Moving Target Defense Approach to Detecting Stuxnet-like Attacks

Jue Tian, Rui Tan, *Senior Member, IEEE,* Xiaohong Guan, *Fellow, IEEE,* Zhanbo Xu, *Member, IEEE* and Ting Liu, *Member, IEEE*

***Abstract*—Recent cybersecurity incidents such as Stuxnet and Irongate alert us to the threats faced by critical cyber-physical systems. These attacks compromise the control signals to push the system to unsafe regions and meanwhile, inject fake sensor measurements to cover the ongoing attack. Detecting these *Stuxnet-like* (SL) attacks still remains an open research issue. This paper analyzes the taxonomy, construction, and implication of SL attacks in CPS control loops. We propose to apply the *moving target defense* (MTD) approach that actively changes the system configuration to detect SL attacks, since these attacks are generally constructed based on the knowledge about the system's configuration. We analyze the basic conditions for MTD to be successful. Finally, as a case study, we apply MTD for the secondary voltage control of power grids and present simulation results based on the IEEE 39-bus test system under realistic settings.**

***Index Terms*—Cyber-physical system security, Stuxnet, moving target defense, stealthy attack.**

I. Introduction

Cyber-physical systems (CPSes), such as power grids, autonomous vehicles, medical systems, are often safety-critical in that their failures would cause severe consequences including losses of life. In the last decades, CPSes are increasingly adopting modern information and communication technologies (ICTs) [1]. While the ICTs improve the system efficiency, they may also make the systems more vulnerable to cyber attacks launched by malicious insiders or hostile national rivals and therefore cause serious consequences. Such threats have been alerted by recent cybersecurity incidents, e.g., the Stuxnet worm against nuclear facilities, Irongate against industrial control systems, and BlackEnergy trojan against power plants. In particular, the Stuxnet and Irongate share a similar attack strategy, i.e., the attacker injects the malicious control commands to the actuators and meanwhile, corrupt the sensor readings to cover the ongoing attack. We define this class of coordinated integrity attacks on the control and sensor data as *Stuxnet-like* (SL) attacks, which will be the focus of this paper. Despite some existing forensic analysis of SL attacks [2], systematic countermeasures against the attacks have not received extensive research.

This paper studies the detection of the SL attacks. The controller of a CPS often adopts an anomaly detector to check the consistency between the transmitted control signals and the received sensor measurements. In particular, as the CPS generally follows known dynamics over time, the anomaly detector can leverage on the temporal correlations between the series of control signals and sensor measurements to improve the anomaly detection performance. However, with the capability of corrupting both the control and sensor data, the attackers can craft a series of fake sensor measurements that match the system dynamics given the original control signals sent from the controller. In this way, the anomaly detector cannot detect any deviation from the system dynamics. We note that a recent work [3] has studied how the false data injection (FDI) attacks bypass the bad data detection (BDD) of power grids' state estimation (SE). However, the SL attack studied in this paper is fundamentally different from the FDI against SE's BDD, in that SL attack bypasses the temporal-correlation-based check over control/sensor data time series, whereas the FDI bypasses the spatial-correlation-based check over one-shot sensor measurements. Thus, existing FDI detection approaches are not applicable to SL attack detection.

The attackers targeting critical infrastructures are often highly crafty, resourceful, and able to design the attacks based on extensive knowledge about the system. To well address such attackers, we consider the attackers who know the system topology, operation mechanism, and can compromise all the control and sensor data transmitted in the communication network. Under this highly adversary setting, the system operator may completely lose the awareness of the system state. As shown in this paper, the attacker can easily bypass the detection strategies that are based on the measurements only, e.g., dissipativity-theoretic fault detector [4] and CUSUM-based χ^2 detector [5]. To preclude and/or detect the SL attacks, a possible approach is to ensure the integrity of the control commands and sensor measurements. However, such data integrity ensurance is often very costly.

In this paper, we propose to apply the *moving target*

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org.

This work was supported by National Key R&D Program of China (2018YFB0803501), National Natural Science Foundation of China (61772408, U1766215, U1736205, 61721002, 61632015), the Fundamental Research Funds for the Central Universities, and a Start-up Grant at Nanyang Technological University.

J. Tian is with Systems Engineering Institute, MOE KLINNS Lab, Xi'an Jiaotong University, China. Part of this work was completed while Jue Tian was a visiting student and then a research intern at School of Computer Science and Engineering, Nanyang Technological University (e-mail: juetian@sei.xjtu.edu.cn).

R. Tan is with School of Computer Science and Engineering, Nanyang Technological University, Singapore (email: tanrui@ntu.edu.sg).

X. Guan is with Systems Engineering Institute, MOE KLINNS Lab, Xi'an Jiaotong University, China and also with the Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, China (email: xhguan@sei.xjtu.edu.cn).

Z. Xu and T. Liu are with Systems Engineering Institute, MOE KLINNS Lab, Xi'an Jiaotong University, China (email: zbxu@sei.xjtu.edu.cn, tliu@sei.xjtu.edu.cn).

defense (MTD) approach to detect stealthy SL attacks. MTD, originally proposed to enhance network security [6], actively changes the configuration of the system such that the attackers' knowledge about the system is always outdated. This increases the barriers for the attackers to launch targeted attacks. In this paper's context, the MTD changes the CPS' configuration to invalidate the attackers' knowledge about the system that is used to craft the SL attacks. Specifically, the MTD can perturb the control and/or the sensing units actively, e.g., through adjusting the units' parameters or gains. This idea can be easily implemented on existing physical units including the *current transformer* (CT) and the *potential transformer* (PT), etc. For instance, the turns ratio of CT can be achieved by modifying the primary circuits through the CT's window [7]. Moreover, recent ICTs, such as flexible manufacturing and multi-sensor information fusion can also be leveraged to implement MTD. Thus, we envision that MTD can be implemented readily in many rapidly evolving CPSes. Our previous work [8] studied the *hidden MTD* against the FDI attacks that compromise measurement signal. It showed that MTD's completeness (i.e., the ability to detect all FDI attacks) and stealthiness (i.e., the ability of being undetected by the attacker) are two conflicting goals. However, the hidden MTD does not address the much stronger SL attacks.

In this paper, we make the following contributions to understand and defend SL attacks using MTD:

First, we analyze the properties of SL attacks. Specifically, based on a general temporal-correlation-based anomaly detector, we analyze the necessary and sufficient condition for the SL attacks to achieve stealthiness. Based on this condition, we investigate the construction strategies of SL attacks, which can be classified into *measurement independent stealthy attack* (MISA) and *measurement dependent stealthy attack* (MDSA). For MISA, the attackers use the full knowledge of the target system to construct SL attacks that are completely stealthy to the anomaly detector, whereas the MDSA leverages the eavesdropped measurements to reduce the attack's reliance on the knowledge about the target system but with reduced stealthiness consequently.

Second, we design MTD against the above different types of SL attacks. We show that, by perturbing the control units only, the MTD can deal with MISAs, whereas the MTD that perturbs the sensing units only can preclude any stealthy MISAs and MDSAs considered in this paper. Moreover, we study several important aspects of MTD's implementation, including the basic prerequisites of the system, the selection of MTD's parameters and the implementation overhead.

In this paper, we demonstrate using our MTD design to protect a real-world CPS, the secondary voltage control (SVC) in power grids. Simulations based on the IEEE 39-bus test system confirm our analytical results.

The rest of this paper is organized as follows. Section II presents preliminaries. Section III describes the SL attacks. Section IV designs MTD against SL attacks. Section V presents simulation results. Section VI discusses a matrix estimation attack against MTD and proposes countermeasures. Section VII reviews related work. Section VIII concludes.

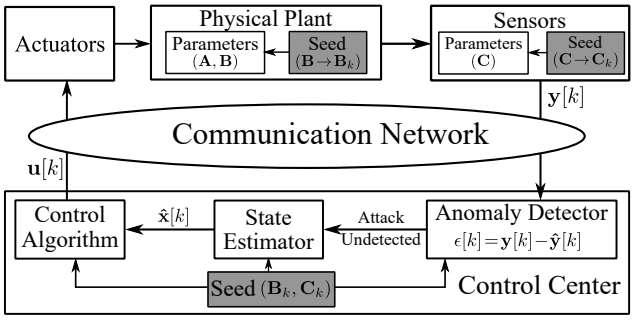


Fig. 1. Block diagram of the CPS with MTD approach. The blocks belonging to MTD are shaded.

