Coding Schemes for Securing Cyber-Physical Systems Against Stealthy Data Injection Attacks

Fei Miao, *Student Member, IEEE,* Quanyan Zhu, *Member, IEEE,*
Miroslav Pajic, *Member, IEEE,* and George J. Pappas, *Fellow, IEEE*.

***Abstract*— This paper considers a method of coding the sensor outputs in order to detect stealthy false data injection attacks. An intelligent attacker can design a sequence of data injection to sensors and actuators that pass the state estimator and statistical fault detector, based on knowledge of the system parameters. To stay undetected, the injected data should increase the state estimation errors while keep the estimation residues small. We employ a coding matrix to change the original sensor outputs to increase the estimation residues under intelligent data injection attacks. This is a low cost method compared with encryption schemes over all sensor measurements in communication networks. We show the conditions of a feasible coding matrix under the assumption that the attacker does not have knowledge of the exact coding matrix. An algorithm is developed to compute a feasible coding matrix, and, we show that in general, multiple feasible coding matrices exist. To defend against attackers who estimates the coding matrix via sensor and actuator measurements, time-varying coding matrices are designed according to the detection requirements. A heuristic algorithm to decide the time length of updating a coding matrix is then proposed.**

***Index Terms*—Coding, detection, feasible coding matrix, stealthy data injection attacks, time-varying coding, state estimator.**

I. Introduction

Cyber-physical systems (CPSs) integrate computation and communications to interact with physical processes. Many applications are considered as CPSs, including high confidence medical devices, energy conservation, environmental control, and safety critical infrastructures–such as water supply systems, electric power, and communication systems [2]. Therefore, security is a critical aspect of these systems, and CPSs involve additional challenges in control layer. The problem of secure control is defined, and reasons for mechanisms of information security, sensor network security alone are not sufficient for the security of CPSs are analyzed [3]. The key challenges of CPSs securities are summarized in [4].

Novel attack-detection algorithms in cyber security area can be designed, by understanding how attacks affect state estimation and control of the system. Two algorithms to maximize the utility of encrypted devices placed to increase system security are proposed to reduce the cost of communication cost in power grids [5]. Tools are developed to protect state-estimation components from stealthy attacks from an intelligent attacker with a partial model of the system [6].

Researchers have explored fault detection, isolation and reconfiguration (FDIR) methods to ensure systems' safety and robustness [7]. Although active techniques have been designed to tackle various types of attacks, fundamental limitations still exist [8]. With a limited number of sensor and actuator compromised by the attacker, i.e., some elements of the injection vector is restricted to be zero, resilient state estimators have been designed by previous work. Fawzi et al. propose estimation and control schemes of noise free linear systems [9]. Pajic et al. present a robust state estimation method in presence of attacks to no more than half of the sensors for systems with noise and modeling errors [10]. In contrast, we examine a different case where the attacker can inject an arbitrary vector to the communication between sensors and the estimator/detector/controller component, thus no element of the injection vector is constrained to be zero.

The monitoring system can detect malicious behaviors in general. Coding and decoding schemes to estimate the state of a scalar stable stochastic linear system with noisy measurements are designed in [11]. A distributed methodology for detecting and isolating multiple sensor faults in interconnected CPS is proposed in [12]. A class of false data injection attacks against state estimators in power grid is analyzed in [13]. Sequential detection techniques of sensor networks are discussed in [14]. Miao et al. design stochastic game approaches for replay attacks detections [15] and secure control of CPSs [16].

However, with knowledge of the system model, an intelligent cyber attacker is able to carefully design a data injection sequence, such that the state estimation error increases without triggering the alarm of the monitor [17], [18]. Manandhar et al. design the Euclidean detector to overcome the limitation of χ^2 detector for fault detection in smart grid [19]. However, the design of Euclidean detector is based on the voltage signal model of smart grid and whether it works for a general linear system model has not been shown yet. In this work, we consider the detection problem of false data injection attacks for a general linear system model. To address the computational overhead of encryptions on embedded architectures [20], we propose an

This material is based on research sponsored by DARPA under agreement number FA8750-12-2-0247. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. This work was also supported in part by NSF CNS-1505701, CNS-1505799 grants, and the Intel-NSF Partnership for Cyber-Physical Systems Security and Privacy. Part of the results in this work appeared at the 53rd Conference on Decision and Control, Los Angeles, CA, USA, December 2014 [1].

F. Miao and G. J. Pappas are with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA 19014. Q. Zhu is with the Department of Electrical and Computer Engineering, New York University, Brooklyn, NY, USA 11201. M. Pajic is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA 27708. Email: {miaofei, pappasg}@seas.upenn.edu,{quanyan.zhu}@nyu.edu, {miroslav.pajic}@duke.edu

alternative low cost method to code the sensor measurements for detection. With the coding scheme, no additional detector is required for the system to detect stealthy data injected by an attacker with the knowledge of system model. Compared with error-correcting coding schemes [21], the sensor outputs coding approaches proposed in this work aim to change the value transmitted over the communication channel instead of correcting errors on bit level. Moreover, the coding scheme proposed in this work does not require additional bits for each plaintext message of the sensor measurements, while an encryption method introduces communication overhead for each sensor message transmitted in the communication channel [22]. We assume that the coding matrix is distributed between sensors and the estimator/detector of the system correctly like an secret encryption key [23], and measurement of individual sensor is not corrupted before coded. With the coding matrix, the values sent over the communication channel are changed, without additional bits for encryption overhead [21], and the scheme is low-cost compared with the scheme of encrypting all sensor outputs.

The contributions of this work are summarized as follows:

1) The main contribution of this work is a low cost method of coding sensor outputs to detect stealthy false data injection attacks. We show that the system can detect the original stealthy sensor injections by coding the sensor outputs according to certain conditions.
2) We also design an algorithm to compute such coding matrices, and show that in general, multiple feasible coding matrices exist.
3) When the attacker can estimate the coding scheme according to several measurements of sensor and actuator values, we show that it is difficult to get the exact coding matrix in general. Moreover, in this case, the system can either change a new coding matrix or randomly use a set of coding matrices within a time length before the attacker has enough measurements for a good estimation. We design a heuristic algorithm to decide the time length of updating a coding matrix.

The paper is organized as follows. In Section II we describe the system and attack models. The conditions that a feasible coding matrix should satisfy are presented in Section III. An algorithm to find a feasible coding matrix based on rotation matrix is developed in Section IV. A time-varying coding scheme is designed in Section V. Section VI shows illustrative examples. Conclusions are given in Section VII.

II. System And Attack Model

We will introduce a discrete-time linear time-invariant (LTI) system model, a data injection attack model, and the attacked system model in this section. The system architecture is shown in Figure 1.

A. Linear system model

Assume that the CPS is composed of a discrete time LTI system with the following form:

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

where $x_k \in \mathbb{R}^n$ is the system state vector, $u_k \in \mathbb{R}^m$ is the control input, and $y_k \in \mathbb{R}^p$ is the sensor observations at time k. We do not have specific restrictions for the linear control input u_k here, since the choice of a linear controller does not affect the detection of false data injection, and we will explain the reason later. We assume that $w_k \sim N(0, Q)$ and $v_k \sim N(0, R)$ are identical independent (i.i.d.) Gaussian noises.

The optimal Kalman filter used to estimate state $\hat{x}_{k|k}$ is:

$$\begin{aligned}
&\hat{x}_{0|-1} = 0, \quad P_{0|-1} = \Theta, P_{k+1|k} = AP_kA^T + Q,\\
&K_{k+1} = P_{k+1|k}C^T(CP_{k+1|k}C^T + R)^{-1},\\
&P_{k+1} = (I - K_{k+1}C)P_{k+1|k},\\
&z_{k+1} = y_{k+1} - C(A\hat{x}_k + Bu_k),\\
&\hat{x}_{k+1|k} = A\hat{x}_k + Bu_k, \quad \hat{x}_{k+1} = \hat{x}_{k+1|k} + K_kz_{k+1}.
\end{aligned}$$

Under the assumption that (A, B) is stabilizable, (A, C) is detectable, we get a steady state Kalman filter, with the error covariance matrix P and Kalman gain matrix K:

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

Without attacks, the estimation residue z_k follows a Gaussian distribution $N \sim (0, CPC^T+R)$. Define the quantities g_k as $g_k = z_k^T P^{-1} z_k$, where P is the error covariance matrix of Kalman filter, then g_k satisfies a χ^2 distribution with p degrees of freedom. A χ^2 failure detector considers the standardized residue sequence $\eta_k = P^{-\frac{1}{2}} z_k$ for a monitoring system, and assumes that there exists a δ_η such that $\lim_{k\to\infty} \|E\eta_k\| \leq \delta_\eta$. We denote α as the threshold for detecting a fault, meaning that the alarm is triggered when $g_k > \alpha$.

