_{i\to\infty} \|\mathcal{C}e + \mathcal{D}\zeta_i\|_\infty \leq 1$$

Hence, ζ^* is the required vector for (41). As a result, we only need to prove that $\{\zeta_i\}$ converges or at least contains a converging subsequence. We will prove that by contradiction. Suppose $\{\zeta_i\}$ does not contains any converging subsequence. Due to Bolzano Weierstrass Theorem, $\{\zeta_i\}$ must be unbounded. Again, by Bolzano Weierstrass theorem, there exists $\zeta_{i_1}, \zeta_{i_2}, \ldots$, such that

$$\lim_{j\to\infty} \|\zeta_{i_j}\|_\infty = \infty, \zeta^0 = \lim_{j\to\infty} \zeta_{i_j}/\|\zeta_{i_j}\|_\infty.$$

We now have

$$\lim_{j\to\infty} \frac{\|\mathcal{C}e + \mathcal{D}\zeta_{i_j}\|_\infty}{\|\zeta_{i_j}\|_\infty} = \|\mathcal{D}\zeta^0\|_\infty \leq \lim_{j\to\infty} 1/\|\zeta_{i_j}\|_\infty = 0.$$

Hence $\mathcal{D}\zeta^0 = 0$. Pick an arbitrary $\alpha \in \mathbb{R}$ and $l \in \mathbb{N}$. There exists an N such that if $j \geq N$, then

$$\frac{\alpha}{\|\zeta_{i_j}\|_\infty} \in [-1, 1], \text{ and } i_j \geq l + 1.$$

As a result,

$$\frac{\alpha}{\|\zeta_{i_j}\|_\infty} (\mathcal{A}e + \mathcal{B}\zeta_{i_j}) \in S_{i_j} \subseteq S_l.$$

Take the limit on the left side and use the fact that S_l is closed, we have

$$\alpha\mathcal{B}\zeta^0 \in S_l, \forall \alpha.$$

As a result, *span*$(\mathcal{B}\zeta^0) \in S_i$. Therefore, $\mathcal{A}_i\mathcal{B}\zeta^0 = 0$ for all i, which implies that for any $\alpha \in \mathbb{R}$,

$$\mathcal{A}e + \mathcal{B}(\zeta_{i_j} - \alpha\zeta^0) \in S_{i_j}, \\ \|\mathcal{C}e + \mathcal{D}(\zeta_{i_j} - \alpha\zeta^0)]\|_\infty \leq 1.$$

Therefore, the fact that ζ_{i_j} goes to infinity contradicts the minimality of ζ_{i_j}, which completes the proof.

□

Now we are ready to prove Lemma 1.

Proof. Similar to the proof of Theorem 1, we can prove that the reachable region of Δe_k is contained in $\alpha\mathcal{S}$, where $\alpha > 0$ is a constant. As a result, $\mathcal{S}$ is unbounded. By

Lemma 2, we know that there exists a subspace $\mathcal{V} \subseteq \mathcal{S}$. Moreover we can assume $\mathcal{V}$ is maximal subspace contained in $\mathcal{S}$ due to Lemma 2(5).

Now pick an arbitrary vector e in $\mathcal{V}$. We know there exists $\zeta_k \in \mathbb{R}^p$, $k \in \mathbb{N}$ such that

$$\mathcal{A}(ke) + \mathcal{B}\zeta_k \in \mathcal{S}, \, \|\mathcal{C}(ke) + \mathcal{D}\zeta_k\|_\infty \leq 1.$$

We will pick such ζ_k with minimal norm. By similar argument as in Lemma 2, we can prove that $\sup_k \|\zeta_k\|/k$ must be finite. By Bolzano Weierstrass theorem, there exists $\zeta_{i_1}, \zeta_{i_2}, \ldots$, such that

$$\lim_{j\to\infty} \zeta_{i_j}/i_j = \zeta^*.$$

Similar to Lemma 2, we can also prove that

$$\|\mathcal{C}e + \mathcal{D}\zeta^*\| = 0.$$

and

$$span(\mathcal{A}e + \mathcal{B}\zeta^*) \subseteq \mathcal{S}.$$

Thus, $span(\mathcal{A}e + \mathcal{B}\zeta^*) \subseteq \mathcal{V}$ due to the maximality of $\mathcal{V}$.

Now suppose $e_0, \ldots, e_j$ form a basis for $\mathcal{V}$. For every e_i, there exists ζ_i^*, such that $\mathcal{A}e_i + \mathcal{B}\zeta_i^* \in \mathcal{V}$ and $\mathcal{C}e_i + \mathcal{D}\zeta_i^* = 0$. Hence, we could find a matrix $\mathcal{L}$, such that $\zeta_i^* = \mathcal{L}e_i$, for all e_i, which implies that $(\mathcal{A} + \mathcal{B}\mathcal{L})\mathcal{V} \subseteq \mathcal{V}$ and $(\mathcal{C} + \mathcal{D}\mathcal{L})\mathcal{V} = 0$. Therefore, there exists $v \in \mathcal{V}$, such that v is the eigenvector of $\mathcal{A} + \mathcal{B}\mathcal{L}$, and $(\mathcal{C} + \mathcal{D}\mathcal{L})v = 0$. □