II. Preliminaries

In this section, we describe a CPS model, a dynamic state estimator, and an anomaly detector under the general settings. The analytical results of this paper are based on these general models. Due to space limitations, the instantiated models are introduced in Appendix A[1], which will be used in the simulations in Section V. The notation convention in this paper is as follows. Take the letter x as an example. $\mathbf{X}$ denotes a matrix; $\mathbb{X}$ represents a set; $\mathbf{x}$ denotes a column vector; $\mathbf{x}[k]$ represents the sample of the signal $\mathbf{x}$ in the kth time period.

A. CPS Model

As illustrated in Fig. 1, the CPS control loop consists of the *control center*, *physical plant*, *actuators* and *sensors* [9], [10]. In this paper, we model the CPS dynamics by the following discrete-time linear time-invariant (LTI) model. This LTI model, which ignores the system nonlinearities, has been proven useful in studying the stability, faults, and attacks in power networks [11], [12], sensor networks [13], and building networks [14], etc.

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] + \mathbf{w}[k], \tag{1}$$

$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{v}[k], \tag{2}$$

where $\mathbf{x}[k] \in \mathbb{R}^n$ and $\mathbf{y}[k] \in \mathbb{R}^m$ respectively denote the system state and sensor measurement; $\mathbf{u}[k] \in \mathbb{R}^l$ is the control signal; $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times l}$, and $\mathbf{C} \in \mathbb{R}^{m\times n}$ are the state transit matrix, control matrix, and measurement matrix, respectively; $\mathbf{w}[k] \in \mathbb{R}^n$ and $\mathbf{v}[k] \in \mathbb{R}^m$ respectively denote the process and sensor noises. We assume that $\mathbf{w}[k]$ and $\mathbf{v}[k]$ follow the zero-mean multivariate Gaussian distributions with $\mathbf{Q}$ and $\mathbf{R}$ as the covariance matrices, respectively. The control signal $\mathbf{u}[k]$ and the sensor measurement $\mathbf{y}[k]$ are transmitted through a communication network.

B. Dynamic State Estimator and Anomaly Detector

This section describes the general dynamic state estimator and anomaly detector that will be used in our analysis.

The dynamic state estimator estimates the system state based on the measurement. Moreover, to better deal with measurement noises, the estimator can use historical estimated

[1]Due to space limitations, all appendixes are omitted and can be found in the supplementary file of this paper.

system states, control signals, and sensor measurements to estimate the current system state. Specifically, in the $(k+1)$th time period, the dynamic state estimator computes the estimated state, denoted by $\hat{\mathbf{x}}[k+1]$, as

$$\hat{\mathbf{x}}[k+1] = L_1(\hat{\mathbf{X}}[k], \mathbf{U}[k], \mathbf{Y}[k+1]), \tag{3}$$

where $\hat{\mathbf{X}}[k] = [\hat{\mathbf{x}}[k] \ ... \ \hat{\mathbf{x}}[0]] \in \mathbb{R}^{n\times(k+1)}$, $\mathbf{U}[k] = [\mathbf{u}[k] \ ... \ \mathbf{u}[0]] \in \mathbb{R}^{l\times(k+1)}$, $\mathbf{Y}[k] = [\mathbf{y}[k] \ ... \ \mathbf{y}[0]] \in \mathbb{R}^{m\times(k+1)}$. Note that the $L_1(\cdot)$ is an abstract function and our analysis does not depend on the specific form of this function. To properly initialize the dynamic state estimator, we assume that the system operator knows the initial system state $\mathbf{x}[0]$. Based on the estimated state, the dynamic state estimator can predict the sensor measurement that can be used for the anomaly detection:

$$\hat{\mathbf{y}}[k+1] = L_2(\hat{\mathbf{X}}[k], \mathbf{U}[k], \mathbf{Y}[k]), \tag{4}$$

where $\hat{\mathbf{y}}[k+1]$ denotes the predicted measurement in the $(k+1)$th time period and $L_2(\cdot)$ is an abstract function.

The anomaly detector detects various data faults or FDI attacks. The main components include the residual generation using the model, the signature generation via statistical testing, and the signature analysis [15]. This paper adopts a state-based residual defined as [15]

$$\boldsymbol{\epsilon}[k+1] \triangleq \mathbf{y}[k+1] - \hat{\mathbf{y}}[k+1]. \tag{5}$$

The statistical-isolability-based anomaly detector [15] yields a negative detection result if $-\boldsymbol{\epsilon}_0 \preceq \boldsymbol{\epsilon} \preceq \boldsymbol{\epsilon}_0$ and a positive detection result otherwise, where $\boldsymbol{\epsilon} \preceq \boldsymbol{\epsilon}_0$ means that each element of $\boldsymbol{\epsilon}$ is no greater than the corresponding element of $\boldsymbol{\epsilon}_0$; $\boldsymbol{\epsilon}_0$ is a predefined vector of small positive values that ensures a certain alpha level of detection.

To ensure stable and safe operation of the CPS, the control algorithm computes the control signal to maintain the system state around a desired target state $\mathbf{x}_0$. We consider a generic control algorithm as follows:

$$\mathbf{u}[k+1] = L_3(\hat{\mathbf{X}}[k+1], \mathbf{U}[k], \mathbf{Y}[k+1], \mathbf{x}_0), \tag{6}$$

where $L_3(\cdot)$ is an abstract function.

III. Stuxnet-like Attacks

This paper considers the FDI attacks that aim to subvert the safe operation of the CPS by tampering with the control signal and sensor measurement in the communication network. This section analyzes the system dynamics in the presence of attack. The results show that, for the FDI attack to be stealthy to the anomaly detector, the attack must obtain write access to both the control signal and sensor measurement, which is referred to as Stuxnet-like attack in this paper. Then, we create the taxonomy of the SL attacks that we will address using the MTD approach in Section IV. This section ends with the analysis of SL attack's practical implementation.

A. Threat Model and System Dynamics under FDI Attack

This section describes the threat model and analyzes the system dynamics under attack. To preserve the generality of the analysis, we assume that the attackers have the following two capabilities:

- Capability 1: The attackers can compromise all the control and sensor data transmitted in the communication network. This capability will be used in all the analyses and discussions in this paper.
- Capability 2: The attackers know the system topology and the operation mechanism. This capability will be mainly used in the SL attack construction in Section III-C.

Appendix B provides justification for the above attacker's capabilities.

Now, we analyzes the system dynamics under attack. We use the symbols defined in Section II to denote the quantities in the absence of attack. In the presence of attack, our notation convention is as follows. We omit the time index for conciseness. Denote by $\mathbf{x}_a$ and $\mathbf{y}_a$ the *actual* system state and sensor measurement, respectively. Denote by $\hat{\mathbf{x}}_m$, $\hat{\mathbf{y}}_m$, $\boldsymbol{\epsilon}_m$, and $\mathbf{u}_m$ the control center's estimated state, predicted sensor measurement, residual error, and the determined control signal, respectively. Denote by $\mathbf{u}_a$ the actual control signal received by the actuators, which has been contaminated by the attack. Specifically, $\mathbf{u}_a = \mathbf{u}_m + \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^l$ denotes the malicious data injected into the control signal $\mathbf{u}_m$ from the control center. Denote by $\mathbf{y}_m$ the sensor measurement received by the control center, which has been contaminated by the attack. Specifically, $\mathbf{y}_m = \mathbf{y}_a + \mathbf{b}$, where $\mathbf{b} \in \mathbb{R}^m$ denotes the malicious data injected into the actual sensor measurement $\mathbf{y}_a$.

In the presence of attack, the system dynamics is as follows:

$$\mathbf{x}_a[k+1] = \mathbf{A}\mathbf{x}_a[k] + \mathbf{B}\mathbf{u}_a[k] + \mathbf{w}[k], \tag{7}$$

$$\mathbf{y}_a[k] = \mathbf{C}\mathbf{x}_a[k] + \mathbf{v}[k]. \tag{8}$$

The compromised sensor measurement will affect the system state estimation process in Eqs. (3) and (4), the anomaly detection in Eq. (5), and the control algorithm in Eq. (6):

$$\hat{\mathbf{x}}_m[k+1] = L_1(\hat{\mathbf{X}}_m[k], \mathbf{U}_m[k], \mathbf{Y}_m[k+1]), \tag{9}$$

$$\hat{\mathbf{y}}_m[k+1] = L_2(\hat{\mathbf{X}}_m[k], \mathbf{U}_m[k], \mathbf{Y}_m[k]), \tag{10}$$

$$\boldsymbol{\epsilon}_m[k+1] = \mathbf{y}_m[k+1] - \hat{\mathbf{y}}_m[k+1], \tag{11}$$

$$\mathbf{u}_m[k+1] = L_3(\hat{\mathbf{X}}_m[k+1], \mathbf{U}_m[k], \mathbf{Y}_m[k+1], \mathbf{x}_0), \tag{12}$$

where $\hat{\mathbf{X}}_m[k] = [\hat{\mathbf{x}}_m[k] \ ... \ \hat{\mathbf{x}}_m[0]]$; $\mathbf{Y}_m[k] = [\mathbf{y}_m[k] \ ... \ \mathbf{y}_m[0]]$; $\mathbf{U}_m[k] = [\mathbf{u}_m[k] \ ... \ \mathbf{u}_m[0]]$. Note that $\hat{\mathbf{y}}_m[0] = \hat{\mathbf{y}}[0]$.