B. False data injection attack model

The system model under sensor data injection attack is described as (2)

$$\begin{aligned}
x'_{k+1} &= Ax'_k + B(u'_k + u^a_k) + w_k,\\
y'_k &= Cx'_k + y^a_k + v_k,
\end{aligned} \tag{2}$$

where $y^a_k \in \mathbb{R}^p$, $u^a_k \in \mathbb{R}^m$ are arbitrary vectors injected to sensor outputs, actuator inputs by the attacker at time k respectively. When $u^a_k = 0$, only sensor values are changed by the attacker. Assume the adversary has knowledge of the system model described in Section II-A, and is able to inject data over communication network between sensors and the estimator/detector/controller.

Without attack, according to the system dynamics and the definition of Kalman filter, the estimation error is

$$e_k \triangleq x_k - \hat{x}_k,$$

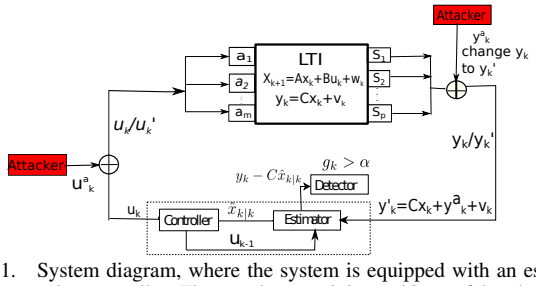


Figure 1. System diagram, where the system is equipped with an estimator, a detector and a controller. The attacker can inject arbitrary false data vector y^a_k to sensor outputs and u^a_k to actuator inputs.

$$e_{k+1} = (A - KCA)e_k - Kv_k + (I - KC)w_k.$$

When matrix $(A - KCA)$ is stable and $\mathbb{E}w_k = \mathbb{E}v_k = 0$, the expectation of estimation error converges to 0 with a static Kalman filter, i.e., $\lim_{k\to\infty} \mathbb{E}[e_k] \to 0$. Meanwhile, the residual z_k stays in the subspace that does not trigger the alarm with a high probability.

To illustrate how the sensor injection sequence y_k^a will affect the estimation and monitoring system, we examine how the estimation error and residue will change with y_k^a. Denote the estimation residuals of attacked system as

$$z'_k = y'_{k+1} - C(A\hat{x}'_k + Bu'_k),$$

where $\hat{x}'_k$ is the state estimation of the compromised system. Similarly, we define the estimation error under attack as

$$e'_k \triangleq x'_k - \hat{x}'_k,$$

The probability that the sensor injection sequence $y_k^a,\ k = 0, 1, \dots$ is detectable is given by

$$Pr(g'_k = (z'_k)^T P^{-1} z'_k > \alpha \text{ for any } k).$$

The difference between the normal and the compromised systems can be captured by:

$$\Delta e_k \triangleq e'_k - e_k, \quad \Delta z_k \triangleq z'_k - z_k. \tag{3}$$

The dynamics of the above difference vectors satisfy

$$\begin{aligned} \Delta e_{k+1} &= (A - KCA)\Delta e_k - Ky^a_{k+1} + (B - KCB)u^a_k, \\ \Delta z_{k+1} &= CA\Delta e_k + y^a_k + CBu^a_k, \end{aligned} \tag{4}$$

Hence the difference vectors between normal and compromised systems, $\Delta z_k(y^a, u^a), \Delta e_k(y^a, u^a)$, are functions of the injection sequences $y^a \triangleq (y^a_0, y^a_1, \dots)$, $u^a \triangleq (u^a_0, u^a_1, \dots)$. To simplify the notations, we concisely denote these vectors as $\Delta z_k, \Delta e_k$, respectively.

The objectives of the attacker include increasing the estimation error e'_k without triggering the alarm, and destabilizing the system with infinite state estimation error e'_k in the long run. Note that these types of attacks on control systems have been illustrated in the recent years. For instance, the estimated trajectories of Unmanned Ground Vehicle (UGV) [10] and Unmanned Aerial Vehicle (UAV) navigation systems [18] under stealthy data injection attacks (e.g., by GPS spoofing) deviate from the actual trajectories of the autonomous vehicles before being detected. Thus the attacker's objective is equivalent to increasing $\|\Delta e_k\|_2$ (the difference between estimation error of the normal and compromised systems) to infinity without increasing $\|\Delta z_k\|_2$ much as time goes by. Since computing the detecting statistic of compromised system g'_k is to integrate a Gaussian distribution on an ellipsoid, the stealthy requirement can be approximated by keeping $\|z'_k\|_2$ small. Residues of the normal system z_k are bounded, and the attacker should keep the change of residues bounded make the injection stealthy. It means the following inequality should hold

$$\|\Delta z_k\|_2 \leq M, \tag{5}$$

where M is a residue norm change threshold designed by the attacker. The compromised estimation residue should be close to that of the normal system, to deceive the monitoring system. [1] When y_k^a can be an arbitrary vector, a necessary and sufficient condition for a stealthy injection y_k^a that can increase $\|e'_k\|_2$, $\|x'_k\|_2$ to infinity while keep $\|z'_k\|_2$, $\|\Delta z_k\|_2$ bounded is derived in [18], [17]. The condition that $Cv \in span(I)$, i.e., there exists y^* satisfying $y^* = Cv$ is always satisfied by the attack model (2). Hence, we have the following proposition.

Proposition 1. *There exists a stealthy sequence $y_k^a, k = 0, 1, \dots,$ given the attacked system model* (2)*, if and only if matrix A has an unstable eigenvalue λ and the corresponding eigenvector v, such that $v \in span(Q_{oa})$, where Q_{oa} is the controllability matrix associated with the pair $(A-KCA, K)$.*

III. Coding Sensor Outputs For Detecting Stealth Sensor Data Injection

Existing statistical detectors, active monitor schemes (design some additive control input u_k^d) and fault detection filters have limitations, that even actuators are not compromised, they cannot detect stealthy sensor data injection attacks. It is necessary to design some inexpensive techniques to compensate for the vulnerability of the system under intelligent sensor data injection attacks. It has been shown that by only compromising sensors, attackers can induce infinite estimation error without being detected under monitoring systems like a χ^2 detector [18]. Therefore, we first discuss the case of stealthy sensor false data injection attacks in this section.

A. Limitations of existing approaches

The limitation of active monitor approach: Under the assumption that actuators work appropriately for the attacked system (2), the challenge here is whether adding u_k^d to the pre-designed linear control input u_k (such as optimal LQG control) can help to detect stealthy sensor data injections. For instance, consider a new control input

$$\tilde{u}_k = u_k + u_k^d, \tag{6}$$

where u_k^d is some random authentication signal or a constant value. It is worth noting that active monitor approaches do not help for detecting sensor data injection attacks described in model (2) .

Lemma 1. *There exists no active monitor in the form* (6) *that can increase the detection probability of a stealthy sensor data injection sequence, for the system* (1) *equipped with a Kalmen Filter and a χ^2 detector.* □

Proof: We denote the difference between estimation residual and estimation error of the normal and compromised system for the system with the controller (6) as $\Delta\tilde{z}_{k+1}$ and $\Delta\tilde{e}_{k+1}$, respectively. By the definition of $\Delta\tilde{z}_{k+1}$ and $\Delta\tilde{e}_{k+1}$ and a similarly calculation process to get (4), we have $\Delta\tilde{e}_k = \Delta e_k$, and $\Delta\tilde{z}_{k+1} = CA\Delta\tilde{e}_k + y^a_{k+1} + CBu^a_k$. Any additional control input u_k^d will be eliminated by the deduction of $\tilde{z}_{k+1}$ and $\tilde{z}'_{k+1}$ to get Δz_{k+1}. The active control input does

[1] The relation between the scale or norm of the injection sequence and the alarm trigger threshold α is shown in Theorem 1 in [18].

not increase the norm of $\Delta\tilde{z}_{k+1}$ compared with Δz_{k+1}, which means there exists no linear form of $\tilde{u}_k$ as described above that can increase $\|\Delta z_{k+1}\|_2$ under y_k^a for the system (2). ■

The limitations of active monitors for a unified LTI model are explained in Theorem 4.7 of [8] [2]. From this perspective, different linear controllers are equivalent under stealth sensor data injection attacks, and we do not restrict the controller model for designing our detection techniques.

The limitation of fault detection filter: Besides Kalman filter, observer-based fault detection filters for LTI systems with unknown error have been developed. The design requirements usually include robustness to unknown inputs and sensitivity to faults. Such filters generate a different residue from z_k of Kalman filter. Consider the following form of residual generator and residual evaluator (including a threshold and a decision logic unit, see [25] for details) [25]:

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Bu_k + H(y_k - \hat{y}_k), \\ \hat{y}_k &= C\hat{x}_k, \quad r_k = V(y_k - \hat{y}_k), \end{aligned} \tag{7}$$

where $\hat{x}_k \in \mathbb{R}^n$ and $\hat{y}_k \in \mathbb{R}^p$ represent the state and output estimation vectors, respectively, and r_k is the residual signal. This fault detector shares the same limitation with Kalman filter, i.e., the intelligent sensor data injection attack is stealth for the filter described as (7), since the residue is still observer based difference between y_k and $\hat{y}_k$.

