Assessing and Mitigating Impact of Time Delay Attack: Case Studies for Power Grid Controls

Xin Lou, Cuong Tran, Rui Tan, *Senior Member, IEEE,* David K.Y. Yau, *Senior Member, IEEE,* Zbigniew T. Kalbarczyk, *Member, IEEE,* Ambarish Kumar Banerjee, Prakhar Ganesh

***Abstract*—Due to recent cyber attacks on various cyber-physical systems (CPSes), traditional isolation based security schemes in the critical systems are insufficient to deal with the smart adversaries in CPSes with advanced information and communication technologies (ICTs). In this paper, we develop real-time assessment and mitigation of an attack's impact as a system's built-in mechanisms. We study a general class of attacks, which we call *time delay attack*, that delays the transmissions of control data packets in the CPS control loops. Based on a joint stability-safety criterion, we propose the attack impact assessment consisting of (i) a machine learning (ML) based safety classification, and (ii) a tandem stability-safety classification that exploits a basic relationship between stability and safety, namely that an unstable system must be unsafe whereas a stable system may not be safe. In this assessment approach, the ML addresses a state explosion problem in the safety classification, whereas the tandem structure reduces false negatives in detecting unsafety arising from imperfect ML. We apply our approach to assess the impact of the attack on power grid automatic generation control, and accordingly develop a two-tiered mitigation that tunes the control gain automatically to restore safety where necessary and shed load only if the tuning is insufficient. We also apply our attack impact assessment approach to a thermal power plant control system consisting of two PID control loops. A mitigation approach by tuning the PID controller is also proposed. Extensive simulations based on a 37-bus system model and a thermal power plant control system are conducted to evaluate the effectiveness of our assessment and mitigation approaches.**

***Index Terms*—Power Grid Control, Cyber-physical system, Delay attack, Stability, Safety, Machine learning**

I. Introduction

By integrating modern information and communication technologies (ICTs), critical systems (e.g., power grids and advanced manufacturing facilities) are transforming into cyber-physical systems (CPSes). However, whereas ICTs can improve system performance, it also incurs cybersecurity risks. To date, the security of these systems has largely relied on isolation from public networks through air gaps and firewalls. However, the isolation is questionable, due to insiders [1] and stepping stone attacks [2]. For instance, the Dragonfly attack against power grids [3] compromised a third-party virtual private network (VPN) software vendor, and then used the result as a stepping stone for intruding into the grids. Once attackers breach the isolation, they can launch powerful data integrity attacks similar to Stuxnet [4].

Motivated by the above security incidents, this paper studies the assessment and mitigation of the impact of an important and general class of attacks, which we call the *delay attack*, on a CPS that employs closed-loop control [5], [6], [7]. The attack maliciously delays transmissions of control packets without tampering with the data content. Since CPS control often has stringent timeliness requirements, the attack can undermine system performance severely and even cause catastrophic safety incidents. Compared with data tampering that needs to break non-trivial cryptographic protection, the delay attack can be implemented more simply using compromised routers to increase the communication latency. Hence, it is an important threat that requires immediate attention. However, whereas the attack can be readily detected by trustworthy synchronization of the clocks of coordinating CPS devices [1], [8] and subsequent verification of packet timestamps, assessing and mitigating its impact in real time are challenging due to the complexity of typical real-world cyber-physical control systems.

In this paper, we propose a joint stability-safety criterion for assessing and mitigating an attack's impact using a data-driven method. *Stability* and *safety* concern a system's ability to keep its state fluctuations *bounded* and *within a prescribed range*, respectively, in the presence of exogenous disturbances that have *bounded magnitudes*. As disturbances (e.g., sensor noises and system input changes) are inevitable, stability is a basic requirement that must be met by any CPS. Otherwise, the system may experience unacceptable state divergence following a disturbance. Besides stability, however, CPS must further operate within its engineered safety limits. For instance, a $60\,\text{Hz}$ power grid must maintain its frequency within a range of about $59.5\,\text{Hz}$ to $60.5\,\text{Hz}$; otherwise, generators/loads may trip automatically causing blackouts. Thus, real-time knowledge of the system's stability and safety is critical. Based on this knowledge, if a delay attack is assessed to destabilize the

Manuscript received June 8, 2019; revised September 13, 2019.