B. Stealthy FDI Attack

To be effective, the FDI attack needs to be stealthy to the anomaly detector. Otherwise, its impact can be mitigated or the attack may be isolated completely. We formally define the stealthiness of the FDI attack as follows.

Definition 1. *An FDI attack is stealthy if* $\boldsymbol{\epsilon}_m[k] = \boldsymbol{\epsilon}[k]$, $\forall k$.

Since the anomaly detection is made based on the residual error (i.e., $\boldsymbol{\epsilon}$ in the absence of attack and $\boldsymbol{\epsilon}_m$ in the presence of attack), from Definition 1, in the presence of attack, the detection results will be same as those in the absence of attack.

We have the following lemma that gives the necessary and sufficient condition for the attack stealthiness.

Lemma 1. *An FDI attack is stealthy if and only if* $\mathbf{y}_m[k] = \mathbf{y}[k]$, $\forall k$.

The sufficiency of Lemma 1 is intuitive, since the sensor measurement is the only input to the control center. If the compromised sensor measurement after the onset of the attack is same as the sensor measurement without attack, the control center will act exactly the same as in the absence of attack. Thus, the attack is stealthy. Regarding the necessity of Lemma 1, as the sensor measurement is the only input to the control center, if $\mathbf{Y}_m[k] = \mathbf{Y}[k]$, we can derive $\hat{\mathbf{y}}_m[k+1] = \hat{\mathbf{y}}[k+1]$. Since the attack is stealthy, according to Eqs. (5) and (11), we have $\mathbf{y}_m[k+1] = \mathbf{y}[k+1]$. Therefore, we can prove $\mathbf{y}_m[k] = \mathbf{y}[k]$, $\forall k$, through mathematical induction. The complete proof of Lemma 1 can be found in Appendix C.

The attackers' write accesses to the communicated data in the CPS are essential to launching the attack. Now, we analyze the needed write access for an FDI attack to be stealthy. We consider three cases.

1) Both the control signal and the sensor measurement are compromised (i.e., Stuxnet-like attack): From Lemma 1, a stealthy FDI attack can be designed as follows. In the kth time period, the attackers inject an arbitrary malicious data $\mathbf{a}[k]$ into the control signal $\mathbf{u}_m[k]$, and then tamper with the sensor measurement in the next time period by $\mathbf{b}[k+1] = \mathbf{y}[k+1] - \mathbf{y}_a[k+1]$. Thus, the sensor measurement received by the control center is $\mathbf{y}_m[k+1] = \mathbf{y}_a[k+1] + \mathbf{b}[k+1] = \mathbf{y}[k+1]$, satisfying the stealthiness condition in Lemma 1.

The above attack behaves like the Stuxnet worm that obtained the write accesses to both the control signal and sensor measurement of the centrifuges in Iran's nuclear plants. The worm intercepted the control commands from the supervisory control and data acquisition (SCADA) software and sent malicious commands to the field programmable logic controllers (PLCs). Meanwhile, to hide the ongoing anomaly, the worm replayed normal sensor measurements to the SCADA software.

2) Only the control signal is compromised (i.e., $\mathbf{b}[k] = \mathbf{0}$, $\forall k$*):* Suppose that the attackers tamper with the control signal from the kth time period. Thus, $\mathbf{x}_a[k] = \mathbf{x}[k]$ and $\mathbf{u}_m[k] = \mathbf{u}[k]$. From Eqs. (1), (2), (7), and (8), the sensor measurement in the $(k+1)$th time period is $\mathbf{y}_m[k+1] = \mathbf{y}_a[k+1] = \mathbf{y}[k+1] + \mathbf{CBa}[k]$. From Lemma 1, the attack $\mathbf{a}[k]$ is stealthy if and only if $\mathbf{a}[k] \in \ker(\mathbf{CB})$, where $\ker(\cdot)$ denotes the kernel space of a matrix. In addition, by defining $\mathbf{e}[k+1] = \mathbf{x}[k+1] - \mathbf{x}_a[k+1]$ (i.e., the deviation of actual system state due to the attack), we have $\mathbf{y}_m[k+1] - \mathbf{y}[k+1] = -\mathbf{Ce}[k]$ and the condition for the stealthy attack is $\mathbf{e}[k] \in \ker(\mathbf{C})$. The feasibility of this type of stealthy attack, called *zero state inducing* (ZSI) attack [13], is conditioned that the matrix $\mathbf{C}$ is not fully column-ranked. Thus, by designing or tuning the system to ensure that the $\mathbf{C}$ has full column rank, the attackers cannot achieve stealthiness by compromising the control signal only.

3) Only the sensor measurement is compromised (i.e., $\mathbf{a}[k] = \mathbf{0}$, $\forall k$*):* Suppose that the attackers tamper with the sensor measurement from the kth time period, i.e., $\mathbf{b}[k] \neq \mathbf{0}$. Thus, $\mathbf{y}_m[k] = \mathbf{y}_a[k] + \mathbf{b}[k] = \mathbf{y}[k] + \mathbf{b}[k]$. From Lemma 1, the attack is not stealthy at its onset time.

In summary, the SL and ZSI attacks are the only two types of stealthy FDI attacks. We can easily nullify the stealthiness of the ZSI attack by designing the system to have fully column-ranked $\mathbf{C}$ or $\mathbf{CB}$. In contrast, the stealthiness of the SL attack imposes no special conditions on the system model. Additional attack detection mechanisms must be developed to address the SL attack, which is the subject of Section IV.

C. SL Attack Taxonomy

From the analysis in Section III-B1, the attackers need to compute $\mathbf{y}[k+1]$ (i.e., the sensor measurement as in the absence of attack) and accordingly the injection $\mathbf{b}[k+1]$ based on the observed $\mathbf{y}_a[k+1]$, to ensure the attack's stealthiness. This section analyzes the SL attack taxonomy in terms of the approaches to computing $\mathbf{y}[k+1]$ and $\mathbf{b}[k+1]$.

SL attacks can be divided into two categories: *measurement independent stealthy attack* and *measurement dependent stealthy attacks*. The construction of MISA does not depend on any sensor measurement, but it needs the knowledge of the system model. In contrast, the construction of MDSA is based on the intercepted sensor measurement and it needs only partial or even no knowledge of the system model. In this paper, we consider two representative MDSAs, namely, *control scaling attack* and *measurement replay attack*. The former scales the control signal; the latter retains the control center's understanding on the system state, which is actually out of date. In what follows, we analyze these attacks separately.

1) MISA: MISA can be constructed based on the knowledge of the system model, i.e., $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$.

Definition 2. *MISA injects an arbitrary* $\mathbf{a}[k]$ *into the control signal and* $\mathbf{b}[k] = -\sum_{s=1}^{k} \mathbf{CA}^{s-1}\mathbf{Ba}[k-s]$ *into the sensor measurement.*

The stealthiness of MISA can be verified as follows. From Lemma 1, for a stealthy attack, we have $\mathbf{y}_a[k] = \mathbf{y}[k] + \sum_{s=1}^{k} \mathbf{CA}^{s-1}\mathbf{Ba}[k-s]$. Thus, by setting $\mathbf{b}[k] = -\sum_{s=1}^{k} \mathbf{CA}^{s-1}\mathbf{Ba}[k-s]$, we have $\mathbf{y}_m[k] = \mathbf{y}[k]$ and the attack is stealthy. The meaning of $\mathbf{b}[k]$ is as follows. By defining $\mathbf{e}[k] = \mathbf{x}[k] - \mathbf{x}_a[k]$ (i.e., the deviation of actual system state due to the attack), the $\mathbf{b}[k]$ given by Definition 2 can be written as $\mathbf{b}[k] = \mathbf{Ce}[k]$, where $\mathbf{e}[k] = \mathbf{Ae}[k-1] + \mathbf{Ba}[k-1]$. In other words, MISA computes the system state deviation due to the injection $\mathbf{a}[k-1]$ and hides the deviation by injecting into the sensor measurement. We note that because $\mathbf{b}[k]$ removes only the deviation projected by the deterministic system matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, the anomaly detector's residual error $\boldsymbol{\epsilon}_m$ will equal the residual error $\boldsymbol{\epsilon}$ purely caused by the process and measurement noises in the absence of attack. Thus, MISA is stealthy from Definition 1 in the presence of random process and measurement noises.

