Preprints of the 20th World Congress
The International Federation of Automatic Control
Toulouse, France, July 9-14, 2017

Reachable Sets of Hidden CPS Sensor Attacks: Analysis and Synthesis Tools

Carlos Murguia * **Nathan van de Wouw** ** **Justin Ruths** ***

* *Singapore University of Technology and Design, Center for Research in Cyber Security* (iTrust). *e-mail: murguia_rendon@sutd.edu.sg*

** *Eindhoven University of Technology, Mechanical Engineering Department, Eindhoven, The Netherlands;*
Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands;
Department of Civil, Environmental and Geo-Engineering, University of Minnesota, Minneapolis, USA.
e-mail: n.v.d.wouw@tue.nl

*** *University of Texas at Dallas, Departments of Mechanical and Systems Engineering. e-mail: jruths@utdallas.edu*

Abstract: For given system dynamics, control structure, and fault/attack detection procedure, we provide mathematical tools–in terms of Linear Matrix Inequalities (LMIs)–for characterizing and minimizing the set of states that sensor attacks can induce in the system while keeping the alarm rate of the fault detector sufficiently close to its false alarm rate in the attack-free case. This quantifies the attack's potential impact when it is constrained to stay hidden from the detector. Simulation results are presented to illustrate the performance of our tools.

Keywords: cyber physical systems; security; stochastic systems; model-based fault detectors; reachable sets.

1. INTRODUCTION

During the past decades, scientific and technological advances have greatly improved the performance of control systems. From heating/cooling devices in our homes, to cruise-control in our cars, to robotics in manufacturing centers. However, these new technologies have also led to vulnerabilities of some our most critical infrastructures – e.g., power, water, transportation. Advances in communication and computing have given rise to adversaries with enhanced and adaptive capabilities. Depending on their resources, attackers may deteriorate the functionality of systems even while remaining undetected. Therefore, designing efficient attack detection schemes and attack-robust control systems is of key importance for guaranteeing the safety and proper operation of critical systems. Tools from sequential analysis and fault detection have to be adapted to deal with the systematic, strategic, and persistent nature of attacks. These new challenges have attracted the attention of many researchers in the control and computer science communities, see e.g., (Cárdenas et al., 2011; Pasqualetti et al., 2013; Mo et al., 2010; Kwon et al., 2013; Miao et al., 2014; Bai et al., 2015), and references therein.

This paper addresses the problem of characterizing the impact of sensor attacks on Linear Time-Invariant (LTI) stochastic systems when fault detection techniques are deployed for attack detection, see, e.g., (Chen and Patton, 1999; Kyriakides and Polycarpou, 2015; Pasqualetti et al., 2013; Cárdenas et al., 2011). The main idea behind fault detection theory is the use of an *estimator* to forecast the evolution of the system dynamics. If the difference between measurements and the estimation is larger than expected, there may be a fault in or an attack on the system. The complete fault detection scheme consists of two parts: the *estimator* and a *change detection procedure* (used to decide whether the estimator and the system are sufficiently different to declare the presence of faults/attacks). We use *observers*, (Luenberger, 1966; Nijmeijer and Mareels, 1997), as estimators; and the *chi-squared procedure* for change detection (Gustafsson, 2000).

The main contribution of the paper is a set of mathematical tools for *quantifying* and *minimizing* the impact of sensor attacks on the system dynamics. This effect depends on where, when, and how, the attack occurs in the system. To capture this, we model attacks as additive perturbations affecting sensors measurements (Kyriakides and Polycarpou, 2015; Pasqualetti et al., 2013; Cárdenas et al., 2011). These perturbations are propagated to the system dynamics through output-based controllers. To quantify the effect of attacks, we need to introduce some measure of impact. However, because malicious adversaries may launch any arbitrary attack, we need a measure which can

⋆ This work was supported by the National Research Foundation (NRF), Prime Minister's Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-40) and administered by the National Cybersecurity R&D Directorate.

capture all possible states that the attacker can induce in the system, given how it accesses the dynamics (i.e., through the control scheme by tampering with sensor measurements). We propose to use the *reachable set* of the attack (Boyd et al., 1994) as our measure of impact.

We remark that all detectors produce false alarms – due to the stochastic nature of the system and measurement noise. We refer to attacks that are able to maintain the alarm rate of the detector equivalent to its attack-free false alarm rate as *hidden attacks* (since they make the behavior of the detector of the attacked system indistinguishable from its behavior without attacks). Contrast hidden attacks with *zero alarm attacks*, which ensure that no alarms are raised during an attack by maintaining the detection statistic just beneath its detection threshold, see, e.g., (Murguia and Ruths, 2016a,b; Giraldo et al., 2016; Cardenas et al., 2009). In this work, we characterize the reachable sets that *hidden attacks* can induce in the system. We refer to these sets as the *hidden reachable sets* of the attack sequence. In general, it is intractable to compute these sets exactly. Instead, for given system dynamics, control structure, and attack detection procedure, we derive *ellipsoidal bounds* on the hidden reachable sets using Linear Matrix Inequalities (LMIs), (Boyd et al., 1994). We provide synthesis tools for minimizing these bounds (minimizing thus the hidden reachable sets) by properly redesigning controllers and detectors.

There are a few results in this direction already; chiefly the work in (Mo and Sinopoli, 2016), where a recursive algorithm to compute ellipsoidal inner and outer bounds of hidden reachable sets is provided. We assert our analysis tools are more constructive and easier to implement, which we achieve by reformulating the problem of computing the ellipsoidal bounds as a convex optimization problem in terms of LMIs. Another aspect that takes our work beyond the analysis results of (Mo and Sinopoli, 2016) is that we also provide synthesis tools for minimizing the hidden reachable sets by redesigning detectors and controllers.