Xin Lou is with Illinois at Singapore e-mail: (lou.xin@adsc-create.edu.sg). Cuong Tran is with Vietnam Posts and Telecommunications Group. He was with Singapore University of Technology and Design (e-mail: bambootran89@gmail.com). Rui Tan is with Nanyang Technological University, Singapore (e-mail: tanrui@ntu.edu.sg). David K.Y. Yau is with Singapore University of Technology and Design (e-mail: david_yau@sutd.edu.sg). Zbigniew T. Kalbarczyk is with University of Illinois at Urbana-Champaign, USA (e-mail: kalbarcz@illinois.edu). Ambarish Kumar Banerjee is with Indian Institute of Technology Bhubaneswar, India (e-mail:akb14@iitbbs.ac.in). Prakhar Ganesh is with Illinois at Singapore e-mail: (prakhar.g@adsc-create.edu.sg).

A preliminary version of this work appeared in The 10th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS 2019). This research was supported in part by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, in part by the Energy Innovation Research Programme (EIRP, Award Nos. NRF2014EWTEIRP002-026 and NRF2017EWTEP003-061), administered by the Energy Market Authority (EMA), in part by Device and System-level Detection and Identification of IoT Attacks, funded by SUTD-ZJU IDEA programme, award number SUTD-ZJU (VP) 201805, and in part by an NTU Start-up Grant.

system or push it into an unsafe region, attack mitigation must be initiated to regain the system's stability and safety.

This paper considers linear time-invariant (LTI) systems that can characterize a wide range of real-world cyber-physical systems. From control theory, an LTI system's stability depends on the system model only. Accordingly, analytical solutions for LTI system stability have been proposed [9], [10], [11]. As the stability analysis must be carried out continually, its real-time performance is a concern. However, some latest analytical methods [11], [10] are only effective for small-scale systems, e.g., load frequency control limited to one or two areas [11]. In contrast, the safety of a system depends on its future transient trajectory, which presents various challenges. Simulating the transient trajectory of a complex system may be too slow for detecting and reacting to its impending unsafety. An alternative approach is to run offline simulations comprehensively to understand the system's safety proactively [12], ahead of actual operations. However, as the trajectory depends on the initial system state, enumerating all the possible states in a continuous value domain is generally impossible. For instance, for an n-bus power system, whose system state dimension is n, its total number of discretized states is m^n, where m is the number of quantization steps for the state variable corresponding to each bus. The value of n for practical systems can be in the hundreds, making the enumeration computationally infeasible.

To address the above challenges, we propose a novel delay-attack impact assessment that features (i) a machine learning (ML) based safety classification, and (ii) a tandem stability-safety classification structure. First, to avoid the exponential complexity of enumerating all the system states, we adopt a Monte Carlo method to randomly sample the state space and run offline transient simulations to generate safety labels for the samples. These samples and their labels are used to train an ML model that can classify the safety of a live system based on its real-time conditions. The ML-based online safety classification is made fast enough to ensure the timeliness of the impact assessment. Second, we leverage a basic relationship between stability and safety to design the tandem structure, so that it classifies the system's stability first and then its safety only if stability is indicated. As the stability classification is simpler, faster, and more accurate than the safety assessment, the tandem structure can reduce (i) false negatives in the unsafety detection due to the ML's inaccuracy, and (ii) overall execution time for the attack impact assessment since the safety classification can be skipped for a system determined to be unstable.

This paper applies the proposed assessment approach to two real-world CPSes: power grid automatic generation control (AGC) [13] and power plant control (PPC). The goal of AGC is to maintain the grid frequency at a standard nominal value (e.g., 60 Hz) in the presence of load changes as primary exogenous disturbances. As the AGC's control signals are transmitted over communication networks, the delay attack is an important concern. We report extensive simulations using PowerWorld [14], an industry-strength power system simulator used by actual grid operators. The results show that the AGC's stability depends on the delay and the total load only, whereas its safety additionally depends on the load changes and detailed distribution of the load among the load buses. The boundary of the stable region can be obtained easily via a small set of offline simulations, while a joint application of the Monte Carlo method and the extreme learning machine (ELM) [15] is used to learn the safety boundary to manage the aforementioned state explosion problem with respect to the number of buses and possible load distributions. Furthermore, we use the achieved stability-safety classification to develop a two-tier mitigation of the attack's impact. The mitigation regains the stability and safety of the AGC whenever needed, by tuning the AGC gain whenever possible and resorting to shedding load whenever the gain tuning is insufficient. Moreover, we also apply our assessment approach to a PPC system, which consists of two proportional-integral-derivative (PID) control loops, where the PID controller is a feedback control loop mechanism widely used in the industrial control systems. This PPC system simulates a thermo power plant in the power grid. To verify the stability-safety classification and the mitigation approach, we generate offline training data using the PPC model to learn the stability and safety boundaries by ELM and also evaluate the accuracy of the trained ELM model. Moreover, we also propose the mitigation approach by tuning the controller in the PPC to mitigate the impact of the attack.

