On Applying Fault Detectors against False Data Injection Attacks in Cyber-Physical Control Systems

Quyen Dinh Vu[1] Rui Tan[2*] David K. Y. Yau[1,3]
[1]Singapore University of Technology and Design [2]Nanyang Technological University, Singapore
[3]Advanced Digital Sciences Center, Illinois at Singapore

***Abstract*—Much recent work has applied existing fault detectors against attacks in cyber-physical control systems. The results demonstrate effectiveness in detecting simplistic attacks that cause fault-like disruptions. However, they do not address motivated and knowledgeable attackers who craft attacks using knowledge of the system including its method of detecting attacks. In this paper, we analyze the conditions for an attacker to bypass a dissipativity-theoretic fault detector adopted in the prior work. We show that the attacker can use a quadratic programming solver to efficiently compute false data injection attacks to bypass the detector. We show further that, by applying an OR gate to fuse binary detection results from a number of the detectors, with carefully chosen parameters, we can achieve an integrated detector bank that *cannot* be bypassed by an attacker, if the attacker can tamper with either the sensor or control data of the system. For an n-dimensional linear time-invariant system, the number of needed fault detectors is $O(n!)$. This number can be dramatically reduced to $O(n)$ under a realistic assumption that the system has converged before the attack starts. Simulations for voltage control based on an IEEE 39-bus power system model validate our analysis.**

I. Introduction

Many critical infrastructures (e.g., power grid, water treatment, and railway systems) are evolving into cyber-physical systems by adopting information and communication technologies for control and situation awareness. The added cyber components, however, can make them vulnerable to potentially devastating cyber-attacks launched by insiders or resourceful foes. For instance, in 2000, a disgruntled former employee of a waste water service company used his insider access to compromise the company's supervisory control and data acquisition (SCADA) system, causing 800,000 liters of untreated sewage to contaminate connected water systems over several weeks [1]. Recent Dragonfly virus [2] and Stuxnet worm [3] attacks bypass air gaps first, then penetrate corporate networks via stolen credentials and zero-day exploits. The Stuxnet finally disrupts control systems that interact directly with nuclear centrifuges. Because of the stealthiness of these attacks, system operators often have little knowledge about them until severe physical damage has already occurred [3].

In this paper, we investigate early detection of a broad class of data integrity attacks called *false data injection* (FDI) against a *cyber-physical control system* (CPCS). The FDI attacks tamper with the system's sensor or control data transmitted on a cyber-plane, and aim to mislead the system to unsafe states and cause physical damage. The detection of FDI attacks is challenging because the detector typically needs to understand physical semantics of the sensor and/or control data being monitored. To address the challenge, there is growing momentum [4]–[9] to apply existing control system fault detectors,[1] designed based on understood physical semantics of the data, against FDI attacks. The rationale is that an FDI attack and a fault may both cause similar observable effects, such as discrepancy between the system state seen and that predicted from an *a priori* system model. An implicit but crucial assumption underlying this approach, however, is that the attacker cannot or will not attempt deliberately to conceal their actions, such that they will cause fault-like disruptions. Examples of simplistic fault-like attacks include: set a signal to its maximum or minimum [4], inject ramps, surges, and random noises [5]–[7]. The attacks considered in [8], [9] are designed based on system dynamics only and can be detected by certain fault detectors.

However, real-world attackers against critical infrastructures are often smart and they can optimize against a chosen target. Their strategies can be guided by knowledge about the system including its defense mechanisms deployed. The knowledge can be obtained in practice by malicious insiders, long-term data exfiltration [2], or social engineering against employees, contractors, or vendors of the infrastructure [3]. Thus, it is imperative to provide fundamental understanding on the usefulness and limitations of fault detectors in security incidents caused by knowledgeable and strategic attackers. In this paper, we follow Kerckhoffs's principle to consider an attacker who has accurate knowledge of the targeted system including its method of detecting attacks. We will analyze whether and how this *knowledgeable attacker* can bypass a fault detector to launch a stealthy attack. Guided by the analysis, we seek to strengthen existing fault detectors to ensure detection and impose limits on what the attacker can do.

To be applicable to real-world systems of non-trivial complexity, in this paper we study an advanced fault detector that is based on a dissipativity-theoretic property of linear time-invariant (LTI) systems [10], [11]. This detector has received growing research interest, due to its robustness in that it does

[*] Part of this work was completed while Rui Tan was with Advanced Digital Sciences Center, Illinois at Singapore.

[1] A fault refers to accidental disruption of sensor/control data due to natural malfunction of system component(s).

not require a detailed and accurate system model [12]–[14]. The detector has been applied to detect simplistic fault-like attacks [4], but its performance against attackers under the Kerckhoffs's setting is hitherto unknown. We derive closed-form conditions for bypassing the detector under different settings of the attacker's access to sensor or control data. We show that the attacker can use an efficient quadratic programming solver to compute FDI attacks to bypass the fault detector and mislead the control system to unsafe states. We further show that, by applying an OR gate to fuse the binary detection results from a number of the dissipativity-based fault detectors, the integrated detector bank *cannot* be bypassed even in the Kerckhoffs's setting, provided that we take care to select the detectors' parameters, and that the attacker can tamper with either the sensor or control data only. For an n-dimensional LTI CPCS, the number of needed fault detectors is $O(n!)$. We show that this number can be dramatically reduced to $O(n)$ if, before the attack starts, the system has been regulated to operate around a target state within a certain error bound. This reduction renders the detector bank feasible for a wide range of real-world LTI CPCSes that aim naturally to maintain the system state at a certain nominal value (e.g., 50 Hz frequency for a power grid).

To illustrate our analysis, we use a real-world CPCS – voltage control in power grids – as a case study. Simulations based on an IEEE 39-bus power system model validate our analysis. In particular, the results show that if a dissipativity-based fault detector is applied without modifications, an attacker tampering with the readings of four voltage meters can deviate the bus voltages to unsafe levels within just one control cycle, without triggering the detector. Hence, the proposed detector bank is needed to identify the attack successfully.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents preliminaries of the problem setup. Section IV states the research problem. Section V derives bypass conditions for the dissipativity-based fault detector. Section VI analyzes a new design of detector bank to ensure detection. Section VII presents our simulation results. Section VIII concludes.

II. Related Work

In a CPCS, the communication network for transmitting sensor/control data can experience faults such as data delays, corruptions, and losses. Various fault detection approaches have been proposed to deal with this problem [12]–[17]. They can be broadly divided into two categories: observer-based and dissipativity-based. The principle of the observer-based approach is to detect an observation's deviation from its predicted value based on historical data and a known system model [15]–[17]. Recently, the dissipativity-based approach has gained research attention [12]–[14], because it does not require a detailed and accurate system model that can be hard to obtain. Rather, it only needs three *energy functions* that are "summaries" of the model. Recent work has also shown that this approach is computationally efficient [12], [13].

Because of increasing reports on attacks against cyber-physical infrastructures, the security of CPCSes has attracted much interest. For instance, FDI attacks against a chemical reactor and water supply SCADA systems and their physical impacts are studied in [5], [6]. To detect attacks, several studies apply observer-based [5]–[9] and dissipativity-based [4] fault detectors. The observer-based detectors employed differ mainly in how they evaluate discrepancy, e.g., sequential change [5] and variable threshold-based test [8]. In [4], the authors apply a dissipativity-based fault detector to detect several types of attacks, and demonstrate its effectiveness using a case study of robotic arm velocity control. All of this prior work [4]–[9] does not address knowledgeable attackers.

Several studies have pointed out the vulnerabilities of existing fault detectors. In [18], the authors derive conditions for FDI attacks to bypass a *bad data detection*, which is an observer-based fault detector, in state estimation for a power grid. In [19], the authors construct a model checker to search for FDI attacks that will increase the electricity generation cost by a specified percentage. Kwon et al. [20] study FDI attacks that can bypass a Kalman filter-based residual checker (an observer-based fault detector) and analyze their impacts on the system state. These studies focus on analyzing the bypass conditions and the impact of attacks. Effective exploitation of existing fault detectors to make them work in the Kerckhoffs's setting is still lacking.