2. SYSTEM DESCRIPTION & ATTACK DETECTION

We study LTI stochastic systems of the form:

$$\begin{cases} x(t_{k+1}) = Fx(t_k) + Gu(t_k) + v(t_k), \\ y(t_k) = Cx(t_k) + \eta(t_k), \end{cases} \tag{1}$$

with sampling time-instants $t_k, k \in \mathbb{N}$, state $x \in \mathbb{R}^n$, measured output $y \in \mathbb{R}^m$, control input $u \in \mathbb{R}^l$, matrices F, G, and C of appropriate dimensions, and i.i.d. multivariate zero-mean Gaussian noises $v \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ with covariance matrices $R_1 \in \mathbb{R}^{n \times n}$, $R_1 \geq 0$ and $R_2 \in \mathbb{R}^{m \times m}$, $R_2 \geq 0$, respectively. The initial state $x(t_1)$ is assumed to be a Gaussian random vector with covariance matrix $R_0 \in \mathbb{R}^{n \times n}$, $R_0 \geq 0$. The processes $v(t_k)$, $k \in \mathbb{N}$ and $\eta(t_k)$, $k \in \mathbb{N}$ and the initial condition $x(t_1)$ are mutually independent. It is assumed that (F, G) is stabilizable and (F, C) is detectable. At the time-instants $t_k, k \in \mathbb{N}$, the output of the process $y(t_k)$ is sampled and transmitted over a communication network. The received output $\bar{y}(t_k)$ is used to compute control actions $u(t_k)$ which are sent back to the process, see Fig. 1. The complete control-loop is assumed to be performed instantaneously, i.e., the sampling, transmission, and arrival time-instants are supposed to be equal. In this paper, we focus on attacks on sensor measurements. That is, in between transmission and reception of sensor data, an attacker may replace the signals coming from the sensors to the controller, see Fig. 1. After each transmission and reception, the attacked output $\bar{y}$ takes the form:

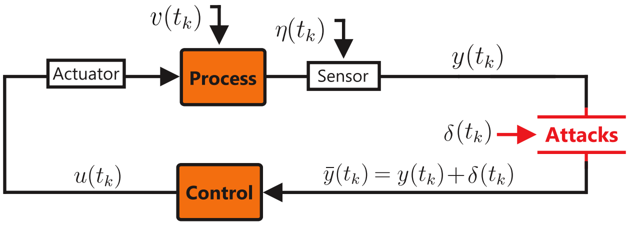

Figure 1. Cyber-physical system under sensor attacks.

$$\bar{y}(t_k) := y(t_k) + \delta(t_k) = Cx(t_k) + \eta(t_k) + \delta(t_k), \tag{2}$$

where $\delta(t_k) \in \mathbb{R}^m$ denotes *additive sensor attacks*. Denote $x_k := x(t_k)$, $u_k := u(t_k)$, $v_k := v(t_k)$, $\bar{y}_k := \bar{y}(t_k)$, $\eta_k := \eta(t_k)$, and $\delta_k := \delta(t_k)$. Using this new notation, the attacked system is written in the following compact form:

$$\begin{cases} x_{k+1} = Fx_k + Gu_k + v_k, \\ \bar{y}_k = Cx_k + \eta_k + \delta_k. \end{cases} \tag{3}$$

2.1 Observer

To estimate the state of the process, we use the observer

$$\hat{x}_{k+1} = F\hat{x}_k + Gu_k + L\big(\bar{y}_k - C\hat{x}_k\big), \tag{4}$$

with estimated state $\hat{x}_k \in \mathbb{R}^n$, $\hat{x}_1 = E[x(t_1)]$, where $E[\,\cdot\,]$ denotes expectation, and observer gain matrix $L \in \mathbb{R}^{n \times m}$. Define the estimation error $e_k := x_k - \hat{x}_k$. Given the system dynamics (3) and the observer (4), the estimation error is governed by the following difference equation

$$e_{k+1} = \big(F - LC\big)e_k + v_k - L\eta_k - L\delta_k. \tag{5}$$

The pair (F, C) is detectable; hence, the observer gain L can be selected such that $(F - LC)$ is Schur. Moreover, under detectability of (F, C), the covariance matrix $P_k := E[e_k e_k^T]$ converges to steady state (in the absence of attacks) in the sense that $\lim_{k\to\infty} P_k = P$ exists, see Aström and Wittenmark (1997). For $\delta_k = \mathbf{0}$ and given L (such that $(F - LC)$ is Schur), it can be verified that the asymptotic covariance matrix $P = \lim_{k\to\infty} P_k$ is given by the solution P of the following Lyapunov equation:

$$(F - LC)P(F - LC)^T - P + R_1 + LR_2L^T = \mathbf{0}, \tag{6}$$

where $\mathbf{0}$ denotes the zero matrix of appropriate dimensions. It is assumed that the system has reached steady state before an attack occurs.

2.2 Residuals and Hypothesis Testing

Attacks can be regarded as intensionally induced faults in the system. Then, it is reasonable to use existing fault detection techniques to identify sensor attacks. The main idea behind fault detection theory is the use of an estimator to forecast the evolution of the system. If the difference between what it is measured and the estimation is larger than expected, there may be a fault in or attack on the system. Although the notion of residuals and model-based detectors is now routine in the fault detection literature, the primary focus has been on detecting and isolating faults with *specific structures* (e.g., constant biases in sensor measurements or random faults in sensors and actuators following specific distributions). Now, in the context of an intelligent adversarial attacker, new challenges arise to understand the effect that an adaptive intruder can have on the system without being detected. In this paper, we use the observer introduced in the previous section as our estimator. Define the *residual sequence* $r_k, k \in \mathbb{N}$, as

$$r_k := \bar{y}_k - C\hat{x}_k = Ce_k + \eta_k + \delta_k, \tag{7}$$

which evolves according to the difference equation:

$$\begin{cases} e_{k+1} = \big(F - LC\big)e_k + v_k - L\eta_k - L\delta_k, \\ \quad r_k = Ce_k + \eta_k + \delta_k. \end{cases} \tag{8}$$

If there are no attacks, the mean of the residual is

$$E[r_{k+1}] = CE[e_{k+1}] + E[\eta_{k+1}] = \mathbf{0}_{m\times 1}, \tag{9}$$

and its asymptotic covariance matrix is given by

$$\Sigma := E[r_{k+1}r_{k+1}^T] = CPC^T + R_2. \tag{10}$$

It is assumed that $\Sigma \in \mathbb{R}^{m\times m}$ is positive definite. For this residual, we identify two hypotheses to be tested: $\mathcal{H}_0$ the *normal mode* (no attacks) and $\mathcal{H}_1$ the *faulty mode* (with faults/attacks). Then, we have

$$\mathcal{H}_0 : \begin{cases} E[r_k] = \mathbf{0}_{m\times 1}, \\ E[r_k r_k^T] = \Sigma, \end{cases} \quad \mathcal{H}_1 : \begin{cases} E[r_k] \neq \mathbf{0}_{m\times 1}, \text{or} \\ E[r_k r_k^T] \neq \Sigma, \end{cases}$$

where $\mathbf{0}_{m\times 1}$ denotes an m-dimensional vector composed of zeros only. In this manuscript, we use the chi-squared procedure for examining the residual and subsequently distinguishing between $\mathcal{H}_0$ and $\mathcal{H}_1$.

