Modeling and Mitigating Impact of False Data Injection Attacks on Automatic Generation Control

Rui Tan, *Member, IEEE,* Hoang Hai Nguyen, Eddy. Y. S. Foo, *Student Member, IEEE,* David K. Y. Yau, *Senior Member, IEEE,* Zbigniew Kalbarczyk, *Member, IEEE*, Ravishankar K. Iyer, *Fellow, IEEE*, Hoay Beng Gooi, *Senior Member, IEEE*

***Abstract*—This paper studies the impact of false data injection attacks on automatic generation control (AGC), a fundamental control system used in all power grids to maintain the grid frequency at a nominal value. Attacks on the sensor measurements for AGC can cause frequency excursion that triggers remedial actions such as disconnecting customer loads or generators, leading to blackouts and potentially costly equipment damage. We derive an attack impact model and analyze an *optimal attack*, consisting of a series of false data injections, that minimizes the remaining time until the onset of disruptive remedial actions, leaving the shortest time for the grid to counteract. We show that, based on eavesdropped sensor data and a few feasible-to-obtain system constants, the attacker can learn the attack impact model and achieve the optimal attack in practice. This paper provides essential understanding on the limits of physical impact of false data injections on power grids, and provides an analysis framework to guide the protection of sensor data links. For countermeasures, we develop efficient algorithms to detect the attack, estimate which sensor data links are under attack, and mitigate attack impact. Our analysis and algorithms are validated by experiments on a physical 16-bus power system testbed and extensive simulations based on a 37-bus power system model.**

***Index Terms*—Power grid, automatic generation control, false data injection, cyber security.**

I. Introduction

POWER grids maintain operation by various closed-loop control systems. Being at the interface between cyberspace intelligence and physical infrastructures, these control systems become attractive targets for cyber-attackers who aim at causing service outage and infrastructural damage. Recent high-profile intrusions such as the Stuxnet [1] and Dragonfly [2], [3] have alerted us to a general class of integrity attacks called *false data injection* (FDI) [4]. The Stuxnet worm attacked nuclear centrifuges by injecting false control commands and forging normal system states. Its design and architecture are not domain-specific [1]; they could be readily customized against other systems like power grids. Similarly, in Dragonfly, the attacker was able to gain access to power grid control systems. More generally, insider attacks are well documented [5] that occurred on critical infrastructures and produced severe consequences. Hence, research must address strong adversaries who are quite knowledgeable about their target control systems and have the ability to eavesdrop on and tamper with real-time data in the control loops.

In this paper, we study FDI attacks that corrupt real-time data in the feedback loop of *automatic generation control* (AGC) [6], a fundamental control system used in all power grids to maintain the grid frequency at its nominal value (50 or 60 Hz). In today's power grids, the real-time data for AGC is transmitted using computer networks, which may provide convenient venues for cyber-attackers to launch attacks. Moreover, AGC is an attractive target, because a successful FDI attack against AGC can cause catastrophic consequences. In a grid, imbalance between power generation and consumption will lead to deviation of the grid frequency from its nominal value. AGC maintains the grid frequency by adjusting the output power of generators based on measurements collected from sensors distributed in the grid. The grid frequency under AGC control is a *safety-critical global parameter* of the grid. A frequency deviation caused by an attack will propagate to the entire grid and trigger remedial actions such as disconnecting generators or customer loads. Such unscheduled actions may cause equipment damage and cascading failures leading to massive blackouts. Moreover, AGC is a highly automated system that requires minimal supervision and intervention by human operators. Once compromised, it may cause the grid frequency to deviate quickly.

Given its credibility and severe consequences, FDI against AGC has attracted initial research attention [7], [8], [9], [10], [11]. However, these studies were conducted in a constrained adversarial setting, by assuming that the attacker will follow limited predefined templates, such as injections of signal scaling, ramps, surges, and random noises [7], [10], [11], sign flip [11], and constant or random packet delays [8], [9]. Instead of following any prescribed templates, resourceful real-world attackers targeting critical infrastructures are likely

This work was supported in part under the Energy Innovation Research Programme (EIRP, Award No. NRF2014EWTEIRP002-026) administrated by the Energy Market Authority (EMA) of Singapore, in part by a Start-up Grant at NTU, 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), in part by U.S. Department of Energy under grant DOE-DE-OE0000780 (NETL), and in part by U.S. National Science Foundation under grant CNS 13-14891. The EIRP is a competitive grant call initiative driven by the Energy Innovation Programme Office, and funded by the National Research Foundation (NRF) of Singapore. The authors acknowledge Xinshu Dong for constructive discussions during the development of this work.

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

H. H. Nguyen, Z. Kalbarczyk, and R. K. Iyer are with Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL, USA (e-mail: hnguye11@illinois.edu; kalbarcz@illinois.edu; rkiyer@illinois.edu).

E. Y. S. Foo and H. B. Gooi are with School of Electrical and Electronic Engineering, NTU (e-mail: eddyfoo@ntu.edu.sg; ehbgooi@ntu.edu.sg).

D. K. Y. Yau is with Singapore University of Technology and Design and Advanced Digital Sciences Center, Illinois at Singapore (e-mail: david_yau@sutd.edu.sg).

to be strategic, and their tactics can adapt during attacks. For example, a preliminary phase of the attack may be designed to uncover system configurations and surveil real-time data to design FDIs that, in subsequent phases, will cause the largest frequency deviation. However, a basic understanding of such strategic AGC attacks that aim to maximize their physical impact is still lacking.

To advance our understanding, in this paper we study strategic attackers and analyze an *optimal attack* in which FDIs on sensor measurements for AGC mislead the grid frequency to exceed certain safety-critical thresholds within the shortest time, without tripping at any integrity checks on the sensor data. Such an attack leaves the shortest time for the grid to counteract before costly and possibly errant remedial actions must kick in. Understanding the optimal attack under various constraints on the attacker's capability (e.g., the number of sensor data links that he can compromise) provides practical insights on strengthening the security of AGC. For instance, we can assess which sensor data links should receive the highest priority for protection, so that the grid frequency can be kept within a safe region until an ongoing attack is detected and isolated. Note that in this paper we focus on FDI attacks against sensor data needed for the AGC. However, our analysis can be readily extended to address FDI attacks on other data types such as AGC commands sent to generators.

Our contributions in this paper are in answering the following three fundamental research questions.

First, how to formulate the optimal attack against the AGC? Based on a classical AGC model in power engineering, we derive a closed-form Laplace-domain model for the impact of a series of FDIs on the grid frequency. To the best of our knowledge, we provide a first rigorous analysis of this problem. Based on a time-domain counterpart of the derived model, we develop an efficient linear programming algorithm to compute the optimal attack.

Second, is the optimal attack achievable by the attacker? We answer this question in two respects: (i) knowledge needed about the grid to guide the attack, and (ii) computational overhead to craft the attack. For (i), our analysis shows that it is feasible for the attacker to learn the attack impact model stealthily, based on eavesdropped sensor data and a few system constants that are either public knowledge or can be obtained in an advanced persistent threat (APT) scenario (e.g., via social engineering against employees of the grid operator). Then, the attacker can use the learned model to compute the optimal attack. For (ii), the issue is whether the attacker can act fast enough to compute the optimal attack in real time, since a next attack step depends on the immediate past grid state. Our analysis shows that, if the attacker has sufficient parallel computing resources, the optimal attack computation has a polynomial time complexity of $\mathcal{O}(t^2 \cdot m)$, where t is the time for the grid frequency to cross the safety-critical thresholds and m is the number of sensors that characterizes the size of the grid. Our measurements on a working system prototype further confirm the feasibility of the attack. These results suggest the importance of understanding optimal FDI attacks and their implications.

Third, how does the defender detect, identify, and mitigate the attack? It is challenging to distinguish an attack from natural disturbances based on untrusted sensor data. We also show that the attack identification that pinpoints the compromised sensor data links is a combinatorial optimization problem. We develop efficient attack detection and identification algorithms. Our simulations for a case study system show that we can detect an optimal attack once launched, and shrink the solution space that contains the correct attack identification result by 99.7%, after 20 seconds from the onset of the attack. Moreover, we propose an approach that leverages the redundancy of measurement to mitigate the impact of an identified attack.

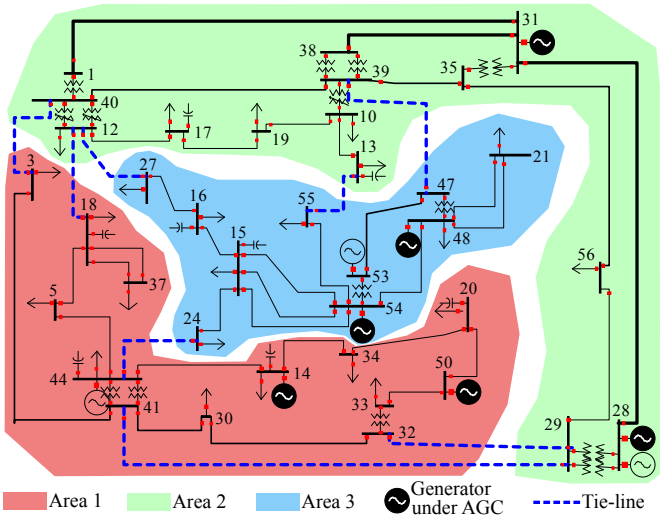


Fig. 1. A three-area 37-bus power grid. (Average line capacity: 160 MVA; total load: about 800 MW.)

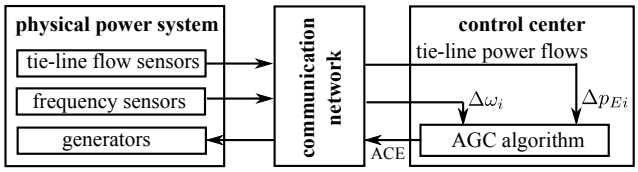
To validate and illustrate our analysis, we conduct extensive PowerWorld [12] simulations based on a 37-bus power system model. We compare the impact caused by the optimal attack and two limited attacks of random and surge injections [7]. We show that the limited attacks are ineffective because their effects can be corrected by the feedback control loop. Moreover, we conduct real experiments on a physical 16-bus power system testbed equipped with a 13.5 kVA generator and a variable load to demonstrate the achievability of the optimal attack in practice.

The balance of the paper is organized as follows. Section II presents preliminaries and related work. Section III defines the FDI attack. Sections IV, V, and VI address the three research problems that we outlined above. Sections VII and VIII present our PowerWorld simulations and testbed experiments, respectively. Section X concludes.

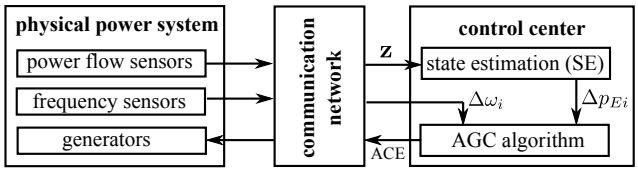
II. Preliminaries and Related Work

A. Preliminaries

AGC is a secondary control system that tunes the setpoints of generators' primary control systems to maintain the grid frequency at its nominal value. For trading purposes, AGC also maintains the power export (or import) of each *area*, which is part of a grid and typically operated by a utility company [6]. A transmission line connecting two buses belonging to two areas is called a *tie-line*. Fig. 1 illustrates a three-area grid



(a) Legacy scheme, where tie-line power flow measurements are directly used as input to AGC.



(b) A new scheme considered in this paper, where state estimation is used to improve the tie-line power flow measurements for AGC.

Fig. 2. Two AGC schemes considered in this paper.

TABLE I
SUMMARY OF NOTATIONS*

Symbol	Definition	Symbol	Definition
$\Delta\omega_i$	grid freq. deviation	$\Delta\omega$	avg grid freq. deviation
ϵ_L	$\Delta\omega$ lower bound	ϵ_U	$\Delta\omega$ upper bound
Δp_{Ei}	power export deviation	ACE_i	area control error
α_i, β_i	AGC algorithm constants	m	number of sensors
$\mathbf{z}$	measurement vector	N	number of areas
$\mathbb{W}$	corruptible $\mathbf{z}$ element indices	$\mathbf{F}$	measurement matrix of SE
H	regression horizon of Eq. (4)	L	number of tie-lines
$\mathbf{a}$	FDI attack vector	$\mathbf{c}$	injected SE error
$\mathbf{a}_{\min}$	lower bound for $\mathbf{a}$	$\mathbf{a}_{\max}$	upper bound for $\mathbf{a}$
Δp_i	change of load	ℓ_{ij}	tie-line from area i to j
$\Delta\mathbf{p}$	$= [\Delta p_1, \ldots, \Delta p_N]^\intercal$	$\mathbf{u}_h$, $\mathbf{v}_h$	coefficients in Eq.(4)
Δp_{ij}	power flow deviation of ℓ_{ij}	$\mathbf{\Psi}$, $\mathbf{\Lambda}$, $\mathbf{\Phi}$	parameters of Eq. (3)
$\mathbf{T}$	$(\mathbf{Tz})[i]$ is a tie-line flow	Δp_{Mi}	mechanical power change
$G_i(s)$	speed controller transfer func.	$T_i(s)$	turbine transfer function
K_i, R_i	generator constants	D_i	load-damping constant
M_i	total generator inertia	$\boldsymbol{\theta}$	$= \frac{1}{N} \cdot [1, 1, \ldots, 1]^\intercal$
$\mathbf{P}$	$(\mathbf{Pz})[i]$ is load of area i	$e(k)$	frequency prediction error

* Subscript i refers to area i. "Tie-line" refers to virtual tie-line.

with 37 buses,[1] where the dotted lines represent the tie-lines. The power export of an area, i.e., the total power transmitted over all the tie-lines from the area, is maintained at a scheduled value by AGC. For the ith area, based on measured deviations of the grid frequency and the power export from their nominal values (denoted by $\Delta\omega_i$ and Δp_{Ei}), the *area control error* (ACE) is $\mathrm{ACE}_i = \alpha_i \cdot \Delta p_{Ei} + \beta_i \cdot \Delta\omega_i$, where α_i and β_i are constants. AGC adjusts the input mechanical power setpoints of the generators to maintain the ACE for each area at zero [6]. We note that both the frequency deviation and the power export deviation are important to AGC, and they jointly determine the ACE. We refer to [14] for detailed analysis on the recommended settings for the weights α and β. Usually, only a subset of the generators are under AGC. The ACE is updated every *AGC cycle* that is typically two to four seconds [6], and sent to generators to determine their setpoints.