III. Preliminaries

Sections III-A and III-B describe a general CPCS model and the dissipativity-based fault detector, respectively. Section III-C presents a case study of power grid voltage control. The notational convention in this paper is as follows.[2] Take the letter x as an example. $\mathbf{X}$ denotes a matrix; $\mathbf{x}$ denotes a column vector; $x[k]$ denotes the kth sample of a time-domain signal x that is sampled periodically; $\mathbb{R}^{p\times q}$ denotes the set of real $p \times q$ matrices; $\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is an identity matrix.

A. CPCS Model

As illustrated in Fig. 1, we consider a CPCS that consists of a *physical plant*, a *cyber controller*, *sensors*, and *actuators*. The system's physical dynamics is described by the following widely adopted discrete-time LTI model:

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k], \quad (1)$$

$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k], \quad (2)$$

where $\mathbf{x}[k] \in \mathbb{R}^{n\times 1}$ and $\mathbf{y}[k] \in \mathbb{R}^{m\times 1}$ are the *state* of the physical plant and the *measurement* of the sensors at time instant k, respectively; $\mathbf{u}[k] \in \mathbb{R}^{l\times 1}$ is the *control signal* determined by the cyber controller and sent to the actuators to affect the state; $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times l}$, and $\mathbf{C} \in \mathbb{R}^{m\times n}$ are constant matrices. A control algorithm often determines $\mathbf{u}[k]$ from $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\{\mathbf{y}[k], \mathbf{y}[k-1], \ldots\}$. The $\mathbf{y}$ and $\mathbf{u}$ are transmitted over a communication network. In practice,

[2] A table summarizing the notation used in this paper can be found in an extended version of this paper [21].

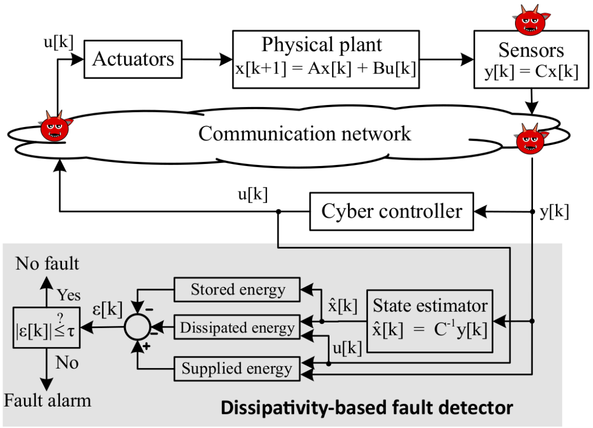


Fig. 1. CPCS model and dissipativity-based fault detector. The $\mathbf{C}^{-1}$ in the fault detector represents the state estimation process. Devils represent potential FDI attacks. The cyber controller and the fault detector are secured, while their input and output data may have been changed by the attacker.

a CPCS is subject to exogenous disturbances and sensor measurement noises. Although the models in Eqs. (1) and (2) do not explicitly capture these factors, we will discuss their impact on our analysis where applicable.

B. Dissipativity-Based Fault Detector

The dissipativity-based fault detector is based on three *energy functions* [11], i.e., *supply* $\omega(\mathbf{y}[k], \mathbf{u}[k])$, *storage* $v(\mathbf{x}[k])$, and *dissipation* $d(\mathbf{x}[k], \mathbf{u}[k])$. They are given by

$$\omega(\mathbf{y}[k], \mathbf{u}[k]) = \mathbf{y}^\intercal[k]\mathbf{Q}\mathbf{y}[k] + 2\mathbf{y}^\intercal[k]\mathbf{S}\mathbf{u}[k] + \mathbf{u}^\intercal[k]\mathbf{R}\mathbf{u}[k],$$
$$v(\mathbf{x}[k]) = \mathbf{x}^\intercal[k]\mathbf{P}\mathbf{x}[k],$$
$$d(\mathbf{x}[k], \mathbf{u}[k]) = (\mathbf{L}\mathbf{x}[k] + \mathbf{W}\mathbf{u}[k])^\intercal(\mathbf{L}\mathbf{x}[k] + \mathbf{W}\mathbf{u}[k]),$$

where $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{R}$, $\mathbf{P}$, $\mathbf{L}$, and $\mathbf{W}$ are constant matrices of proper dimensions. Starting from time instant K_0, the *supplied energy*, *stored energy*, and *dissipated energy* up to time instant k, are respectively given by $E_{\text{supplied}}[k] = \sum_{i=K_0}^{k} \omega(\mathbf{y}[i], \mathbf{u}[i])$, $E_{\text{stored}}[k] = V(\mathbf{x}[k+1]) - V(\mathbf{x}[K_0])$, and $E_{\text{dissipated}}[k] = \sum_{i=K_0}^{k} d(\mathbf{x}[i], \mathbf{u}[i])$. The *energy balance error* $\epsilon[k]$ is defined by $\epsilon[k] = E_{\text{supplied}}[k] - E_{\text{stored}}[k] - E_{\text{dissipated}}[k]$. An energy balance property for a class of LTI systems called *QSR-dissipative* systems is restated as the following lemma.

Lemma 1 ([11]). *If there exist matrices* $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{R}$, $\mathbf{L}$, $\mathbf{W}$, *and a positive definite matrix* $\mathbf{P}$ *such that* $\mathbf{A}^\intercal\mathbf{P}\mathbf{A} - \mathbf{P} = \mathbf{C}^\intercal\mathbf{Q}\mathbf{C} - \mathbf{L}^\intercal\mathbf{L}$, $\mathbf{A}^\intercal\mathbf{P}\mathbf{B} = \mathbf{C}^\intercal\mathbf{S} - \mathbf{L}^\intercal\mathbf{W}$, *and* $\mathbf{B}^\intercal\mathbf{P}\mathbf{B} = \mathbf{R} - \mathbf{W}^\intercal\mathbf{W}$, *the system described by Eqs. (1) and (2) is QSR-dissipative and*

$$\epsilon[k] = 0, \quad \forall k > K_0.$$

The energy balance property in Lemma 1 has been leveraged to construct fault detectors [12]–[14]. The bottom part of Fig. 1 illustrates such a fault detector. Specifically, a detector is characterized by a sextuple $\langle \mathbf{Q}, \mathbf{S}, \mathbf{R}, \mathbf{L}, \mathbf{W}, \mathbf{P} \rangle$ that satisfies the conditions in Lemma 1. Once a fault occurs, the conditions in Lemma 1 do not hold any more and thus $\epsilon[k] \neq 0$ after the fault. To avoid excessive false alarms caused by exogenous disturbances and random measurement noises, $|\epsilon[k]|$ can be compared with a positive threshold τ to make fault detection decisions. In Section VII, we will discuss the setting of τ.

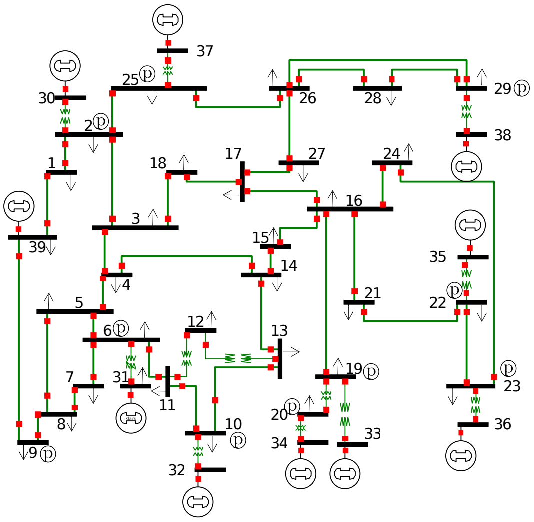


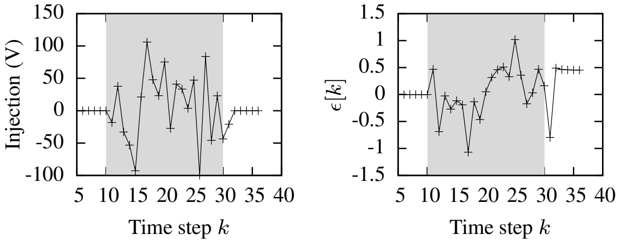
Fig. 2. One-line diagram of the IEEE 39-bus system. The buses labeled with ⓟ are pilot buses.