2.3 Distance Measure and Chi-squared Procedure

The input to any detection procedure is a *distance measure* $z_k \in \mathbb{R}$, $k \in \mathbb{N}$, i.e., a measure of how deviated the estimator is from the sensor measurements. We employ distance measures any time we test to distinguish between $\mathcal{H}_0$ and $\mathcal{H}_1$. The chi-squared procedure uses a quadratic form as distance measure to test for substantial variations in the covariance of the error between the measured output and the estimate. Consider the residual sequence r_k, (8), and its covariance matrix Σ, (10).

Chi-squared procedure:

$$\text{If} \quad z_k := r_k^T \Sigma^{-1} r_k > \alpha, \quad \tilde{k} = k. \tag{11}$$

Design parameter: threshold $\alpha \in \mathbb{R}_{>0}$.
Output: alarm time(s) $\tilde{k}$.

Thus, the procedure is designed so that alarms are triggered if z_k exceeds the threshold α. The normalization by Σ^{-1} makes setting the value of the threshold α system independent. This quadratic expression leads to a sum of the squares of m normally distributed random variables which implies that the distance measure z_k follows a chi-squared distribution with m degrees of freedom, see, e.g., Ross (2006) for details.

2.4 False Alarms

The occurrence of an alarm in the chi-squared when there are no attacks to the CPS is referred to as a false alarm. Operators need to tune this false alarm rate depending on the application. To do this, the threshold α must be selected to fulfill a *desired false alarm rate* $\mathcal{A}^*$. Let $\mathcal{A} \in [0,1]$ denote the *false alarm rate* of the chi-squared procedure defined as the expected proportion of observations which are false alarms, i.e., $\mathcal{A} := \text{pr}[z_k \geq \alpha]$, where $\text{pr}[\cdot]$ denotes probability, see van Dobben de Bruyn (1968) and Adams et al. (1992).

Proposition 1. [Murguia and Ruths (2016b)]. Assume that there are no attacks on the system and consider the chi-squared procedure (11) with residual $r_k \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and threshold $\alpha \in \mathbb{R}_{>0}$. Let $\alpha = \alpha^* := 2P^{-1}(\frac{m}{2}, 1-\mathcal{A}^*)$, where $P^{-1}(\cdot,\cdot)$ denotes the inverse regularized lower incomplete gamma function (see Ross (2006)), then $\mathcal{A} = \mathcal{A}^*$.

2.5 Output Feedback Controller

We consider observer-based output feedback controllers of the form:

$$u_k := K\hat{x}_k, \tag{12}$$

where $\hat{x}_k \in \mathbb{R}^n$ is the state of the observer (4) and $K \in \mathbb{R}^{l\times n}$ denotes the control matrix. The pair (F, G) is stabilizable; hence, the matrix K can be selected such that $(F+GK)$ is Schur. The closed-loop system (3),(4),(12) can be written in terms of the estimation error $e_k = x_k - \hat{x}_k$:

$$\begin{cases} x_{k+1} = (F + GK)x_k - GKe_k + v_k, \\ e_{k+1} = \big(F - LC\big)e_k + v_k - L\eta_k - L\delta_k. \end{cases} \tag{13}$$

Note that the attack sequence δ_k directly affects the estimation error dynamics, whereas the effect of the attack on the system dynamics is through the interconnection term GKe_k due to the control structure.

3. HIDDEN REACHABLE SETS

In this section, we provide tools for *quantifying* (for given L and K) and *minimizing* (by selecting L and K) the impact of the attack sequence δ_k on both the estimation error and the state of the system when the chi-squared procedure is used for attack detection. We are interested in attacks that can change the false alarm rate $\mathcal{A}$ of the detector by a small amount, say $\epsilon \in \mathbb{R}_{>0}$, i.e., $\bar{\mathcal{A}} < \mathcal{A} + \epsilon$, where $\bar{\mathcal{A}}$ denotes the alarm rate under the attacker's action. This class of attacks is what we refer to as *hidden attacks*. Here, we characterize *ellipsoidal bounds* on the set of states that hidden attacks can induce in the system. In particular, we provide tools based on Linear Matrix Inequalities (LMIs) for computing ellipsoidal bounds on the *reachable set* of the attack sequence δ_k given the system dynamics, the control strategy, the chi-squared procedure, and the bias ϵ on the false alarm rate $\mathcal{A}$.

Define the stacked noise vector $\omega_k := (v_k^T, \eta_k^T)^T$. Following the approach in Mo and Sinopoli (2016), we rewrite the estimation error as $e_k = e_{k,\omega_k} + e_{k,\delta_k}$, where e_{k,ω_k} denotes the part of e_k that is driven by noise and e_{k,δ_k} is the part driven by the attack sequence. Similarly, write the state of the system as $x_k = x_{k,\omega_k} + x_{k,\delta_k}$ and the residual as $r_k = r_{k,\omega_k} + r_{k,\delta_k}$. Using this new notation, because the system and the observer are linear, we can write the closed-loop dynamics (13) as follows:

$$\begin{cases} x_{k+1,\omega_k} &= (F+GK)x_{k,\omega_k} - GKe_{k,\omega_k} + v_k, \\ e_{k+1,\omega_k} &= (F-LC)e_{k,\omega_k} - L\eta_k + v_k, \\ r_{k,\omega_k} &= Ce_{k,\omega_k} + \eta_k, \end{cases} \tag{14}$$

$$\begin{cases} x_{k+1,\delta_k} &= (F+GK)x_{k,\delta_k} - GKe_{k,\delta_k}, \\ e_{k+1,\delta_k} &= (F-LC)e_{k,\delta_k} - L\delta_k, \\ r_{k,\delta_k} &= Ce_{k,\delta_k} + \delta_k, \end{cases} \tag{15}$$