Tie-line power flow sensors that measure Δp_{Ei} can be noisy and faulty. *State estimation* (SE) [4] can reduce measurement noise and detect faulty sensor data. However, due to limited compute capability in the past, legacy power grids often execute SE at five-minute (or longer) intervals and apply it for the tertiary control of economic dispatch, rather than the secondary control of AGC executed at seconds intervals. Thus, as illustrated in Fig. 2(a), in legacy systems, the measurements from the frequency sensors and the tie-line power flow sensors are directly used by the AGC algorithm to compute the ACE. Recently, high-performance computing has significantly reduced the execution time of SE [15] and made it feasible for AGC. In this paper, we consider a new scheme that pipelines SE and AGC to address the latest idea that SE can enhance AGC's reliability [16]. As illustrated in Fig. 2(b), the measurements from the power flow sensors including the tie-line flow sensors will be first processed by the SE program to reduce measurement noise and remove faulty data. SE is executed every AGC cycle and the Δp_{Ei} from the SE result is used to compute ACE. However, it is easy to remove SE from our analysis if we need to model the legacy systems as shown in Fig. 2(a) faithfully, which will be discussed in Section III-A in detail. The simulations in Section VII consider both schemes shown in Figs. 2(a) and 2(b).

We consider the DC state estimation (DCSE), which accounts for active power only. Note that AGC aims to control active power only as well. Although DCSE is less accurate than the AC state estimation (ACSE) that considers both active and reactive power, DCSE is much faster and more robust than ACSE due to the linearization to the power flow model. Let $\mathbf{z} = [z_1, z_2, \ldots, z_m]^\intercal$ denote all the power flow sensors' measurements. The grid state, denoted by $\mathbf{x} = [x_1, \ldots, x_n]^\intercal$, consists of voltage angles of all the buses. The relationship between $\mathbf{z}$ and $\mathbf{x}$ is $\mathbf{z} = \mathbf{Fx} + \mathbf{e}$, where $\mathbf{F}$ is the *measurement matrix* and $\mathbf{e}$ is the noise. DCSE estimates $\mathbf{x}$ as $\hat{\mathbf{x}} = (\mathbf{F}^\intercal\mathbf{V}\mathbf{F})^{-1}\mathbf{F}^\intercal\mathbf{V}\mathbf{z}$, where $\mathbf{V}$ is a weight matrix. The power export deviation Δp_{Ei} can be computed from an improved measurement vector $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. The DCSE's *bad data detection* (BDD) raises an alarm if $\|\mathbf{z} - \hat{\mathbf{z}}\|_2$ is greater than a threshold [4].

Fig. 2 overviews the AGC. A *control center* of the area i collects transmission line power flow measurements and frequency measurements from distributed sensors. If SE is not employed, the control center computes ACE_i directly from the tie-line power flow and frequency measurements. Otherwise, it estimates the grid state $\hat{\mathbf{x}}$ using SE and computes ACE_i based on the frequency measurement and the Δp_{Ei} obtained from $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. Then, the control center transmits the ACE_i to the generators. This process is performed every AGC cycle. If the attacker can corrupt the power flow measurements, the control center will compute wrong ACEs. As a result, the generators will wrongly adjust their input mechanical power, leading to grid frequency deviations. Table I summarizes the notations used in this paper.

B. Related Work

As discussed in Section I, existing studies on the security of AGC [7], [8], [9], [10], [11] adopt limited attack templates that cannot well characterize real-world attackers. Reachability algorithms have been used to check the existence of a series

[1]We use the 37-bus model in Fig. 1 as a case study system throughout this paper. Its scale generally represents small-/mid-scale grids. According to our rough count based on a grid topology database [13], a major fraction of 130 national grids consist of less than 37 buses.

of FDI attacks that will lead to the breach of a safety condition [17], [18]. In contrast to *qualitative* reachability analysis, we compute the minimum time until the grid frequency deviates to an unacceptable value, which provides a *quantitative* vulnerability metric in a worst-case sense. Moreover, the studies in [17], [18] fall short of developing countermeasures such as attack detection and identification. Post-contingency power system safety under tie-line and frequency control has been studied in [19]. However, it does not specifically address power grid cybersecurity.

Liu et al. [4] analyze the conditions for FDI attacks on the sensor measurement $\mathbf{z}$ to bypass the BDD of SE. Specifically, if an attacker adds an *attack vector* $\mathbf{a} = \mathbf{Fc}$ to $\mathbf{z}$, where $\mathbf{c}$ is an arbitrary vector, the BDD cannot detect the attack and the grid state will be estimated wrongly as $\hat{\mathbf{x}} + \mathbf{c}$. Hendrickx et al. [20] show that the problem of minimizing the number of non-zeros in $\mathbf{a}$ is NP-hard. The minimization result can be used to identify attack-sensitive sensor links and guide the allocation of protection resources. Liu et al. [21] propose an FDI attack detector by fusing the heterogeneous data from the power system and the information network. The FDI attacks can mislead grid operations. Rahman et al. [22] construct a model checker to search for attack vectors that can increase the grid's generation cost by a specified percentage. The physical impact of FDI attacks has received little attention. In this paper, we analyze this impact in terms of disruptions of the grid frequency.

Beyond power grids, the security of a broader class of cyber-physical systems has received increasing attention. Amin et al. [23] perform threat assessment of water supply SCADA systems. Cárdenas et al. [24] study the impact of attacks on a chemical reactor process control system. The optimal attack analysis approach advanced in this paper can likewise be applied to other cyber-physical control systems besides AGC. In [25], [26], fundamental limits of secure SE, as well as attack detection and identification, are studied under a general linear control system model. They consider arbitrary FDI attacks on the sensor and control data. However, they fall short of analyzing the attacks' optimality.

In addition to FDI attacks, physical attacks against power equipment and attacks that cause contingencies of substations and transmission lines are also studied. Li et al. [27] apply an autoregression-based quickest detection approach to detect deviations of the voltage signal from its nominal sinusoid waveform, which can be caused by physical attacks to power equipment. In contrast to the heuristic autoregression model in [27], the attack detection model developed in this paper stems from a dynamic model of AGC in power engineering. Zhu et al. propose a set of power grid vulnerability assessment approaches when a subset of substations and transmission lines fail synchronously [28] or sequentially [29] due to attacks. Different from our work that investigates the attacker's strategy to inject false data to cause control failures, the vulnerability assessment studies in [28], [29] assume failures of transmission lines and substations in the first place.

Our prior work [30] presented the formulation of optimal attack against AGC and investigated the knowledge needed about the grid to guide the attack. Based on [30], we make the following two novel contributions in this paper. First, we analyze and evaluate the time complexity for the attacker to compute the optimal attack in real time. Second, we develop algorithms to detect attack, identify which sensor data links are under attack, and mitigate the attack impact. Extensive simulations are conducted to evaluate these countermeasures.

III. Attack Model

A. Attacker's Constraints

In this paper, we focus on a general class of FDI attacks on the power flow sensor measurement vector $\mathbf{z}$, which can be achieved by compromising physical sensors, sensor data communication links, and data processing programs at the control center. Hacking geographically distributed physical sensors is tedious and hard to coordinate. Although compromising computer programs at the strongly protected control center is not impossible given existing similar attacks [1], [2], targeting the sensor data links may pose a lower bar for the attacker. To be cost effective, power grids often leverage existing network infrastructures (e.g., those leased from third-party service providers) and set up virtual private networks (VPNs) as logically isolated channels to collect data from the distributed sensors [10], [31]. However, such software-based protection cannot guarantee security, because of pervasive software vulnerabilities. For instance, by exploiting the Heartbleed bug, the attacker may be able to obtain uninterrupted read and write access to a sensor data link protected by SSL. The attacker can also launch stepping stone attacks and compromise the VPN software providers first as in the Dragonfly attacks [2]. With these abilities, the attacker can mount the attack at a few central spots of the communication network to tamper with the data from many sensors.

ACE signals and frequency measurements are two other important data streams in AGC's control loop. The data links from the control center to the generators for transmitting ACE signals are usually well protected (e.g., by physically isolated cables) because of their limited quantity. For instance, in Fig. 1, at most nine links to the generators need to be protected, whereas there are 81 sensors feeding the SE and AGC. The grid frequency is a global parameter of the grid. Its measurements by remote sensors can be easily verified by frequency sensors inside the secured control center. Thus, the FDI attacks on the frequency measurements can be easily detected and isolated. These observations motivate us to focus on FDI attacks on power flow measurements in $\mathbf{z}$. In AGC, the controls of frequency and area power exports are coupled, in that the overall area control error is computed as the weighted sum of the frequency and area power export control errors. When the frequency measurements are intact, wrong area power export deviation information due to the FDI attacks on power flow measurements will lead to wrong area control errors, which will mislead the generators to generate wrong amounts of power and result in frequency deviations.

For an FDI attack on $\mathbf{z}$ to be stealthy, it needs to bypass the BDD of SE. Moreover, the grid operator may apply other data quality checks on $\mathbf{z}$. For instance, $\mathbf{z}$ should not change significantly over a short time period. Intuitively, if

each element of the FDI attack vector $\mathbf{a}$ is bounded around zero, these data quality checks, designed to be insensitive to natural random noises in $\mathbf{z}$, will not be alerted. In this paper, we consider an FDI attack that satisfies the following two assumptions:

(1) Attack's stealthiness: There exist constant vectors $\mathbf{a}_{\min}$ and $\mathbf{a}_{\max}$ where $\mathbf{a}_{\min} \preceq \mathbf{0} \preceq \mathbf{a}_{\max}$, such that for any FDI attack vector $\mathbf{a}$, the compromised measurement vector, i.e., $\mathbf{z}+\mathbf{a}$, can pass all the data quality checks if

$$\mathbf{a} = \mathbf{F}\mathbf{c} \quad \text{and} \quad \mathbf{a}_{\min} \preceq \mathbf{a} \preceq \mathbf{a}_{\max}, \tag{1}$$

where $\mathbf{c}$ is an arbitrary vector and $\mathbf{a} = \mathbf{F}\mathbf{c}$ is the bypass condition of DCSE's BDD [4]. Note that $\mathbf{x} \preceq \mathbf{y}$ means that each element of $\mathbf{x}$ is no greater than the corresponding element of $\mathbf{y}$. We assume that the attacker knows $\mathbf{F}$, $\mathbf{a}_{\min}$, and $\mathbf{a}_{\max}$ to construct attack vectors satisfying Eq. (1). Otherwise, the compromised measurement vectors will be discarded and the injected data will not enter the control loop. In this paper we focus on FDIs that can enter the control loop. Note that, to address ACSE, the $\mathbf{a} = \mathbf{F}\mathbf{c}$ in Eq. (1) should be replaced by the bypass condition of ACSE's BDD (e.g., [32]).

(2) Attacker's access to sensor measurements in $\mathbf{z}$: We assume that the attacker has read access to the power flow measurements in $\mathbf{z}$. The attacker has write access to a subset of the elements in $\mathbf{z}$. Denote by $\mathbb{W}$ the set of indices of $\mathbf{z}$ elements writable by the attacker and $\mathbf{a}[j]$ the jth element of an attack vector $\mathbf{a}$. Thus, the attack vector $\mathbf{a}$ is subject to

$$\mathbf{a}[j] = 0, \quad \forall j \notin \mathbb{W}. \tag{2}$$

Our formulation of the optimal attack will incorporate Eqs. (1) and (2) as constraints for the attacker. From [4], the attacker generally needs to tamper with a number of power flow measurements (not just the tie-line power flow measurements) to satisfy $\mathbf{a} = \mathbf{F}\mathbf{c}$. Thus, pipelining SE and AGC increases the requirement to launch the attack. Legacy power grids do not apply SE for AGC. By ignoring the condition $\mathbf{a} = \mathbf{F}\mathbf{c}$ in Eq. (1), our analysis in this paper addresses legacy grids faithfully. In such case, by tampering with the tie-line power flow measurements only, the attacker can mislead AGC.