2) MDSA: MDSA leverages the intercepted sensor measurements to reduce the reliance on the knowledge about the system model in the attack construction. We note that, since the crafted injection $\mathbf{b}$ is based on the sensor measurements that contain noises, it does not exactly remove the deviation

projected by the system matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. As a result, MDSA nearly achieves stealthiness. In this paper, to simplify the discussion, we assume that the system is noiseless (i.e., $\mathbf{w}[k] = \mathbf{0}$ and $\mathbf{v}[k] = \mathbf{0}$) for all MDSA-related analysis. Under this assumption, MDSA achieves the stealthiness as defined in Definition 1. This simplification helps understand the essence of MDSA. Section V will evaluate the impact of the noises on our analysis via simulations. In the following, we define the control scaling and measurement replay attacks.

Definition 3. *The control scaling attack injects* $\mathbf{a}[k] = \lambda_k \mathbf{u}[k]$ *into the control signal, where* $\lambda_k \in \mathbb{R}$ *and* $\lambda_k \neq -1$. *Accordingly, the attack injects* $\mathbf{b}[k+1] = -\frac{\lambda_k}{1+\lambda_k}(\mathbf{y}_a[k+1] - \mathbf{C}\mathbf{A}\mathbf{C}^{+}\mathbf{y}_a[k]) + \mathbf{C}\mathbf{A}\mathbf{C}^{+}\mathbf{b}[k]$ *into the sensor measurement, where* $\mathbf{C}^{+}$ *denotes the generalized inverse of* $\mathbf{C}$*, i.e.,* $\mathbf{C}^{+} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}$.

The stealthiness of the scaling attack can be verified by mathematical induction. Due to space limitation, the proof of the scaling attack's stealthiness can be found in Appendix D.

The construction of the control scaling attack uses $\mathbf{A}$, $\mathbf{C}$, and the actual sensor measurement $\mathbf{y}_a$. The intuition of the attack is as follows. A precondition for the scaling attack is that $\mathbf{C}$ has full column rank. Thus, the system state of each time period can be estimated from the sensor measurement directly. Since the impact of the malicious injection into the control signal is proportional to that of the original control signal, combined with the system transit matrix $\mathbf{A}$ and the system state, the attack can remove a corresponding portion from the sensor measurement to achieve stealthiness.

Definition 4. *Suppose that a system has converged to the desired target state* $\mathbf{x}_0$ *in the* t*th time period. The measurement replay attack injects an arbitrary* $\mathbf{a}[k]$ *into the control signal and replays the historical sensor measurements from the* t*th to the* k*th time period, i.e.,* $\mathbf{b}[k+1] = \mathbf{y}[s] - \mathbf{y}_a[k+1], t \leq s \leq k$.

After convergence, the control signal will remain unchanged to maintain the system state, i.e., $\mathbf{u}[k] = \mathbf{u}[t]$ and $\mathbf{x}[k] = \mathbf{x}_0$, $\forall k \geq t$. Since $\mathbf{y}_m[k+1] = \mathbf{y}_a[k+1] + \mathbf{b}[k+1] = \mathbf{y}[s] = \mathbf{C}\mathbf{x}[s] = \mathbf{C}\mathbf{x}[k+1] = \mathbf{y}[k+1]$, $\forall k \geq t$, the attack is stealthy according to Lemma 1. The intuition of the attack is as follows. For a converged system, the control center would expect unchanged system states. Thus, the attackers can simply replay the historical sensor measurements after convergence to achieve stealthiness. The replay attack does not need any system knowledge. By monitoring the control signal and the sensor measurement, the attackers can judge whether the system has converged and launch the attack accordingly.

In addition to the SL attacks discussed above, several attacks, such as denial-of-service (DoS) attack and false data injection attack, are also studied in the literature. We note that these attacks are not stealthy in the dynamic state estimator and anomaly detector, and thus not the focus of this paper. Due to space limitation, these attacks are discussed in Appendix E.

D. SL Attack Implementation

This section discusses the implementation of the three SL attacks (i.e., the MISA, the scaling attack and the replay attack) described in Section III-C.

To launch stealthy SL attacks, the attackers will firstly obtain the read and write access of both the control signals and sensor measurements, according to the analysis in Section III-B. For a traditional system using a centralized control theme, this can be achieved by intruding into several critical routers and inject malware close to the control center or directly intruding into the control center (just as the Stuxnet). Then, the attackers will design the SL attack scheme accordingly. The attack schemes can be divide into two types: online and off-line modes. If the attackers can transmit real-time attack commands, and synchronously corrupt the sensor measurements, all the three SL attacks can be launched online. This can be more easily achieved for a small-scale system. For a large-scale networked system, transmitting real-time attack commands becomes more challenging, due to the unreliable network links and the unresponsive machines. However, we cannot rule out this possibility. In addition, according to Definition 2, 3 and 4, we can easily derive that only the MISA and the replay attack can be launched in an off-line mode. The detailed analysis can be founded in Appendix F. Appendix F also provides an illustrating example of MISA.

IV. MTD against SL Attacks

In general, MTD actively introduces controlled changes to a system to increase uncertainty and complexity for the attackers. As the construction of the SL attacks depends on the system matrices explicitly (in MISA and control scaling attack) or implicitly (in measurement replay attack), this section investigates whether MTD is effective in detecting the SL attacks. Section IV-A describes the MTD approach in CPS; Section IV-B derives the analytical conditions that MTD can always detect the SL attacks defined in Section III-C.

A. MTD in CPS

The proposed MTD actively perturbs the control matrix $\mathbf{B}$ or the measurement matrix $\mathbf{C}$. Thus, the $\mathbf{B}$ or $\mathbf{C}$ become time-varying and we denote by $\mathbf{B}_k$ and $\mathbf{C}_k$ the respective matrices in the kth time period. The perturbation is performed every time period. Our analysis in Section IV-B will show the necessity of this per-period perturbation. Note that we choose not to perturb the state transit matrix $\mathbf{A}$, because otherwise the physical plant is to be changed during the whole control process, which significantly increases implementation cost and also may introduce new risks to the system (e.g., instability). In contrast, perturbations to $\mathbf{B}$ or $\mathbf{C}$ can be implemented purely in the digital space. For instance, perturbing $\mathbf{B}$ can be achieved by a hybrid controller that can actively update the feedback gain (i.e., $\mathbf{H}$ in Appendix A) of the control units [16], or by adjusting the gains for the actuation system; perturbing $\mathbf{C}$ can be achieved by adjusting the sensors' gains (e.g., the transformation ratio of CT and PT).

In MTD, the control center needs the latest $\mathbf{B}_k$ or $\mathbf{C}_k$. Communicating the $\mathbf{B}_k$ or $\mathbf{C}_k$ from the field to the control center is not advisable since the communication network is subjected to the FDI attack. Instead, as illustrated by the shaded blocks in Fig. 1, common seeds can be used to generate the $\mathbf{B}_k$ or $\mathbf{C}_k$ in the field and the control center. We

assume that the common seeds are shared symmetric keys that are unknown to the attackers, such that the attackers cannot estimate/predict the current/future $\mathbf{B}_k$ or $\mathbf{C}_k$. We note that the traditional method, e.g., the encryption-and-decryption, is not applicable in this paper. Except for the MTD, the detection of stealthy attacks on CPS against the attackers who can obtain the system model requires at least one secure communication channel between the system operator and the plant [17], [18], [19]. Since this paper considers the scenario where an attacker can compromise all the control and sensor data transmitted in the communication network, the attackers may also obtain the keys of the encryption-and-decryption method, and thus subvert it. If the attackers launch the MISA or control scaling attack, they will need the system matrices. We assume that the attackers can obtain the true $\mathbf{A}$ (since our MTD does not perturb $\mathbf{A}$) and out-of-date control and/or measurement matrices denoted by $\mathbf{B}_\mu$ and $\mathbf{C}_\nu$, respectively, where $\mu < k$ and $\nu < k$. In Section VI, we will discuss how to ensure that the attackers cannot obtain the latest $\mathbf{B}_k$ and $\mathbf{C}_k$. We note that the attackers should not use random matrices for $\mathbf{B}$ and $\mathbf{C}$, because otherwise the injections will mostly be detected by the anomaly detector.