and the distance measure as $z_k = ||\Sigma^{-\frac{1}{2}}(r_{k,\omega_k} + r_{k,\delta_k})||^2$, where $\Sigma^{-\frac{1}{2}}$ denotes the symmetric squared root matrix of Σ^{-1}. Note that, in the absence of attacks, r_{k,ω_k} and r_k have the same asymptotic distribution. Hence, the contribution of attacks to the alarm rate of the detector is solely determined by r_{k,δ_k} generated by (15). Moreover, using the triangle inequality, we can write the following

$$z_k = ||\Sigma^{-\frac{1}{2}}(r_{k,\omega_k} + r_{k,\delta_k})||^2 \leq (||\Sigma^{-\frac{1}{2}} r_{k,\omega_k}|| + ||\Sigma^{-\frac{1}{2}} r_{k,\delta_k}||)^2;$$

then, if the attack sequence is restricted to satisfy

$$||\Sigma^{-\frac{1}{2}} r_{k,\delta_k}||^2 = ||\Sigma^{-\frac{1}{2}}(Ce_{k,\delta_k} + \delta_k)||^2 \leq \kappa, \ \forall\, k \in \mathbb{N}, \tag{16}$$

for some $\kappa \in \mathbb{R}_{>0}$, it is intuitive to think that $\bar{\mathcal{A}} < \mathcal{A} + \epsilon$ for some $\epsilon \in \mathbb{R}_{>0}$, i.e., the alarm rate under the attacker's action, $\bar{\mathcal{A}}$, is biased from the false alarm rate, $\mathcal{A}$, by ϵ. This observation is formally stated in the following theorem, which is slightly modified from Mo and Sinopoli (2016).

Theorem 1. [Mo and Sinopoli (2016)]. Consider the chi-squared procedure (11) with threshold $\alpha \in \mathbb{R}_{>0}$. Let inequality (16) be satisfied for some $\kappa \in (0, \sqrt{\alpha})$; then

$$\bar{\mathcal{A}} \leq 1 - P\left(\tfrac{m}{2}, \tfrac{(\sqrt{\alpha}-\kappa)^2}{2}\right), \tag{17}$$

where $P(\cdot,\cdot)$ denotes the regularized lower incomplete gamma function (see Ross (2006)). Moreover

$$1 - \lim_{\kappa \to 0^+} P\left(\tfrac{m}{2}, \tfrac{(\sqrt{\alpha}-\kappa)^2}{2}\right) = \mathcal{A}. \tag{18}$$

Therefore, by selecting κ sufficiently small, the attacker can make $\bar{\mathcal{A}}$ arbitrarily close to $\mathcal{A}$. This complicates the operator's task to distinguish between the attacked system and the system without attacks. The set of feasible attack sequences that the opponent can launch while satisfying (17) can be written as the following constrained control problem on δ_k:

$$\left\{ \delta_k \in \mathbb{R}^m \;\middle|\; \begin{array}{l} x_{k+1,\delta_k} = (F+GK)x_{k,\delta_k} - GKe_{k,\delta_k}, \\ e_{k+1,\delta_k} = (F-LC)e_{k,\delta_k} - L\delta_k, \\ ||\Sigma^{-\frac{1}{2}}(Ce_{k,\delta_k} + \delta_k)||^2 \leq \kappa, \ \forall\, k \in \mathbb{N}, \end{array} \right\}. \tag{19}$$

We are interested in the state trajectories that the attacker can induce in the system restricted to satisfy (19). To this end, we introduce the notion of a *hidden reachable set*, $\mathcal{R}_\kappa$, defined as follows.

$$\mathcal{R}_\kappa := \left\{ x_{k,\delta_k}, e_{k,\delta_k} \in \mathbb{R}^n \;\middle|\; \begin{array}{l} x_{1,\delta_k} = e_{1,\delta_k} = \mathbf{0}_{n\times 1}, \\ \delta_k, x_{k,\delta_k}, e_{k,\delta_k} \text{ satisfy (19)}, \end{array} \right\}.$$

In general, it is analytically intractable to compute $\mathcal{R}_\kappa$ exactly. Instead, using LMIs, for some positive definite matrices $\mathcal{P}_e, \mathcal{P}_x \in \mathbb{R}^{n\times n}$, we derive *outer ellipsoidal bounds* of the form $\mathcal{E}_e = \{e_{k,\delta_k} \in \mathbb{R}^n | e_{k,\delta_k}^T \mathcal{P}_e e_{k,\delta_k} \leq 1\}$ and $\mathcal{E}_x = \{x_{k,\delta_k} \in \mathbb{R}^n | x_{k,\delta_k}^T \mathcal{P}_x x_{k,\delta_k} \leq 1\}$ containing $\mathcal{R}_\kappa$. Note that the e_{k,δ_k} dynamics in (19) does not depend on x_{k,δ_k}. Thus, we first compute $\mathcal{E}_e$ and then we analyze how it propagates to $\mathcal{E}_x$. The following result is used to compute these ellipsoids.

Lemma 1. [That et al. (2013)]. Let V_k be a positive definite function, $V_1 = 0$, and $\zeta_k^T \zeta_k \leq \kappa \in \mathbb{R}_{>0}$. If there exists a constant $a \in (0,1)$ such that

$$V_{k+1} - aV_k - \frac{1-a}{\kappa}\zeta_k^T \zeta_k \leq 0, \forall\, k \in \mathbb{N}, \tag{20}$$

then, $V_k \leq 1$.

Define $\zeta_k := \Sigma^{-\frac{1}{2}}(Ce_{k,\delta_k} + \delta_k)$, then, from (19), we can write the hidden reachable set of the estimation error, $\mathcal{R}_e$, as follows.

$$\mathcal{R}_e = \left\{ e_{k,\delta_k} \in \mathbb{R}^n \;\middle|\; \begin{array}{l} e_{k+1,\delta_k} = Fe_{k,\delta_k} - L\Sigma^{\frac{1}{2}}\zeta_k, \\ e_{1,\delta_k} = \mathbf{0}, \, \zeta_k^T \zeta_k \leq \kappa, \, \forall\, k \in \mathbb{N}, \end{array} \right\}. \tag{21}$$

Note that if for some $k = k^*$, $e_{k^*,\delta_k} \neq 0$ and $\rho[F] > 1$, where $\rho[\cdot]$ denotes spectral radius, then $||e_{k,\delta_k}||$ diverges to infinity as $k \to \infty$ for any non-stabilizing ζ_k. That is, $\mathcal{R}_e$ is unbounded if the system is open-loop unstable. If $\rho[F] \leq 1$, then $||e_{k,\delta_k}||$ may or may not diverge to infinity depending on algebraic and geometric multiplicities of the eigenvalues with unit modulus of F (a known fact from stability of LTI systems), see Aström and Wittenmark (1997) for details.