To simplify the exposition, we assume that the measurement noise $\mathbf{e}$ is negligible (i.e., $\mathbf{e} = \mathbf{0}$). Under this assumption, we can verify that the improved measurement vector by SE, i.e., $\hat{\mathbf{z}}$, is the same as the possibly compromised measurement vector. In other words, an attack vector injected into the raw measurement vector is not altered by the SE. When $\mathbf{e}$ is not negligible, we can address the alteration by replacing all the $\mathbf{a}$ in the rest of this paper with $\mathbf{F}(\mathbf{F}^\intercal\mathbf{V}\mathbf{F})^{-1}\mathbf{F}^\intercal\mathbf{V}\mathbf{a}$, where the latter is the altered attack vector. Note that, whether $\mathbf{e}$ is negligible or not, the original attack vector $\mathbf{a}$ needs to satisfy Eqs. (1) and (2).

B. Attacker's Objective

The attacker's objective is to cause unsafe frequency deviations. As the grid frequency is a *safety-critical global parameter* of the grid, an unsafe frequency deviation caused by the attack will propagate to the entire grid and trigger remedial actions such as disconnecting generators or customer loads. The disconnected region of customers will be severely affected in that they will be totally without power for some period of time. These customers may suffer significant loss of comfort or money, and their personal safety could be endangered in certain situations. In particular, the unscheduled generator/load disconnection actions may cause permanent equipment damage and cascading failures leading to massive blackouts. Thus, the economic loss and societal impact of a successful attack against AGC can be tremendous, whereas the FDI attacks on the power flow measurements that are both technically and economically possible as implied by the recent high-profile intrusions such as Stuxnet [1] and Dragonfly [2].

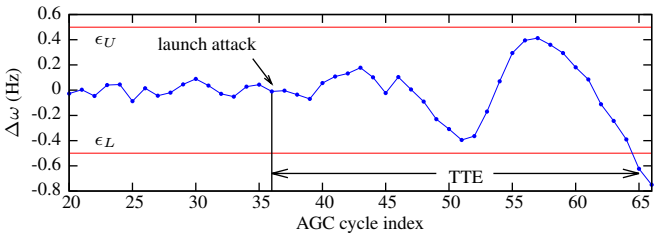


Fig. 3. Illustration of TTE. The attacker launches an attack sequence from the 36th AGC cycle and the frequency goes out of the safety region (-0.5 Hz, 0.5 Hz) at the 65th AGC cycle. The TTE for this attack sequence is then $65 - 36 = 29$ AGC cycles.

Because of the constraints in Eqs. (1) and (2) and the system's inertia, the attacker may not be able to cause an unsafe frequency deviation in a single AGC cycle. To overcome the system inertia, he can continuously launch FDI attacks over multiple AGC cycles to achieve an unsafe frequency deviation, where each FDI attack vector is subjected to the constraints in Eqs. (1) and (2). Specifically, if the current AGC cycle index is k and the attacker launches a series of FDI attacks from the $(k+1)$th to the $(k+h)$th AGC cycle with attack vectors $\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}$, the sequence $\{\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}\}$ is called an *attack sequence*. The following metric characterizes the effectiveness of a certain attack sequence.

Time-to-emergency (TTE): Given a safety region (ϵ_L, ϵ_U) where $\epsilon_L < 0 < \epsilon_U$, TTE is the time from the onset of an attack sequence to the first time instant when the average frequency deviation of all the areas, denoted by $\Delta\omega$, is out of (ϵ_L, ϵ_U).

The thresholds ϵ_L and ϵ_U can be set to those for triggering remedial actions. For example, we can set $\epsilon_L = -0.5$ Hz and $\epsilon_U = 0.5$ Hz [6]. This setting is used for all the simulations and testbed experiments in this paper. Fig. 3 illustrates the TTE. We now formally define the *optimal attack sequence*.

Optimal attack sequence: Suppose the attacker launches a series of FDI attacks from the $(k+1)$th AGC cycle. An optimal attack sequence $\{\mathbf{a}_{k+1}, \mathbf{a}_{k+2}, \ldots\}$ is the attack sequence that minimizes the TTE given a safety region (ϵ_L, ϵ_U), where each FDI attack vector $\mathbf{a}$ in the attack sequence satisfies Eqs. (1) and (2).

In this paper, we use the terms *optimal attack sequence* and *optimal attack* interchangeably. The optimality of attack is defined in terms of TTE, because TTE is a key metric to

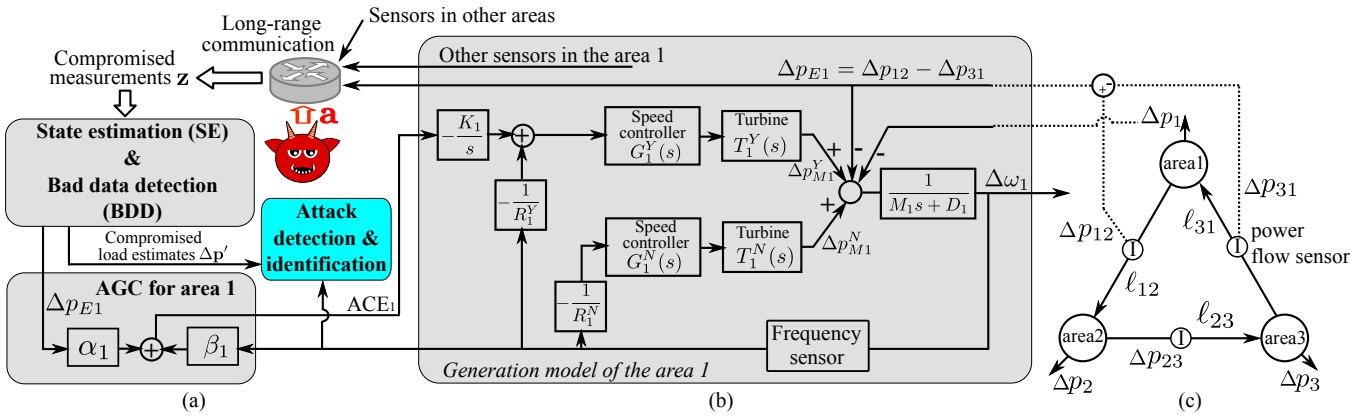


Fig. 4. (a) SE, BDD, AGC programs, attack detection and identification; (b) Block diagram of the generation model for the area 1; (c) Virtual tie-lines of the three-area grid in Fig. 1. Notation explanation: Δp_{ij} is the deviation of the power flow over ℓ_{ij} from its scheduled value; $G_i(s)$ and $T_i(s)$ are transfer functions of the speed controller and the turbine of a generator, respectively; Δp_{Mi} is change of input mechanical power; gain K_i; droop constant R_i; total generator inertia M_i; load-damping constant D_i; superscripts 'Y' and 'N' modify the symbols for the generators under and out of AGC, respectively.

assess the effectiveness of an attack and to guide the design of countermeasures. Specifically, from the attacker's perspective, the optimal attack leaves the shortest time for the grid to counteract. Therefore, an intelligent attacker will design the attack sequence to minimize TTE. From the defender's perspective, the minimum TTE is an important system resilience metric, since the attack detection and isolation need time, as shown in Section VI. For instance, the defender can assess which sensor data links should receive the highest priority for protection, so that the minimum TTE is larger than the time delay in detecting and isolating an ongoing attack.

IV. Optimal Attack Sequence

This section derives the models of the impact of an attack sequence on the grid frequency, which include a Laplace-domain model in Section IV-A and a time-domain approximation model in Section IV-B. Based on the attack impact models, in Section IV-C, we present an algorithm to compute the optimal attack sequence that minimizes the TTE.

A. Laplace-Domain Attack Impact Model

In this section, we establish the mathematical relationship between the attack sequence and the grid frequency. To the best of our knowledge, this is a first rigorous analysis of this problem. To derive the relationship, we extend Fig. 2(b) as Fig. 4 to incorporate more details of AGC. Several symbols in Fig. 4 are defined as follows. For an N-area grid, denote by ℓ_{ij} a *virtual tie-line* from area i to area j. The power flow over ℓ_{ij} is the sum of power flows over all the real tie-lines from area i to area j. For instance, Fig. 4(c) illustrates the virtual tie-lines of the three-area grid in Fig. 1. Denote by $\Delta\omega_i$ and Δp_i the frequency deviation and the change of load in area i, respectively; $\Delta\omega$ the average of the frequency deviations of all the areas; $\Delta\mathbf{p} = [\Delta p_1, \ldots, \Delta p_N]^\intercal$. Suppose there are a total of L virtual tie-lines. Let $\mathbf{T}$ represent an $L \times m$ matrix (m is the number of power flow sensors) that consists of -1, 0, and 1, and aggregates the real tie-line power flows in $\mathbf{z}$ as virtual tie-line power flows. That is, an element of $\mathbf{Tz}$ is the power flow over a virtual tie-line. Following existing approaches [6], we model the two sets of generators under and out of AGC in an area as two *virtual generator*s, respectively.[2] Fig. 4(b) shows a block diagram of a widely adopted Laplace-domain model [6] for the two virtual generators. Other symbols in Fig. 4 are briefly explained in the figure caption.

From a control-theoretic perspective, in the presence of FDI attacks, an AGC system can be viewed as an open-loop system with the load change $\Delta\mathbf{p}$ and the FDI attack vector $\mathbf{a}$ as the inputs, and the frequency deviation $\Delta\omega$ and the area power export deviations as the outputs. In this section, we treat $\mathbf{a}$ as a vector of continuous-time variables. Denote by s the Laplace coordinates and $\widetilde{x}$ the Laplace transform of x. Based on the model in Fig. 4, the output $\Delta\omega$ is given by the following equation (a detailed derivation is omitted due to space limitations and can be found in [33]):

$$\widetilde{\Delta\omega} = \boldsymbol{\theta}^\intercal \boldsymbol{\Phi}^{-1} \widetilde{\Delta\mathbf{p}} + \boldsymbol{\theta}^\intercal \boldsymbol{\Phi}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Psi} \mathbf{T} \widetilde{\mathbf{a}}, \tag{3}$$

where $\boldsymbol{\theta} = \frac{1}{N} \cdot [1, 1, \ldots, 1]^\intercal \in \mathbb{R}^{N\times 1}$; $\boldsymbol{\Lambda} = \mathrm{diag}(s \cdot n_1 - 1, \ldots, s \cdot n_N - 1)$ and the expression of n_i will be presented in Section V-A2 when used; $\boldsymbol{\Psi}$ is an $N \times L$ matrix consisting of -1, 0, and 1; $\boldsymbol{\Phi}$ is an $N \times N$ matrix and its elements are expressions of the generators' transfer functions (i.e., $G_i^N(s)$, $G_i^Y(s)$, $T_i^N(s)$, and $T_i^Y(s)$). Note that the ACE signal has been considered in the derivation of Eq. (3). The parameters for computing ACE, i.e., α and β, are contained in the matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$. As the detailed expressions of $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ are not used in this paper, they are omitted but can be found in [33]. Eq. (3) is a complicated model in the Laplace domain. As discussed in next section, it is intractable to compute the optimal attack sequence based on Eq. (3) and simplification of Eq. (3) will be needed. However, as analyzed in Section V, Eq. (3) is key for the attacker to learn the attack impact models stealthily and achieve the optimal attack.

[2]We assume that all generators in the power grid are synchronous generators. Thus, the two virtual generators in Fig. 4(b) are synchronous generators. Although our analysis does not address renewable energy sources (RES) that often employ asynchronous generators, we will discuss how to extend our analysis and algorithm to address RES wherever applicable.

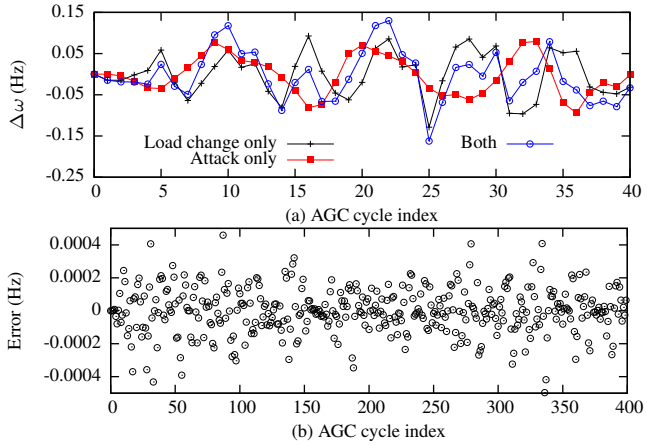


Fig. 5. Validating additive property of Eq. (3).