The dissipativity-based detector can yield good fault detection performance under realistic settings. For instance, our numerical results [21] show that, compared with several residual-based fault detectors (i.e., threshold tests based on the ℓ_2-norm or each element of the residual $\mathbf{x}[k+1] - \mathbf{A}\mathbf{x}[k] - \mathbf{B}\mathbf{u}[k]$), the dissipativity-based fault detector can achieve a better receiver operating characteristic (ROC) curve in the presence of random measurement noises and inaccuracy of the system model (i.e., $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$). Note that, for a system without a detailed system model, the energy functions can be directly learned from sensor/control data [12]. However, as this paper studies the vulnerability of the fault detector under the Kerckhoffs's setting, our analysis assumes that the system model is known.

C. A Case Study – Voltage Control in Power Grids

Although all the analytic results in this paper are based on the general CPCS model in Section III-A, we employ a real-world CPCS – voltage control – as a case study. A power grid consists of a number of *buses* connected with transmission lines. For instance, Fig. 2 shows a one-line diagram of the IEEE 39-bus system. Maintaining the bus voltages at nominal values is a basic control objective. Bus voltage deviations will cause power device trips, equipment damage, and even widespread loss of power. At a generator bus (i.e., a bus connected with a generator, such as Bus 30 and Bus 39 in Fig. 2), the voltage can be controlled by the generator. The voltage control maintains the voltages of selected non-generator buses at nominal values by adjusting the generator output voltages [22]. These selected non-generator buses are called *pilot buses*, which are often chosen by the system operator in terms of criticality of voltage regulation. In Fig. 2, the buses labeled with ⓟ represent pilot buses.

The LTI modelling of voltage control is as follows. At the kth time instant, the state $\mathbf{x}[k]$ is a vector of the pilot bus



(a) An element of the injection. (b) Energy balance error $\epsilon[k]$.

Fig. 3. Energy balance error due to random additive injections to measurement $\mathbf{y}$ of the voltage control system. Attack happens during the shaded period.

voltages and the control signal is $\mathbf{u}[k] = \mathbf{v}[k] - \mathbf{v}[k-1]$, where $\mathbf{v}$ is a vector of generator output voltages. Under these definitions, the model in Eq. (1) with $\mathbf{A} = \mathbf{I}_n$ is an approximation of the system dynamics [22]. As the state $\mathbf{x}$ can be directly measured by voltage meters at the pilot buses, the measurement matrix $\mathbf{C}$ in Eq. (2) is an identity matrix and $\mathbf{y}[k] = \mathbf{x}[k]$. Let $\mathbf{x}_0$ denote a vector of nominal voltages of the pilot buses. A major exogenous disturbance to the system is the changing reactive power draw of loads [22]. From control theory, for a constant $\alpha \in (0, 1)$, if the voltage control algorithm satisfies $\mathbf{Bu}[k] = \alpha(\mathbf{x}_0 - \mathbf{x}[k])$, the system is bounded-input bounded-output stable. This control algorithm is adopted in real systems [23] and also used in this paper.

IV. Problem Statement

A. Cybersecurity Threats and FDI Attack Model

In many large-scale critical infrastructures, the use of shared communication media to transmit sensor/control data introduces significant cybersecurity risks to CPCSes. Take voltage control as an example. To be scalable and cost effective, power grids often leverage existing network infrastructures and set up virtual private networks (VPNs) as logically isolated channels to collect measurements from voltage meters distributed over a vast area [24], [25]. However, such a software-based protection cannot guarantee the security of meter data links, because of pervasive software vulnerabilities (e.g., the Heartbleed bug [26] of OpenSSL-based VPNs). In this paper, we focus on FDI attacks that tamper with the measurement $\mathbf{y}$ or the control signal $\mathbf{u}$. As illustrated in Fig. 1, an FDI attack can be launched by compromising the sensors and the communication network. In this paper, we assume that the cyber controller and fault/attack detector are not compromised. We refer to [27] for a study on detecting malicious control signals from a possibly compromised cyber controller.

B. Dissipativity-Based FDI Attack Detection

At first glance, the dissipativity-based fault detector described in Section III-B can detect FDI attacks, since malicious changes to the measurement $\mathbf{y}$ and/or the control signal $\mathbf{u}$ can invalidate the energy balance condition in Lemma 1. For instance, Fig. 3(b) shows the energy balance error due to random additive injections to the measurement $\mathbf{y}$ of the voltage control system in Fig. 2, where Fig. 3(a) shows the trace of an element of the injection. We can see that this attack can be detected according to the significant non-zero energy balance errors. It is also known that the fault detector can identify several other types of attacks [4]. However, in this paper, we pose the following additional question: If an attacker possesses full knowledge of the system (i.e., $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$) and the dissipativity-based fault detector (i.e., $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{R}$, $\mathbf{L}$, $\mathbf{W}$, and $\mathbf{P}$), as well as the historical measurements and control signals, can they bypass the detector if they can tamper with a certain subset of $\mathbf{u}$'s and/or $\mathbf{y}$'s components? Our analysis in Section V answers the question in the positive. Thus, a natural follow-up question is: Can we design a new attack detector based on the dissipativity principle, such that the attacker cannot bypass the detector? This is the subject of Section VI.

V. Bypassing a Dissipativity-Based Fault Detector

This section investigates the vulnerabilities of a single dissipativity-based fault detector. We derive expressions of the energy balance error in the presence of FDI attacks and investigate efficient algorithms to bypass the detector.

A. Energy Balance Error in the Presence of FDI Attacks

We first consider the case that both the measurement $\mathbf{y}$ and the control signal $\mathbf{u}$ are compromised. Suppose that the attacker launches attack from time instant K_1 to K_2, inclusively. At the time instant $k \in [K_1, K_2]$, the compromised control signal, denoted by $\tilde{\mathbf{u}}[k]$, is $\tilde{\mathbf{u}}[k] = \mathbf{u}[k] + \mathbf{a}[k]$, where $\mathbf{a}[k]$ is the malicious injection. This compromised control signal $\tilde{\mathbf{u}}$ will be applied to the physical plant. The compromised measurement, denoted by $\tilde{\mathbf{y}}[k]$, is $\tilde{\mathbf{y}}[k] = \mathbf{y}[k] + \mathbf{Ce}[k] = \mathbf{C}(\mathbf{x}[k] + \mathbf{e}[k])$, where $\mathbf{x}[k]$ is the true system state, $\mathbf{Ce}[k]$ is the injection on the measurement $\mathbf{y}[k]$, and $\mathbf{e}[k]$ is a change to the estimated system state due to the injection to the measurement. As illustrated in Fig. 1, for $k \in [K_1, K_2]$, the inputs to the fault detector are the compromised measurement $\tilde{\mathbf{y}}[k]$ and the true control signal $\mathbf{u}[k]$. Based on $\tilde{\mathbf{y}}[k]$, the state estimator in the fault detector (cf. Fig. 1) will wrongly estimate the system state as $\tilde{\mathbf{x}}[k] = \mathbf{x}[k] + \mathbf{e}[k]$. We define the following functions:

$$
\begin{aligned}
f(\mathbf{x}, \mathbf{e}) &= -(\mathbf{e} + 2\mathbf{x})^\intercal \mathbf{Pe}, \\
g(\mathbf{x}, \mathbf{u}, \mathbf{a}) &= -(2\mathbf{Ax} + 2\mathbf{Bu} + \mathbf{Ba})^\intercal \mathbf{PBa}, \\
h(\mathbf{x}, \mathbf{u}, \mathbf{e}) &= (2\mathbf{Ax} + 2\mathbf{Bu} + \mathbf{Ae})^\intercal \mathbf{PAe} + f(\mathbf{x}, \mathbf{e}).
\end{aligned}
$$

The energy balance error in the presence of FDI attacks is given by the following lemma. The proof is omitted due to space constraints and can be found in [21].

Lemma 2. *When both the control signal $\mathbf{u}$ and the measurement $\mathbf{y}$ are compromised from time instant K_1 to K_2, inclusively, the energy balance error computed by the fault detector is $\epsilon[k] = f(\mathbf{x}[k+1], \mathbf{e}[k+1]) + \sum_{i=K_1}^{k} g(\mathbf{x}[i], \mathbf{u}[i], \mathbf{a}[i]) + \sum_{i=K_1}^{k} h(\mathbf{x}[i], \mathbf{u}[i], \mathbf{e}[i])$, $\forall k \in [K_1, K_2]$.*

By setting $\mathbf{e}[k] = \mathbf{0}$ or $\mathbf{a}[k] = \mathbf{0}$, we have the following two corollaries of Lemma 2 for the cases that either the control signal or the measurement is compromised, respectively.