Theorem 2. For given F, observer gain L, and residual covariance matrix Σ, consider the set $\mathcal{R}_e$ in (21). If there exists a positive definite matrix $\mathcal{P}_e \in \mathbb{R}^{n\times n}$ and $a \in (0,1)$ satisfying the following matrix inequality:

$$\begin{bmatrix} a\mathcal{P}_e & F^T\mathcal{P}_e & \mathbf{0} & \mathbf{0} \\ \mathcal{P}_e F & \mathcal{P}_e & -\mathcal{P}_e L\Sigma^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & -\Sigma^{\frac{1}{2}}L^T\mathcal{P}_e & \frac{1-a}{\kappa}I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}; \tag{22}$$

then, $\mathcal{R}_e \subseteq \mathcal{E}_e$, i.e., the hidden reachable set is contained in the ellipsoid $\mathcal{E}_e = \{e_{k,\delta_k} \in \mathbb{R}^n | e_{k,\delta_k}^T \mathcal{P}_e e_{k,\delta_k} \leq 1\}$.

Proof: For a positive definite matrix $\mathcal{P}_e \in \mathbb{R}^{n\times n}$, consider the function $V_k := e_{k,\delta_k}^T \mathcal{P}_e e_{k,\delta_k}$, then, from (21), inequality (20) takes the form:

$$\begin{aligned} -\vartheta^T &\begin{bmatrix} a\mathcal{P}_e - F^T\mathcal{P}_e F & F^T\mathcal{P}_e L\Sigma^{\frac{1}{2}} \\ \Sigma^{\frac{1}{2}}L^T\mathcal{P}_e F & \frac{1-a}{\kappa}I - \Sigma^{\frac{1}{2}}L^T\mathcal{P}_e L\Sigma^{\frac{1}{2}} \end{bmatrix} \vartheta \\ &=: -\vartheta^T \mathcal{Q}_e \vartheta \leq 0, \end{aligned}$$

where $\vartheta := (e_{k,\delta_k}^T, \zeta_k^T)^T$. The above inequality is satisfied if and only if $\mathcal{Q}_e \geq \mathbf{0}$. Matrix $\mathcal{Q}_e$ can be written as the Schur complement of a higher dimensional matrix $\mathcal{Q}_e'$; hence, $\mathcal{Q}_e \geq \mathbf{0} \Leftrightarrow \mathcal{Q}_e' \geq \mathbf{0}$, i.e.,

$$\mathcal{Q}_e \geq \mathbf{0} \Leftrightarrow \mathcal{Q}_e' := \begin{bmatrix} a\mathcal{P}_e & \mathbf{0} & F^T\mathcal{P}_e & \mathbf{0} \\ \mathbf{0} & \frac{1-a}{\kappa}I & -\Sigma^{\frac{1}{2}}L^T\mathcal{P}_e & \mathbf{0} \\ \mathcal{P}_e F & -\mathcal{P}_e L\Sigma^{\frac{1}{2}} & \mathcal{P}_e & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}. \tag{23}$$

Finally, inequality (22) follows from (23) after the congruence transformation:

$$T := \begin{bmatrix} I & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I & \mathbf{0} \\ \mathbf{0} & I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix}.$$

The assertion follows now from Lemma 1. ■

The result in Theorem 2 provides a tool for computing ellipsoidal bounds on $\mathcal{R}_e$. To make the bounds most useful, we next construct ellipsoids with minimal volume, i.e., the tightest possible ellipsoid bounding $\mathcal{R}_e$. In this case, we minimize $\det \mathcal{P}_e^{-1}$ subject to (22) (because $\det \mathcal{P}_e^{-1}$ is proportional to the volume of $e_{k,\delta_k}^T \mathcal{P}_e e_{k,\delta_k} = 1$). This is formally stated in the following corollary of Theorem 2, see Boyd et al. (1994) for further details.

Corollary 1. For given matrices (F, L, Σ) and $a \in (0,1)$, the solution $\mathcal{P}_e$ of the following convex optimization:

$$\begin{cases} \min\limits_{\mathcal{P}_e} \quad -\log\det\mathcal{P}_e, \\ \text{s.t. } \mathcal{P}_e > 0 \text{ and (22)}, \end{cases} \tag{24}$$

minimizes the volume of the ellipsoid $\mathcal{E}_e$ bounding $\mathcal{R}_e$.

See Lofberg (2004) for an example of how to solve (24) using YALMIP.

As we now move toward designing L to minimize the ellipsoids, we note that as $||L|| \to 0$, the volume of the ellipsoid $\mathcal{E}_e$ goes to zero because the attack-dependent term in (21), $L\Sigma^{\frac{1}{2}}\zeta_k$, vanishes. To make this concrete, without any other considered criteria, the observer gain leading to the minimum volume ellipsoid is trivially given by $L = \mathbf{0}$. While this is effective at eliminating the impact of the attacker, it implies that we discard the observer altogether and, therefore, forfeit any ability to build a reliable estimate of the system. If we impose a performance criteria that the observer must satisfy in the attack-free case (e.g., convergence speed, noise-output gain, or minimum asymptotic variance), it has to be added into the minimization problem (24) so as to minimize the volume of $\mathcal{E}_e$ while still achieving the observer performance in the attack-free case. For completeness, in the following proposition, we provide an LMI criteria for ensuring that the H_∞ gain from the noise to the residual r_k in (8) is less than or equal to some $\gamma \in \mathbb{R}_{>0}$. Then, using this criteria and Theorem 2, we provide a synthesis tool for minimizing $\mathcal{E}_e$ while ensuring a desired H_∞ performance in the attack-free case.