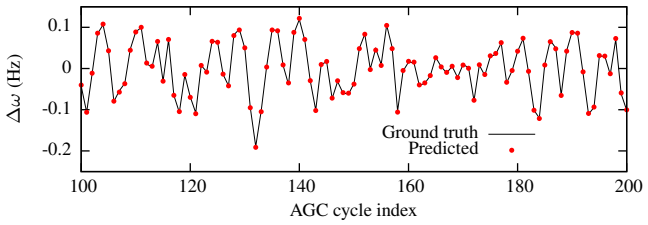
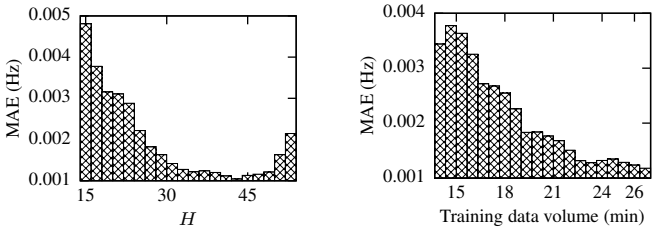


Fig. 6. Predicted $\Delta\omega$ based on regression. Prediction horizon H is 34; mean absolute prediction error is 0.0014 Hz.



(a) MAE vs. H (training data volume: 400 AGC cycles)

(b) MAE vs. training data volume ($H = 43$)

Fig. 7. Prediction error under various settings.

B. Regression-Based Attack Impact Model

As TTE is a time-domain metric, it is intractable to find a fastest attack sequence using Eq. (3) that is expressed in the Laplace domain. Thus, we need to convert Eq. (3) to an equivalent time-domain model. However, the inverse Laplace transform of Eq. (3) is a set of extremely complex differential equations, especially when the generators' transfer functions $G_i(s)$ and $T_i(s)$ are complex. Even if the inverse Laplace transform can be discretized, an exhaustive search may be the only viable solution to the TTE minimization. The high compute overhead will render the optimal attack computationally impractical. This section proposes a linear regression model based on a key observation from Eq. (3).

From Eq. (3), $\Delta\mathbf{p}$ and $\mathbf{a}$ produce *additive* impacts on $\Delta\omega$. From the linearity principle of Laplace transform, this additive property also holds in the time domain. To validate this, we conduct simulations using PowerWorld [12], a high-fidelity power system simulator. For the grid in Fig. 1, we run simulations driven by randomly generated traces for $\Delta\mathbf{p}$ and $\mathbf{a}$. The trace for $\Delta\mathbf{p}$ is generated by scaling the steady-state load of each load bus by a zero-mean Gaussian random variable of standard deviation 0.02 per unit (p.u.), while each element of $\mathbf{a}$ is randomly and uniformly sampled from $[-5\,\text{MW}, 5\,\text{MW}]$. Fig. 5(a) plots $\Delta\omega$ when the simulation is driven by the $\Delta\mathbf{p}$ trace only, the $\mathbf{a}$ trace only, and both traces. Fig. 5(b) plots the difference between the third curve and the sum of the first two curves in Fig. 5(a). The errors are two orders of magnitude lower than $\Delta\omega$ in Fig. 5(a).

Based on the additive property, we propose an attack impact model based on linear regression. Denote $\Delta\omega(k)$, $\Delta\mathbf{p}_k$, and $\mathbf{a}_k$ the grid frequency deviation, the load change vector, and the attack vector in the kth AGC cycle, respectively. The model is

$$\Delta\omega(k) = \sum_{h=0}^{H-1} \mathbf{u}_h^\intercal \Delta\mathbf{p}_{k-h} + \mathbf{v}_h^\intercal \mathbf{T}\mathbf{a}_{k-h}, \tag{4}$$

where H is the horizon of the regression, $\mathbf{u}_h \in \mathbb{R}^{N\times 1}$ and $\mathbf{v}_h \in \mathbb{R}^{L\times 1}$ are the coefficients that "encode" the coefficients $\boldsymbol{\theta}^\intercal\boldsymbol{\Phi}^{-1}$ and $\boldsymbol{\theta}^\intercal\boldsymbol{\Phi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Psi}$ in Eq. (3). Eq. (4) preserves the additive property of Eq. (3). Fig. 6 shows the trace of $\Delta\omega$ predicted from a trained regression model and the ground truth in the presence of load fluctuations and random FDI attacks. We can see that the model accurately predicts $\Delta\omega$. Fig. 7 shows the evaluation results for the regression model's accuracy in terms of mean absolute error (MAE). From Fig. 7(a), the MAE is convex against the regression horizon H, which is consistent with the intuition that older system states are less correlated with the current system state. The setting $H = 43$ yields the smallest MAE. In Fig. 7(b), the MAE decreases with the training data volume, which is also consistent with intuition. Moreover, from Fig. 7, the MAE is on the order of $0.001\,\text{Hz}$, which is insignificant compared with the natural fluctuations of the grid frequency on the order of $0.1\,\text{Hz}$.

C. Optimal FDI Attack Sequence

Based on Eq. (4), we develop an algorithmic formulation of an optimal FDI attack sequence that minimizes the TTE. Suppose $l \in \mathbb{Z}$ and $k \in \mathbb{Z}$ are the onset time of the attack and the current AGC cycle index, respectively, where $l \le k$. From Eq. (4), the frequency deviation in the $(k+h)$th AGC cycle is predicted by

$$\Delta\omega(k{+}h) = \begin{bmatrix} \mathbf{u}_{H-1} \\ \vdots \\ \mathbf{u}_{h+k-l+1} \\ \mathbf{u}_{h+k-l} \\ \vdots \\ \mathbf{u}_h \\ \mathbf{u}_{h-1} \\ \vdots \\ \mathbf{u}_0 \end{bmatrix}^\intercal \begin{bmatrix} \Delta\mathbf{p}_{k-H+h+1} \\ \vdots \\ \Delta\mathbf{p}_{l-1} \\ \Delta\mathbf{p}_l \\ \vdots \\ \Delta\mathbf{p}_k \\ \Delta\hat{\mathbf{p}}_{k+1} \\ \vdots \\ \Delta\hat{\mathbf{p}}_{k+h} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_{H-1} \\ \vdots \\ \mathbf{v}_{h+k-l+1} \\ \mathbf{v}_{h+k-l} \\ \vdots \\ \mathbf{v}_h \\ \mathbf{v}_{h-1} \\ \vdots \\ \mathbf{v}_0 \end{bmatrix}^\intercal \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{T}\mathbf{a}_l \\ \vdots \\ \mathbf{T}\mathbf{a}_k \\ \mathbf{T}\mathbf{a}_{k+1} \\ \vdots \\ \mathbf{T}\mathbf{a}_{k+h} \end{bmatrix}, \tag{5}$$

where $\Delta\hat{\mathbf{p}}_{k+1}, \ldots, \Delta\hat{\mathbf{p}}_{k+h}$ are the forecast load changes; $\mathbf{a}_l, \ldots, \mathbf{a}_k$ are the past attack vectors; $\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}$ are the future attack vectors to be optimized. If the attacker

Algorithm 1 To compute the optimal attack sequence.

Input: $\{\Delta\mathbf{p}_i | i \in [k-H+1, k]\}$, $\{\mathbf{a}_l, \ldots, \mathbf{a}_k\}$, $\{\mathbf{u}_h, \mathbf{v}_h | h \in [0, H-1]\}$, $h_{\max}$
Output: The attack sequence that minimizes the TTE

1: $h = 1$
2: **loop**
3: $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\} = \arg\max_{\{\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}\}} \Delta\omega(k+h)$ subject to that $\mathbf{a}_{k+i}$ satisfies Eqs. (1) and (2), $\forall i \in [1, h]$
4: compute $\Delta\omega^*(k+h)$ using $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\}$ and Eq. (5)
5: **if** $\Delta\omega^*(k+h) \geq \epsilon_U$ **then** return $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\}$
6: $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\} = \arg\min_{\{\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}\}} \Delta\omega(k+h)$ subject to that $\mathbf{a}_{k+i}$ satisfies Eqs. (1) and (2), $\forall i \in [1, h]$
7: compute $\Delta\omega^*(k+h)$ using $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\}$ and Eq. (5)
8: **if** $\Delta\omega^*(k+h) \leq \epsilon_L$ **then** return $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h}\}$
9: $h = h + 1$
10: **if** $h > h_{\max}$ **then** return with no solution
11: **end loop**

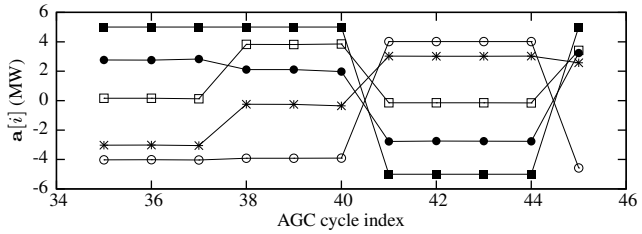


Fig. 8. Five elements of the optimal attack vectors.

has no access to the load forecast, he can set $\Delta\hat{\mathbf{p}}_{k+1} = \ldots = \Delta\hat{\mathbf{p}}_{k+h} = 0$. We propose Algorithm 1 to compute an attack sequence. Specifically, for each h starting from one, Algorithm 1 maximizes and minimizes $\Delta\omega(k+h)$ subject to the stealthiness and write access constraints in Eqs. (1) and (2), and stops once $\Delta\omega(k+h)$ exits the safety region defined by ϵ_U and ϵ_L. The parameter $h_{\max}$ is the stop condition for h, i.e., if $h > h_{\max}$, the algorithm will return with no solution. We have the following proposition.

Proposition 1. *Modulo the approximation error of Eq. (5), the solution computed by Algorithm 1 is the optimal attack sequence.*

Proof. The optimality of the solution given by Algorithm 1 can be proved by contradiction as follows. Suppose the solution given by Algorithm 1, denoted by $\{\mathbf{a}^*_{k+1}, \ldots, \mathbf{a}^*_{k+h^*}\}$, is not optimal and there exists a shorter attack sequence $\{\mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h'}\}$ where $h' < h^*$ such that $\Delta\omega(k+h') \notin (\epsilon_L, \epsilon_U)$. This supposition contradicts the fact that Algorithm 1 cannot find an attack sequence such that $\Delta\omega(k+h) \notin (\epsilon_L, \epsilon_U)$ and thus does not return when $h = h'$. □

Fig. 8 shows the time series of five elements of the attack vector $\mathbf{a}$ computed using Algorithm 1 for the three-area grid

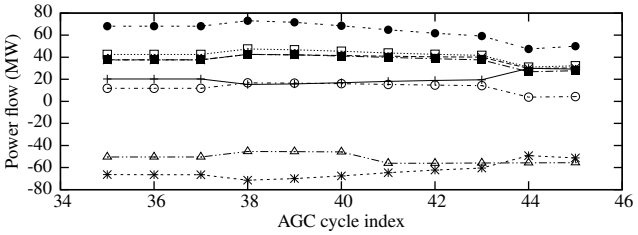


Fig. 9. The compromised tie-line power flows during attack.

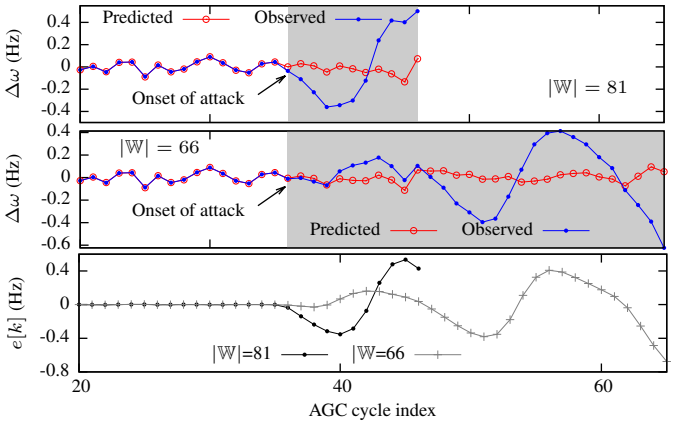


Fig. 10. Two examples of the effects of an optimal attack. The grid is under attack during the shaded periods. We stop a simulation once the frequency deviation exceeds the safety region $(-0.5\,\text{Hz}, 0.5\,\text{Hz})$. The curves labeled "Predicted" are the $\Delta\hat{\omega}(k)$ used to detect attacks as discussed in Section VI-A.

in Fig. 1, when the attacker has write access to all the 81 sensor data links. Each element of $\mathbf{a}_{\min}$ and $\mathbf{a}_{\max}$ is $-5\,\text{MW}$ and $5\,\text{MW}$, respectively. We can see that the attack vector changes over time. Fig. 9 shows the compromised measurements of the power flows through the eight tie-lines shown in Fig. 1 during the attack. Table II summarizes the profiles of the compromised tie-line power flows shown in Fig. 9. During the attack, the maximum deviation seen by the system operator for a single tie-line is $20.66\,\text{MW}$. Note that it is possible for the system operator to suspect an attack from the fluctuations of the tie-line power flows. Our aim, however, is to go fundamentally further by providing a reliable and automated method of attack detection, which is the subject of Section VI. The curve labeled 'Observed' in the top part of Fig. 10 shows the trajectory of $\Delta\omega$ when the attacker injects the attack sequence in Fig. 8. The safety condition defined by $\epsilon_L = -0.5\,\text{Hz}$ and $\epsilon_U = 0.5\,\text{Hz}$ is breached after 10 AGC cycles from the onset of the attack. We can see that the optimal attack sequence first misleads the system to reduce the grid frequency and then leverages the system's response to the frequency reduction to achieve an overshoot that breaches the safety condition. The bottom part of Fig. 10 shows the result when the attacker has write access to 66 sensor measurements. As now fewer measurements can be tampered with, the attacker takes a longer time to breach the safety condition. The optimal attack sequence exhibits a similar strategy, i.e., it leverages the system's response to achieve oscillation and overshoot.