B. Coding sensor outputs to detect stealth data injection

Since existing monitoring systems cannot detect intelligent false data injection attacks, and encryption method has a constraint of significant computation overhead, we propose a design of *coding the sensor outputs* to detect stealth sensor data injection attacks. An intelligent attacker designs the sequence y_k^a carefully to keep the change of residue $\|\Delta z_k\|_2 \leq M$, where M is a constant. Thus, the objective of a detecting approach is equivalent to increasing $\|\Delta z_k\|_2$ as fast as possible under a stealthy data injection sequence, and $\|\Delta z_k\|_2$ should increase to infinity as time goes to infinity.

The necessary and sufficient conditions for stealth false sensor data injection in Corollary 1 assume that the attacker knows (A, B, C, K). Parameters A and B are related to physical dynamics that may not be altered, while C is related to the sensor measurements, corresponding specific physical states. Without changing the physical setup, we still can manipulate the sensor outputs. To violate the attacker's design, we consider the method of transforming sensor outputs as shown in Figure 2–instead of sending the output vector $\mathbf{y}_k = Cx_k + v_k$ to the estimator/controller/detector, sensors transmit the value

$$Y_k = \Sigma(Cx_k + v_k), C \in \mathbb{R}^{p\times n}, \tag{8}$$

where $\Sigma \in \mathbb{R}^{p\times p}$ is an invertible matrix. We assume that the measurement of individual sensor is not corrupted yet before coding, and injection sequence appears in the communication between sensors to the estimator/controller/detector. One can think of Σ as an inexpensive code, and compare Σ with an encryption key. By encrypting only the coding matrix channel once, the coding approach saves encryption cost compared with encrypting all sensor outputs for every time k.

We assume that the attacker does not know the matrix Σ at least before estimating the matrix based on knowledge of matrices (A, B, C, K) and sensor and actuator values, since the coding matrix Σ is not fixed by the physical model of the system and can be time-varying, calculated in polynomial time when a new coding matrix is needed (an algorithm will be proposed in the next section). We assume that the attacker cannot access the coding matrix directly if he/she only applies eavesdropping techniques to the unencrypted communication channel and the process of distributing Σ is protected. We will propose a time-varying coding scheme later in this work. When the attacker has designed a sequence of stealthy attack signal y_k^a for the original system without the knowledge of the coding matrix Σ, the false sensor value after coding changes to:

$$Y_k' = \Sigma C x_k + y_k^a + \Sigma v_k. \tag{9}$$

Since Σ is an invertible matrix, when the state estimator receives Y_k or Y_k', the encoded packet is decoded as

$$\tilde{y}_k' = \Sigma^{-1}Y_k' = y_k + \Sigma^{-1}y_k^a, \tag{10}$$

And we still use the same Kalman filter and χ^2 detector on the decoded sensor outputs. Similar as the definitions of Δe_k and Δz_k (4) for sensor outputs before coding, we define $\Delta e_k'$ and $\Delta z_k'$ as the change of state estimation and residue for coded sensor outputs without attack (8) and under attack (9), respectively. With Σ, a stealth data injection designed for (1) (with parameters (A, B, C, K)), $\|\Delta z_k'\|_2$ increases to infinity as $k \to \infty$ under certain conditions. In the following theorem, we show the sufficient conditions that Σ should satisfy for any stealth sequence of $y_k^a, k = 0, 1, \ldots$ that satisfies Theorem 1.

Theorem 1. *Given an attacked system model* (2)*, assume that* (A, C) *is detectable,* $u_k^a = 0$*, and the attacker designs a sequence of sensor data injection* $y_k^a, k = 0, 1, \ldots$*, based on one unstable eigenvector* $v \in span(Q_{oa})$*, where* Q_{oa} *is the controllability matrix associated with the pair* $(A - KCA, K)$*. If there exists an invertible matrix* Σ*, and the direction of* ΣCv *is not the same with that of* Cv*, i.e.,*

$$\frac{(Cv)'\Sigma Cv}{\|\Sigma Cv\|_2\|Cv\|_2} \neq 1, \tag{11}$$

then after injecting y_k^a *the estimation residue change satisfies* $\lim_{k\to\infty}\|\Delta z_k'\|_2 \to \infty$*, by coding sensor outputs* (8) *with* Σ*.* □

[2] A different case when adding exogenous Gaussian distribution control input can detect replay attacks is discussed in [24].

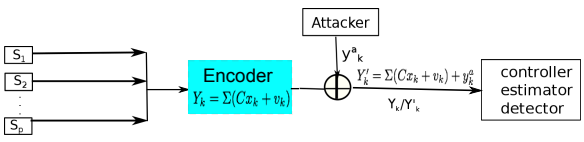


Figure 2. System diagram when coding sensor outputs with a matrix Σ that satisfies the conditions of Theorem 1. The attacker can inject arbitrary false data vector y_k^a to sensor outputs.

Proof: Given a system under data injection attacks as (2), we assume that the system has one unstable eigenvector v with corresponding eigenvalue λ. According to the definition in equation (3), the dynamics of $\Delta e_k, \Delta z_k$ satisfy (4) with $u_k^a = 0$. For coded sensor outputs (9), after decoding

$$\begin{aligned} \Delta e'_{k+1} &= (A - KCA)\Delta e'_k - K\Sigma^{-1} y^a_{k+1}, \\ \Delta z'_{k+1} &= CA\Delta e'_k + \Sigma^{-1} y^a_{k+1}, \end{aligned} \tag{12}$$

The proof of *Theorem* 1 in [18] shows that under a stealthy sensor data injection sequence, the only component of Δe_k that goes to infinity eventually depends on the unstable eigenvector, denoted as $c_k v, \lim_{k\to\infty} c_k = \infty$, and Δe_k can be decomposed as $\Delta e_k = c_k v + \epsilon_{1k}, \|\epsilon_{1k}\|_2 \leq M_1$.

To keep Δz_k bounded as $k \to \infty$, any stealthy injection sequence y_k^a must satisfy

$$y^a_{k+1} = -c_k \lambda C v + \epsilon_{2k}, \|\epsilon_{2k}\|_2 \leq M_2, k = 0, 1, 2, \ldots, \tag{13}$$

where M_2 is a constant such that $\|\Delta z_k\|_2 \leq M$ for all k.

We assume that the attacker does not know Σ, and designs an injection sequence for the original system (1) as described in (13). Similarly as Δe_k, the only component of $\Delta e'_k$ that can goes to infinity is $c_k v$, since matrix A is not changed by the coding matrix Σ. However, with any y_k^a in (13), $\Delta z'_k$ can be decomposed as

$$\Delta z'_k = c_k \lambda (Cv - \Sigma^{-1} Cv) + \epsilon_{3k}, k = 0, 1, 2, \ldots, \tag{14}$$

where ϵ_{3k} is a bounded vector components of $\Delta z'_k$. When Σ satisfies equation (11), $\Sigma Cv - Cv \neq 0$. With $c_k \to \infty$, $\|\Delta z'_k\| \to \infty$ as $k \to \infty$. ∎

We call a matrix Σ that satisfies the conditions of Theorem 1 a feasible coding matrix. Theorem 1 shows that even the attacker knows system parameters (A, B, C, K), without changing the physical structure or altering A, B, we can utilize the sensor data to get different residues for detecting. Leveraging sensor outputs is the key reason to detect a stealth sensor data injection. It is worth noting that here we do not constrain specific structure of the matrix Σ besides conditions in Theorem 1. For an LTI system, ΣC is simply a linear transform of the original sensor measurement. When A has several unstable eigenvectors satisfying Corollary 1, the following lemma extends the result of Theorem 1.

Lemma 2. *Given an attacked system* (2) *with* (A, C) *detectable and a set of unstable eigenvectors* $v_1, \ldots, v_u \in span(Q_{oa})$*, where* Q_{oa} *is the controllability matrix associated with the pair* $(A - KCA, K)$*, if* Σ *is an invertible matrix, and*

$$\frac{(C\tilde{v})' \Sigma C \tilde{v}}{\|\Sigma C \tilde{v}\|_2 \|C\tilde{v}\|_2} \neq 1, \tag{15}$$

for any linear combinations of $v_1, \ldots, v_u$ *–* $\tilde{v}$*, then* Σ *is a feasible coding matrix to increase* $\|\Delta z'_k\|_2$ *for any stealth data injection to attacked system* (2). □

Proof: When matrix A has a set of unstable eigenvectors $v_1, \ldots, v_u$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_u$, similar as the proof of Theorem 1, a stealthy injection sequence takes the form

$$y^a_k = \sum_{i=1}^{u} c_{ik} \lambda_i C v_i + \epsilon_{2k}, \|\epsilon_{2k}\|_2 \leq M_2, k = 0, 1, 2, \ldots,$$

and the change of residual is defined as

$$\begin{aligned} \Delta z'_k &= \sum_{i=1}^{u} c_{ik} \lambda_i (C v_i - \Sigma^{-1} C v_i) + \epsilon_{3k}, \\ &= C(\sum_{i=1}^{u} c_{ik} \lambda_i v_i) - \Sigma^{-1} C (\sum_{i=1}^{u} c_{ik} \lambda_i v_i), k = 0, 1, 2, \ldots. \end{aligned}$$

Hence, we consider $\tilde{v} = \sum_{i=1}^{u} c_{ik} \lambda_i v_i$ as a linear combination of all the unstable eigenvectors, the conclusion holds with the coding matrix Σ satisfying all the constraints. ∎

Remark 1. *When the attacker is able to learn* Σ *by analyzing sensor outputs and actuator inputs, the system can send a new* Σ *before the attacker figures out the current applied coding matrix. The process of learning* Σ *from the perspective of an attacker will be discussed in Section V.* □

C. When sensor and actuator packets are both injected

We will derive the condition for a feasible coding matrix when the attacker can mount deception attacks to both sensor packets and actuator packets.