B. Design of MTD against SL Attacks

This section analyzes the design of MTD to negate the stealthiness of the SL attacks defined in Section III-C. We assume the attackers start injecting into the control signal and sensor measurement from the kth and $(k+1)$th time period, respectively. Our analysis focuses on this attack onset time period. We consider the different SL attacks separately.

1) MTD against MISA: In the $(k+1)$th time period, we have $\boldsymbol{\epsilon}_m[k+1]-\boldsymbol{\epsilon}[k+1] = \mathbf{y}_m[k+1]-\mathbf{y}[k+1] = \mathbf{y}_a[k+1]+\mathbf{b}[k+1]-\mathbf{y}[k+1] = (\mathbf{C}_{k+1}\mathbf{B}_k-\mathbf{C}_\nu\mathbf{B}_\mu)\mathbf{a}[k]$. If MTD ensures $\ker\{\mathbf{C}_{k+1}\mathbf{B}_k - \mathbf{C}_\nu\mathbf{B}_\mu\} = \{\mathbf{0}\}$, i.e., $(\mathbf{C}_{k+1}\mathbf{B}_k - \mathbf{C}_\nu\mathbf{B}_\mu)$ has full column rank, there exist no injections $\mathbf{a}[k]$ and $\mathbf{b}[k+1]$ to ensure the attack stealthiness defined in Definition 1. We consider two approaches to achieve the full column rank:

- Perturb $\mathbf{C}$ only: By ensuring that $(\mathbf{C}_{k+1} - \mathbf{C}_\nu)\mathbf{B}$ has full column rank, there exist no stealthy MISAs. In addition, from Section III-C1, the deviation of the actual system state due to the attack, i.e., $\mathbf{e}[k+1]$, can be calculated exactly. Then, we have $\boldsymbol{\epsilon}_m[k+1]-\boldsymbol{\epsilon}[k+1] = -(\mathbf{C}_{k+1}-\mathbf{C}_\nu)\mathbf{e}[k+1]$. Thus, alternatively, by ensuring that $(\mathbf{C}_{k+1} - \mathbf{C}_\nu)$ has full column rank, there exist no stealthy MISAs.
- Perturb $\mathbf{B}$ only: By ensuring that $\mathbf{C}(\mathbf{B}_k - \mathbf{B}_\mu)$ has full column rank, there exist no stealthy MISAs.

2) MTD against control scaling attack: In the $(k+1)$th time period, we have $\boldsymbol{\epsilon}_m[k+1] = \frac{\lambda_k}{1+\lambda_k}(\mathbf{C}_\nu\mathbf{A}\mathbf{C}_\nu^+\mathbf{y}[k] - \mathbf{C}_{k+1}\mathbf{A}\mathbf{C}_k^+\mathbf{y}[k]) = \frac{\lambda_k}{1+\lambda_k}(\mathbf{C}_\nu\mathbf{A}\mathbf{C}_\nu^+\mathbf{C}_k - \mathbf{C}_{k+1}\mathbf{A})\mathbf{x}[k]$. We consider two approaches:

- Perturb $\mathbf{C}$ only: We assume that the attackers know $\mathbf{x}[k]$ since they can infer it using $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{u}$ according to Eq. (1). Thus, we rewrite $\mathbf{b}[k+1]$ in Definition 3 as $\mathbf{b}[k+1] = -\frac{\lambda_k}{1+\lambda_k}(\mathbf{y}_a[k+1]-\mathbf{C}_\nu\mathbf{A}\mathbf{x}[k])$ and $\boldsymbol{\epsilon}_m[k+1] = \frac{\lambda_k}{1+\lambda_k}(\mathbf{C}_\nu - \mathbf{C}_{k+1})\mathbf{A}\mathbf{x}[k]$. With any non-zero system state, by ensuring that $(\mathbf{C}_\nu-\mathbf{C}_{k+1})\mathbf{A}$ has full column rank, there exist no stealthy control scaling attacks.
- Perturb $\mathbf{B}$ only: Since $\boldsymbol{\epsilon}_m[k+1]$ does not involve $\mathbf{B}$, perturbing $\mathbf{B}$ only is unable to detect the attack.

3) MTD against measurement replay attack: We aim to detect the attack without affecting the already converged system state. Suppose that in the $(k+1)$th time period, the attackers replay the previous measurements, i.e., $\mathbf{y}_m[k+1] = \mathbf{y}[s]$, where $s \in [t,k]$. Thus, $\boldsymbol{\epsilon}_m[k+1] = \mathbf{y}_m[k+1] - \mathbf{y}[k+1] = (\mathbf{C}_s - \mathbf{C}_{k+1})\mathbf{x}_0$, $s \in [t,k]$. We consider two approaches:

- Perturb $\mathbf{C}$ only: For any non-zero converged system state $\mathbf{x}_0$, by ensuring that $(\mathbf{C}_s-\mathbf{C}_{k+1})$ has full column rank, $\forall s \in [t,k]$, there exist no stealthy measurement replay attacks.
- Perturb $\mathbf{B}$ only: As $\boldsymbol{\epsilon}_m[k+1]$ does not involve $\mathbf{B}$, there does not exist a $\mathbf{B}$ perturbation to facilitate the attack detection.

We note that the MTD must be executed every time period to ensure that any measurement replay attack can be detected. As a counterexample, if the MTD is executed every two time periods and $\mathbf{C}_{2k+1} = \mathbf{C}_{2k}$, $\forall k$, the attackers can replay the sensor measurement $\mathbf{y}[2k]$ in the $(2k+1)$th time period to achieve stealthiness.

Now, we summarize the analytical results for the three attacks. By perturbing $\mathbf{B}$ only, the MTD can deal with MISAs only. For the MTD that perturbs $\mathbf{C}$, a condition that ensures no stealthy MISAs, control scaling, and measurement replay attacks is: $(\mathbf{C}_\nu - \mathbf{C}_{k+1})$ has full column rank, $\forall \nu \in [0,k]$, and $\mathbf{A}$ has full rank. Note that $\mathbf{C}_\nu$ represents any out-of-date measurement matrix that the attackers obtain. Since the ν is undetermined to the defenders, we should consider any historical $\mathbf{C}_\nu$, i.e., $\forall \nu \in [0,k]$. Table I summaries the analytical results of this section.

C. Implementation Considerations

This section discusses several considerations in implementing the MTD approach presented in Section IV-B.

First, cyber-physical systems are often safety-critical. Thus, it is always desirable and imperative to improve the security of safety-critical systems. Note that the implementation overhead of MTD to enhance the system security is not high, since the physical plant and sensors generally have computation capability and the proposed MTD can be implemented purely in the digital space. Moreover, because the system operation cost mainly depends on the system state and the control signal, it remains nearly unchanged under the MTD that perturbs the measurement matrix only. Thus, it is wise to deploy the proposed MTD to deal with various attacks.

Second, only the systems with $m \geq n$ can achieve fully column-ranked $(\mathbf{C}_\nu - \mathbf{C}_{k+1})$. In practice, such systems generally have fully column-ranked $\mathbf{C}$, which is also the condition to prevent the ZSI attack introduced in Section III-B2. Thus, to counteract the ZSI and SL attacks, $m \geq n$ is necessary.

Third, we discuss how to construct $\mathbf{C}_{k+1}$ to ensure that $(\mathbf{C}_\nu - \mathbf{C}_{k+1})$ has full column rank, $\forall \nu \in [0,k]$. In fact, the selection of $\mathbf{C}_{k+1}$'s elements under MTD is essential. Denote by $c_{m,ij}$ an element of the measurement matrix under MTD, which is in the ith row and jth column. Denote by $\mathbb{M}_r$ and $\mathbb{M}_c$ the sets of the row and column indices of all $c_{m,ij}$, respectively, i.e., $i \in \mathbb{M}_r, j \in \mathbb{M}_c, \forall c_{m,ij}$. We can derive that $|\mathbb{M}_r| \geq n$ and $|\mathbb{M}_c| = n$ is necessary to guarantee the above

TABLE I
CONDITIONS FOR MTD TO ENSURE NO STEALTHY ATTACKS.