Algorithm 1 is based on the analysis in previous sections that does not address renewable energy sources (RES) like wind and solar generators. With low RES penetration, its generation fluctuation can be regarded as *negative* load change and the attacker can still optimize his attack using Algorithm 1 if he can access the past and predicted RES generation. However, a high RES penetration may invalidate the steady-state assumptions of the AGC model [6]. The modeling of the grid frequency dynamics in the presence of high RES penetration is still being studied [34], [35] and further study is needed to understand its impact on our analysis.

TABLE II
PROFILES OF THE COMPROMISED TIE-LINE POWER FLOWS SHOWN IN FIG. 9

Tie-line (bus# - bus#)	3 - 40	12 - 18	32 - 29	29 - 41	12 - 27	13 - 55	39 - 47	24 - 44
Power flow before attack (MW)	20.33	37.56	66.38	42.52	37.72	11.84	68.12	-50.40
Maximum power flow deviation during attack (MW)	9.69	7.72	17.15	11.42	10.85	7.94	20.66	5.43

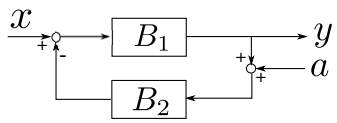


Fig. 11. A barebone example that illustrates a basic challenge of passive monitoring approach to learn attack impact model.

V. ACHIEVING OPTIMAL ATTACK

Through analysis, this section discusses whether and how an attacker can achieve the optimal attack. We show that, with a modest amount of prior knowledge and sufficient computation power, the attacker may learn the attack impact models and use Algorithm 1 to strategize his attack beyond the random or heuristic attacks studied in prior work [7], [8], [9], [10].

A. Learning Attack Impact Models

A model in either Eq. (3) or Eq. (4) is a prerequisite for computing the optimal attack sequence using Algorithm 1. However, such detailed models that describe the system dynamics may not be readily available. This is mainly because the real-time AGC control does not rely on these models. In this section, we discuss two approaches, *active probing* and *passive monitoring*, for the attacker to learn these models, starting from a modest amount of feasible-to-obtain prior knowledge about the grid. The former approach launches FDI attacks of small magnitudes to learn the model in Eq. (4), while the latter learns the model in Eq. (3) by passively eavesdropping on sensor data without actually tampering with them. Apparently, the latter approach is more stealthy.

1) Active Probing: The attacker injects a series of attack vectors of small magnitudes that satisfy the constraints in Eqs. (1) and (2) and cause grid frequency fluctuations similar to those caused by natural demand fluctuations, so that these small "probes" will neither alert the grid operator nor damage anything. For instance, in Fig. 5, the random FDIs of limited magnitudes introduce little changes to $\Delta\omega$. Meanwhile, the attacker keeps track of $\Delta\mathbf{p}$ and $\Delta\omega$. After accumulating enough data, he can apply linear regression to learn the model in Eq. (4). The attacker can treat $\mathbf{v}_h^\intercal\mathbf{T}$ in Eq. (4) as a single row vector. Thus, prior knowledge of $\mathbf{T}$ is not needed. Section VII will evaluate this approach.

2) Passive Monitoring: Based on passively eavesdropped sensor measurements only, we can learn the coefficient $\mathbf{u}_h$ in Eq. (4), but not $\mathbf{v}_h$. Thus, we fall back on the Laplace-domain model in Eq. (3), which preserves additional information about the coefficient of $\mathbf{a}$. Before presenting details of the passive monitoring approach, we use a barebone example to illustrate a basic challenge of the approach and a key to its success. Fig. 11 shows an abstract feedback system with scalar input x and output y, unknown scalar gains B_1 and B_2, and malicious injection a on the measurement of y. We can derive $y = \frac{B_1}{1+B_1B_2}x - \frac{B_1B_2}{1+B_1B_2}a$. Based on passively eavesdropped traces of x and y, the attacker can estimate the value of $\frac{B_1}{1+B_1B_2}$. However, he cannot estimate the individual values of B_1 and B_2, and thus cannot derive the coefficient for a, i.e., $-\frac{B_1B_2}{1+B_1B_2}$. But if he has additional prior information about B_1 and B_2, e.g., $B_1 = B_2$, he may be able to estimate B_1 and B_2, and derive the coefficient for a. For the more complex AGC system, we have the following proposition.

Proposition 2. *If the attacker knows the generator inertia M_i and the load-damping constant D_i in Fig. 4(b), the weights α_i and β_i of the AGC algorithm in Fig. 4(a), and $\mathbf{T}$ in Eq. (3), and he can eavesdrop on the time series of load change Δp_i, virtual tie-line power flow deviation Δp_{ij}, and frequency deviation $\Delta\omega_i$ for each area, he can apply system identification techniques to learn the attack impact model in Eq. (3).*

The proof, which provides a detailed learning procedure, is omitted here due to space limitations and can be found in [33]. Now, we discuss how the attacker can obtain the constants and time series data required by Proposition 2. In the second assumption of the attack model in Section III-A, we assume that the attacker can obtain the time series of $\mathbf{z}$ that contains Δp_i and Δp_{ij} for each area. He can also obtain the time series of $\Delta\omega_i$ by using his own frequency sensors plugged into any power outlets in the areas. The parameters M_i, D_i, α_i, β_i, and $\mathbf{T}$ are basic grid information. The attacker can launch data exfiltration attacks such as in the initial phase of the Dragonfly attack [3] to obtain them. The attacker can also try other ways that may be easier. The grid operator periodically estimates M_i and D_i, and uses them to configure various algorithms [6]. The attacker can steal their values by insiders or social engineering against employees of the grid. As defined in Section IV-A, $\mathbf{T}$ is a matrix that aggregates the real tie-line power flows in $\mathbf{z}$ as virtual tie-line power flows. It can be easily derived from the grid's topology graph (e.g., Fig. 1), which can be public knowledge or commercially available. For instance, an open database [13] provides the topology graphs of about 130 national grids. In [28], the authors were able to purchase the topological data of the Bay Area power grid in North America. The settings for α_i and β_i can also be public knowledge [36].

We conduct PowerWorld simulations for the grid in Fig. 1 and apply the passive monitoring procedure detailed in [33], where the elements of $\boldsymbol{\theta}^\intercal\boldsymbol{\Phi}^{-1}$ and $\boldsymbol{\Lambda}$ in Eq. (3) are identified as fourth- and second-order polynomial fractions of s, respectively. Fig. 12 shows the $\Delta\omega$ predicted using the learned model and the ground truth in the presence of random FDI attacks without load fluctuations. Thus, it specifically evaluates the performance of the learned model in characterizing the attack impact. The model is learned under different AGC

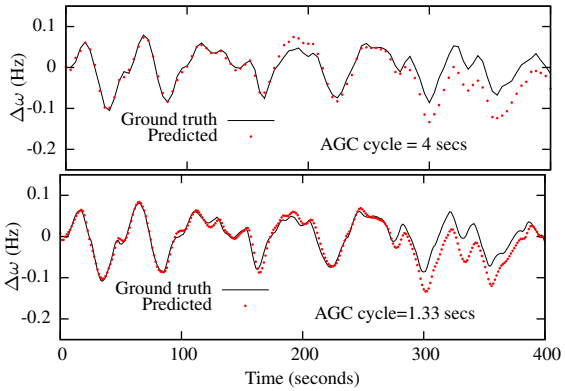
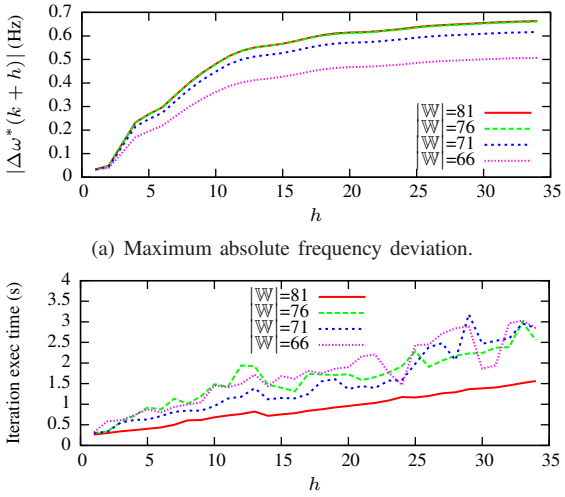


Fig. 12. Prediction using the model in Eq. (3) learned by passive monitoring under two settings of the AGC cycle (no load changes; training data length is 26.6 minutes).



(a) Maximum absolute frequency deviation.

(b) Execution time of an iteration of Algorithm 1.

Fig. 13. Results and execution time of Algorithm 1.

cycle lengths of 4 seconds and 1.33 seconds. The training data collection takes 26.6 minutes. Under both settings, the mean absolute errors, which are 0.021 Hz and 0.015 Hz, are comparable. This result shows the robustness of the approach to the AGC cycle length within its typical range (two to four seconds). Although the prediction error of this approach is higher than that of the active probing approach, which is on the order of 10^{-3} Hz as shown in Section VII, its performance is satisfactory when the prediction horizon is not long (e.g., 200 seconds).

As Algorithm 1 is based on the regression model in Eq. (4), the attacker can use the learned Laplace-domain model to generate simulated traces of $\Delta\omega$, $\Delta\mathbf{p}$, and $\mathbf{a}$ to train the regression model. Then, he can use Algorithm 1 to compute the optimal attack.

3) Discussion: It is not trivial to learn the attack impact model, and care is needed to obtain the required prior information, choose proper orders for the transfer functions, and prevent overfitting. However, these tasks are certainly within reach of skillful attackers. In Section VIII, we demonstrate an oracle implementation of the passive monitoring approach on a physical power system testbed. The evaluation results indicate its feasibility for real-world power grids.

B. Compute Overhead

From Algorithm 1, the optimal attack sequence depends on load changes in the immediate past. Thus, the attacker needs to complete the execution of Algorithm 1 within one AGC cycle before launching the attack. After the launch, he can rerun Algorithm 1 every AGC cycle to update the attack sequence to adapt to the latest load changes. Thus, the execution time of Algorithm 1 should be shorter than an AGC cycle. To assess the feasibility of this speed, we now discuss the time complexity of Algorithm 1.

The optimization problems in Line 3 and Line 6 of Algorithm 1 are linear programming problems that can be solved in polynomial time. For instance, in our simulations, we use a CVX solver of $\mathcal{O}(h^2)$ complexity, where h is the iterator in Algorithm 1. As all the iterations of Algorithm 1 are independent and hence parallelizable, the compute time of a parallel implementation of Algorithm 1 is polynomial in the minimum TTE (MTTE), i.e., the maximum value of the iterator h. The scale of the grid also affects the compute overhead. The complexity of Eq. (4) is $\mathcal{O}(m)$, where m is the number of sensors that characterizes the scale of the grid. As Eq. (4) is computed once in each step of the linear programming, the compute time of a parallel implementation of Algorithm 1 based on the CVX solver is $\mathcal{O}(\text{MTTE}^2 \cdot m)$.

We implement Algorithm 1 in MATLAB and time the execution of each iteration on a quad-core 2.8 GHz Xeon CPU for the grid in Fig. 1. Fig. 13(a) plots the maximum absolute frequency deviation (i.e., $|\Delta\omega^*(k+h)|$ in Line 4 and Line 7 in Algorithm 1) over the iterator h. We can see that $|\Delta\omega^*(k+h)|$ increases with h, which means that a longer attack sequence can cause a larger frequency deviation. Moreover, the observation that $|\Delta\omega^*(k+h)|$ increases with the number of compromised sensor links (i.e., $|\mathbb{W}|$) is consistent with intuition. Fig. 13(b) plots the execution time of an iteration versus h, which shows polynomial trends. As the longest execution time of an iteration is shorter than three seconds if the MTTE is shorter than 35 AGC cycles, using a parallelized Algorithm 1, the attacker can solve the optimal attack sequence within an AGC cycle. The compute time can be further reduced by using faster CPUs.

VI. Attack Detection, Identification, and Mitigation

In this section, we present an attack detection approach. Then, we discuss an attack identification approach that estimates the attacker's write access $\mathbb{W}$. As illustrated in Fig. 4(a), the input to our attack detection and identification approaches includes the measured frequency deviation $\Delta\omega$ and the estimated load changes that may be compromised due to the FDI attack. Note that as ACE is computed based on the estimated load changes, including it in the inputs to the attack detection does not increase the information available for the detection. Finally, we discuss approaches to mitigate an detected/identified attack.