Corollary 1. *When only the control signal* $\mathbf{u}[k]$ *is compromised from time instant* K_1 *to* K_2*, inclusively,* $\epsilon[k] = \sum_{i=K_1}^{k} g(\mathbf{x}[i], \mathbf{u}[i], \mathbf{a}[i])$*,* $\forall k \in [K_1, K_2]$.

Corollary 2. *When only the measurement* $\mathbf{y}[k]$ *is compromised from time instant* K_1 *to* K_2*, inclusively,* $\epsilon[k] = f(\mathbf{x}[k+1], \mathbf{e}[k+1]) + \sum_{i=K_1}^{k} h(\mathbf{x}[i], \mathbf{u}[i], \mathbf{e}[i])$*,* $\forall k \in [K_1, K_2]$.

B. Bypassing a Dissipativity-Based Fault Detector

Based on the analytic expressions for energy balance error in Section V-A, we now derive bypass conditions for a single dissipativity-based fault detector, as well as efficient algorithms for the attacker to find attack vectors satisfying the conditions.

1) Case 1: Only the control signal is compromised: We assume that the attacker can read the measurement and the control signal, but can change the control signal only. From Corollary 1, a sufficient and necessary condition for the injection $\mathbf{a}[k]$ to bypass the fault detector is $g(\mathbf{x}[k], \mathbf{u}[k], \mathbf{a}[k]) = 0$, $\forall k \in [K_1, K_2]$. Explicitly, the condition is

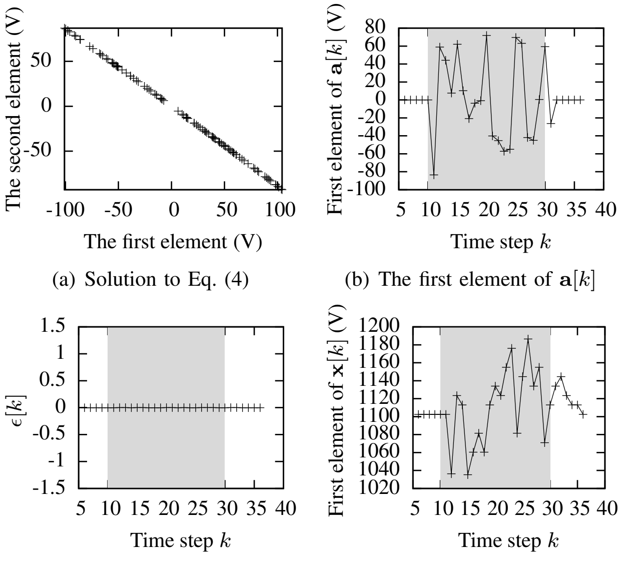
$$(2\mathbf{A}\mathbf{x}[k]+2\mathbf{B}\mathbf{u}[k]+\mathbf{B}\mathbf{a}[k])^\intercal \mathbf{P}\mathbf{B}\mathbf{a}[k]=0,\ \forall k\in[K_1, K_2]. \quad (3)$$

In practice, the attacker may have limited write access to $\mathbf{u}$. The bypass condition in Eq. (3) can be updated to address the case that the attacker can tamper with a subset of $\mathbf{u}$'s components only. Specifically, denote by $i_1, i_2, ..., i_a$ the indices of $\mathbf{u}$'s components that can be tampered with, by $\bar{\mathbf{B}}$ a matrix formed by the i_1-th, i_2-th, ..., and i_a-th columns of $\mathbf{B}$, and by $\bar{\mathbf{a}}$ a vector formed by the i_1-th, i_2-th, ..., and i_a-th components of $\mathbf{a}$. As the other components of $\mathbf{a}$ have to be zeros, we have $\mathbf{B}\mathbf{a}[k] = \bar{\mathbf{B}}\bar{\mathbf{a}}[k]$ and Eq. (3) can be updated as

$$(2\mathbf{A}\mathbf{x}[k]+2\mathbf{B}\mathbf{u}[k]+\bar{\mathbf{B}}\bar{\mathbf{a}}[k])^\intercal \mathbf{P}\bar{\mathbf{B}}\bar{\mathbf{a}}[k]=0,\ \forall k\in[K_1, K_2]. \quad (4)$$

As Eq. (4) is underdetermined, it can have an infinite number of solutions. In practice, in addition to the dissipativity-based fault detector, the system may check the range of data. Thus, to avoid being detected, the attacker needs to find solutions to Eq. (4) while not triggering those data range checks. From Eq. (4), a sufficient bypass condition is $2\mathbf{A}\mathbf{x}[k] + 2\mathbf{B}\mathbf{u}[k] + \bar{\mathbf{B}}\bar{\mathbf{a}}[k] = \mathbf{0}$, yielding a closed-form solution $\bar{\mathbf{a}}[k] = (\bar{\mathbf{B}}^\intercal\bar{\mathbf{B}})^{-1}\bar{\mathbf{B}}^\intercal(-2\mathbf{A}\mathbf{x}[k] - 2\mathbf{B}\mathbf{u}[k])$. However, this solution may easily violate the data range checks. For instance, for the voltage control system in Fig. 2, a component of $\mathbf{a}$ given by this solution can be up to 2 kV, significantly exceeding the nominal bus voltage of 1.05 kV. Thus, the attack vector $\mathbf{a}$ may not pass the data range checks.

Thus, the attacker should fully explore the attack vector space given by Eq. (4). An exhaustive search may be too slow to complete within one control cycle when the dimension of $\bar{\mathbf{a}}[k]$ is high. We now discuss an efficient approach for the attacker to quickly find a feasible solution. Define $\bar{g}(\mathbf{x}, \mathbf{u}, \bar{\mathbf{a}}) = (2\mathbf{A}\mathbf{x} + 2\mathbf{B}\mathbf{u} + \bar{\mathbf{B}}\bar{\mathbf{a}})^\intercal\mathbf{P}\bar{\mathbf{B}}\bar{\mathbf{a}}$. The attacker solves the following minimization problem: $\bar{\mathbf{a}}[k] = \mathrm{argmin}_{\bar{\mathbf{a}}}\, \bar{g}(\mathbf{x}[k], \mathbf{u}[k], \bar{\mathbf{a}})^2$ subject to other known constraints such as the aforementioned data range checks. As $\bar{g}(\mathbf{x}[k], \mathbf{u}[k], \bar{\mathbf{a}})^2$ is quadratic, the above problem can be efficiently solved using a quadratic programming solver, if other known constraints can be represented in linear forms (which is true for data range checks).



(a) Solution to Eq. (4)

(b) The first element of $\mathbf{a}[k]$

(c) Energy balance error $\epsilon[k]$

(d) The first element of $\mathbf{x}[k]$

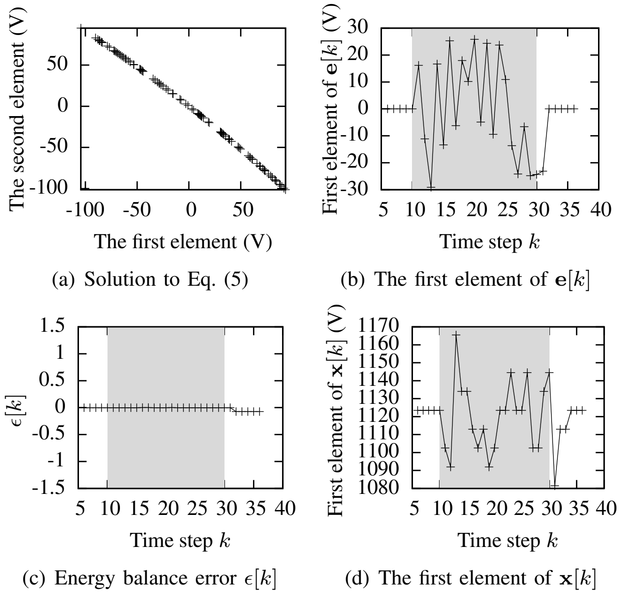
Fig. 4. Numerical results for the voltage control system in Fig. 2 when only control signal is compromised. Attack happens during the shaded time period.