Our prior work [16] presented the case study of our assessment and mitigation approaches in AGC. Based on the work in [16], we make the following new contributions: 1) we add simulations in Section VII to study how the trajectory of the load change may affect the system's stability and safety. In practice, the load change may take time. For instance, the customers' solar power generation may change due to the movement of clouds and the ensuing load changes may take tens of seconds. Thus, we conduct new simulations to understand the impact of load change trajectories on stability and safety of AGC of power grid. 2) we add section Section VIII to present the second case study of applying our assessment and mitigation approaches to PPC. We propose the ELM-based stability and safety assessment approach as well as the mitigation approach by tuning the PID controller in PPC. We add new simulations using Modelica simulator to investigate the performance of our approaches. 3) we add new diagrams, redraw some of the diagrams and improve the presentation of several sections to make the paper clearer.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents preliminaries and a motivating example. Section IV overviews our approach. Section V and Section VI present the attack impact assessment and mitigation approaches, respectively. Section VII and Section VIII present the extensive evaluation results for AGC and PPC. Section IX concludes this paper.

II. Related work

Power system stability and safety classifications are often studied separately in the literature. In [9], Lyapunov stability theory and linear matrix inequalities are used to estimate delay margins. In [17], the stability of a system is classified

based on its energy accumulated during a certain time period. Traditional safety classification methods often analyze post-contingency power flows [18]. They use active power [18], [19] or composite indices based on various physical parameters [19] to classify the safety. However, the high computational overhead of these approaches makes them unsuitable for real-time classification [20], [21].

To reduce the computational overhead of real-time classification, recent studies apply ML (e.g., decision tree [22], support vector machine (SVM) [21], and artificial neural network (ANN) [20], [23]) to classify a power system's stability [21], [22] and safety with respect to certain contingencies [23], based on measured physical conditions of the system. In [21], a trained SVM classifies the power system's stability by using phasor measurement unit data. The SVM must be retrained if the system condition has changed significantly. The ANN model in [23] takes the system loading as input to rank the severity of the contingency in question, in terms of a composite performance index. However, all these studies do not address the emergent concern of cybersecurity.

Power grid cybersecurity has received increasing research. Chen et al. [5] study the impact on voltage and angle transient stability of data tampering attacks against voltage support devices. They do not address attack mitigation. An analytical solution has been proposed [11] for computing delay margins for the stability of a load frequency control system. However, their approach can only deal with small (e.g., one- or two-area) systems. Zhang et al. [10] propose closed-form expressions for evaluating delay-dependent stability in power grids for the load frequency control. Similarly, this approach is limited to small systems (e.g., less than three generation units in each control area) due to the limitations of current solvers.

Existing research on the cybersecurity of AGC has mainly focused on false data injection (FDI) attacks [24], [25], [26], where the attacker tampers with sensor and/or control data in the AGC control loop. Specifically, reachability analysis has been used [24], [25] to analyze the safety impacts of cyber-attacks against a two-area system. Rather than qualitative reachability analysis, a quantitative analysis of the minimum time until the system is unsafe has also been applied [26]. FDI attacks rely on an adversary's non-trivial ability to corrupt data. In contrast, this paper considers the easier and thus arguably more important attack of maliciously delaying data packets between communicating system components. In [27], [28], the authors show how the delay attack can impact the AGC's stability. In [29] and the reference therein, different schemes are proposed to characterize and eliminate the instability caused by delay attacks in load frequency control. However, they do not consider the more subtle but equally critical property of safety.

III. Stability and Safety under Delay Attack

This section defines stability and safety, as well as our threat model. Then, we use a simple control system to illustrate the impacts of the delay attack on the stability and safety.