Attack type	Information needed by attackers	Condition for MTD to ensure no stealthy attacks	
		Perturbing $\mathbf{C}$	Perturbing $\mathbf{B}$
MISA	$\mathbf{A}, \mathbf{B}, \mathbf{C}$	$\text{rank}((\mathbf{C}_{k+1} - \mathbf{C}_\nu)\mathbf{B}) = l$ or $\text{rank}(\mathbf{C}_{k+1} - \mathbf{C}_\nu) = n$	$\text{rank}(\mathbf{C}(\mathbf{B}_k - \mathbf{B}_\mu)) = l$
Scaling attack	$\mathbf{A}, \mathbf{C}$, and sensor measurements	$\text{rank}((\mathbf{C}_\nu - \mathbf{C}_{k+1})\mathbf{A}) = n$	Ineffective
Replay attack	sensor measurements	$\text{rank}(\mathbf{C}_s - \mathbf{C}_{k+1}) = n, \forall s \in [t, k]$	Invalid
Summary		$\text{rank}(\mathbf{C}_\nu - \mathbf{C}_{k+1}) = n, \forall \nu \in [0, k]$, $\text{rank}(\mathbf{A}) = n$	Detect MISA only

condition. In addition, for the MTD that satisfies $|\mathbb{M}_r| \geq n$ and $|\mathbb{M}_c| = n$, if any $c_{m,ij}$ varies randomly, the probability to get a $\mathbf{C}_{k+1}$ with fully column-ranked $(\mathbf{C}_\nu - \mathbf{C}_{k+1})$ is high.

V. CASE STUDY: SECONDARY VOLTAGE CONTROL

Many power grid control loops can be described using the LTI model mentioned in Section II-A. This section uses the *secondary voltage control* (SVC) as a case study. Besides, we adopt the Kalman filter (KF) and the χ^2 detector as the instantiated dynamic state estimator and anomaly detector, respectively, which are introduced in Appendix A.

A. Secondary Voltage Control and Simulation Settings

Maintaining the buses' voltages at their nominal values is a fundamental control task of any power grid. The generators' primary voltage controls maintain their output voltages at the setpoints via the excitation systems. The SVC maintains the voltages at selected non-generator buses called *pilot buses* by adjusting the voltage output setpoints of the generators. SVC can be modeled as follows. In the kth time period, the system state $\mathbf{x}[k]$ is the vector of the pilot bus voltages and the control signal $\mathbf{u}[k] = \mathbf{v}_G[k] - \mathbf{v}_G[k-1]$, where $\mathbf{v}_G$ denotes the vector of the generator output voltages. SVC can be modeled using Eqs. (1) and (2) [20], where $\mathbf{A}$ is an identity matrix. As the pilot bus voltages can be measured directly, $\mathbf{C} = \mathbf{I}_n$. Denote by $\mathbf{x}_0$ the vector of the desired pilot bus voltages. For control theory, if the voltage control algorithm satisfies $\mathbf{B}\mathbf{u}[k] = \beta(\mathbf{x}_0 - \mathbf{x}[k])$, where $\beta \in (0, 1)$, the system is bounded-input bounded-output stable. The actual state $\mathbf{x}[k]$ is often estimated from the noisy measurements $\mathbf{y}[k]$. In practice, the estimated state $\hat{\mathbf{x}}[k]$ is used to compute the control signal [20]: $\mathbf{B}\mathbf{u}[k] = \beta(\mathbf{x}_0 - \hat{\mathbf{x}}[k])$. We note that the above LTI model is an approximation to the actual system dynamics. The modeling inaccuracy can be captured by the process noise in Eq. (1).

We simulate the SVC based on the IEEE 39-bus system model that consists of 39 buses and 10 generators (i.e., $l = 10$). We choose 10 buses as the pilot buses. Thus, $m = n = 10$. We estimate the control matrix $\mathbf{B}$ from simulation data traces generated using PowerWorld, an industry-class high-fidelity power system simulator. While our analysis assumes a noise-less system, our simulations generate zero-mean process and measurement noises using covariance matrices of $\mathbf{Q} = \mathbf{R} = 0.003^2\mathbf{I}_n$ as the default setting. The KF-based state estimator and the χ^2 anomaly detector introduced in Section II are used. We set $\beta = 0.5$ for the control algorithm. We set $\alpha = 0.99$, i.e., the false alarm rate of each time period is 0.01.

B. Simulations and Results

We conduct four sets of simulations.

1) Effectiveness of various detectors against MISA: We evaluate three detection approaches including the MTD proposed in Section IV-B, a baseline approach called *partial MTD*, and the watermarking approach proposed in [9], [19].

- The MTD approach perturbs $\mathbf{C}$ only. Specifically, in the kth time period, the MTD applies a measurement matrix of $\mathbf{C}_k = \text{diag}(c_{k1}, c_{k2}, ..., c_{kn})$, where c_{ki} is uniformly and randomly sampled from $[1/(1 + d_m), 1 + d_m]$, where d_m characterizes the magnitude of MTD. By default, we set $d_m = 1$. The $\mathbf{C}_k$ satisfies all the conditions in Table I.
- The partial MTD applies $\mathbf{C}_k = \text{diag}(c_{k1}, c_{k2}, ..., c_{kn})$, where $c_{ki} = 1$ for the voltage sensors not used for MTD and the other c_{ki}'s are uniformly and randomly sampled from $[1/(1 + d_m), 1 + d_m]$. We assume that the attackers know which voltage sensors are not used for MTD. Thus, they can construct stealthy SL attacks by enforcing $\mathbf{x}[k](i) = \mathbf{x}_a[k](i)$, where $\mathbf{x}[k](i)$ and $\mathbf{x}_a[k](i)$ represent the voltages of any bus i monitored by a sensor used for MTD, in the absence and presence of the attack, respectively.
- The watermarking approach [9], [19] uses the actuation system to add a private excitation $\Delta\mathbf{u}[k]$ to the control signal. The excitation generation algorithm is known to the control center but unknown to the attackers. The control signal executed by the actuators is $\mathbf{u}_a[k] + \Delta\mathbf{u}[k]$. We generate $\mathbf{B}\Delta\mathbf{u}[k]$ from a zero-mean Gaussian with the same covariance as the process and measurement noises.

In each simulation, the MISA is launched from the first time period. Each element of the injection $\mathbf{a}$ in Definition 2 is sampled from a uniform distribution $\mathcal{U}(-d_a, d_a)$, where d_a characterizes the attack magnitude. In this set of simulations, we set $d_a = 0.04$. Note that in the simulation, the normal control commands are always within $[-0.1, 0.1]$, and thus the attack vector is negligible and may significantly affect the system state. For a time period, the attack detection probability and false alarm rate are assessed as the probabilities that the detector makes at least one positive detection decision from the first to the current time period, in the presence and absence of attack, respectively. Fig. 2 shows the attack detection probability by various detection approaches over time. The number in the legend of Fig. 2 is the ratio of the voltage sensors that are not used for MTD out of totally n voltage sensors used for SVC. In the first time period, our MTD approach achieves a detection probability of about 0.5. This is because that the χ^2 detector adopts a high detection threshold to ensure a high alpha level of 0.99 given the high process and measurement noise levels. The high detection threshold results in a relatively low detection probability. Nevertheless, after four time periods, our MTD approach achieves extremely high detection probabilities. In contrast, the partial MTD's detection probabilities remain low. This is because that the

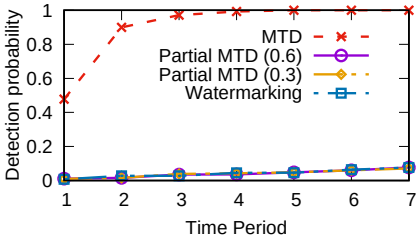


Fig. 2. MISA detection probability by various detection approaches in different time periods.

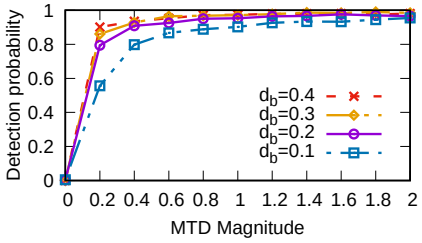


Fig. 3. Control scaling attack detection probability under various MTD and attack magnitudes.

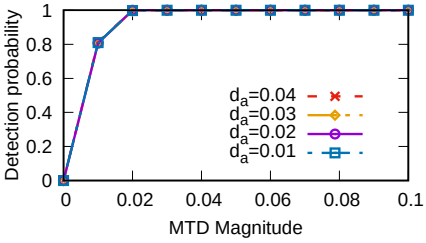


Fig. 4. Measurement replay attack detection probability under various MTD and attack magnitudes.

attackers can always find non-zero MISA injections while enforcing the stealthiness condition $\mathbf{x}[k](i) = \mathbf{x}_a[k](i)$ for all buses monitored by the sensors used for MTD. Note that the partial MTD's detection probabilities are close to its false alarm rate of $1 - \alpha^k$, where $\alpha = 0.99$. In other words, the partial MTD approach cannot well discriminate the residual errors caused by the process/measurement noises and the attacks. The watermarking approach has similarly low detection probabilities. This is because that MISA only removes the effect of the injection on the control signal by tampering with the sensor measurement, while the watermark still remains in the compromised sensor measurement.