We now illustrate the above analysis using numerical results for the voltage control system in Fig. 2. As there are ten generators, the dimension of $\mathbf{u}$ is ten. We assume that the attacker can only tamper with the first two elements of $\mathbf{u}$. We implement the above minimization approach using the SLSQP solver [28], which takes 0.1 seconds on a laptop computer to converge to a solution. Each point in Fig. 4(a) is a solution of $\bar{\mathbf{a}}[k]$ when the solver is initialized with a random seed. Exhaustive search also yields a similar result. From the figure, we can see that, even if the attacker can only tamper with two out of ten components of $\mathbf{u}$, they can find many solutions to bypass the fault detector. Figs. 4(b) to 4(d) show the simulation results over time when the attacker can tamper with all the components of $\mathbf{u}$. Specifically, the figures show the first element of $\mathbf{a}$ computed by the solver, the energy balance error computed by the fault detector, and an element of the system state (i.e., a bus voltage). We can see that the attack can cause a voltage deviation of up to 80 V from the nominal value of 1.05 kV, while keeping zero energy balance errors. This deviation is 7.6% of the nominal voltage, exceeding a basic requirement of 7% in power grids [29].

2) Case 2: Only the measurement is compromised: We assume that the attacker can read the measurement and the control signal, but can change the measurement only. From Corollary 2, a sufficient and necessary condition for the injection $\mathbf{C}\mathbf{e}[k]$ to bypass the fault detector is

$$f(\mathbf{x}[k],\mathbf{e}[k]) + (2\mathbf{A}\mathbf{x}[k-1] + 2\mathbf{B}\mathbf{u}[k-1] + \mathbf{A}\mathbf{e}[k-1])^\intercal\mathbf{P}\mathbf{A}\mathbf{e}[k-1] = 0, \quad \forall k \in [K_1, K_2]. \quad (5)$$

We note that the above condition contains both $\mathbf{e}[k-1]$ and $\mathbf{e}[k]$ and it is initialized to $\mathbf{e}[K_1 - 1] = 0$. Similar to Case 1,



(a) Solution to Eq. (5)

(b) The first element of $\mathbf{e}[k]$

(c) Energy balance error $\epsilon[k]$

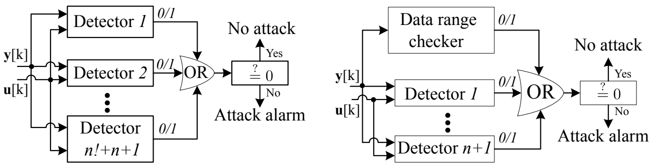
(d) The first element of $\mathbf{x}[k]$

Fig. 5. Numerical results for the voltage control system in Fig. 2 when only measurement is compromised. Attack happens during the shaded time period.

an attack vector $\mathbf{e}[k]$ can be found by solving the following minimization problem: $\mathbf{e}[k] = \operatorname{argmin}_{\mathbf{e}}(f(\mathbf{x}[k], \mathbf{e}) + (2\mathbf{A}\mathbf{x}[k-1] + 2\mathbf{B}\mathbf{u}[k-1] + \mathbf{A}\mathbf{e}[k-1])^\intercal \mathbf{P}\mathbf{A}\mathbf{e}[k-1])^2$, subject to other known constraints such as the data range checks. If these constraints can be represented in linear form, a quadratic programming solver can be used.

We now discuss the bypass condition when a certain subset of $\mathbf{y}$'s components can be compromised. Our analysis shows that it is difficult to derive such a condition under the general CPCS model in Eqs. (1) and (2). However, we can derive it for a class of systems with $\mathbf{C} = \mathbf{I}_n$, which is presented as follows. (Note that this limitation does not apply to the other analytic results in this paper.) Denote by $i_1, i_2, ..., i_e$ the indices of $\mathbf{y}$'s components that can be tampered with, by $\bar{\mathbf{P}}$ a matrix formed by the i_1th, i_2th, ..., i_eth (called "corresponding" for short) columns of $\mathbf{P}$, by $\bar{\mathbf{A}}$ a matrix formed by the corresponding columns of $\mathbf{A}$, by $\hat{\mathbf{P}}$ a matrix formed by the corresponding rows of $\bar{\mathbf{P}}$, by $\bar{\mathbf{e}}$ a vector formed by the corresponding components of $\mathbf{e}$. As the other components of $\mathbf{e}$ have to be zeros, we have $\mathbf{e}^\intercal\mathbf{P}\mathbf{e} = \bar{\mathbf{e}}^\intercal\hat{\mathbf{P}}\bar{\mathbf{e}}$, $\mathbf{P}\mathbf{e} = \bar{\mathbf{P}}\bar{\mathbf{e}}$, and $\mathbf{A}\mathbf{e} = \bar{\mathbf{A}}\bar{\mathbf{e}}$. By denoting $\bar{f}(\mathbf{x}, \bar{\mathbf{e}}) = \bar{\mathbf{e}}^\intercal\hat{\mathbf{P}}\bar{\mathbf{e}} + 2\mathbf{x}^\intercal\bar{\mathbf{P}}\bar{\mathbf{e}}$, the attacker can find an attack vector by solving the following minimization problem: $\bar{\mathbf{e}}[k] = \operatorname{argmin}_{\bar{\mathbf{e}}}(\bar{f}(\mathbf{x}[k], \bar{\mathbf{e}}) + (2\mathbf{A}\mathbf{x}[k-1] + 2\mathbf{B}\mathbf{u}[k-1] + \bar{\mathbf{A}}\bar{\mathbf{e}}[k-1])^\intercal\mathbf{P}\bar{\mathbf{A}}\bar{\mathbf{e}}[k-1])^2$, subject to other constraints such as the data range checks.

We apply the above minimization approach to the voltage control system, where the attacker can tamper with the first two elements of $\mathbf{e}$. Note that the precondition of $\mathbf{C} = \mathbf{I}_n$ holds for voltage control. Each point in Fig. 5(a) is a solution when the SLSQP solver is initialized with a random seed. We can see that the attacker can find many solutions to bypass the fault detector. Fig. 5(b) to 5(d) show the simulation results over time when the attacker can tamper with all the components of $\mathbf{e}$. Specifically, the figures show the first element of $\mathbf{e}$ computed by the solver, the energy balance error computed by the fault detector, and an element of the system state (i.e., a bus voltage). From Fig. 5(d), we can see that the attack causes a voltage deviation of up to 45 V, while keeping zero energy balance errors during the attack period (marked by shaded areas in Fig. 5). In Section VII, we will investigate the maximum voltage deviations caused by attacks on the measurement under various settings.



(a) Detector bank in Theorem 1. A total of $n! + n + 1$ dissipativity-based detectors are used.

(b) Detector bank in Theorem 2. A data range checker and a total of $n+1$ dissipativity-based detectors are used.

Fig. 6. Detector banks for detecting FDI attacks on either control signal or sensor measurement. The one in (b) can be applied to δ-converged systems.

3) Case 3: Both the control signal and the measurement are compromised: Our analysis shows that the attacker can compute the injections every time step to bypass a dissipativity-based fault detector with any parameters (i.e., $\mathbf{P}$ and other matrices). Specifically, at time instant k, the attacker injects an arbitrary $\mathbf{a}[k]$ into the control signal. Then, they choose $\mathbf{e}[k+1] = \mathbf{A}\mathbf{e}[k] - \mathbf{B}\mathbf{a}[k]$ such that the compromised measurement is consistent with the previous true control signal that is an input to the fault detector. In other words, since the attacker can manipulate both the control signal and the measurement, they can create an "illusion" that is consistent with the system model in Eq. (1). Therefore, the fault detector under any parameter setting cannot detect the attack.

VI. Dissipativity-Based FDI Attack Detection

In this section, we study the design of a dissipativity-based FDI attack detector that cannot be bypassed by the attacker who can tamper with either the control signal or the measurement, even if they possess full knowledge of the system and the attack detector. Our new approach applies multiple dissipativity-based detectors and uses an OR-rule to fuse their detection results, i.e., the existence of an attack is assumed if any detector raises an alarm. We call such a structure *detector bank*, which is illustrated in Fig. 6(a). Under the OR fusion rule, if we design these detectors such that the intersection of the attack vector solution spaces defined by their bypass conditions is empty, the attacker cannot bypass the detector bank. Thus, the OR fusion rule is a natural choice. In the following, Section VI-A presents the design of the detector bank and shows that its complexity is $O(n!)$, where n is the dimension of the system state. Our further analysis in Section VI-B shows that, for a system that has converged, assisted with a data range checker for the system state, the detector bank has complexity that can be reduced to $O(n)$.