Proposition 2. For given matrices (F, C, L), if there exists a positive definite matrix $\mathcal{P}_e \in \mathbb{R}^{n\times n}$ and $\gamma \in \mathbb{R}_{>0}$ satisfying the following matrix inequality:

$$\begin{bmatrix} \mathcal{P}_e & \mathbf{0} & \mathbf{0} & (F-LC)^T\mathcal{P}_e & C^T \\ \mathbf{0} & \gamma^2 I & \mathbf{0} & -L^T\mathcal{P}_e & I \\ \mathbf{0} & \mathbf{0} & \gamma^2 I & \mathcal{P}_e & \mathbf{0} \\ \mathcal{P}_e(F-LC) & -\mathcal{P}_e L & \mathcal{P}_e & \mathcal{P}_e & \mathbf{0} \\ C & I & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \tag{25}$$

then, the H_∞ gain from the noise $\omega_k = (\eta_k^T, v_k^T)^T$ to the residual $r_k = C_{e_k} + \eta_k$ of the estimation error dynamics (8) is less than or equal to γ.

The proof of Proposition 2 is omitted here due to the page limit. However, this is a standard result and details about the proof can be found in, e.g., Scherer and Weiland (2000) and references therein.

Remark 1. Note that the attack sequence δ_k enters the estimation error dynamics in the same manner as the sensor noise η_k (see (8)). It follows that, in this particular configuration, minimizing the influence of sensor noise (e.g., by using H_∞ techniques) would also reduce the effect of sensor attacks on the estimation error dynamics. This would tend to reduce the size of the hidden reachable sets but it would not necessarily lead to minimal ones. See Figure 5 in Section 4.

In the following corollary of Theorem 2 and Proposition 2, we formulate the optimization problem for designing the observer gain L such that the volume of the ellipsoid $\mathcal{E}_e$ is minimized and a desired H_∞ performance is achieved in the attack-free case.

Corollary 2. For given (F, C, Σ), $a \in (0,1)$, and $\gamma \in \mathbb{R}_{>0}$, if there exist matrices $\mathcal{P}_e \in \mathbb{R}^{n\times n}$ and $M \in \mathbb{R}^{n\times m}$ solution to the following convex optimization:

$$\begin{cases} \min\limits_{\mathcal{P}_e, M} \quad -\log\det\mathcal{P}_e, \\ \text{s.t. } \mathcal{P}_e > 0, \begin{bmatrix} a\mathcal{P}_e & F^T\mathcal{P}_e & \mathbf{0} & \mathbf{0} \\ \mathcal{P}_e F & \mathcal{P}_e & -M\Sigma^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & -\Sigma^{\frac{1}{2}}M^T & \frac{1-a}{\kappa}I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \text{ and} \\ \begin{bmatrix} \mathcal{P}_e & \mathbf{0} & \mathbf{0} & F^T\mathcal{P}_e - C^T M^T & C^T \\ \mathbf{0} & \gamma^2 I & \mathbf{0} & -M^T & I \\ \mathbf{0} & \mathbf{0} & \gamma^2 I & \mathcal{P}_e & \mathbf{0} \\ \mathcal{P}_e F - MC & -M & \mathcal{P}_e & \mathcal{P}_e & \mathbf{0} \\ C & I & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \end{cases} \tag{26}$$

then, the observer gain $L = \mathcal{P}_e^{-1}M$ minimizes the volume of the ellipsoid $\mathcal{E}_e$ bounding $\mathcal{R}_e$ and guarantees that the H_∞ gain from the noise ω_k to the residual r_k of (8) is less than or equal to γ in the attack-free case.

Proof: This follows from Theorem 2 and Proposition 2 and the linearizing change of variables $M = \mathcal{P}_e L$. ■

Next, once we have an ellipsoid $\mathcal{E}_e$ such that $\mathcal{R}_e \subseteq \mathcal{E}_e$, from (19), we can write the attacker's reachable states, $\mathcal{R}_x$, as:

$$\mathcal{R}_x = \left\{ x_{k,\delta_k} \in \mathbb{R}^n \;\middle|\; \begin{aligned} x_{k+1,\delta_k} &= (F+GK)x_{k,\delta_k} \\ &\quad -GK\bar{\mathcal{P}}_e^{-1}\xi_k, \\ x_{1,\delta_k} &= \mathbf{0},\, \xi_k^T\xi_k \leq 1,\, \forall\, k \in \mathbb{N}, \end{aligned} \right\}. \tag{27}$$

where $\xi_k := \bar{\mathcal{P}}_e e_{k,\delta_k}$ and the matrix $\bar{\mathcal{P}}_e \in \mathbb{R}^{n\times n}$ is such that $\mathcal{P}_e = \bar{\mathcal{P}}_e^T\bar{\mathcal{P}}_e$ (Cholesky factorization). Then, analogous to Theorem 2, we have the following result for computing ellipsoidal bounds on $\mathcal{R}_x$.

Theorem 3. For given matrices (F, G), controller gain K, and positive definite matrix $\mathcal{P}_e$, consider the set $\mathcal{R}_x$ in (27). If there exists a positive definite matrix $\mathcal{P}_x \in \mathbb{R}^{n\times n}$ and $b \in (0,1)$ satisfying the following matrix inequality:

$$\begin{bmatrix} b\mathcal{P}_x & (F+GK)^T\mathcal{P}_x & \mathbf{0} & \mathbf{0} \\ \mathcal{P}_x(F+GK) & \mathcal{P}_x & -\mathcal{P}_x GK\bar{\mathcal{P}}_e^{-1} & \mathbf{0} \\ \mathbf{0} & -(GK\bar{\mathcal{P}}_e^{-1})^T\mathcal{P}_x & (1-b)I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & I \end{bmatrix} \geq \mathbf{0}, \tag{28}$$

then, $\mathcal{R}_x \subseteq \mathcal{E}_x$, i.e., the hidden reachable set is contained in the ellipsoid $\mathcal{E}_x = \{x_{k,\delta_k} \in \mathbb{R}^n | x_{k,\delta_k}^T \mathcal{P}_x x_{k,\delta_k} \leq 1\}$.

The proof of Theorem 3 follows the same lines as the proof of Theorem 2 and it is omitted here. As before, if an ellipsoid with minimal volume is required, we minimize $\det\mathcal{P}_x^{-1}$ subject to (28). This is formally stated in the following corollary of Theorem 3.