A. System Model and Definitions of Stability and Safety

We consider a discrete-time CPS control system. Time is divided into slots. It can be any time unit according to the physical system settings. A *controller* collects measurements by the *sensors* in a *plant* (i.e., the physical system) and sends control commands to the *actuators*, which may change the state of the plant to maintain it at a given setpoint. The system is subjected to various disturbances, such as measurement noises, actuation biases, setpoint changes, etc. We adopt a bounded-input, bounded-output (BIBO) stability criterion:

Definition 1: A system is *BIBO-stable* if its state remains bounded while it experiences bounded disturbances.

We note that there are other stability definitions, e.g., asymptotic stability [30]. A system is asymptotically stable if for any positive ϵ, there exists a positive δ such that for any initial state of the system $x(0)$, the system's asymptotic equilibrium $\lim_{t\to\infty} x(t)$ satisfies $||x(t) - \lim_{t\to\infty} x(t)|| < \epsilon$, $\forall t \geq 0$, where $||x(0) - \lim_{t\to\infty} x(t)|| < \delta$. An asymptotically stable system is also BIBO-stable. Thus, BIBO stability is more basic and it is widely adopted in research on CPS control. For instance, the IEEE/CIGRE joint task force defines power system stability based on the BIBO concept [31]. In this paper, by *stability* we mean BIBO stability unless otherwise stated. Stability is a mandatory property for CPS design and operations. We adopt the following safety definition.

Definition 2: A system is *safe* if its state remains within a specified range while it experiences disturbances of magnitudes no larger than specified values.

Safety is naturally a key concern of system operators, because devices are designed to function properly only within specified ranges. Crossing these ranges may damage the devices or cause system failures. From Definitions 1 and 2, note that stability describes a *qualitative* "bounded" nature of the system state, whereas safety additionally imposes a *quantitative* range of the bounds. Thus, stability is a more basic requirement in that an unstable system must be unsafe, but a stable system may not be safe. This relationship between the two different properties of a system will be exploited in Section V to improve the performance (e.g., accuracy and timeliness) of the attack impact assessment for both the properties.

B. Threat Model

The delay attack is formally described as follows. Let $w[t]$ denote packetized control data generated and transmitted by the controller in the t^{th} time slot. The transmissions of the packets are maliciously delayed by τ time slots. Thus, in the $[t+\tau]^{\text{th}}$ time slot, the data $w[t]$ arrives at the actuator. Note that τ is an integer since the actuator operates in discrete time. The delay attack does not tamper with the content of the transmitted data. As Section I discusses, it can be launched through a compromised router. Note that the delay τ can also include the natural communication latency. Different from the well-known distributed denial-of-service (DDoS) attack, where the target is to overload the system with superfluous requests, delay attack is not to overload the system but using the outdated command to degrade the system performance and even cause damage to the system. Actually if we consider the extreme case of delay attack, i.e., the delay is infinite, delay

attack becomes DDoS attack. Note that DDoS attack can be easily detected. Differently, the delay attack is stealthy. In this paper, we assume that τ is a constant during the attack process. The results of this paper provide a baseline for understanding the more complicated situation where the attacker introduces time-varying delays. The extension of our study to address time-varying delay is left to future work.

In this paper, we assume that the clocks of the controller and the actuator are synchronized. Thus, if the controller adds a timestamp t to the transmitted data $w[t]$, the actuator can easily measure the delay τ introduced by the attack. The measured τ is used as an input to the attack impact assessment and mitigation. We note that secure clock synchronization techniques [1] can be used to ensure trustworthy measurements of τ. The scenario in which τ is unknown to the actuator (e.g., due to disrupted clock synchronization between the controller and actuator) is left to future work.

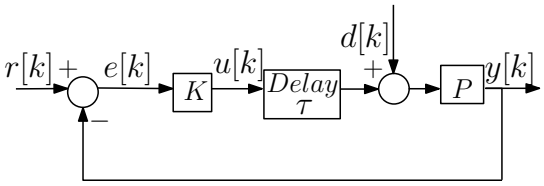


Fig. 1. A closed-loop control system.