A. Design of Detector Bank

From Lemma 1, each detector is characterized by six matrices satisfying three equality conditions. We propose an

Algorithm 1 Construct a dissipativity-based detector.

```
Input: System model $\langle \mathbf{A}, \mathbf{B}, \mathbf{C} \rangle$, a positive definite matrix $\mathbf{P}$
Output: A dissipativity-based detector $\langle \mathbf{Q}, \mathbf{S}, \mathbf{R}, \mathbf{L}, \mathbf{W}, \mathbf{P} \rangle$
1: Choose a real number $\delta$ larger than all eigenvalues of $\mathbf{B}^\intercal\mathbf{P}\mathbf{B}$, $\mathbf{R} = \delta\mathbf{I}_n$ ensures that all eigenvalues of $(\mathbf{R} - \mathbf{B}^\intercal\mathbf{P}\mathbf{B})$ are positive
2: Use Cholesky decomposition to find $\mathbf{W}$ such that $\mathbf{W}^\intercal\mathbf{W} = \mathbf{R} - \mathbf{B}^\intercal\mathbf{P}\mathbf{B}$
3: Choose a symmetric matrix $\mathbf{S}$ and find $\mathbf{L}$ to meet $\mathbf{L}^\intercal\mathbf{W} = (-\mathbf{A}^\intercal\mathbf{P}\mathbf{B} + \mathbf{C}^\intercal\mathbf{S})$
4: $\mathbf{Q} = (\mathbf{C}\mathbf{C}^\intercal)^{-1}\mathbf{C}(\mathbf{A}^\intercal\mathbf{P}\mathbf{A} - \mathbf{P} + \mathbf{L}^\intercal\mathbf{L})\mathbf{C}^\intercal(\mathbf{C}\mathbf{C}^\intercal)^{-1}$
```

algorithm that computes $\mathbf{Q}$, $\mathbf{S}$, $\mathbf{R}$, $\mathbf{L}$, and $\mathbf{W}$ based on a given positive definite matrix $\mathbf{P}$, such that they satisfy those conditions. The algorithm is given by Algorithm 1 and the correctness of its output can be verified by checking against the equality conditions in Lemma 1. As a result, under this algorithm, each detector can be characterized by a single matrix $\mathbf{P}$ and the problem of designing a detector bank is reduced to designing a set of positive definite $\mathbf{P}$ matrices. This reduction makes the problem tractable. Let $\mathcal{P} = \{\mathbf{P}_i | i = 1, 2, \ldots\}$ denote the set of $\mathbf{P}$ matrices characterizing the dissipativity-based detectors in a detector bank. The following theorem gives a $\mathcal{P}$ such that the attacker cannot bypass the detector bank.

Theorem 1. *For a CPCS with n-dimensional state, any FDI attack on either the control signal $\mathbf{u}$ or the measurement $\mathbf{y}$ cannot bypass a detector bank $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$ at the onset of attack, where $\mathcal{P}_1 = \mathbf{I}_n \cup \{\mathrm{diag}(\pi_1, \ldots, \pi_n) | \forall(\pi_1, \ldots, \pi_n)$ is a permutation of $(2, 1, 1, \ldots, 1)\}$ and $\mathcal{P}_2 = \{\mathbf{I}_{pc} + 2\mathbf{I}_n | \forall \mathbf{I}_{pc}$ is a column permutation of $\mathbf{I}_n\}$.*

The proof of Theorem 1 can be found in Appendix A. Fig. 6(a) illustrates the detector bank given by Theorem 1. As the cardinality of $\mathcal{P}_1$ and $\mathcal{P}_2$ is $(n+1)$ and $n!$, respectively, the complexity of the detector bank is $O(n!)$. We now illustrate Theorem 1 using an example. When $n = 3$, $\mathcal{P}_1 = \{\mathrm{diag}(1,1,1), \mathrm{diag}(2,1,1), \mathrm{diag}(1,2,1), \mathrm{diag}(1,1,2)\}$,

$$\mathbf{I}_{pc} \in \left\{ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \right\}.$$

We note that the detector bank given by Theorem 1 is a *sufficient* design to nullify the intersection of the attack vector spaces defined by the fault detectors' bypass conditions. Other designs with fewer needed fault detectors may exist. This is mainly because we use Algorithm 1 to reduce the design space.

B. Detector Bank with a Data Range Checker

The $O(n!)$ complexity of the detector bank given by Theorem 1 will lead to prohibitive computational and storage overhead when n is large. As discussed in Section V-B, the system can also apply data range checks that impose additional constraints for the attacker in finding attack vectors. This will potentially reduce the number of needed dissipativity-based detectors. In this section, we consider a CPCS that has converged to its nominal state before the attack starts. We can then apply a data range checker for the system state to detect FDI attacks. We first define a *δ-converged* CPCS as follows.

Definition 1. *A CPCS is said δ-converged if $\frac{|x - x_0|}{|x_0|} < \delta$ in the absence of FDI attacks, where x is any element of the system state with x_0 as its nominal value and δ is a positive constant.*

We define a data range checker for a δ-converged system.

Definition 2. *For a δ-converged system, after receiving the possibly compromised measurement $\mathbf{y}$, the data range checker first estimates the system state $\mathbf{x}$ based on $\mathbf{y}$ and raises an attack alarm if $\frac{|x - x_0|}{|x_0|} \geq \delta$ for any system state element x.*

With the above definitions, we have the following theorem. The proof is omitted here due to space constraints and can be found in [21].

Theorem 2. *For a δ-converged CPCS with $0 < \delta < 1$, a detector bank formed by the data range checker and a set of dissipativity-based fault detectors given by $\mathcal{P}_1$ that is defined in Theorem 1 can detect any FDI attack on either the control signal or the measurement at the onset of the attack.*

Fig. 6(b) illustrates the detector bank given by Theorem 2. Its cardinality is $O(n)$. It can therefore scale well with the dimension of the system state. We note that for many real systems, δ is much smaller than the upper bound of one required by Theorem 2. For instance, $\delta \leq 0.07$ is a basic requirement for voltage control in power grids [29].

C. Discussion

At the onset of an attack on the measurement $\mathbf{y}$, the proposed detector banks can detect the attack once the detector banks receive $\mathbf{y}$. Attack response strategies (e.g., switch to a model-driven control algorithm [7]) can be activated to avoid affecting the physical plant. However, an attack on the control signal $\mathbf{u}$ at its onset will directly affect the plant (cf. Fig. 1). Although it will be detected after the detector banks receive the affected $\mathbf{y}$, to avoid damage before detection of the attack, the actuators can check $\mathbf{u}$ using heuristics (e.g., range checks).

VII. Simulations

A. Simulation Methodology and Settings

To illustrate our analysis, we conduct simulations using PowerWorld [30] for the voltage control system based on the power system model in Fig. 2. PowerWorld is a high-fidelity power system simulator widely used in the power industry. According to Section III-C, the $\mathbf{B}$ in Eq. (1) is the only parameter of the LTI model for voltage control. We estimate $\mathbf{B}$ by linear regression using data traces of $\mathbf{x}[k+1] - \mathbf{x}[k]$ and $\mathbf{u}[k]$ obtained in a PowerWorld simulation. Our evaluation shows that the model error, which is defined as $\|\mathbf{x}[k+1] - \mathbf{x}[k] - \mathbf{B}\mathbf{u}[k]\|_{\ell_2}$, is just about $4\,\mathrm{V}$, where the

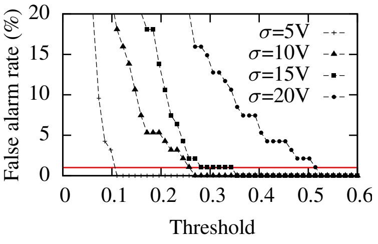


Fig. 7. False alarm rate vs. detection threshold under different noise standard deviations.

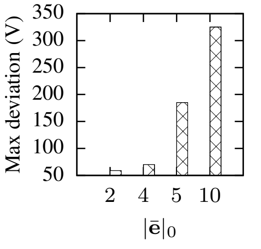
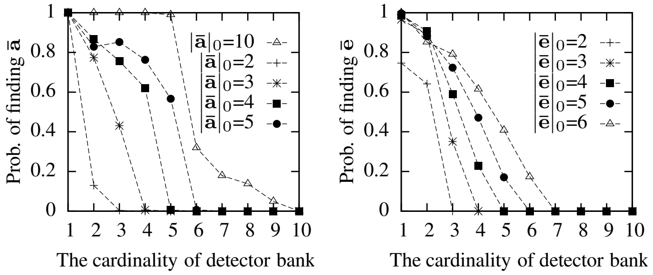


Fig. 8. Maximum voltage deviation vs. number of compromised $\mathbf{y}$ components (σ=20V).