Corollary 3. For given $(F, G, K, \mathcal{P}_e)$ and $b \in (0,1)$ the solution $\mathcal{P}_x$ of the following convex optimization:

$$\begin{cases} \min\limits_{\mathcal{P}_x} \quad -\log\det\mathcal{P}_x, \\ \text{s.t. } \mathcal{P}_x > 0 \text{ and (28)}, \end{cases} \tag{29}$$

minimizes the volume of the ellipsoid $\mathcal{E}_x$ bounding $\mathcal{R}_x$.

Remark 2. Similar to the case with the observer gain L and $\mathcal{E}_e$, note that as $||K|| \to 0$, the volume of $\mathcal{E}_x$ goes to

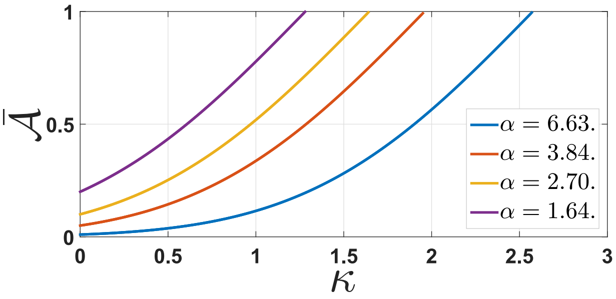

Figure 2. Alarm Rate $\bar{\mathcal{A}}$ under hidden attacks for different values of the chi-squared threshold α.

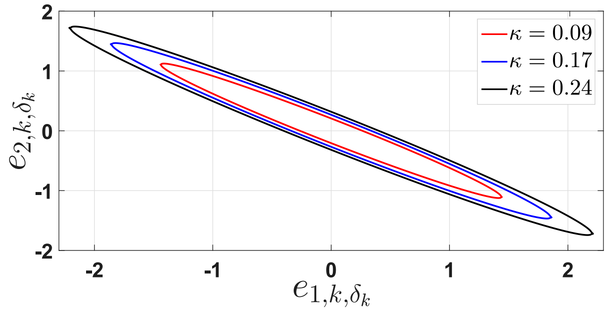

Figure 3. Ellipsoid $\mathcal{E}_e$ for different values of κ.

zero because the term $GK\bar{\mathcal{P}}_e^{-1}\zeta_k$ in (27) vanishes. This implies that without any other performance specification, the control gain leading to the minimum volume ellipsoid is trivially given by $K = \mathbf{0}$. However, as we have done in Corollary 2 for designing L subject to satisfying some H_∞ performance, we should impose a performance criteria to be satisfied by the controller in the attack-free case. This must be added to the minimization problem (29) so as to minimize the volume of $\mathcal{E}_x$ while guaranteeing the controller performance in the attack-free case. However, for the sake of briefly, we do not include a result considering this case.

4. SIMULATION EXPERIMENTS

Consider the closed-loop system (3),(4),(12) with matrices:

$$\begin{cases} F = \begin{pmatrix} 0.84 & 0.23 \\ -0.47 & 0.12 \end{pmatrix}, G = \begin{pmatrix} 0.07 \\ 0.23 \end{pmatrix}, C = (1 \ \ 0), \\ K = (-1.85 \ \ -0.96), L = \begin{pmatrix} 1.16 \\ -0.69 \end{pmatrix}, \\ R_1 = \begin{pmatrix} 0.45 & -0.11 \\ -0.11 & 0.20 \end{pmatrix}, R_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad R_2 = 1, \\ \Sigma = 3.26. \end{cases} \tag{30}$$

Using Proposition 2, the observer gain L is designed such that the H_∞ gain from the noise to the residual r_k of (8) is less than or equal to $\gamma = 1.86$ in the attack-free case. For $\alpha = \{6.63, 3.84, 2.70, 1.64\}$ and corresponding $\mathcal{A} = \{0.01, 0.05, 0.10, 0.20\}$, as a function of κ, Figure 2 depicts the upper bound on the alarm rate $\bar{\mathcal{A}}$, in (17), under hidden attacks. For a false alarm rate $\mathcal{A} = 0.10$ ($\alpha = 2.70$), we select $\kappa = \{0.09, 0.17, 0.24\}$ which leads to, correspondingly, $\mathcal{A} \leq \{0.12, 0.14, 0.16\}$, i.e., increments of 2% on $\mathcal{A}$. For these values of κ and $a = 0.65$ (this value of a leads to minimal volume ellipsoids), in Figure 3, we depict the ellipsoidal bounds $\mathcal{E}_e$ on the hidden reachable sets $\mathcal{R}_e$ obtained using Theorem 2 and Corollary 1. Next, Figure 4 shows the corresponding ellipsoidal bounds $\mathcal{E}_x$ on $\mathcal{R}_x$ obtained through Theorem 3 and Corollary 3. Finally, for $\kappa = 0.24$ ($\mathcal{A} \leq 0.16$), using Corollary 2, we redesign the observer gain L to minimize the volume of $\mathcal{E}_e$ while maintaining the H_∞ performance of $\gamma = 1.86$. The obtained optimal ellipsoidal bounds, $\mathcal{E}_e$ and $\mathcal{E}_x$, are depicted in Figure 5 and Figure 6 for the optimal observer gain $L = (0.1272, -0.0160)^T$. In light of Remark 1, for comparison, we compute the gain L leading to the minimal H_∞ gain, γ, from the noise to the residual r_k in (8). The obtained gain is $L = (0.4812, -0.2936)^T$ leading to $\gamma = 1.2518$. Figure 5 depicts the ellipsoidal bound obtained using this L. Note that even though the obtained ellipsoid is smaller than the one obtained with the original L, it is still bigger than the optimal one obtained using Corollary 2.

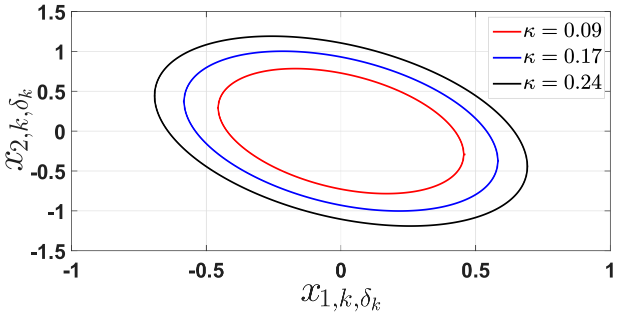

Figure 4. Ellipsoid $\mathcal{E}_x$ for different values of κ.