A. Detection of Attack and Onset Time

Our attack detector checks the consistency between the observed frequency deviation (denoted by $\Delta\omega(k)$) and the predicted frequency deviation (denoted by $\Delta\hat{\omega}(k)$). The frequency deviation is predicted using the first term of the right-hand side of Eq. (4) based on the observed load change vectors. Specifically, $\Delta\hat{\omega}(k) = \sum_{h=0}^{H-1} \mathbf{u}_h^\intercal \Delta\mathbf{p}'_{k-h}$, where $\Delta\mathbf{p}'$ is the load change vector computed from the possibly compromised power flow measurements. The detection is based on a metric $e(k) = \Delta\omega(k) - \Delta\hat{\omega}(k)$. In the absence of an attack, $\Delta\mathbf{p}' = \Delta\mathbf{p}$ and $e(k)$ is the error of the regression model, which is close to zero. Let $\mathbf{P}$ represent an $N \times m$ matrix that consists of -1, 0, and 1, and aggregates the load measurements in $\mathbf{z}$ as area-wise loads. In the presence of an attack, the load change vector used to predict the frequency deviation is also contaminated by the attack vector. That is, $\Delta\mathbf{p}' = \Delta\mathbf{p} + \mathbf{P}\mathbf{a}$. Thus, the defender's prediction of frequency deviation is

$$\Delta\hat{\omega}(k) = \sum_{h=0}^{H-1} \mathbf{u}_h^\intercal \Delta\mathbf{p}'_{k-h} = \sum_{h=0}^{H-1} \mathbf{u}_h^\intercal \left(\Delta\mathbf{p}_{k-h} + \mathbf{P}\mathbf{a}_{k-h}\right).$$

The actual observed frequency deviation is given by Eq. (4). Thus, $e(k) = \Delta\omega(k) - \Delta\hat{\omega}(k) = \sum_{h=0}^{H-1} (\mathbf{v}_h^\intercal \mathbf{T} - \mathbf{u}_h^\intercal \mathbf{P})\mathbf{a}_{k-h}$. In summary, with $\mathcal{H}_0$ and $\mathcal{H}_1$ denoting the case of attack absence and presence, respectively, we have

$$\begin{cases} e(k)|\mathcal{H}_0 \simeq 0; \\ e(k)|\mathcal{H}_1 = \sum_{h=0}^{H-1} (\mathbf{v}_h^\intercal \mathbf{T} - \mathbf{u}_h^\intercal \mathbf{P})\mathbf{a}_{k-h}. \end{cases} \tag{6}$$

Thus, we apply a threshold-based attack detector. Specifically, the detector makes a positive detection decision if $|e(k)| \geq \eta$ and a negative decision otherwise, where η is a constant threshold. The value of η can be set to be the maximum prediction error of the attack impact model in Eq. (4) in the absence of an attack.

If the attacker knows the details of the above detector, he can add additional constraints, i.e., $e(k+j)|\mathcal{H}_1 = 0$ for $1 \leq j \leq h$, when solving the optimization problems in Line 3 and Line 6 of Algorithm 1. By doing so, the attacker can bypass the detector. In other words, the attacker creates an illusion that the frequency deviation is caused by natural load changes. This is consistent with the concept of *undetectable attack* in [25]. However, in practice, it is often difficult to satisfy the assumption of an omniscient attacker [25] to achieve the undetectable attack. Specifically, for the proposed attack detection approach, the coefficients $\mathbf{v}_h$ and $\mathbf{u}_h$ as well as the horizon H learned by the attacker and the defender based on different data sets will not be exactly the same. As a result, if the defender can keep these coefficients confidential, an attack computed with the aforementioned additional constraints based on the attacker's $\mathbf{v}_h$, $\mathbf{u}_h$, and H will also cause considerable $e(k)$. In this paper, we primarily focus on the FDI attacks that are stealthy to the SE's BDD and other data quality checks only.

We now present a numeric example of the attack detection. Fig. 10 shows $\Delta\hat{\omega}(k)$, $\Delta\omega(k)$, and $e(k)$ over time k when the attacker with different write access constraints (81 and 66 compromised sensor data links) launches the optimal attack sequence to the grid shown in Fig. 1. In the bottom part of Fig. 10, the $|e(k)|$ is $0.036\,\text{Hz}$ and $0.019\,\text{Hz}$ right after the onset of the attack, under the two write access constraints, respectively. As the prediction error of the model in Eq. (4) is in the order of $0.001\,\text{Hz}$, the attack and its onset time can be easily detected.

B. Attack Identification

On detection of an attack, it is desirable for the defender to identify the attack, i.e., to estimate which sensor data links have been compromised (i.e., $\mathbb{W}$). The identification result can be used to guide a response such as attack mitigation discussed in Section VI-C. Our attack identification approach assumes that the attacker adopts the optimal attack sequence, since we have shown in Section V that the attacker can achieve the optimal attack. Our attack identification approach includes two steps. First, we develop an algorithm for the defender to accurately recover the optimal attack sequence given a candidate $\mathbb{W}$. Second, our approach tries a number of $\mathbb{W}$ candidates and identify the one that produces a $\Delta\omega$ trace closest to the observed $\Delta\omega$ trace. If the $\mathbb{W}$ candidates include the true $\mathbb{W}$, the identified $\mathbb{W}$ is the true $\mathbb{W}$ since they should produce the same attack sequence and $\Delta\omega$ trace. We will discuss how to generate the $\mathbb{W}$ candidates at the end of this section. The details of the two steps are given as follows.

Attack sequence recovery given a candidate $\mathbb{W}$: For the defender, the true load change $\Delta\mathbf{p}$ is unknown because the load change measurement vector has been contaminated by the unknown attack vector. From Eq. (4), $\Delta\mathbf{p}$ must be accurately estimated to recover $\mathbf{a}$ from the observed $\Delta\omega$. To address this, the defender can replace $\Delta\mathbf{p}_i$, $l \leq i \leq k$, in Eq. (5) with $\mathbf{P}(\mathbf{z}'_i - \mathbf{a}_i)$, where l is the detected attack onset time and $\mathbf{z}'_i$ is the compromised measurement vector observed by the defender in the ith AGC cycle. The defender runs Algorithm 1 every AGC cycle, where the variables for the constrained optimization problems in Line 3 and Line 6 of Algorithm 1 are $\{\mathbf{a}_l, \ldots, \mathbf{a}_k, \mathbf{a}_{k+1}, \ldots, \mathbf{a}_{k+h}\}$. If the attacker adopts the optimal attack, the attacker and the defender run the same algorithm, i.e., Algorithm 1. As a result, the defender can recover the optimal attack sequence.

Identification of the true $\mathbb{W}$: We present a method for the defender to identify the true $\mathbb{W}$ from a set of candidates (denoted by $\mathcal{W}$). The defender computes the optimal attack sequence for every candidate $\hat{\mathbb{W}} \in \mathcal{W}$ and the corresponding trajectory of $\Delta\omega$ from the detected attack onset time using Eq. (5). He then selects one or multiple candidates $\hat{\mathbb{W}} \in \mathcal{W}$ that produce past $\Delta\omega$ traces (i.e., $\{\Delta\omega(i)|l \leq i \leq k\}$) closest to the observed trace. For instance, by using the mean absolute error as the matching metric, our simulations in Section VII show that, from 16 candidate $\mathbb{W}$, this approach can correctly identify $\mathbb{W}$ from the second AGC cycle since the onset of the attack. Depending on the defender's computing resources, this approach can scale with the size of the candidate set $\mathcal{W}$. Evaluation results with a much larger candidate set can be found in Section VII.

We now discuss how to define the $\mathbb{W}$ candidates. When each of the m sensor links can be compromised, it is computation-prohibitive to consider all 2^m combinations of them as the

$\mathbb{W}$ candidates. In practice, the system operator may have knowledge about the weakness of the SCADA system. Such knowledge can be leveraged to identify a subset of vulnerable sensor links to significantly reduce the $\mathbb{W}$ candidates. In particular, if SE is not used to preprocess the input data for AGC, the attacker should focus on the tie-line flow measurements only since the FDI attack on other measurements will not affect AGC. As there are often a limited number of tie-lines (e.g., eight tie-lines in Fig. 1), iterating all combinations of tie-lines as the $\mathbb{W}$ candidates will not introduce significant computation overhead. If SE and BDD are used with AGC, the attacker needs to address more constraints and compromise more sensor links as shown in Section VII, which may significantly increase the $\mathbb{W}$ candidates to be considered. In our future work, we will study how to prune the sensor links considered for attack identification. A possible approach is to evaluate the impact of each individual compromised sensor link on the grid frequency and only consider those links that are most effective in reducing TTE.

C. Attack Mitigation

We now discuss several approaches to mitigate the impact of a detected attack. If the attack identification approach in Section VI-B is not employed, a possible mitigation approach is to stop using the compromised power export deviation estimate to compute the ACE. Instead, the ACE is computed based on the intact frequency deviation measurement only, i.e., $\text{ACE}_i = \beta_i \cdot \Delta\omega_i$. Thus, the AGC falls back to regulating frequency only. In addition, Sridhar et al. [7] propose to use forecast load, rather than measured load, to drive the AGC. This approach can prevent the grid frequency from reaching unsafe thresholds quickly, leaving time for the system operator to further counteract. If a detected attack can be identified, we can isolate the attack using an attack-mitigating SE program as follows. Denote by $\mathring{\mathbf{z}}$ a vector consisting of all intact elements of $\mathbf{z}$, $\mathring{\mathbf{F}}$ a matrix consisting of the rows of the original measurement matrix $\mathbf{F}$ that correspond to the intact $\mathbf{z}$ elements, and $\mathring{\mathbf{V}}$ a matrix consisting of the rows and columns of the original weight matrix $\mathbf{V}$ that correspond to the intact $\mathbf{z}$ elements only. The attack-mitigating SE estimates the state as $\hat{\mathbf{x}} = (\mathring{\mathbf{F}}^\intercal \mathring{\mathbf{V}} \mathring{\mathbf{F}})^{-1} \mathring{\mathbf{F}}^\intercal \mathring{\mathbf{V}} \mathring{\mathbf{z}}$. It leverages the fact that the original SE is an over-determined problem and therefore, removing several compromised elements in $\mathbf{z}$ may not affect the estimation much. The improved measurement vector including the elements corresponding to the compromised sensors is $\hat{\mathbf{z}} = \mathbf{F}\hat{\mathbf{x}}$. The virtual tie-line flows $\mathbf{T}\hat{\mathbf{z}}$ are then used to compute ACEs for AGC control. In Section VII, we will evaluate the impact of the number of the compromised sensor links on the accuracy of attack-mitigating SE.

VII. Simulations

A. Simulation Settings

To validate our analysis and compare the optimal attack with prior limited attacks, we conduct PowerWorld [12] simulations based on the three-area 37-bus model in Fig. 1. Default settings include: AGC cycle length is four seconds; $\epsilon_L = -0.5$ Hz and $\epsilon_U = 0.5$ Hz; all the sensor measurements are writable to the attacker; each element of $\mathbf{a}_{\min}$ and $\mathbf{a}_{\max}$ is -5 MW and 5 MW, respectively; for all the areas, $\alpha_i = 12$, $\beta_i = 100$ MW/Hz, and the AGC gain $K_i = 10^{-4}$. As the focus of this paper is to study how to push $\Delta\omega$ to ϵ_L or ϵ_U in the shortest time, we stop a simulation once $\Delta\omega$ goes out of (ϵ_L, ϵ_U). Remedial programs like load shedding can be integrated with our simulations, but they are beyond the present scope of our analysis.

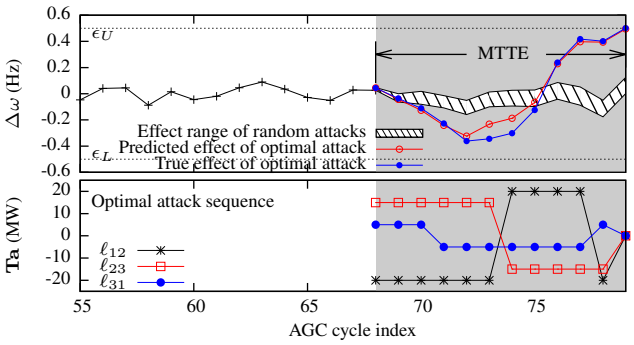


Fig. 14. Optimal attack sequence vs. random attack sequence. The grid is under attack during the shaded period.

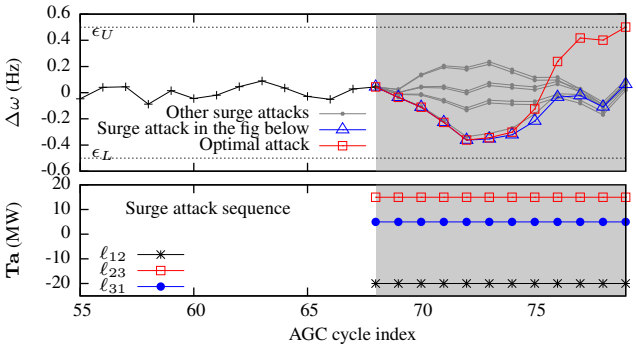


Fig. 15. Optimal attack sequence vs. surge attack sequence. The grid is under attack during the shaded period.