Theorem 2. *Given an attacked system model* (2)*, assume that the attacker designs a sequence of stealthy sensor and actuator data injection* $(y_k^a, u_k^a), k = 0, 1, \ldots,$ *that* u_k^a *is bounded and drives the estimation error to infinity* $\lim_{k\to\infty} \|\Delta e_k\|_2 \to \infty$*. If there exists an invertible matrix* Σ *such that* $y_k^a - \Sigma^{-1} y_k^a \neq 0$ *for any* y_k^a*, then after injecting* (y_k^a, u_k^a) *the estimation residue change satisfies* $\lim_{k\to\infty} \|\Delta z'_k\|_2 \to \infty$*, by coding sensor outputs* (8) *with* Σ*.* □

Proof: The dynamics of change of estimation error, residuals between the normal and compromised system is described as (4), where y_k^a, u_k^a is the injected sequence to sensor and actuator packets, respectively. Since $\|\Delta z_{k+1}\|_2 \leq M$ for all $k = 0, 1, \ldots,$ any pair of (y^a_{k+1}, u_k^a) must satisfy

$$y^a_{k+1} = -CA\Delta e_k - CBu_k^a + \epsilon_k, \ \|\epsilon_k\|_2 \leq M. \tag{16}$$

For bounded u_k^a, the injection sequence satisfies that $\lim_{k\to\infty} \|y_k^a\|_2 \to \infty$ to make sure $\lim_{k\to\infty} \|\Delta e_k\|_2 \to \infty$. When coded sensor values are injected as (9), and the estimator decodes the value as

$$\tilde{y}'_k = \Sigma^{-1} Y'_k = Cx_k + v_k + \Sigma^{-1} y_k^a,$$

the coded system with the original design of Kalman filter is equivalent to be injected by a sequence of pair $(\Sigma^{-1} y^a_{k+1}, u_k^a)$. It is worth noting that the actuator data is not coded, and u_k^a keeps the same for both the original and coded system. The dynamics of the change of estimation error, residuals between the normal and compromised coded system are as following

$$\begin{aligned} \Delta e'_{k+1} &= (A - KCA)\Delta e'_k - K\Sigma^{-1} y^a_{k+1} + (B - KCB)u_k^a, \\ \Delta z'_{k+1} &= CA\Delta e'_k + \Sigma^{-1} y^a_{k+1} + CBu_k^a. \end{aligned}$$

Without loss of generality, we assume that $\Delta e_0 = 0$, then

$$\Delta e_k = \sum_{j=1}^{k} (A - KCA)^{k-j} (-K y_j^a + (B - KCB) u^a_{j-1}),$$

$$\Delta e'_k = \sum_{j=1}^{k} (A - KCA)^{k-j}(-K\Sigma^{-1}y^a_j + (B - KCB)u^a_{j-1}).$$

Plug in the expression of $\Delta e'_k$ in the equation of $\Delta z'_{k+1}$, with $CBu^a_k = -y^a_{k+1} - CA\Delta e_k + \epsilon_k$, we have

$$\begin{aligned}\Delta z'_{k+1} =& CA\sum_{j=1}^{k}(A - KCA)^{k-j}(-K\Sigma^{-1}y^a_j \\ &+ (B - KCB)u^a_{j-1}) + \Sigma^{-1}y^a_{k+1} + CBu^a_k \\ =& CA\sum_{j=1}^{k}(A - KCA)^{k-j}K(I - \Sigma^{-1})y^a_j \\ &+ (\Sigma^{-1} - I)y^a_{k+1} + \epsilon_k.\end{aligned} \tag{17}$$

Hence, for $\Sigma \neq I$, $\lim_{k\to\infty} \|y^a_{k+1}\|_2 \to \infty$, we have $\lim_{k\to\infty} \|\Delta z'_k\|_2 \to \infty$ for $\Delta z'_k$ defined in (17). ■

IV. Algorithm to Compute A Coding Matrix

In this section we propose an algorithm to compute a set of feasible coding matrices for the case there exists a sequence of sensor data injections to cause unbounded state estimation error, i.e., the system has unstable eigenvectors of A.

The coded sensor values should increase the difference between estimation residue of the normal and attacked system – $\|\Delta z'_k\|_2$ as $k \to \infty$, which is equivalent to keep $\|Cv - \Sigma^{-1}Cv\|_2$ or $\|C\tilde{v} - \Sigma^{-1}C\tilde{v}\|_2$ for multiple unstable eigenvectors nonzero, by the proof of Theorem 1 and Lemma 2. The system satisfies that (A, C) is detectable, then with an invertible coding matrix Σ and the decoded sensor value $\tilde{y}'_k$ defined in (10), $\tilde{y}'_k = y_k$ when $y^a_k = 0$. Hence, the state estimator still converges to the true state without attacks and the coding scheme does not sacrifice the performance of state estimator.

For multiple unstable eigenvectors, when we do not know the exact linear combination result of $\tilde{v}$ applied by the attacker to design the injection sequence, we can not guarantee that Σ works for the exact injected sequence y^a_k by finding a feasible coding matrix with respect to a specific vector v. According to Theorem 1 and Lemma 2, the coding matrix should work for any possible injection sequence y^a_k designed based on unstable eigenvectors of the system matrix A. Hence, we consider to find a coding matrix based on the concept of a rotation matrix without specific knowledge about the value of injected data to sensors.

Definition 1. *A Givens rotation is a $n\times n$ rotation matrix, with 1's on the diagonal, 0's elsewhere, except the intersections of the ith and jth rows and columns corresponding to a rotation in the (i, j) plane in n dimensions. It takes the following form*

$$G(i,\ j,\ \theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}, \tag{18}$$

where $c = \cos\theta,\ s = \sin\theta$. □

The product $G(i, j, \theta)x$ represents a counterclockwise rotation of the vector $x \in \mathbb{R}^p$ in the i, j plane of θ radians. Hence, only the i-th and j-th elements of x will be changed. Given system model (1), there are multiple ways to choose a rotation matrix as a coding matrix in general. If a rotation matrix can guarantee that the direction of any possible stealthy injection is changed, it must rotate all nonzero elements in the vector space

$$span(Cv_1, \ldots, Cv_u) \tag{19}$$

The following algorithm provides a design process of a rotation matrix given system matrix A.

Algorithm 1 : Compute a feasible coding matrix Σ

Input: System model parameters A, C, unstable eigenvalues and eigenvectors $\lambda_i, v_i,\ i = 1, \ldots, u$ of A.
Initialization: Calculate vectors $Cv_i \in \mathbb{R}^p$ for all unstable eigenvectors $i = 1, \ldots, u$. Construct the standard basis $(e_{p_1},\ e_{p_2}, \ldots, e_{p_l})$, $e_{p_j} \in \mathbb{R}^p$ for the vector space defined as (19), where $1 \leq p_1 < p_2 \cdots < p_l \leq p$, and e_{p_j} is a vector with the p_j-th element as 1 and all the other elements as 0. Define rotation step as $r = 1$, uncovered unstable dimension set as $S = \{p_1, p_2, \ldots, p_l\}$.
Iteration: When $S \neq \phi$
If more than two elements are left in the set S: randomly picking up a rotation radian $\theta \in (0, \frac{\pi}{2}]$, rotation dimension $p_i, p_j \in S$, let $S = S \setminus \{p_i, p_j\}$;
Else: randomly picking a rotation radian $\theta \in (0, \frac{\pi}{2}]$ with uniform distribution, rotation dimension $p_i \in S$, $p_j \in \{1, \ldots, p\}$ and $p_j \neq p_i$, let $S = S \setminus \{p_i\}$.
Get the rotation matrix $G_r = G(p_i, p_j, \theta)$ as defined in (18). Let $r = r + 1$.
Return: A feasible transform matrix $\Sigma = G_1 G_2 \ldots G_r$.