(a) Control signal compromised.

(b) Measurement compromised.

Fig. 9. Probability of finding an attack vector.

nominal voltage is 1.05 kV. Thus, the LTI model accurately characterizes the system dynamics.

B. Simulation Results

1) Impact of measurement noise: As discussed in Section III-B, in the presence of measurement noise, the energy balance error should be compared with a non-zero threshold τ to avoid excessive false alarms. This set of simulations evaluates the false alarm rate versus the threshold under different measurement noise levels. Specifically, we simulate the power system under voltage control for 10,000 time steps in the absence of FDI attacks. For each time step, a random noise vector sampled from a zero-mean normal distribution of standard deviation σ is added to the measurement $\mathbf{y}$. For a given threshold τ, the false alarm rate is the ratio of the time steps with $|\epsilon[k]| \geq \tau$. Fig. 7 shows the results. We can see that, to enforce a false alarm rate upper bound of 1% (represented by a horizontal line), we need to apply different thresholds for different noise levels. Such thresholds are also used in the following sets of simulations.

2) Maximum attack impact: In this set of simulations, only a single dissipativity-based fault detector is used and the attacker tampers with the sensor measurement $\mathbf{y}$. We evaluate the maximum voltage deviation at any pilot bus caused by an FDI attack at its onset, while the energy balance error computed by the fault detector is below the threshold τ that ensures a false alarm rate of 1%. We find the maximum voltage deviation by evaluating a large number of attack vectors given by the minimization approach in Section V-B2 that is initialized with many random seeds. Fig. 8 shows the maximum voltage deviations when the attacker can tamper with different numbers of $\mathbf{y}$'s components (denoted by $|\bar{\mathbf{e}}|_0$). Consistent with intuition, a larger voltage deviation will result if more of $\mathbf{y}$'s components are compromised. In particular, if four components of $\mathbf{y}$ are compromised, the maximum voltage deviation is 70 V, i.e., 6.7% of the nominal voltage, almost reaching the safety margin of 7% [29].

3) Difficulty of finding attack vector: In this set of simulations, we evaluate the difficulty for the attacker to compute an attack vector under different settings of the detector bank cardinality (i.e., the number of dissipativity-based detectors). Specifically, we incrementally include a detector within $\mathcal{P}_1$ given by Theorem 1 into an evaluated detector bank. For each cardinality setting, we solve the energy balance error minimization problem with random seeds for 10,000 times for the system at a particular time instant. For some seeds, we cannot find a valid attack vector because the algorithm does not converge. Thus, we use the probability of finding a valid attack vector to characterize the difficulty of interest. Fig. 9(a) shows this probability versus the cardinality of the detector bank when the attacker can compromise different numbers of the control signal's components (denoted by $|\bar{\mathbf{a}}|_0$). The probability reduces to zero when the cardinality is ten, which is below and near the needed cardinality from Theorem 2 (i.e., eleven). Fig. 9(a) also shows that if the attacker can compromise more components of $\mathbf{u}$, it is easier for them to find an attack vector. Fig. 9(b) shows the results when the attacker can compromise different numbers of the measurement's components (denoted by $|\bar{\mathbf{e}}|_0$). The results are similar to Fig. 9(a). As the system is δ-converged before the onset of the attack, an injection must be bounded to be stealthy to the data range checker defined in Section VI-B. As a result, the attacker may not always succeed in finding an attack vector, especially when it can only compromise a limited number of the signal components. For instance, as shown in Fig. 9(b), when the attacker can compromise only two components of the measurement (i.e., $|\bar{\mathbf{e}}|_0 = 2$) and a single detector is used, the probability of finding a valid attack vector is 75%.

VIII. Conclusion

In this paper, we analyze the bypass conditions for a dissipativity-based fault detector that is applied to detect attacks. Based on the analysis, we develop detector banks that cannot be bypassed even in the Kerckhoffs's setting, provided that the attacker can tamper with either the sensor or control data of an LTI CPCS. We also analyze the complexity of the detector banks. Our results provide general insights into understanding the deficiency of directly applying existing fault detectors to identify attacks. They may need to be hardened to defeat knowledgeable attackers. Future attempts of applying existing fault detectors in a security context should keep this observation in mind.

Acknowledgment

The authors acknowledge Hoang Hai Nguyen for constructive discussions during the development of this work. This work was supported in part by the National Research

Foundation (NRF), Prime Minister's Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCR-NCR001-31) and administered by the National Cybersecurity R&D Directorate, in part by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR), and in part by a Start-up Grant at NTU.

References

[1] U.S. Department of Homeland Security, "Insider threat to utilities," https://info.publicintelligence.net/DHS-InsiderThreat.pdf.

[2] "Hackers infiltrated power grids," http://on.recode.net/1FpKP7Y.

[3] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, 2011.

[4] E. Eyisi and X. Koutsoukos, "Energy-based attack detection in networked control systems," in *HiCoNS*, 2014.

[5] A. A. Cardenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *AsiaCCS*, 2011.

[6] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Cyber security of water scada systems part ii: Attack detection using enhanced hydrodynamic models," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 5, 2013.

[7] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," *IEEE Trans. Smart Grid*, vol. 5, no. 2, 2014.

[8] V. L. Do, L. Fillatre, and I. Nikiforov, "A statistical method for detecting cyber/physical attacks on scada systems," in *IEEE Conf. Control Applications (CCA)*, 2014.

[9] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Allerton*, 2012.

[10] J. C. Willems, "Dissipative dynamical systems," *European Journal of Control*, vol. 13, no. 2, 2007.

[11] G. C. Goodwin and K. S. Sin, *Adaptive filtering prediction and control*. Courier Dover Publications, 2013.

[12] D. Theilliol, H. Noura, D. Sauter, and F. Hamelin, "Sensor fault diagnosis based on energy balance evaluation: Application to a metal processing," *ISA transactions*, vol. 45, no. 4, 2006.

[13] W. Chen, S. Ding, A. Khan, and M. Abid, "Energy based fault detection for dissipative systems," in *Conf. Control & Fault-Tolerant Syst.*, 2010.

[14] H. Yang, V. Cocquempot, and B. Jiang, "Fault tolerance analysis for switched systems via global passivity," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 55, no. 12, 2008.

[15] R. Patton and J. Chen, "Observer-based fault detection and isolation: robustness and applications," *Control Eng. Practice*, vol. 5, no. 5, 1997.

[16] Y. Wang, H. Ye, S. X. Ding, G. Wang, and D. Zhou, "Residual generation and evaluation of networked control systems subject to random packet dropout," *Automatica*, vol. 45, no. 10, 2009.

[17] X. Wan, H. Fang, and S. Fu, "Observer-based fault detection for networked discrete-time infinite-distributed delay systems with packet dropouts," *Applied Mathematical Modelling*, vol. 36, no. 1, 2012.

[18] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *CCS*, 2009.

[19] M. A. Rahman, E. Al-Shaer, and R. G. Kavasseri, "A formal model for verifying the impact of stealthy attacks on optimal power flow in power grids," in *ICCPS*, 2014.

[20] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *ACC*, 2013.

[21] http://publish.illinois.edu/resilient-grid/files/2016/01/fdtechreport.pdf.

[22] M. D. Ilic, X. Liu, G. Leung, M. Athans, C. Vialas, and P. Pruvot, "Improved secondary and new tertiary voltage control," *IEEE Trans. Power Syst.*, vol. 10, no. 4, 1995.

[23] J. P. Paul and J. Y. Leost, "Improvements of the secondary voltage control in france," in *IFAC Symposium*, 1986.

[24] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu, "Cyber-physical security testbeds: Architecture, application, and evaluation for smart grid," *IEEE Trans. Smart Grid*, vol. 4, no. 2, 2013.

[25] S. Sridhar and G. Manimaran, "Data integrity attacks and their impacts on scada control system," in *IEEE PES General Meeting*, 2010.

[26] "The heartbleed bug," http://heartbleed.com.