B. Simulation Results

1) Effectiveness of optimal attack sequence: We use Algorithm 1 to compute the optimal attack sequence, where all the 81 components of the attack vector $\mathbf{a}$ are writable by the attacker. To simplify illustration, in the bottom part of Fig. 14, we show the traces of the malicious injections to the three virtual tie-line power flow measurements. That is, the figure shows $\mathbf{Ta}$ over time. The top part of Fig. 14 shows the trajectory of $\Delta\omega$ when the attacker injects the optimal attack sequence. It also shows the trajectory of $\Delta\omega$ predicted by the attacker at the 68th AGC cycle, which well matches the true attack effect. As $\Delta\omega$ hits the ϵ_U threshold at the 78th AGC cycle, the minimum TTE (MTTE) is 10 AGC cycles. Note that, as indicated in the caption of Fig. 1, the total load of the simulated grid is about 800 MW. From the bottom part of Fig. 14, the range of the resulted injections on the virtual tie-line power flow measurements is $[-20\,\text{MW}, 20\,\text{MW}]$ only. This simulation result shows that small errors in tie-line power flow measurement or estimation with respect to the total load

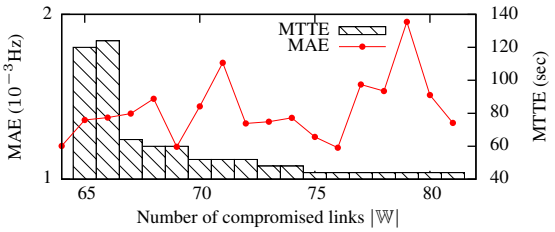


Fig. 16. Impact of write access constraint with SE.

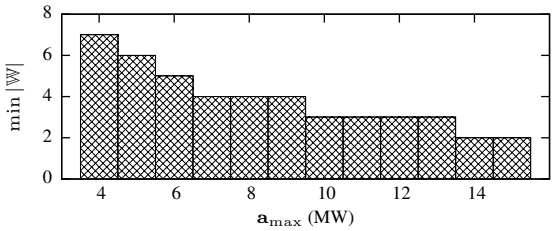


Fig. 17. Minimum $|\mathbb{W}|$ yielding MTTE $<$ 2 minutes.

can lead to significant frequency deviations. This is consistent with the understanding that the area power export control is a major component of AGC.

We employ two baseline attack approaches that are consistent with the two limited attack templates studied in [7]. The first baseline, *random attack*, uniformly and randomly generates an attack vector every AGC cycle from the feasible space defined by the constraints in Eqs. (1) and (2). The top part of Fig. 14 shows the range of $\Delta\omega$ caused by 2,800 random attack sequences. We can see that the random attack cannot push $\Delta\omega$ beyond either ϵ_U or ϵ_L within MTTE. The second baseline, *surge attack*, minimizes or maximizes each component of $\mathbf{Ta}$ under the constraint $\mathbf{a}_{\min} \preceq \mathbf{a} \preceq \mathbf{a}_{\max}$. Thus, there are a total of $2^3 = 8$ surge attack sequences for the three virtual tie-lines. For instance, the bottom part of Fig. 15 shows a surge attack sequence. The top part of Fig. 15 shows the trajectory of $\Delta\omega$ under all the eight surge attack sequences and the optimal attack. The surge attack cannot breach the safety condition within MTTE. The ineffectiveness of the random and surge attacks is due to the AGC's ability to correct the frequency deviations caused by these restricted attacks. To breach the safety limit, the attacker needs to strategically design his injections based on knowledge of the system dynamics.

2) Impact of write access constraint with SE: Fig. 16 shows the mean absolute error (MAE) of the model in Eq. (4) learned by the active probing approach versus the number of sensor data links writable by the attacker (i.e., $|\mathbb{W}|$). We can see that the attacker's model accuracy is insensitive to the write access constraint. Note that the learning and testing phases are subject to the same write access constraint. This result implies that overfitting does not occur when the attacker compromises more sensor data links and needs to learn more parameters. This is mainly due to the linearity of the attack impact as described in Eq. (4). Fig. 16 also shows the MTTE from a particular attack onset time versus $|\mathbb{W}|$. The decreasing trend is consistent with the intuition that a less constrained attacker can

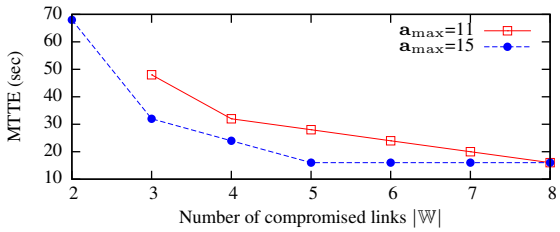


Fig. 18. Impact of $|\mathbb{W}|$ on MTTE without SE.

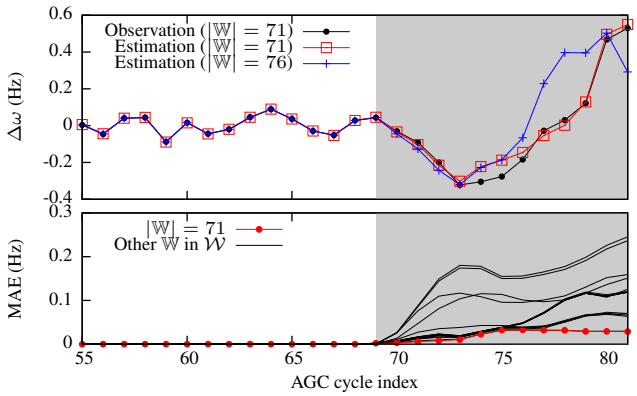


Fig. 19. An example of identifying $\mathbb{W}$. The grid is under attack during the shaded period.

cause a larger impact. From the figure we can see that if the attacker can compromise more than 65 out of 81 sensor links, the optimal attack can cause an unsafe frequency deviation within two minutes. For grids of larger scales (e.g., hundreds of buses), it becomes hard for the attacker to compromise massive sensor links and manipulate the grid frequency. Instead, the attacker may focus on a selected area and aim at increasing the tie-line power flows to breach safety limits. Our attack impact analysis can be easily extended to establish the relationship between the tie-line power flows and the attack vector, which can be used to guide the attack.

3) Minimum write access requirement without SE: This set of simulations does not consider SE and its BDD. Thus, the attacker can just focus on the sensors on the eight real tie-lines shown in Fig. 1. We evaluate the minimum number of tie-line sensors that the attacker needs to compromise in order to trigger remedial actions within two minutes. Fig. 17 shows this minimum number versus the setting of each element of $\mathbf{a}_{\max}$, where $\mathbf{a}_{\min} = -\mathbf{a}_{\max}$. The decreasing trend suggests that, for a more stringent stealthiness constraint (i.e., a smaller $\mathbf{a}_{\max}$), the attacker needs to compromise more sensor data links to achieve a certain TTE. Moreover, by comparing Fig. 16 and Fig. 17, the attacker needs to compromise much fewer sensor links if SE is not employed. This shows that SE effectively raises the bar for the attacker to succeed. Fig. 18 shows the MTTE versus $|\mathbb{W}|$ under two different settings for $\mathbf{a}_{\max}$. Fig. 18 shows that the attacker needs to compromise more sensor data links to reduce the MTTE. This result is similar to the result in Fig. 16.

4) Attack identification with SE: The candidate set of $\mathbb{W}$, i.e., $\mathcal{W}$, consists of 16 $\mathbb{W}$ candidates. The cardinality of the

$\mathbb{W}$ candidates is from 66 to 81 and that of the true $\mathbb{W}$ is 71. The top part of Fig. 19 shows the observed trace of $\Delta\omega$ and the computed traces with two different $\mathbb{W}$ at the 90th AGC cycle. We can see that the $\Delta\omega$ trace computed based on the correct $\mathbb{W}$ well matches the observed trace, which validates our discussion in Section VI-B. The bottom part of Fig. 19 shows the trajectory of matching errors in MAE for different candidates in $\mathcal{W}$. From the second AGC cycle after the onset of the attack, the MAE with the correct $\mathbb{W}$ remains the smallest and the approach presented in Section VI-B can correctly identify $\mathbb{W}$. We also evaluate our approach using a larger candidate set $\mathcal{W}$ consisting of 3,850 candidates. Compared with the true $\mathbb{W}$, there are eight wrong candidates with lower MAEs than the true $\mathbb{W}$ at the 5th AGC cycle after the onset of the attack. In other words, we can shrink the solution space that contains the true $\mathbb{W}$ by 99.7%. During the entire time period of the MTTE, only two wrong $\mathbb{W}$ candidates consistently yield lower MAEs than the true $\mathbb{W}$. In our future work, we will explore more effective matching metrics to improve the $\mathbb{W}$ identification performance.

5) Attack mitigation: This set of simulations evaluate the attack-mitigating SE program discussed in Section VI-C. The evaluation methodology is as follows. In each run, we generate a random subset of n elements out of all the 81 elements of a certain measurement vector $\mathbf{z}$ for the 37-bus system. We assume these n elements are compromised by the attacker and identified by the defender. Based on the remaining $(81-n)$ elements, the attack-mitigating SE program estimates the state as $\hat{\mathbf{x}}$. We adopt the relative error $\|\hat{\mathbf{x}}-\mathbf{x}\|_2/\|\mathbf{x}\|_2$ to characterize the SE error, where $\mathbf{x}$ is the true state. Fig. 20 shows the average relative error in parts-per-million (ppm) over many runs versus the number of removed elements, i.e., n. We implement the matrix inverse $(\mathring{\mathbf{F}}^\intercal\mathring{\mathbf{V}}\mathring{\mathbf{F}})^{-1}$ in the SE using the LU decomposition provided by the uBLAS library of the Boost C++ Libraries. For an ill-conditioned $\mathring{\mathbf{F}}^\intercal\mathring{\mathbf{V}}\mathring{\mathbf{F}}$, the matrix inversion may fail and throws an exception. The average relative error in Fig. 20 only accounts for the successful cases. Fig. 20 also shows the failure probability of the matrix inversion. The results show that, if the attacker can compromise more sensor links, the attack-mitigating SE may fail with a higher probability due to the ill-conditioned $\mathring{\mathbf{F}}^\intercal\mathring{\mathbf{V}}\mathring{\mathbf{F}}$. When n increases from 30 to 44, the failure probability increases sharply from 10% to 96%. When n is larger than 44, the attack-mitigating SE becomes underdetermined because less than 37 intact $\mathbf{z}$ elements are remaining for estimating the 37-dimensional $\mathbf{x}$. As such, the matrix inversion always fails. However, from Fig. 20, we can see that, as long as the matrix inversion is successful, the attack-mitigating SE is highly accurate, with relative SE error below one ppm. The result suggests that, the 37-bus system operator can selectively secure 37 sensor links to ensure a well-conditioned $\mathring{\mathbf{F}}^\intercal\mathring{\mathbf{V}}\mathring{\mathbf{F}}$. By doing so, isolating any other compromised sensor links from the SE introduces small errors to the state estimate $\hat{\mathbf{x}}$, AGC control, and other grid controls based on SE.

VIII. Testbed Experiments

We conduct experiments on a three-phase 16-bus 400 V power system testbed to evaluate the passive monitoring approach presented in Section V-A2. The 16 buses, each installed in a cabinet as shown in Fig. 22(b), are connected to form a ring topology as shown in Fig. 21. Each bus is monitored by a Schneider PowerLogic PM820MG smart meter. A variable load, as shown in Fig. 22(c), is connected to Bus 4 in the system. Its power consumption can be tuned manually using a knob. A 13.5kVA generator, shown in Fig. 22(a), is driven by a motor (which simulates a turbine) and is connected to Bus 10 in the system. The input power of the motor is supplied by a Current Vector Drive (CVD), which communicates with a remote computer. As the testbed is equipped with a single generator only, power engineering researchers can only implement a single-area AGC system. Specifically, the single-area AGC algorithm is implemented using LabVIEW on the computer, which regulates the grid frequency based on the frequency measurement of the smart meter on Bus 4. We note that the frequency measurement is the only input to the single-area AGC algorithm. The LabVIEW program retrieves frequency measurements from the smart meter via Modbus/TCP and sends the ACE to the CVD. Thus, different from the attacks on the power flow measurements described in the previous sections, in this section, we study attacks on the frequency measurements and extend the passive monitoring approach to address this new attack model. The detailed extension is omitted here due to space limitations and can be found in [33]. The extended approach assumes that the attacker knows D, M, β, and can eavesdrop on the measurements of load deviation Δp and frequency deviation $\Delta\omega$. We refer the reader to Section V-A2 for discussions on how the attacker can obtain these system constants and measurements. We note that the purpose of our testbed experiments regarding the FDI attacks on the frequency measurements is to demonstrate the effectiveness of our attack modeling methodology and achievability of the optimal attack, rather than to propose the FDI attack on the frequency measurements as an effective attack method, given that this attack can be easily detected and isolated in practice as discussed in Section III-A.

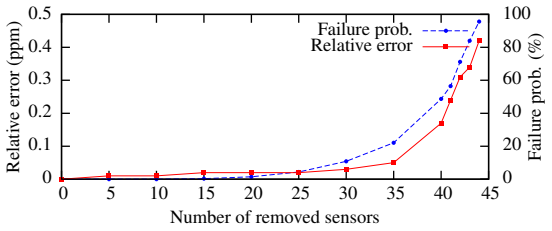


Fig. 20. Relative error and failure probability of attack-mitigating SE.