The existence condition of a feasible coding matrix designed as a rotation matrix is then explained in the following lemma.

Lemma 3. *When the dimension of matrix C of the system* (1) *satisfies that $p \geq 2$, there always exists a feasible givens rotation matrix Σ that satisfies the condition of Theorem 1 or Lemma 2 for the system.* □

Proof: According to the definition of a Givens matrix (18) and the process of calculating a feasible rotation matrix, when $p \geq 2$, we apply Algorithm 1. Since every rotation has an angle $\theta \in (0,\ \frac{\pi}{2}]$ and there are no two rotations in the same plane, vector ΣCv is not in the same direction with Cv. Hence, Algorithm 1 provides a feasible rotation matrix. ■

The coding scheme proposed in this work is a low cost approach from computation perspective. Specifically, the proposed coding scheme requires only $O(n^3+p^3)$ multiplications and additions, where n and p denote the number of plant states and sensors respectively. As we clarify now in the new version of the manuscript (in Section IV), this is significantly lower than the computation cost for even basic encryption and coding schemes that involve computation of highly complex non-linear primitives [26], [20], [21].

The coding scheme proposed in this work is also a low cost approach from communication perspective. The coding scheme proposed in this work does not require additional bits for each plaintext message of the sensor measurements, while an encryption method introduces communication overhead for each sensor message transmitted in the communication channel [22]. The sensor outputs coding approaches proposed in this work aim to change the value transmitted over the communication channel instead of correcting errors on bit level compared with error-correcting additional coding bits [21]. Hence, the communication overhead of the proposed scheme in this work is relatively low.

Remark 2. *The rotation matrix Σ calculated by Algorithm 1 is a sparse matrix in general, since a rotation matrix has many 0 elements, and Algorithm 1 is a polynomial heuristic algorithm. This means the coding process is computationally efficient.* □

For systems with structural constraints, two potential schemes can be considered. One is that the structure of Σ is also limited and we design a coding matrix Σ with an additional constraint that some components Σ must be 0 because of the sparsity of the sensors the system equipped with. Another scheme is distributed coding that multiple coding matrices are applied for the whole system. This is a revenue for future work.

V. Time-Varying Coding Scheme When the Attacker Estimates the Coding Matrix

The coding scheme in this work is effective for the cases that sensor values are not manipulated by the attacker before they are coded by matrix Σ. We also assume that the attacker does not know when the system starts to apply Σ for transforming sensor output values, and aims to inject a stealthy sequence y_k^a to the sensor communication channel with respect to the original system. If the attacker is powerful enough to update the system model and acquire the knowledge of the coding design after some time steps, the system should constantly apply a time-varying coding scheme, and the time length for updating the coding matrix depends on the learning ability of the attacker and detecting requirements of the system.

Each time the system updates the coding matrix, it will cost the attacker some time to figure out the transformed sensor outputs values. Since it is sufficiently fast to compute a feasible transform based on the algorithm, the system can even generate new coding matrices during the running process. Before the attacker learns Σ or the coded observer parameter ΣC, the false data injection sequence is not stealthy for the coded system. We assume that the attacker cannot directly acquire the coding matrix during its communication process, similar as the secrecy requirement of a key for encryption sensor nodes [22], [23]. We assume that the sensors and controller are synchronized, which is a standard assumption in safety-critical control systems. Thus, with the same notion of time, both sensors and the controller can use the same random generator to (re)generate the coding matrix or exploit some of the existing schemes for secret key distribution. In addition, they will be able to synchronously switch from using one matrix to the newly created/obtained ones. Various protocols of key distributions have been proposed according to the properties of the systems [22], [27].

A. The time length an attacker needs to learn Σ

To learn the matrix Σ that distributed secured between sensors and the controller/estimator/detector, we assume that the attacker is able to eavesdrop the sensor outputs and actuator inputs via the communication channel for estimating Σ, instead of directly capturing the matrix Σ. Since y_k^a is designed by the attacker, the sensor information received by the attacker is then the true sensor measurements under the coding scheme $Y_k = \Sigma y_k$. System dynamics from the perspective of an attacker are

$$\begin{aligned} x_k =& A^k x_0 + \sum_{j=0}^{k-1} A^{k-j-1}(Bu_j + w_j), \\ Y_k =& \tilde{C}A^k x_0 + \sum_{j=0}^{k-1} \tilde{C}A^{k-j-1}(Bu_j + w_j) + y_k^a + v_k, \end{aligned} \tag{20}$$

where $\tilde{C} = \Sigma C$. When the attacker does not have any knowledge about the structure of the coding matrix $\Sigma \in \mathbb{R}^{p\times p}$, there are p^2 variables for estimating Σ. Meanwhile, in general initial state x_0 can only be acquired via estimation, and there are n variables additionally in (20). Without loss of generality, we initialize $k = 0$ as the time that attacker starts to observe the system's sensor outputs and actuator inputs to update the knowledge of the system coding scheme. It is worth noting that for designing a sequence of stealthy injection data, the attacker needs to know the model of the system, including the estimator and statistics detector, while the values of sensor outputs or actuator inputs are not necessary for the attacker. When the attacker starts to record sensor and actuator communicational packets at an arbitrary time k, the corresponding system state x_0 can not be directly retrieved by the attacker. Hence, $\Sigma \in \mathbb{R}^{p\times p}$ and $x_0 \in \mathbb{R}^n$ are variables to be estimated.

We examine a simpler case to estimate the the coding matrix first—how many steps of sensor values the attacker need to measure for the following noise-free LTI system

$$x_{k+1} = Ax_k + Bu_k, \quad \bar{y}_k = Cx_k. \tag{21}$$

The sensor outputs coded by Σ at time k are

$$\bar{Y}_k = \tilde{C}A^k x_0 + \sum_{j=0}^{k-1} \tilde{C}A^{k-j-1}Bu_j, \quad \bar{Y}_k \in \mathbb{R}^p. \tag{22}$$

We define the attacker's observation $Y_{\Sigma,N}$ during time $k = 0, 1, \ldots, N$ when the system applies Σ, and the corresponding noise-free measurements $\bar{Y}_{\Sigma,N}$ as

$$Y_{\Sigma,N} = \left[Y_0|Y_1|\cdots|Y_N\right], \quad \bar{Y}_{\Sigma,N} = \left[\bar{Y}_0|\bar{Y}_1|\ldots|\bar{Y}_N\right].$$

The observed sensor values from the perspective of the attacker are bilinear equations with respect to Σ and x_0. Consider the noise-free dynamics of sensor measurements as the following

$$\bar{Y}_{\Sigma,N} = \Sigma C \begin{bmatrix} x_0 & \cdots & A^N x_0 + \sum_{j=0}^{N-1} A^{N-j-1}Bu_j \end{bmatrix}, \tag{23}$$

where $[\bar{Y}_1, \dots, \bar{Y}_N]$, $[u_0, \dots, u_N]$ is eavesdropped by the attacker and (A, B, C) is within the knowledge space of the attacker. To write the above equation as a standard form of bilinear equations regarding to vectors, we denote the coding matrix Σ as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{p1} & \Sigma_{p2} & \cdots & \Sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_p \end{bmatrix},$$

where $\sigma_i \in \mathbb{R}^{1\times p},\ i \in \{1, \dots, p\}$ is the i-th row of matrix Σ. We also vectorize $\bar{Y}_{\Sigma,N} \in \mathbb{R}^{p\times(N+1)}$ $(Y_{\Sigma,N}) \in \mathbb{R}^{p\times(N+1)}$ as $\bar{d} \in \mathbb{R}^{p(N+1)}$ $(d \in \mathbb{R}^{p(N+1)})$

$$\text{vec}(\bar{Y}_{\Sigma,N}) = \begin{bmatrix} [\bar{Y}_0]_1 \\ \vdots \\ [\bar{Y}_0]_p \\ \vdots \\ [\bar{Y}_N]_p \end{bmatrix} = \begin{bmatrix} \bar{d}_1 \\ \vdots \\ \bar{d}_p \\ \vdots \\ \bar{d}_{p(N+1)} \end{bmatrix} \in \mathbb{R}^{p(N+1)}, \quad (24)$$

where $[\cdot]_j$ means the j-the element of a vector, and $\bar{d}_i \in \mathbb{R}$. Then equation (23) can be written as the following $p(N+1)$ equations

$$\begin{aligned} &\sigma_i C x_0 = [\bar{Y}_0]_i = \bar{d}_i, \\ &\sigma_i (CA^k) x_0 + \sigma_i \left(C \sum_{j=0}^{k-1} A^{k-j-1} B u_j \right) = [\bar{Y}_k]_i = \bar{d}_{pk+i}, \end{aligned} \quad (25)$$

In particular, define coefficient matrices

$$T_0 = C,\ T_k = CA^k, S_0 = 0,\ S_k = C \sum_{j=0}^{k-1} A^{k-j-1} B u_j,$$

for $k = 1, \dots, N$. For the case of a noise-free system, the attacker is possible to solve the bilinear problem (25) only after observing enough time steps of $\bar{Y}_k$.