[27] H. Lin, A. Slagell, Z. Kalbarczyk, P. W. Sauer, and R. K. Iyer, "Semantic security analysis of scada networks to detect malicious control commands in power grids," in *ACM SEGS*, 2013.

[28] "Sequential least squares programming," http://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin_slsqp.html.

[29] E. Fumagalli, L. Schiavo, and F. Delestre, *Service quality regulation in electricity distribution and retail*. Springer, 2007.

[30] "Powerworld version 18," http://www.powerworld.com/.

Appendix A: Proof of Theorem 1

Proof. First, we prove that any $\mathbf{P}$ in $\mathcal{P}$ is positive definite. Clearly, all members of $\mathcal{P}_1$ are positive definite, as they are diagonal matrices with positive elements. For any $\mathbf{P}$ in $\mathcal{P}_2$, we now prove $\boldsymbol{\delta}^\intercal(\mathbf{I}_{pc} + 2\mathbf{I}_n)\boldsymbol{\delta} > 0$ for any non-zero vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots, \delta_n]^\intercal$. We have $\boldsymbol{\delta}^\intercal\mathbf{I}_{pc}\boldsymbol{\delta} = \sum_{k=1}^n \delta_k\delta_{i_k}$, where $\{i_1, i_2, ..., i_n\}$ is a permutation of $\{1, 2, ..., n\}$ unique to $\mathbf{I}_{pc}$. As $|\sum_{k=1}^n \delta_k\delta_{i_k}| \leq \sqrt{\sum_{i=1}^n \delta_i^2}\sqrt{\sum_{i=1}^n \delta_{i_k}^2} = \sum_{i=1}^n \delta_i^2 = \boldsymbol{\delta}^\intercal\boldsymbol{\delta}$, $-\boldsymbol{\delta}^\intercal\boldsymbol{\delta}^\intercal < \boldsymbol{\delta}^\intercal\mathbf{I}_{pc}\boldsymbol{\delta} < \boldsymbol{\delta}^\intercal\boldsymbol{\delta}^\intercal$. Thus, $\boldsymbol{\delta}^\intercal(\mathbf{I}_{pc} + 2\mathbf{I}_n)\boldsymbol{\delta} \geq \boldsymbol{\delta}^\intercal\boldsymbol{\delta} > 0$. Hence, $\mathbf{P}$ is positive definite.

Assume the attack's onset time is K_1. We omit K_1 in the following notations, except specified otherwise. Denote by $\epsilon_{\mathbf{P}_i}$ the energy balance error computed using $\mathbf{P}_i$. We now prove that $\exists\mathbf{P}_i \in \mathcal{P}$, $\epsilon_{\mathbf{P}_i} \neq 0$ for any FDI attack on the control signal $\mathbf{u}$. From Corollary 1, $\epsilon_{\mathbf{P}_i} = -(2\mathbf{Ax} + 2\mathbf{Bu} + \mathbf{Ba})^\intercal\mathbf{P}_i\mathbf{Ba}$. Define $\boldsymbol{\theta} = 2\mathbf{Ax} + 2\mathbf{Bu} + \mathbf{Ba}$ and $\boldsymbol{\psi} = \mathbf{Ba}$. We assume that $\boldsymbol{\psi} \neq \mathbf{0}$, because otherwise $\mathbf{x}[K_1+1] = \mathbf{Ax}+\mathbf{Bu}+\mathbf{Ba} = \mathbf{Ax} + \mathbf{Bu}$, i.e., the attack has no effect on the system state. We consider the following cases:

Case 1*: At least two elements of $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_n]^\intercal$ are different and at least two elements of $\boldsymbol{\psi} = [\psi_1, \psi_2, ..., \psi_n]^\intercal$ are different.* Suppose $\theta_1 \neq \theta_2$ and $\psi_1 \neq \psi_2$. There exists a column permutation of $\mathbf{I}_n$ (denoted by $\mathbf{P}_0$) such that $\boldsymbol{\theta}^\intercal\mathbf{P}_0 = [\theta_2, \theta_1, \theta_3, ..., \theta_n]^\intercal$. We now prove $\epsilon_{\mathbf{P}_0+2\mathbf{I}_n}$ or $\epsilon_{\mathbf{I}_n}$ is non-zero by contradiction, where $\mathbf{I}_n \in \mathcal{P}_1$ and $(\mathbf{P}_0 + 2\mathbf{I}_n) \in \mathcal{P}_2$. Assume $\epsilon_{\mathbf{P}_0+2\mathbf{I}_n} = \epsilon_{\mathbf{I}_n} = 0$. Since $\epsilon_{\mathbf{P}_0+2\mathbf{I}_n} = \boldsymbol{\theta}^\intercal\mathbf{P}_0\boldsymbol{\psi} + 2\epsilon_{\mathbf{I}_n} = \theta_2\psi_1 + \theta_1\psi_2 + \sum_{i=3}^n \theta_i\psi_i + 2\epsilon_{\mathbf{I}_n} = 0$ and $\epsilon_{\mathbf{I}_n} = \boldsymbol{\theta}^\intercal\mathbf{I}_n\boldsymbol{\psi} = \sum_{i=1}^n \theta_i\psi_i = 0$, we can derive $(\theta_2 - \theta_1)(\psi_2 - \psi_1) = 0$, which contradicts $\theta_1 \neq \theta_2$ and $\psi_1 \neq \psi_2$.

Case 2*: At least two elements of $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_n]^\intercal$ are different and $\boldsymbol{\psi} = [\psi, \psi, ..., \psi]^\intercal$ with $\psi \neq 0$.* Suppose $\theta_1 \neq \theta_2$. Denote $\mathbf{P}_1 = \mathrm{diag}(2, 1, 1, ..., 1) \in \mathcal{P}_1$ and $\mathbf{P}_2 = \mathrm{diag}(1, 2, 1, 1, ..., 1) \in \mathcal{P}_1$. We now prove $\epsilon_{\mathbf{I}_n}$, $\epsilon_{\mathbf{P}_1}$, or $\epsilon_{\mathbf{P}_2}$ is non-zero by contradiction. Assume $\epsilon_{\mathbf{I}_n} = \epsilon_{\mathbf{P}_1} = \epsilon_{\mathbf{P}_2} = \mathbf{0}$. As $\epsilon_{\mathbf{P}_1} - \epsilon_{\mathbf{I}_n} = \theta_1\psi$ and $\epsilon_{\mathbf{P}_2} - \epsilon_{\mathbf{I}_n} = \theta_2\psi$, we have $\theta_1\psi = 0$ and $\theta_2\psi = 0$, which contradicts $\psi \neq 0$ and $\theta_1 \neq \theta_2$.

Case 3*: $\boldsymbol{\theta} = [\theta, \theta, ..., \theta]^\intercal$ and at least two elements of $\boldsymbol{\psi} = [\psi_1, \psi_2, ..., \psi_n]^\intercal$ are different.* If $\theta \neq 0$, as $\epsilon_{\mathbf{P}} = \boldsymbol{\theta}^\intercal\mathbf{P}\boldsymbol{\psi} = \boldsymbol{\psi}^\intercal\mathbf{P}\boldsymbol{\theta}$, the proof procedure in Case 2 can be applied; otherwise, $\boldsymbol{\theta} = 2\mathbf{Ax} + 2\mathbf{Bu} + \mathbf{Ba} = \mathbf{0}$ and $\mathbf{x}[K_1 + 1] = \mathbf{Ax} + \mathbf{Bu} + \mathbf{Ba} = -(\mathbf{Ax} + \mathbf{Bu})$, i.e., the attack flips the state's sign, which can be easily detected.

Case 4*: $\boldsymbol{\theta}^\intercal = [\theta, \theta, \ldots, \theta]$ and $\boldsymbol{\psi} = [\psi, \psi, \ldots, \psi]^\intercal$ with* $\psi \neq 0$. For $\mathbf{I}_n \in \mathcal{P}_1$, to ensure $\epsilon_{\mathbf{I}_n} = \boldsymbol{\theta}^\intercal\mathbf{I}_n\boldsymbol{\psi} = n\theta\psi = 0$, $\theta = 0$, which flips the state's sign as in Case 3.

The proof for the case that the measurement $\mathbf{y}$ is under attack is similar to the above proof. It is omitted here due to space constraints and can be found in Appendix A of [21]. □