C. Illustration of Stability and Safety with a Simple Control System under Delay Attack

We use the feedback control system in Fig. 1 to illustrate impacts of the attack on stability and safety. The results provide important observations that motivate the design of the attack impact assessment and mitigation approaches. In the absence of the attack, the system dynamics is

$$\begin{aligned} \mathbf{x}[t+1] &= \mathbf{A}\mathbf{x}[t] + \mathbf{B}(\mathbf{u}[t] + \mathbf{d}[t]), \\ \mathbf{y}[t] &= \mathbf{C}\mathbf{x}[t], \quad \mathbf{u}[t] = \mathbf{K}\mathbf{e}[t], \quad \mathbf{e}[t] = \mathbf{r}[t] - \mathbf{y}[t], \end{aligned} \tag{1}$$

where $\mathbf{x}$, $\mathbf{y}$, $\mathbf{d}$, $\mathbf{r}$, $\mathbf{u}$ and $\mathbf{e}$ are the system state, sensor measurement, disturbance, setpoint, control signal, and error signal, respectively; $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are system-specific matrices; $\mathbf{K}$ is a matrix characterizing the control law. Thus, the system employs proportional control. Note that the attack impact assessment and mitigation developed later in this paper do not depend on the control law. In particular, the AGC case study presented in this paper employs proportional-integral (PI) control. In another case study of power plant control, PID control is employed. We consider the delay attack on $\mathbf{u}$, as illustrated in Fig. 1. Because of the attack, the $\mathbf{u}$ in Eq. (1) will be a delayed version $\mathbf{u}[t-\tau]$, which is given by $\mathbf{u}[t-\tau] = \mathbf{K}(\mathbf{r}[t-\tau] - \mathbf{y}[t-\tau] = \mathbf{K}(\mathbf{r}[t-\tau] - \mathbf{C}\mathbf{x}[t-\tau])$. Thus, Eq. (1) becomes

$$\mathbf{x}[t+1] = \mathbf{A}\mathbf{x}[t] - \mathbf{B}\mathbf{K}\mathbf{C}\mathbf{x}[t-\tau] + \mathbf{B}\mathbf{K}\mathbf{r}[t-\tau] + \mathbf{B}\mathbf{d}[t]. \tag{2}$$

By the result in [32], a necessary and sufficient condition for the stability of the discrete time-delay system is that the eigenvalue of the maximal left solvent of $M(\mathbf{X})$, where $M(\mathbf{X}) = \mathbf{X}^{\tau+1} - \mathbf{X}^{\tau}\mathbf{A} - \tilde{\mathbf{B}}$, $\mathbf{X} \in \mathbb{C}^{n\times n}$ and $\tilde{\mathbf{B}} = -\mathbf{B}\mathbf{K}\mathbf{C}$, i.e., the eigenvalue of the solution for $M(\mathbf{X}) = 0$, is less than 1. From the expression of $M(\mathbf{X})$, the eigenvalue of the maximal left solvent only corresponds to the delay length τ and the system specific matrices. In the following, we will use one numerical example in Matlab [33] to explore the system stability and safety. The numeric results in the rest of this section are based on the following settings: $\mathbf{A} = [-1\ -3;\ 3\ -5]$, $\mathbf{B} = [2\ -1;\ 1\ 0]$, $\mathbf{C} = [0.8\ \ 2.4;\ 1.6\ \ 0.8]$, $\mathbf{K} = 2$. Moreover, we measure time in units of slot, which can be translated to actual time in a real system.

1) Impacts of delay on stability and safety: We run time-domain simulations to understand the system's stability and safety under different delays. The system output $\mathbf{y}$ over time under different settings is shown in Fig. 2. Both the delay against $\mathbf{u}$ and the step-change disturbance $\mathbf{d}$ of magnitude of 1.5 are introduced at $t = 50$. In Figs. 2(a) and 2(b), where $\tau = 2$ and $\tau = 3$, the system is convergent and divergent, respectively. The system becomes unstable when we increase the delay to 3 time slots. The safety classification depends on the safe range definition. For example, if we define the safe deviation range of $\mathbf{y}$'s components to be $[-1,\ 1]$, the system in Fig. 2(a) is safe. However, if the safe range is defined to be $[-0.4,\ 0.4]$, the system is unsafe. Thus, even if the system is stable, it can be either safe or unsafe, depending on the given safety conditions and the system's state trajectory.

2) Impacts of disturbance on stability and safety: Since stability is determined by the eigenvalue of the maximal left solvent of $M(\mathbf{X})$ only, it is not affected by the disturbance $\mathbf{d}$, so that $\mathbf{A}$ and $\tilde{\mathbf{B}}$ do not include $\mathbf{d}$. In contrast, as safety depends on the trajectory of $\mathbf{y}$, which depends on $\mathbf{d}$, the magnitude of $\mathbf{d}$ can significantly affect the safety. We now illustrate this observation using Fig. 2(c) that has the same setting as Fig. 2(a) except that the disturbances in Fig. 2(a) and Fig. 2(c) are 1.5 and 30, respectively. Fig. 2(c) shows larger output deviations, which may violate the safety requirement.