Remark 3. *By the property of bilinear equations [28], the attacker needs at least $N \geq max\{n, p\} - 1$ measurements of sensor and actuator values to calculate the exact coding matrix Σ and true initial state x_0 when there is no noise.* □

With noises in practical, we have

$$\begin{aligned} &\sigma_i C x_0 + [v_0]_i = [Y_0]_i = d_i, \\ &\sigma_i (CA^k) x_0 + \sigma_i \left(C \sum_{j=0}^{k-1} A^{k-j-1} (B u_j + w_j) \right) + [v_k]_i \\ =&[Y_k]_i = d_{pk+i}, \quad i = 1, \dots, p,\ k = 1, \dots, N, \end{aligned} \quad (26)$$

where $[Y_k]_i$ and d_i are defined similar as $[\bar{Y}_k]_i$ an $\bar{d}_i$ in vectorization (24). Under the assumption that both w_k, v_k are i.i.d. Gaussian noise, for any k, their expectations satisfy

$$\mathbb{E}\left[\sigma_i (C \sum_{j=0}^{k-1} A^{k-j-1} w_j) + [v_k]_i \right] = 0.$$

Then the noise-free and noisy sensor values satisfy that $\mathbb{E} Y_{\Sigma,N} = \bar{Y}_{\Sigma,N}$.

Hence, when the attacker observes noisy sensor outputs $Y_{\Sigma,N}$, the objective of retrieving the coding matrix Σ without the knowledge of x_0 is equivalent to finding $\sigma_1, \dots, \sigma_p, x_0$ that fit for the noise-free equation set (25). With even Gaussian noise, it becomes difficult to numerically find an exact solution of the true coding matrix, and the problem is then to minimize the total error between the left and right sides of the equations. The problem of estimating Σ, x_0 is described as Problem 1.

Problem 1. *The problem of estimating Σ, x_0 in the minimum mean square error perspective is defined as the following bilinear programming problem*

$$\begin{aligned} \underset{\sigma_1,\dots,\sigma_p,x_0}{\textit{minimize}} \quad & \sum_{k=0}^{N} \sum_{i=1}^{p} \left\| \sigma_i T_k x_0 + \sigma_i S_k - d_{pk+i} \right\|_2 \\ \textit{subject to} \quad & \textit{rank}(\Sigma = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_p \end{bmatrix}) = p. \end{aligned} \quad (27)$$

When there exists an invertible matrix Σ that satisfies the equations defined in (26), the above bilinear optimization problem (27) has an optimal cost 0. However, the optimal solution Σ^* does not need to be the true coding matrix Σ, since there is noise and the objective function of problem (27) does not include noise of each time step.

The rank constraint of problem (27) is non-convex, and in practice the attacker does not know how many measurements return the best estimation before calculating Σ^* given all existing measurements. Hence, we design the following heuristic algorithm for the attacker, which ignores the rank constraint first, and checks whether Σ is full rank every step till a feasible solution is reached.

Algorithm 2 Algorithm of estimating Σ for the attacker

Inputs: System's parameter (A, B, C), design of Kalman Filter K, the threshold α of χ^2 detector, algorithm stopping condition–estimation error ϵ.
Initialization: Initialize the value of estimation error $Er > \epsilon$, and the estimation of coding matrix $\hat{\Sigma}$ as a n identical matrix.
While $Er > \epsilon$ **or** $\hat{\Sigma} = I$**.**
1). Read one new sensor and actuator observation, and update parameters of problem (27);
2). Solve problem (27). If the optimal solutions $\sigma_1^*, \dots, \sigma_p^*$ satisfy the full rank constraints, let Er be the value of the optimal cost, and $\hat{\Sigma} = \Sigma^* = \left[(\sigma_1^*)^T \dots (\sigma_p^*)^T \right]^T$.
Return: Estimation result of $\hat{\Sigma}$.

Remark 4. *It is worth noting that a bilinear equation usually has multiple solutions, and Algorithm 2 returns different optimal solutions $\hat{\Sigma}$ under different sensor and actuator measurements time N. Under this situation, it is not clear for the attacker to decide how many time steps to measure and which optimal solution to choose, even when the optimal cost of problem* (27) *is* 0*. Even for a simple two dimensional system A, multiple solutions exist and do not converge to one estimation after* 20 *steps of measurements, which we will show in simulation.* □

To summarize, there are two main challenges for the attacker to estimate the true coding matrix, the first one is because multiple solutions exist for bilinear equations or bilinear optimization problems. The second one comes from the noise in the communication channel, that even the attacker find a feasible solution to the bilinear equation set (26), it is only an unbiased estimation instead of the true coding matrix.

B. When the estimated $\hat{\Sigma} \neq \Sigma$

After the attacker estimates a coding scheme $\hat{\Sigma}$ and considers it as the true coding matrix the system is applying, the easiest way to keep stealthy is to inject $\hat{\Sigma}y_k^a$, where y_k^a is a stealthy data injection designed for the original system without coding. However, as discussed above, when there exists noise, the attacker is not able to calculate the exact coding matrix the system is applying. When $\hat{\Sigma} \neq \Sigma$, the injection sequence $\hat{\Sigma}y_k^a$ can only extend the time length before detected and cannot pass the detector. Then the system needs to evaluate how long the attacker needs to measure the sensor outputs and how long the attacker can stay stealthy by applying a new injection sequence, in order to decide the time length of changing the coding matrix.

Definition 2. *An estimated coding matrix $\hat{\Sigma}$ calculated by the attacker is called a feasible estimation of Σ that keeps the attacker stealthy for time $k = 0, \ldots, T$ while causing e error, if and only if for all sequence of injections $\hat{\Sigma}y_k^a$ designed by the attacker according to the estimated coding matrix $\hat{\Sigma}$, the dynamics of $\Delta e'_k, \Delta z'_k$ satisfy that*

$$\begin{aligned} &\Delta e'_{k+1} = (A - KCA)\Delta e_k - K\Sigma^{-1}\hat{\Sigma}y_{k+1}^a, \\ &\Delta z'_{k+1} = CA\Delta e_k + \Sigma^{-1}\hat{\Sigma}y_{k+1}^a, \\ &\max_{k=1,\ldots,T} \|\Delta e_k\|_2 \geq e, \ \max_{k=1,\ldots,T} \|\Delta z_k\|_2 \leq M, \end{aligned} \tag{28}$$

where Σ is the true coding matrix the system is applying. □

Define the time length of keeping stealthy with injection sequence $y_k^a, k = 1, \ldots, T_s$, for a system (1) as

$$T_s(y_k^a) = \inf_k \{k : \|\Delta z_k(y_k^a)\|_2 > M\}. \tag{29}$$

The attacker increases the time length of keeping stealthy when $T_s(\hat{\Sigma}y_k^a) > T_s(y_k^a)$. However, the attacker does not have a guarantee about $T_s(\hat{\Sigma}y_k^a)$ without the knowledge of the true coding matrix, since $\Delta z_k(\hat{\Sigma}y_k^a)$ is affected by both Σ and $\hat{\Sigma}$. There exists a trade-off between the time N the attacker takes to measure sensor and actuator values to estimate a better $\hat{\Sigma}$ and the time the attacker starts to apply a new injection sequence $\hat{\Sigma}y_k^a$. If the measuring time N is large, it is possible that the system already triggers the alarm before the attacker successfully recovers the coding scheme. If the attacker does not have enough measurements for a good estimation and then applies the estimated $\hat{\Sigma}$ to design a new injection sequence $\hat{\Sigma}y_k^a$, $T_s(\hat{\Sigma}y_k^a)$ will not be much larger than $T_s(y_k^a)$ and the malicious behavior will still be detected quickly by the system.

It is worth noting that the system can not decide whether Σ is easy to be estimated by the attacker by only checking $\|\Delta z_k(y_k^a)\|_2, k = 1, \ldots, T_s(\hat{\Sigma}y_k^a)$. When $\|\Delta z_k(y_k^a)\|_2$ stays in a small range for a long time and $T_s(\hat{\Sigma}y_k^a)$ is large, the reason may be the original injection sequence y_k^a also has a large time length of keeping stealthy $T_s(y_k^a)$. Define the stealthy time increasing proportion for an estimated $\hat{\Sigma}(N)$ calculated after measuring time N as

$$\alpha(N_\Sigma) = \frac{T_s(\hat{\Sigma}(N)y_k^a) - T_s(y_k^a)}{T_s(y_k^a)}, \tag{30}$$

where $\hat{\Sigma}(N)$ is estimated from N steps measurements of sensor and actuator values. As we will show in Section VI, $\alpha(N_\Sigma)$ increases with an increasing N_Σ for a fixed Σ in general. When the attacker is able to estimate the coding matrix and inject $(\hat{\Sigma}(N)y_k^a)$ to stay stealthy for a longer time, the system needs to apply a new coding matrix before the attacker has enough measurements to estimate an $\hat{\Sigma}(N)$ that reaches the threshold $\tilde{\alpha}(N_\Sigma)$ of the increasing time proportion $\alpha(N_\Sigma)$. From the perspective of the system, a heuristic way to decide the time length N_Σ of changing Σ is as Algorithm 3.