2) Effectiveness of MTD against MDSA: Figs. 3 and 4 show the probabilities in detecting the control scaling and measurement replay attacks by the MTD with different magnitudes. For the control scaling attack, the scaling factor λ_k (cf. Definition 3) is sampled from a uniform distribution $\mathcal{U}(0, d_b)$, where d_b characterizes the attack magnitude. For the measurement replay attack, the injection $\mathbf{a}$ is constructed same as in Section V-B1. From the two figures, when MTD is not applied (i.e., the MTD magnitude is zero), the attacks are stealthy. The detection probability increases with the MTD magnitude, which is consistent with intuition. In Fig. 4, the detection probability is the same for different attack magnitudes (d_a). This is because the detection of the measurement replay attack is irrelevant to how the control signal is tampered with.

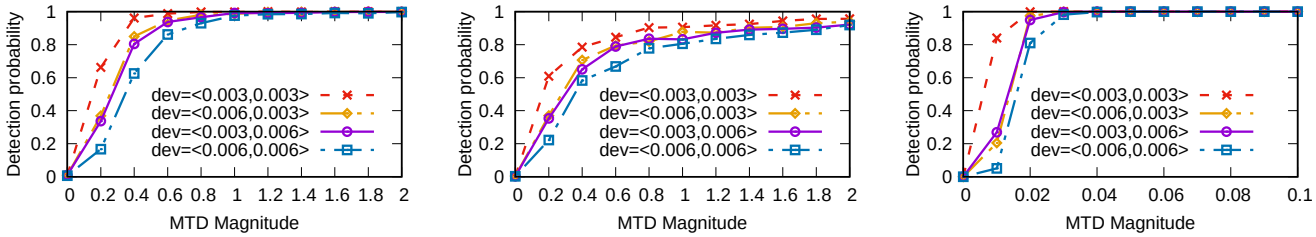
3) Noise sensitivity analysis: Above simulations are conducted under fixed system noises. This section analyzes the detection probability under different process and measurement noise levels. We select four groups of noisy environments: 1) $\mathbf{Q} = \mathbf{R} = 0.003^2\mathbf{I}_n$; 2) $\mathbf{Q} = 0.006^2\mathbf{I}_n, \mathbf{R} = 0.003^2\mathbf{I}_n$; 3) $\mathbf{Q} = 0.003^2\mathbf{I}_n, \mathbf{R} = 0.006^2\mathbf{I}_n$; 4) $\mathbf{Q} = \mathbf{R} = 0.006^2\mathbf{I}_n$. Fig. 5 shows the performance of detecting SL attacks under different noises and MTD magnitudes. From the three figures, for the same MTD magnitude, the SL attack detection probability decreases with the system noises. This is because that the anomaly detector should adopt a higher detection threshold to ensure a certain false alarm rate given a higher process or measurement noise level, which results in a relatively low detection probability. Therefore, to ensure a certain false alarm rate, a system with higher noise level should adopt a higher MTD magnitude. For instance, to detect the MISA with a probability of 0.8, the MTD magnitude should be set to be 0.3 and 0.6 under the first and fourth noise levels, respectively.

4) Effectiveness of the linearized model for SL attack detection using MTD approach: Previous simulations assume that the real system dynamics exactly follow the LTI model. This section discusses the effectiveness of this simplified model for SL attack detection. Note that the state transit matrix $\mathbf{A}$ in the SVC is the identity matrix. That is, the system state will not change regardless of the process noise, and thus the system in each period reaches a steady state. In this set of experiments, we adopt the power flow analysis module (provided by MATPOWER) to determine the real system state, and assume that the power flows in the real system can be characterized by the ac power flow model. Besides, we adopt the following settings: 1) the process noises are only caused by load perturbations, and the load perturbation between adjacent time periods follows the zero-mean Gaussian distribution; 2) the defenders/attackers adopt the linearized model to detect/generate the SL attacks. We respectively consider the detection performance of the linearized model against the three SL attacks.

Fig. 6 shows the receiver operating characteristic (ROC) curves [21] of SL attack detection performance. The curves labeled "MTD" and "Without MTD" refer to the ROC curves for the defenders in detecting the SL attacks when the MTD is used or not, respectively. The curves labeled "AM1" and "AM2" refer to the ROC curves when the attack magnitude is low or high. The concrete settings of the MTD magnitude and the attack magnitude can be seen in the subtitle of each figure. The third and fourth curves in each figure are the results when MTD is not used. From the ROC curve, the attacker's detection performance is poor. For instance, when the false positive rate is 0.5, the true positive rate is also about 0.5. The first and second curves are the results when the MTD is applied. We can see that the MTD significantly improves the defenders' attack detection performance. For instance, the curves with regard to the replay attack nearly passes through the $(0, 1)$ point, i.e., the replay attack will be always detected by the defenders who adopt the MTD approach and appropriately select the detection threshold. Moreover, the detection performance is always better when a greater attack magnitude is adopted, which is consistent with intuition.

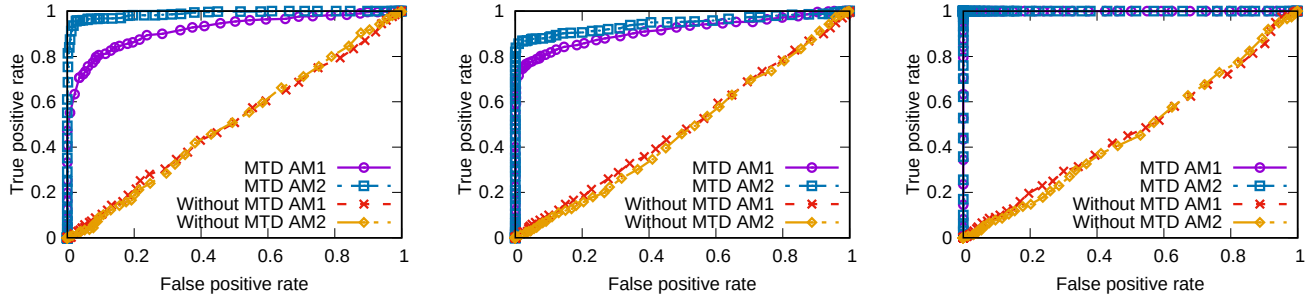
VI. Discussion

Our analysis in Section IV shows that MTD is effective in detecting the SL attacks constructed based on out-of-date control and measurement matrices. This section discusses an attack called *matrix estimation attack* that estimates the latest control and/or measurement matrices from the control signals and sensor measurements. We also propose countermeasures.



(a) MISA ($d_a = 0.2$) (b) Scaling attack ($d_b = 0.01$) (c) Replay attack ($d_a = 0.01$)

Fig. 5. SL attack detection probabilities under different noises and MTD magnitudes.



(a) MISA (MTD: $d_m = 1$, AM1: $d_a = 0.03$, AM2: $d_a = 0.05$) (b) Scaling attack (MTD: $d_m = 1$, AM1: $d_b = 0.02$, AM2: $d_b = 0.04$) (c) Replay attack (MTD: $d_m = 0.05$, AM1: $d_a = 0.02$, AM2: $d_a = 0.04$)

Fig. 6. ROC curves of SL attack detection performance for the linearized model.

We consider the MTD that perturbs the measurement matrix only. We illustrate the matrix estimation attack for a system that adopts a diagonal measurement matrix $\mathbf{C}_k = \text{diag}(c_{k1}, c_{k2}, ..., c_{kn})$. We assume that the attackers have obtained the previous system state $\mathbf{x}_a[k-1]$. Since only the measurement matrix is perturbed, the attackers can estimate $\mathbf{x}_a[k]$ using Eq. (7) and the known $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{u}_a$. Moreover, as $\mathbf{y}_a[k] = \mathbf{C}_k\mathbf{x}_a[k]$, the attackers can estimate the $\mathbf{C}_k$ by $c_{ki} = \frac{y_{ai}[k]}{x_{ai}[k]}$, $1 \le i \le n$, where x_{ai} and y_{ai} respectively denote the ith element of $\mathbf{x}_a$ and $\mathbf{y}_a$.