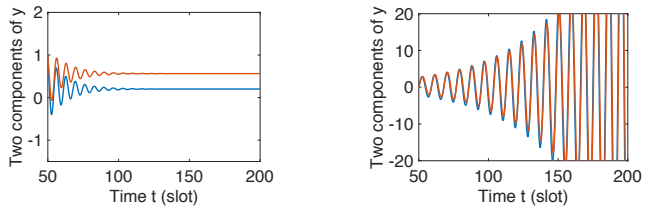
3) Impacts of initial state on stability and safety: As the eigenvalue of the maximal left solvent of $M(\mathbf{X})$ does not depend on the initial system state, the stability does not depend on the initial state. In contrast, since the initial state affects the system trajectory, it affects the system's safety. For instance, Fig. 2(d) has the same setting as Fig. 2(a) except that they have different initial states. The system remains convergent in this case, which generally implies a stable system. However, the output deviation is doubled compared with that of Fig. 2(a), and the larger deviation may violate safety.

In summary, we have these two observations: (i) the delay τ affects both stability and safety, (ii) the safety depends on the disturbance and the system's initial state, while the stability does not. These observations will guide the design of the proposed tandem stability-safety assessment method.

IV. Objective and Approach Overview

A. Objective and Challenges

We aim to develop delay attack impact assessment and mitigation for CPS control. The input for the assessment includes the measured delay τ and the measurements of sensors monitoring the system state. If the system is classified



$\tau = 2,\ d = 1.5,\ \mathbf{x}[0] = [1,\ 1]$ (a)

$\tau = 3,\ d = 1.5,\ \mathbf{x}[0] = [1,\ 1]$ (b)

$\tau = 2,\ d = 30,\ \mathbf{x}[0] = [1,\ 1]$ (c)

$\tau = 2,\ d = 1.5,\ \mathbf{x}[0] = [2,\ 2]$ (d)

Fig. 2. The system output $\mathbf{y}$ under different settings.

unsafe (i.e., it will enter an unsafe region), mitigation actions should be initiated to regain safety.

We face the following main challenges. First, although we can obtain an analytic stability condition for the simple system in Fig. 1(a), it is challenging to obtain similar conditions for real-world complex systems. Second, the safety classification needs the system's trajectory such as those shown in Fig. 2. Although we can use a high-fidelity simulator to predict the trajectory, the transient simulations for complex systems can be too slow for real-time online prediction and control. For instance, a transient simulation for the 37-bus power grid shown in Fig. V-A (see Section V-A) takes 138 s on a 28-core computing server, while the grid under attack takes less than two minutes to cross its safe range (cf. Table II in Section V-B2). Thus, the system will have well entered the unsafe region by the time the transient simulation completes. Third, as locating and removing an ongoing cyber-attack often takes significant time, before the attack is removed, it is critical to tolerate the attack and mitigate its impact by adapting tunable system parameters and settings. However, a model that characterizes the effects of the new parameters and settings on the safety will be needed to determine their suitable values. It is similarly challenging to obtain this model for complex systems.

B. Approach Overview

This section overviews our approach. In every time slot, if the measured total delay τ in transmitting sensor measurements and control commands exceeds a threshold (e.g., the typical communication delay), we execute the attack impact assessment and mitigation pipeline shown in Fig. 3. First, we classify the system's stability. If the system is unstable, which implies that it is unsafe, we initiate mitigation to restore safety; otherwise, we classify the system's safety. If and only if the system is classified unsafe, we initiate mitigation. We now

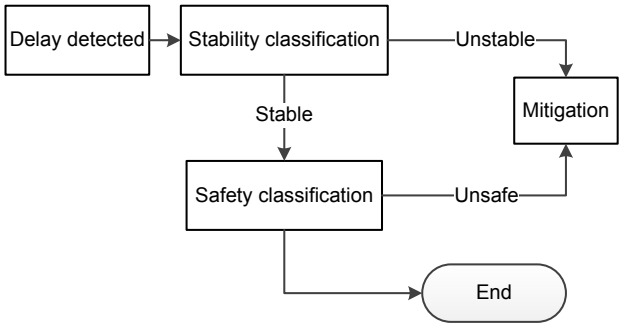


Fig. 3. Attack impact assessment and mitigation pipeline.

discuss the design of the stability and safety classification, as well as the mitigation, that addresses the challenges described in Section IV-A.