Algorithm 3 Heuristic Algorithm for choosing N_Σ

Inputs: Coded system's parameter (A, B, C, K, Σ), χ^2 detector threshold α, time step t_s for increasing N_Σ, threshold proportion $\tilde{\alpha}(N_\Sigma)$.
Initialization: Initialize the value of $\hat{\Sigma}$ as an identical matrix, let $N_\Sigma = 0$, calculate $\alpha(N_\Sigma)$
While $\alpha(N_\Sigma) < \tilde{\alpha}(N_\Sigma)$
1). Estimate $\hat{\Sigma}$ with t_s steps of new sensor and actuator values, and update $\alpha(N_\Sigma)$.
2). Let $N_\Sigma = N_\Sigma + t_s$, and save sensor and actuator values for next iteration.
Return: Measurement time length N_Σ for estimating Σ.

VI. Illustrative Examples

A. Coding scheme detect stealthy data injection

We show the effects of coding sensor outputs by examples of two-dimensional LTI systems. Consider a detectable 2-dimensional linear system with parameters:

$$\mathbf{A} = \begin{bmatrix} 0.8 & 0 \\ 0.5 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 2 & 0.5 \\ 0 & 1 \end{bmatrix}, \mathbf{D} = 0,$$

where A has an unstable eigenvalue $\lambda = 1$ and eigenvector $v = [0\ 1]^T$. One stealth attack sequence is: $y_0^a = [0.0588\ 0.0588]^T$, $y_1^a = [0.1286\ \ -0.9706]^T$, $y_k = y_{k-2}^a - y_0^a, k \geq 2$. Multiple solutions of feasible coding matrices that satisfy Theorem 1 exist in general. For instance, for the above system, $\Sigma_1 = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}$ are both feasible.

Figure 3 shows the comparison result of $\|\Delta z_k\|_2, \|\Delta z'_k\|_2$ when there is injection attacks for the original and coded systems, and $\|\Delta z'_k\|_2$ increases with time k after coded by Σ_1, while without coding $\|\Delta z_k\|_2$ is bounded. Figure 4 shows that for the sensor outputs transformed by Σ_2, $\Delta z'_k$ increases with time k, while the original system Δz_k stays inside a bounded range. For the transformed sensor outputs, the change of the estimation error $\Delta e'_k$ increases even slower than Δe_k under data injection attack as shown in Figure 5. By comparing the change of estimation error Δe_k and $\Delta e'_k$,

we show that estimation error of a coded system does not necessarily increase faster than the original system.

B. When the attacker tries to estimate the coding matrix

In this example, the system applies the coding matrix designed based on Algorithm 1, a scaled rotation matrix $2 * G(1, 2, \frac{\pi}{4})$ with a rotation radian $\theta = \frac{\pi}{4}$ in the $(1, 2)$ plane $\Sigma = \begin{bmatrix} 0.7 & 0.5 \\ -0.5 & 0.7 \end{bmatrix}$. When the attacker estimates Σ according to $N = 20$ steps of sensor and actuator measurements via Algorithm 2, the estimated result is $\hat{\Sigma}$ and the attacker designs a new injection sequence $\hat{\Sigma} y_k^a$ based on $\hat{\Sigma}$

$$\hat{\Sigma} = \begin{bmatrix} 2.80 & -0.15 \\ -0.89 & 0.05 \end{bmatrix}.$$

In Figure 6, we compare the residue change for: the original system under injection sequence y_k^a, the coded system under data injection y_k^a, and the coded system under injection sequence $\hat{\Sigma} y_k^a$. Assume the threshold for $\|\Delta z_k\|_2$ is set as $M = 2$, in Figure 6 we can see that the attack will be detected after injecting a sequence of data y_k^a (designed for the original system) for 12 seconds to the coded system, i.e., $T_s(y_k^a) = 12$. In contrast, $T_s(\hat{\Sigma} y_k^a) = 50$ seconds, however, $\hat{\Sigma}$ is estimated via $N = 20$ seconds of measurements of sensor and actuator values. Hence, the attacker does not have enough time to get such $\hat{\Sigma}$ before being detected.

C. Number of measurements to estimate the coding matrix

Figure 7 shows how the estimation of $\hat{\Sigma}$ changes with the number measurement steps N. In general, when N increases, the difference between $\hat{\Sigma}$ and Σ decreases, and the norm of residue change $\|\Delta z_k\|_2$ increases slower with sensor injection sequence $\hat{\Sigma} y_k^a$. However, as shown in Figure 7, for both $N = 25$ and $N = 200$, $\|\Delta z_k\|_2$ are almost the same, hence, the attacker does not infer a better coding matrix to keep stealthy with a greater measurement time. Comparing the time of keeping stealthy with estimated coding matrix, we have $T_s(\hat{\Sigma} y_k^a) = 20$ for $N = 2$, $T_s(\hat{\Sigma} y_k^a) = 30$ for $N = 5$, and approximately $T_s(\hat{\Sigma} y_k^a) = 51$ for $N \geqslant 25$. From the perspective of the system, if we set the threshold $\tilde{\alpha}(N_\Sigma) = 1.5$ in this case, and $\frac{T_s(\hat{\Sigma}(N) y_k^a) - T_s(y_k^a)}{T_s(y_k^a)} = \frac{30-12}{12} = 1.5$, by the heuristic Algorithm 3, the system can change the coding matrix every 5 seconds.

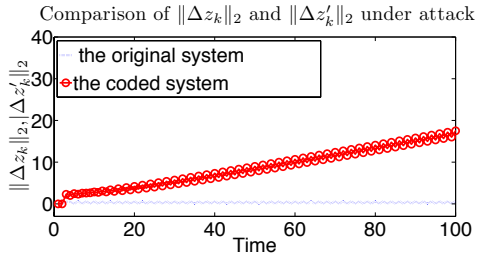


Figure 3. Comparison of $\|\Delta z_k\|_2$ of the original system and $\|\Delta z'_k\|_2$ of the coded system with Σ_1.

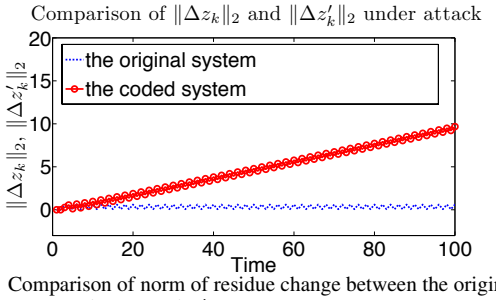


Figure 4. Comparison of norm of residue change between the original system and coded system, Δz_k and $\Delta z'_k$, for Σ_2 that satisfies Theorem 1.

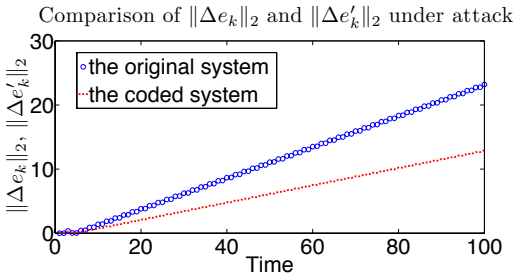


Figure 5. Comparison of norm of estimation error change between the original system and coded system, Δe_k and $\Delta e'_k$, for Σ_2 that satisfies Theorem 1.

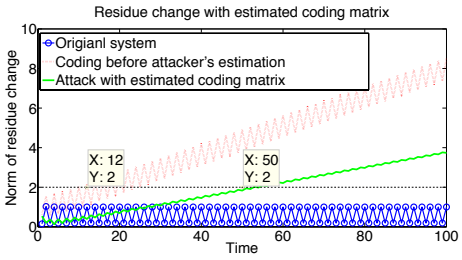


Figure 6. Comparison of norm of estimation residue change between the original system—Δz_k, coded system—$\Delta z'_k$ when the system applies Σ and the attacker injects y_k^a designed for the original system, and the sensor data injection sequence designed with estimated coding matrix— Δz_{sk} when the attacker injects $\hat{\Sigma} y_k^a$ and the true coding matrix is Σ. When $M = 2$, $T_s(y_k^a) = 12$ seconds, $T_s(\hat{\Sigma} y_k^a) = 50$ seconds.