We conduct experiments to validate the extended passive monitoring approach. For this testbed, the constants needed by the attacker are $D = -23$ W/Hz, $M = 2.6$ kJ/Hz, and $\beta = 300$ W/Hz. The AGC cycle length is two seconds. During the attacker's learning phase, we manually tune the load to simulate load fluctuations. To mimic the attacker's eavesdropping, we install Wireshark (a packet sniffer) on the computer running AGC and use it to extract Δp and $\Delta\omega$ from the network traffic. We note that, in practice, the attacker can apply realistic approaches to eavesdrop the measurements. For

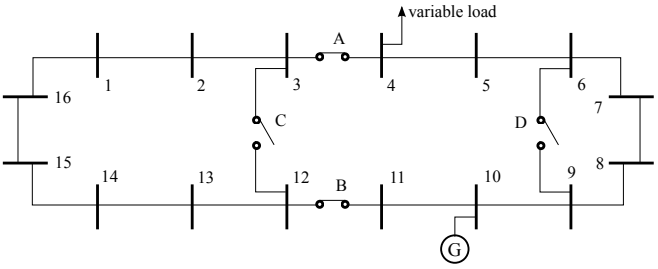
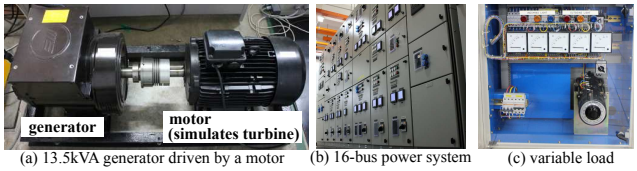


Fig. 21. A 16-bus 400V power system testbed. A 13.5kVA synchronous generator is connected to Bus 10, a variable load is connected to Bus 4, circuit breakers A and B are closed, and circuit breakers C and D are open.



(a) 13.5kVA generator driven by a motor (b) 16-bus power system (c) variable load

Fig. 22. Components of the power system testbed.

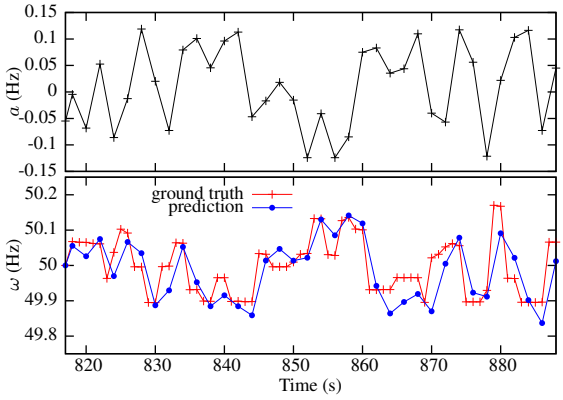


Fig. 23. Top: injection to frequency readings. Bottom: frequency predicted by learned model and ground truth.

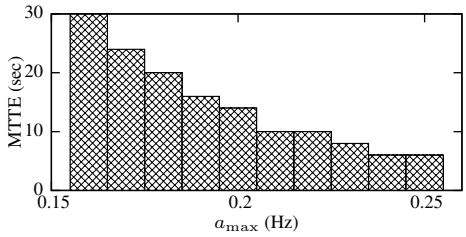


Fig. 24. MTTE vs. $a_{\max}$ setting ($a_{\min} = -a_{\max}$).

instance, as shown in [37], the measurements from an industry-class wireless smart meter can be easily eavesdropped. Using two minutes of eavesdropped data, we follow the extended passive monitoring approach [33] to learn the attack impact model using MATLAB's system identification toolbox. We try different orders for some intermediate transfer functions to be identified and choose the orders that best fit the training data. The resulting transfer function for the FDI to the grid frequency is of the seventh order. We evaluate the learned attack impact model as follows. Using the model, we predict the trajectory of the grid frequency given a random attack sequence of limited magnitude, as shown in the top part of Fig. 23. Then, to mimic the FDI attack, we inject this attack sequence to the real-time frequency measurements in the LabVIEW program that implements the single-area AGC during an experiment. Note that, as our experiments focus on demonstrating the effectiveness of attack modeling and achievability of optimal attack, launching real eavesdropping and FDI attacks by exploiting the vulnerabilities of the testbed construct is non-essential. We limit the magnitude of this test attack sequence to ensure that it will not cause damage to the testbed. The bottom part of Fig. 23 shows our prediction and the observed ground truth. The prediction matches the ground truth well and the mean absolute error of the prediction is $0.036\,\text{Hz}$ only. This suggests that the learned model is accurate.

With the learned model, we compute the optimal attack sequences under different settings for the FDI bound $a_{\max}$, where $a_{\min} = -a_{\max}$. Fig. 24 shows the computed MTTE versus $a_{\max}$. We can see that the MTTEs are below 30 seconds. Such short MTTEs suggest that it is critical to protect the frequency measurements of this testbed. Although we stop the experiment before physical damage happens on the testbed, the demonstrated accuracy of the learned attack impact model substantiates the importance of the optimal attacks in practice.

IX. Discussion

This section discusses more constraints on the attacker and their impact on our analysis. This paper mainly considers the two constraints in Eqs. (1) and (2) that the attacker has to meet. The attack's stealthiness condition in Eq. (1) can be extended to address other BDD forms (e.g., ACSE's BDD) without affecting the formulation of the optimal attack sequence. The generic and abstract expression of the write access condition in Eq. (2) can capture various concrete and practical constraints on the attacker. To be more expressive, we can extend the current write access condition to capture the case where the attacker cannot launch an FDI attack every AGC cycle, due to say the attacker's limited compute capacity to breach the cryptographic protection. This additional constraint can be easily addressed by imposing $\mathbf{a}_i = 0$ in Eq. (5) and Algorithm 1, where i is the AGC cycle index when the attacker cannot launch an FDI attack. This additional constraint will result in longer TTE.

In this paper, we assume that the attacker has read access to the power flow measurements. This assumption is an essential aspect of the Kerckhoffs's principle to consider an attacker who has accurate knowledge and real-time state of the targeted system. It allows the grid operator to assess the minimum TTE and guide the defense with sufficient strength. If the control center does not apply state estimation and BDD for AGC, the attacker will only need the read access to the tie-line power flow measurements. We note that, if the attacker cannot obtain read access to necessary power flow measurements, they will have to fall back to some non-strategic attack methods such as the random and surge attacks evaluated in Section VII.

While the simulations conducted in this paper are based on

the 37-bus system in Fig. 1, the attack impact modeling and the developed countermeasures in this paper are based on a general system model. Thus, our analysis also applies to power grids of larger scales. For large-scale grids, it becomes hard for the attacker to compromise a massive number of sensor links and manipulate the grid frequency. Instead, the attacker may focus on a selected area and aim at increasing the tie-line power flows there to breach safety limits. Independent of whether the attacks can scale up, the proposed countermeasures developed under a strong adversary model will be able to detect and mitigate the attacks should they happen.

X. Conclusion

This paper studied FDI attacks on sensor data for AGC. We derived key attack impact models and showed that the attacker can learn the models based on eavesdropped sensor data and a modest amount of prior knowledge about the grid. Then, the attacker can compute an attack sequence to minimize the remaining time before the grid must initiate costly and disruptive remedial actions such as disconnecting generators and customer loads. We developed efficient algorithms to detect, identify, and mitigate the attack. Our analysis and algorithms are validated by experiments on a physical 16-bus power system testbed and extensive PowerWorld simulations based on a 37-bus power system model.

References

[1] S. Karnouskos, "Stuxnet worm impact on industrial cyber-physical system security," in *IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society*, 2011.

[2] "Hackers infiltrated power grids," 2014, http://on.recode.net/1FpKP7Y.

[3] "The dragonfly attack," 2014, http://rsa.dev.neptuneweb.com/dragonfly-attack/.

[4] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *The 16th ACM Conference on Computer and Communications Security (CCS)*, 2009.

[5] U.S. DHS, "Insider threat to utilities," 2011, https://info.publicintelligence.net/DHS-InsiderThreat.pdf.

[6] P. Kundur, N. J. Balu, and M. G. Lauby, *Power system stability and control*. McGraw-hill New York, 1994.

[7] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 580–591, 2014.

[8] S. Bhowmik, K. Tomsovic, and A. Bose, "Communication models for third party load frequency control," *IEEE Transactions on Power Systems*, vol. 19, no. 1, pp. 543–548, 2004.

[9] K. Tomsovic, D. E. Bakken, V. Venkatasubramanian, and A. Bose, "Designing the next generation of real-time control, communication, and computations for large power systems," *Proceedings of the IEEE*, vol. 93, no. 5, pp. 965–979, 2005.

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

[11] A. Ashok, P. Wang, M. Brown, and M. Govindarasu, "Experimental evaluation of cyber attacks on automatic generation control using a cps security testbed," in *2015 IEEE Power & Energy Society General Meeting*. IEEE, 2015, pp. 1–5.

[12] "PowerWorld," 2016, http://www.powerworld.com/.

[13] "National grid maps," 2016, http://www.geni.org/globalenergy/library/national_energy_grid/.

[14] J. Willems, "Sensitivity analysis of the optimum performance of conventional load-frequency control," *IEEE Transactions on Power Apparatus and Systems*, no. 5, pp. 1287–1291, 1974.

[15] Y. Chen, Z. Huang, Y. Liu, M. J. Rice, and S. Jin, "Computational challenges for power system operation," in *Hawaii International Conference on System Science*, 2012.

[16] S. Grijalva, "Research needs in multi-dimensional, multi-scale modeling and algorithms for next generation electricity grids," 2011, http://1.usa.gov/1VBJAgu.

[17] P. M. Esfahani, M. Vrakopoulou, K. Margellos, J. Lygeros, and G. Andersson, "Cyber attack in a two-area power system: Impact identification using reachability," in *American Control Conference (ACC)*, 2010.

[18] ——, "A robust policy for automatic generation control cyber attack in two area power network," in *49th IEEE Conference on Decision and Control (CDC)*, 2010.

[19] P. Rabbanifar and S. Jadid, "Stochastic multi-objective tie-line power flow and frequency control in market clearing of multi-area electricity markets considering power system security," *IET Generation, Transmission & Distribution*, vol. 8, no. 12, pp. 1960–1978, 2014.

[20] J. M. Hendrickx, K. H. Johansson, R. M. Jungers, H. Sandberg, and K. C. Sou, "Efficient computations of a security index for false data attacks in power networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3194–3208, 2014.

[21] T. Liu, Y. Sun, Y. Liu, Y. Gui, Y. Zhao, D. Wang, and C. Shen, "Abnormal traffic-indexed state estimation: A cyber–physical fusion approach for smart grid attack detection," *Future Generation Computer Systems*, vol. 49, pp. 94–103, 2015.

[22] 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 *ACM/IEEE 5th International Conference on Cyber-Physical Systems (ICCPS)*, 2014.

[23] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Cyber security of water scada systems–part i: analysis and experimentation of stealthy deception attacks," *IEEE Transactions on Control System Technology*, vol. 21, no. 5, pp. 1963–1970, 2013.

[24] A. A. Cárdenas, 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 *6th ACM Symposium on Information, Computer and Communication Security (ASIACCS)*, 2011.

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

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

[27] S. Li and X. Wang, "Cooperative change detection for voltage quality monitoring in smart grids," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 86–99, 2016.

[28] Y. Zhu, J. Yan, Y. Tang, Y. Sun, and H. He, "Joint substation-transmission line vulnerability assessment against the smart grid," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 1010–1024, 2015.

[29] Y. Zhu, J. Yan, Y. Tang, Y. L. Sun, and H. He, "Resilience analysis of power grids under the sequential attack," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2340–2354, 2014.

[30] R. Tan, H. H. Nguyen, E. Y. S. Foo, X. Dong, D. K. Y. Yau, Z. Kalbarczyk, R. K. Iyer, and H. B. Gooi, "Optimal false data injection attack against automatic generation control in power grids," in *7th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2016.

[31] A. Hahn, A. Ashok, S. Sridhar, and M. Govindarasu, "Cyber-physical security testbeds: Architecture, application, and evaluation for smart grid," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 847–855, 2013.

[32] G. Hug and J. A. Giampapa, "Vulnerability assessment of ac state estimation with respect to false data injection cyber-attacks," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1362–1370, 2012.

[33] Technical report of this paper, Tech. Rep., 2016, http://publish.illinois.edu/resilient-grid/files/2016/01/AGC-full.pdf.

[34] Q. Liu and M. D. Ilic, "Enhanced automatic generation control (E-AGC) for future electric energy systems," in *IEEE Power and Energy Society General Meeting*, 2012.

[35] I. D. Margaris, S. A. Papathanassiou, N. D. Hatziargyriou, A. D. Hansen, and P. Sorensen, "Frequency control in autonomous power systems with high wind power penetration," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 2, pp. 189–199, 2012.

[36] "PJM manual 12," 2015, http://www.pjm.com/markets-and-operations/ancillary-services/~/media/documents/manuals/m12.ashx.

[37] T. Liu, Y. Liu, Y. Mao, Y. Sun, X. Guan, W. Gong, and S. Xiao, "A dynamic secret-based encryption scheme for smart grid wireless communication," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1175–1182, 2014.