The above matrix estimation attack is no longer valid if the measurement matrix is not a simple diagonal matrix such that the problem of estimating $\mathbf{C}_k$ from $\mathbf{y}_a[k] = \mathbf{C}_k\mathbf{x}_a[k]$ is underdetermined. Nevertheless, we discuss the following approaches to counteracting the attack. First, we can perturb the matrices $\mathbf{C}$ and $\mathbf{B}$ simultaneously. Thus, the attackers cannot estimate $\mathbf{x}_a[k]$ since $\mathbf{B}$ is unknown. Second, similar to the watermarking approach [19], we can add a controlled noise $\mathbf{q}[k]$ to the sensor measurement, i.e., $\mathbf{y}[k] = \mathbf{C}_k\mathbf{x}[k] + \mathbf{q}[k]$. If this controlled noise is known to the control center (e.g., by common seeding) but unknown to the attackers, it will not affect the control after being removed from $\mathbf{y}[k]$ by the control center and disable the attackers from estimating $\mathbf{C}_k$ accurately even if they have a good estimate of $\mathbf{x}_a[k]$.

The analysis and evaluation of this paper show the effectiveness of the MTD approach in a cyber-physical control system with a communication network that may be compromised. In practice, the communication credential and the seed used by the MTD may be stored in different memory areas. For instance, by exploiting the Heartbleed bug, the attacker may be able to obtain uninterrupted read and write access to a data link protected by SSL. However, the attacker in general cannot access the whole memory spaces of the communicating nodes. We also acknowledge that MTD is not meant to provide perfect security. It is effective when the attacker can compromise the data links only. If the attackers can intrude into physical plant or sensors, they can launch stronger attacks and MTD can no longer protect the system. In this sense, the adoption of MTD will increase the bar for the attackers to launch successful attacks to drive the system into unsafe states.

VII. Related Work

Ensuring the safe operation of CPSes is always critical. Towards the safety threats from the unpredictable and diversified cyber attacks, Cárdenas et al. [22] presented a high-level analysis of the vulnerabilities of the CPS and defined certain attack models. One of the popular attacks is the Denial-of-Service (DoS) attacks, in which the attackers jam the communication channel to disrupt the normal operation of the CPS. Another one is the deception attacks, in which the attackers tamper with the control signals and/or sensor measurements to derail the system's safe operations.

Several integrity attacks that corrupt either the control signals or sensor measurements have received research attention. For instance, Pasqualetti et al. [18] studied the attack detection and identification if the control commands are corrupted. The *zero dynamics attacks*, which are constructed off-line, tamper with the control signals only [13], [23]. Lakshminarayana et al. [24] provided the basic understanding on the impact limit of the attacks on the sensor measurements based on the trade-off between attack impact and stealthiness. Vu et al. [4] studied the detection of the FDI attacks on either the sensor measurements or control signals based on the dissipativity-theoretic fault detector. However, detecting the FDI attacks on both the sensor and control data, i.e., SL attacks, has received limited research,

since the SL attacks are considered to be undetectable by any anomaly detectors [4], [25].

MTD approaches can enhance network security [6], and power grid SE's security [26]. The *physical watermarking* can be considered an MTD approach. The defenders actively inject additional secret excitations into the control signals, expecting to see the corresponding changes in the system outputs. Mo et al. [9] first applied the watermarking approach to detect replay attacks. However, since the additional excitations also induce fluctuations in the system state, the defenders face a trade-off between the attack detection accuracy and system operation performance. In contrast, our MTD approach has no influence on the converged system state. Watermarking approaches were also extended to detect more general data integrity attacks [19]. However, we show through simulations and analysis in Section V-B1 that these watermarking approaches are ineffective in detecting the MISAs. Another MTD approach actively perturbs the system structure. Teixeira et al. [23] applied the MTD concept to reveal zero dynamics attacks. Weerakkody and Sinopoli [27] proposed an MTD approach through the extended dynamic physical plant, while the nominal operation of the original system remains unchanged. However, this approach increases the system manufacturing and operational overheads, and may also induce new risks to the system (e.g., instability). In contrast, our approach is free from these issues.

VIII. Conclusion and Future Work

This paper studies the SL attacks that aim at disrupting the CPS. First, we present the taxonomy of SL attacks, and study the construction and implication of MISA, scaling attack and replay attack. Then, we propose the MTD-based SL attack detection framework, and study the design of MTD against different SL attacks. Lastly, a case study of detecting SL attacks against the SVC in power grids is presented with extensive simulation results under realistic settings.

We now discuss several issues that can be studied in future work. First, the attack identification and isolation through the MTD approach is an interesting issue. On the detection of the SL attack, it is desirable to identify the attacks, i.e., to identify which actuators or sensors have been compromised. According to the attack identification results, the impact of the attacks can be mitigated, or the attacks can be isolated. Pasqualetti et al. [18] showed the fundamental limitations in the attack identification of the defenders. Weerakkody and Sinopoli [28] further presented the superiority of the MTD approach in the identification of the malicious sensors. The identification when both the sensors and the actuators have been compromised (i.e., the SL attack) through the MTD approach can be studied in the future. Second, we wish to distinguish between the malicious attack and system failures in CPS through the MTD approach. The ability to distinguish a system fault and an attack is useful for the system operators to respond to the situation in an appropriate way. However, a fault and an attack are generally indistinguishable in that a fault may be considered as an attack in many cases. Meanwhile, an attack may also act as a critical fault to mislead the control center. However, the ability for an attacker to create an attack acting as a fault requires the knowledge of the system model. By limiting the attackers' knowledge through the MTD approach, a system fault and an attack may become distinguishable.

References

[1] S. K. Khaitan and J. D. Mccalley, "Design techniques and applications of cyberphysical systems: A survey," *IEEE Systems Journal*, 2015.

[2] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *IECON*, 2011.

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Info. & Syst. Security*, vol. 14, no. 1, 2011.

[4] Q. D. Vu, R. Tan, and D. K. Yau, "On applying fault detectors against false data injection attacks in cyber-physical control systems," in *IEEE INFOCOM*, 2016.

[5] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *Control Applications*, 2016.

[6] E. Al-Shaer, Q. Duan, and J. H. Jafarian, "Random host mutation for moving target defense," in *Intl. Conf. Security and Privacy in Commun. Syst.*, 2012.

[7] "Current transformer." [Online]. Available: http://www.idc-online.com/control/Current_Transformer.pdf

[8] J. Tian, R. Tan, X. Guan, and T. Liu, "Enhanced hidden moving target defense in smart grids," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, March 2019.

[9] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *IEEE Allerton Conf. Communication, Control, and Computing*, 2009.

[10] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[11] C. L. Demarco, J. V. Sariashkar, and F. Alvarado, "The potential for malicious control in a competitive power systems environment," in *IEEE International Conf. Control Applications*, 1996.

[12] H. Lin, A. Slagell, Z. T. Kalbarczyk, P. W. Sauer, and R. K. Iyer, "Runtime semantic security analysis to detect and mitigate control-related attacks in power grids," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 163–178, 2018.

[13] Y. Chen, S. Kar, and J. M. Moura, "Dynamic attack detection in cyber-physical systems with side initial state information," *IEEE Transactions on Automatic Control*, 2016.

[14] F. Harirchi, S. Z. Yong, E. Jacobsen, and N. Ozay, "Active model discrimination with applications to fraud detection in smart buildings," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 9527–9534, 2017.

[15] J. J. Gertler, "Survey of model-based failure detection and isolation in complex plants," *IEEE Control Systems Magazine*, vol. 8, no. 6, 1988.

[16] C. Kwon and I. Hwang, "Hybrid robust controller design: Cyber attack attenuation for cyber-physical systems," in *IEEE CDC*, 2013.

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

[18] F. Pasqualetti, F. Drfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Automatic Control*, 2013.

[19] B. Satchidanandan and P. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2016.

[20] J. P. Paul and J. Y. Leost, "Improvements of the secondary voltage control in france," *IFAC Proceedings*, vol. 20, no. 6, 1987.

[21] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[22] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

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

[24] S. Lakshminarayana, T. Z. Teng, D. K. Y. Yau, and R. Tan, "Optimal attack against cyber-physical control systems with reactive attack mitigation," in *ACM Intl. Conf. Future Energy Systems (e-Energy)*, 2017.

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

[26] M. A. Rahman, E. Al-Shaer, and R. B. Bobba, "Moving target defense for hardening the security of the power system state estimation," in *ACM Workshop on Moving Target Defense*, 2014.

[27] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *IEEE CDC*, 2015.

[28] ——, "A moving target approach for identifying malicious sensors in control systems," in *Allerton*, 2016, pp. 1149–1156.