First, since it is difficult to analyze the stability and safety of complex systems, we use a simulation-based approach. We assume that a high-fidelity simulator that can accurately characterize the system dynamics is available. This assumption agrees with practice. For instance, power grid operators generally maintain high-fidelity simulators of their systems to guide design and operations. Using the simulator, we can explore key factors that affect the system's stability and safety.

Second, since the transient simulations, though accurate, are generally too slow for online use, we conduct offline simulations to generate extensive data with appropriate stability and safety labels. The labeled data will be used to characterize the stability and safety boundaries. However, the dependence of safety on the system's initial state, as illustrated in Section III-C3, leads to state explosion if we were to enumerate all the initial states during the generation phase of training data. To deal with this issue, we apply a Monte Carlo method to generate the training data and train an ML model to characterize the safety boundary. The ML model can also be used to guide the search for suitable mitigation actions.

Third, the ML model may introduce the error occasionally in the safety classification. On the other hand, as observed from the case studies in Section III-C and Section V, the stability classification is simpler, faster, and more accurate. Thus, we apply the stability classification first in the overall assessment, so that we can condition the safety classification on the more reliable and faster stability classification result. This conditional sequential strategy reduces the overall classification errors and runtime overheads.

We note that the detailed design of the components shown in Fig. 3 is system specific. However, we believe that the basic design paradigm is applicable to a wide range of CPSes, e.g., the process control in the chemical system and the train control system. In the rest of this paper, we will apply it to two power grid control systems, i.e., automatic generation control (AGC) and power plant control (PPC), and design the domain-specific components. AGC is a fundamental networked control system used in real-world power grids; PPC is a critical control system for power plants.

V. Stability-Safety Assessment for AGC

Since AGC involves long-range communications and its malfunction can cause grid-wide failures and infrastructure damage, it can be an attractive target for attackers. In Section V-A, we present necessary background of the AGC for our discussions. Section V-B presents extensive simulations to understand the AGC's stability and safety under the delay attack. Section V-C applies the proposed tandem stability-safety assessment to the AGC.

A. Background of AGC

AGC maintains the grid frequency at a nominal value (e.g., 60 Hz) by adjusting setpoints of generators. It also maintains the net power interchanges among neighboring areas at scheduled values [13]. Here, an area is a part of the grid and it is usually operated by a utility. Two areas are connected by *tie-lines*. Fig. V-A illustrates a three-area 37-bus system[1], where dotted lines represent the tie-lines. As illustrated in Fig. 5, the AGC, located in the grid control center, receives over a communication network measurements of the deviations of the grid frequency (from the standard frequency) and the ith area's power export from their respective setpoints (which are denoted by $\Delta\omega_i$ and ΔP_{Ei}), and it computes the *area control error* (ACE). The control center sends ACE_i to the area's power plants over the communication network. Each plant applies a PI controller with a gain of k to generate a reference signal for its generator. Specifically, the reference signal is $-k\int \mathrm{ACE}_i(t)\mathrm{d}t$. The above process is repeated every AGC cycle, which is often two to four seconds. The sensor measurements and ACE are transmitted in long-range communication networks that are susceptible to cybersecurity threats. In this paper, we focus on the delay attack against transmissions of ACE signals. However, our approach can be readily applied to delay attacks on sensor measurements, or both ACE signals and sensor measurements. Note that we do not consider the deadband effect in this work, i.e., there is a deadband around the frequency nominal value. However, according to several power system control papers on AGC, e.g., [35] and [36]. the non-linear effect of the deadband can be mitigated with proper design of the AGC controller. Thus, we can extend our work to such kind of AGC controllers to mitigate the deadband effect.

B. AGC's Stability and Safety under Delay Attack

This section presents two extensive simulation studies to investigate how the following factors may affect the AGC's stability and safety: (i) the grid's total load, (ii) the distribution of the load among the load buses, (iii) the change of load, and (iv) the communication delay. We note that the load distribution determines the power system's state, which is often defined as the union of all the buses' voltage phasors.

[1]We use the 37-bus system as a case study throughout this paper. It is a test system [34]. Its scale corresponds to a small-/mid-scale grid in real life. According to our rough count based on a grid topology database (http://bit.ly/2vRH5Nd), a major fraction of 130 national grids consist of fewer than 37 buses.