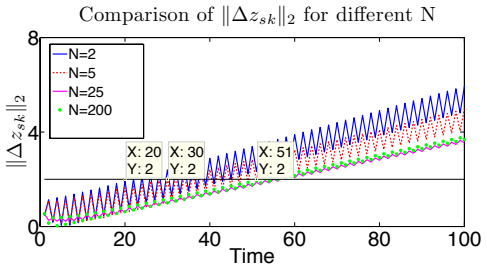


Figure 7. Comparison of norm of estimation residue change when the attacker designs a sensor injection sequence according to $\hat{\Sigma}$ estimated with different measurement number N. When attacker injects $\hat{\Sigma} y_k^a$ and the system applies coding, the detection time, i.e., the time $\|\Delta z_{sk}\|_2 \geqslant 2$ is labeled for different measurement time N: $N = 2, T_s(\hat{\Sigma} y_k^a) = 20$; $N = 5, T_s(\hat{\Sigma} y_k^a) = 30$. For $N = 25$ and $N = 200$, $T_s(\hat{\Sigma} y_k^a)$ are almost the same value: 51.

VII. Conclusion

In this work, we have proposed a method of coding sensor outputs to detect stealthy data injection attacks that designed by an intelligent attacker with system model knowledge. We show the conditions of a feasible coding scheme to detect a stealthy injection sequence with statistical detectors, and develop an efficient algorithm to compute such feasible coding

matrices. The sensor coding scheme is valid for the scenarios where the attacker is capable to estimate the coding matrix via measuring sensor outputs and actuator inputs. Simulation examples show that the adaptive injection sequence designed based on an estimated coding matrix cannot pass the detector without knowledge of the coding matrix applied by the system in general. In the future, we will explore a coding scheme for a system with structural constraints.

REFERENCES

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

[2] K.-D. Kim and P. R. Kumar, "Cyber-physical systems: A perspective at the centennial," *Proceedings of the IEEE*, pp. 1287–1308, 2012.

[3] A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *28th International Conference on Distributed Computing Systems Workshops*, 2008, pp. 495–500.

[4] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Workshop on future directions in cyber-physical systems security*, 2009.

[5] G. Dán and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *2010 First IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2010, pp. 214–219.

[6] A. Teixeira, S. Amin, H. Sandberg, K. Johansson, and S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *2010 49th IEEE Conference on Decision and Control (CDC)*, 2010, pp. 5991–5998.

[7] I. Hwang, S. Kim, Y. Kim, and C. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *Control Systems Technology, IEEE Transactions on*, vol. 18, no. 3, pp. 636–653, 2010.

[8] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, Nov 2013.

[9] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[10] M. Pajic, J. Weimer, N. Bezzo, P. Tabuada, O. Sokolsky, I. Lee, and G. Pappas, "Robustness of attack-resilient state estimators," in *2014 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2014, pp. 163–174.

[11] S. Dey, A. Chiuso, and L. Schenato, "Remote estimation with noisy measurements subject to packet loss and quantization noise," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 3, pp. 204–217, Sept 2014.

[12] V. Reppa, M. Polycarpou, and C. Panayiotou, "Distributed sensor fault diagnosis for a network of interconnected cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 1, pp. 11–23, March 2015.

[13] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*. ACM, 2009, pp. 21–32.

[14] A. Nayyar and D. Teneketzis, "Signaling in sensor networks for sequential detection," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 1, pp. 36–46, March 2015.

[15] F. Miao, M. Pajic, and G. Pappas, "Stochastic game approach for replay attack detection," in *IEEE 52nd Annual Conference on Decision and Control (CDC)*, Dec 2013, pp. 1854–1859.

[16] F. Miao and Q. Zhu, "A moving-horizon hybrid stochastic game for secure control of cyber-physical systems," in *IEEE 53rd Annual Conference on Decision and Control (CDC)*, Dec 2014, pp. 517–522.

[17] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *First Workshop on Secure Control Systems, CPS Week*, 2010.

[18] W. C.Kwon and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *Amerian Control Conference (ACC)*, June 2013.

[19] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 4, pp. 370–379, Dec 2014.

[20] P. Ganesan, R. Venugopalan, P. Peddabachagari, A. Dean, F. Mueller, and M. Sichitiu, "Analyzing and modeling encryption overhead for sensor network nodes," in *2nd ACM International Conference on Wireless Sensor Networks and Applications*, 2003, pp. 151–159.

[21] H. Imai and S. Hirakawa, "A new multilevel coding method using error-correcting codes," *Information Theory, IEEE Transactions on*, vol. 23, no. 3, pp. 371–377, May 1977.

[22] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 733–742, May 1993.

[23] H. Chan and A. Perrig, "Security and privacy in sensor networks," *Computer*, vol. 36, no. 10, pp. 103–105, Oct 2003.

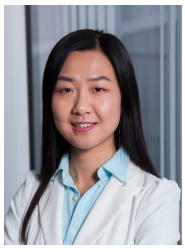
[24] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Contro, and Computing, 47th Annual Allerton Conference on*, 2009, pp. 911–918.

[25] M. Zhong, S. X. Ding, J. Lam, and H. Wang, "An LMI approach to design robust fault detection filter for uncertain LTI systems," *Automatica*, vol. 39, no. 3, pp. 543 – 550, 2003.

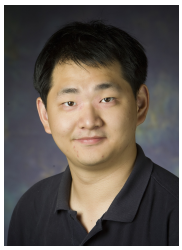
[26] O. Goldreich, *Foundations of Cryptography: Volume 2, Basic Applications*. New York, NY, USA: Cambridge University Press, 2004.

[27] R. C. Merkle, "Protocols for public key cryptosystems," in *Security and Privacy, 1980 IEEE Symposium on*, April 1980, pp. 122–122.

[28] S. Cohen and C. Tomasi, "Systems of bilinear equations," Computer Science Department, Stanford University, Tech. Rep. CS-TR-97-1588, Tech. Rep., 1997.



Fei Miao (S'13) received the B.Sc. degree in Automation from Shanghai Jiao Tong University, Shanghai, China in 2010, and the M.A. degree in Statistics and the Ph.D. degree in Electrical and Systems Engineering from the University of Pennsylvania, Philadelphia, in 2015 and 2016, respectively. Currently, she is a Postdoc Researcher in the Department of Electrical and Systems Engineering at University of Pennsylvania. Her research interests include data-driven real-time control frameworks of large-scale interconnected cyber-physical systems under model uncertainties, and resilient control frameworks to address security issues of cyber-physical systems. She was a Best Paper Award Finalist at the 6th ACM/IEEE International Conference on Cyber-Physical Systems in 2015.



Quanyan Zhu (S'04-M'12) is an assistant professor in the Department of Electrical and Computer Engineering at New York University. He received the B. Eng. in Honors Electrical Engineering with distinction from McGill University in 2006, the M.A.Sc. from University of Toronto in 2008, and the Ph.D. from the University of Illinois at Urbana-Champaign (UIUC) in 2013. From 2013-2014, he was a postdoctoral research associate at the Department of Electrical Engineering, Princeton University. He is a recipient of many awards including NSERC Canada Graduate Scholarship (CGS), Mavis Future Faculty Fellowships, and NSERC Postdoctoral Fellowship (PDF). He spearheaded and chaired INFOCOM Workshop on Communications and Control on Smart Energy Systems (CCSES), Midwest Workshop on Control and Game Theory (WCGT), and 7th Game and Decision Theory for Cyber Security (GameSec). His current research interests include resilient and secure interdependent critical infrastructures, energy systems, cyber-physical systems, and smart cities.

Miroslav Pajic (S06-M13) received the Dipl. Ing. and M.S. degrees in electrical engineering from the University of Belgrade, Serbia, in 2003 and 2007, respectively, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pennsylvania, Philadelphia, in 2010 and 2012, respectively. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Duke University. He also holds a secondary appointment in the Computer Science Department. Prior to joining Duke, Dr. Pajic was a Postdoctoral Researcher in the PRECISE Center, University of Pennsylvania, from 2012-2015. His research interests focus on the design and analysis of cyber-physical systems and in particular real-time and embedded systems, distributed/networked control systems, and high-confidence medical devices and systems. Dr. Pajic received various awards including the 2011 ACM SIGBED Frank Anger Memorial Award, the Joseph and Rosaline Wolf Award for Best Electrical and Systems Engineering Dissertation from Penn, the Best Paper Award at the 2014 ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), and the Best Student Paper award at the 2012 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS).

George J. Pappas (S'90-M'91-SM'04-F'09) received the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, CA, USA, in 1998. He is currently the Joseph Moore Professor and Chair of the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. He also holds a secondary appointment with the Department of Computer and Information Sciences and the Department of Mechanical Engineering and Applied Mechanics. He is a Member of the GRASP Lab and the PRECISE Center. He had previously served as the Deputy Dean for Research with the School of Engineering and Applied Science. His research interests include control theory and, in particular, hybrid systems, embedded systems, cyber-physical systems, and hierarchical and distributed control systems, with applications to unmanned aerial vehicles, distributed robotics, green buildings, and bimolecular networks. Dr. Pappas has received various awards, such as the Antonio Ruberti Young Researcher Prize, the George S. Axelby Award, the Hugo Schuck Best Paper Award, the George H. Heilmeier Award, the National Science Foundation PECASE award and numerous best student papers awards at ACC, CDC, and ICCPS.