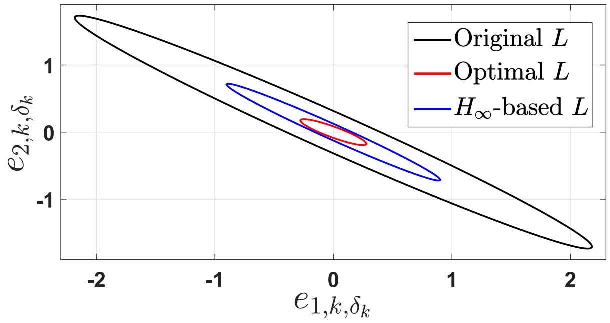

Figure 5. In red, improvement in the hidden reachable set ellipsoidal bound $\mathcal{E}_e$ through application of Corollary 2 to design the optimal observer gain. In blue, ellipsoidal bound obtained for the gain L minimizing the H_∞ gain from the noise to the residual in (8).

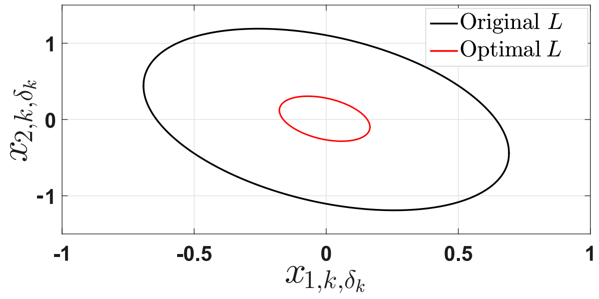

Figure 6. The improvement in the hidden reachable set ellipsoid bound $\mathcal{E}_x$ through application of Corollary 2 to design the optimal observer gain.

5. CONCLUSION

In this paper, for a class of discrete-time LTI systems subject to sensor/actuator noise, we have provided tools for *quantifying* and *minimizing* the negative impact of sensor attacks on the system performance given how the opponent accesses the dynamics (i.e., through the controller by tampering with sensor measurements). We have proposed to use the *reachable set* as a measure of the impact of an attack given a chosen detection method. For given system dynamics, control structure, and attack detection scheme, we have derived ellipsoidal bounds on these reachable sets using LMIs. Then, we have provided synthesis tools for minimizing these bounds (minimizing thus the reachable set) by properly redesigning controllers and detectors.

REFERENCES

Adams, B., Woodall, W., and Lowry, C. (1992). The use (and misuse) of false alarm probabilities in control chart design. *Frontiers in Statistical Quality Control 4*, 155–168.

Aström, K.J. and Wittenmark, B. (1997). *Computer-controlled Systems (3rd Ed.).* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Bai, C.Z., Pasqualetti, F., and Gupta, V. (2015). Security in stochastic control systems: Fundamental limitations and performance bounds. In *American Control Conference (ACC), 2015*, 195–200.

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. (1994). *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics.* SIAM, Philadelphia, PA.

Cárdenas, A., Amin, S., Lin, Z., Huang, Y., Huang, C., and Sastry, S. (2011). Attacks against process control systems: Risk assessment, detection, and response. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, 355–366.

Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A., and Sastry, S. (2009). Challenges for securing cyber physical systems. In *Workshop on Future Directions in Cyber-physical Systems Security.*

Chen, J. and Patton, R.J. (1999). *Robust Model-based Fault Diagnosis for Dynamic Systems.* Kluwer Academic Publishers, Norwell, MA, USA.

Giraldo, J., Cardenas, A., and Quijano, N. (2016). Integrity attacks on real-time pricing in smart grids: Impact and countermeasures. *IEEE Transactions on Smart Grid*, PP.

Gustafsson, F. (2000). *Adaptive Filtering and Change Detection.* John Wiley and Sons, LTD, West Sussex, Chichester, England.

Kwon, C., Liu, W., and Hwang, I. (2013). Security analysis for cyber-physical systems against stealthy deception attacks. In *American Control Conference (ACC), 2013*, 3344–3349.

Kyriakides, E. and Polycarpou, M.M. (eds.) (2015). *Intelligent Monitoring, Control, and Security of Critical Infrastructure Systems*, volume 565 of *Studies in Computational Intelligence.* Springer.

Lofberg, J. (2004). Yalmip : a toolbox for modeling and optimization in matlab. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, 284–289.

Luenberger, D. (1966). Observers for multivariable systems. *IEEE Transactions on Automatic Control*, 11, 190–197.

Miao, F., Zhu, Q., Pajic, M., and Pappas, G.J. (2014). Coding sensor outputs for injection attacks detection. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, 5776–5781.

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

Mo, Y. and Sinopoli, B. (2016). On the performance degradation of cyber-physical systems under stealthy integrity attacks. *IEEE Transactions on Automatic Control*, 61, 2618–2624.

Murguia, C. and Ruths, J. (2016a). Characterization of a cusum model-based sensor attack detector. In *proceedings of the 55th IEEE Conference on Decision and Control (CDC).*

Murguia, C. and Ruths, J. (2016b). Cusum and chi-squared attack detection of compromised sensors. In *proceedings of the IEEE Multi-Conference on Systems and Control (MSC).*

Nijmeijer, H. and Mareels, I.M.Y. (1997). An observer looks at synchronization. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44, 882–890.

Pasqualetti, F., Dörfler, F., and Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58, 2715–2729.

Ross, M. (2006). *Introduction to Probability Models, Ninth Edition.* Academic Press, Inc., Orlando, FL, USA.

Scherer, C. and Weiland, S. (2000). *Linear Matrix Inequalities in Control.* Springer-Verlag, The Netherlands.

That, N.D., Nam, P.T., and Ha, Q.P. (2013). Reachable set bounding for linear discrete-time systems with delays and bounded disturbances. *Journal of Optimization Theory and Applications*, 157, 96–107.

van Dobben de Bruyn, C. (1968). *Cumulative sum tests : theory and practice.* London : Griffin.