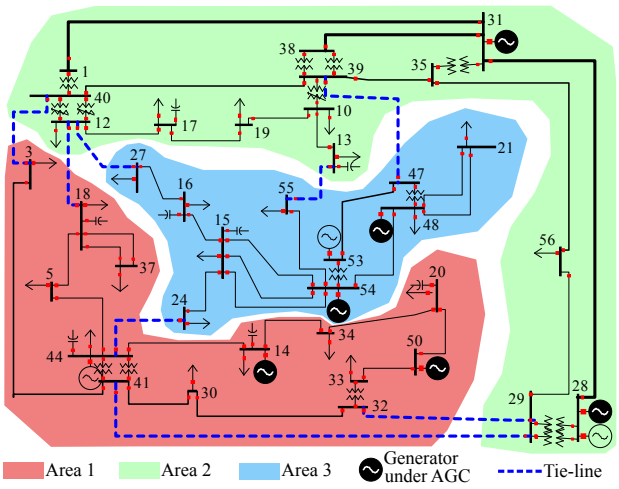


Fig. 4. A three-area 37-bus system (Adopted from [26]). Each area is a part of the grid and operated by a utility. Two areas are connected by tie-lines, i.e., the dashed lines in the figure.

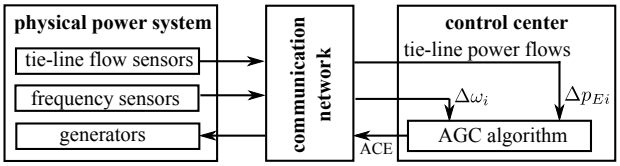


Fig. 5. Overview of AGC (Adopted from [26]).

Thus, the total load can be considered a statistic of the system's initial state. The load change is the primary exogenous disturbance to the AGC. The simulations are conducted using PowerWorld, an industry-strength high-fidelity power system simulator, based on the system model in Fig. V-A. The main simulation settings are: the length of a time slot is 1 s[2]; the length of an AGC cycle is 4 s (see Chapter 11 of [13]); each simulation lasts for 300 s; the delay attack on the ACE signal is launched at $t = 120$ s; the load change occurs at $t = 140$ s.

1) AGC's stability: The stability is assessed by checking the system's convergence, i.e., whether the power system frequency will be convergent in the AGC control. We have the following observations.

AGC's stability depends on the total load: Fig. 6 shows the AGC's stability boundary under different total loads and delays. Each point in the curve represents the maximum delay the system can keep stable under corresponding total load, i.e., for a total load from X-axis, if the delay is larger than the maximum delay in the curve, the system will be unstable. Thus, all these points in the curve form the boundary to separate the stable and unstable regions. In our simulation, a total of 7,900 combinations of the total load and delay are tested. We can see that the total load affects the maximum delay that the system can tolerate to keep stable. For instance, when the total load is 600 MW, the maximum tolerable delay is 6 s. When the total load is 1000 MW, the maximum tolerable delay is 2 s only. Fig. 6 also shows a clear cut boundary

[2] This setting well balances the simulation computation overhead and the fidelity.

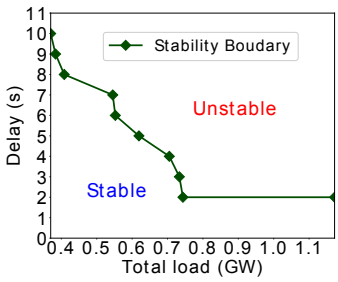


Fig. 6. AGC's stability under different total loads and delays.

TABLE I
MAX TOLERABLE DELAY.

Load change	Total load 715	795	874
–10%	5	3	3
–5%	5	3	3
0	5	3	3
5%	5	3	3
10%	5	3	3

[a]The delays are in seconds.
[b]Total Loads are in MW.

between the stable and unstable regions.

AGC's stability is independent of the detailed load distribution: We fix the total load at 795 MW and distribute it among the load buses randomly. Simulations using 1,000 random load distributions show that the maximum tolerable delay is always 2 s. Under other settings of the total load, the maximum tolerable delay is also a constant over the different load distributions. This gives strong empirical evidence that the AGC's stability is independent of the load distribution. The observation is consistent with the standard practice of analytical modeling of AGC, which considers the total load only but not the load distribution [13].

AGC's stability is independent of load change: Table I shows the maximum tolerable delay under different settings of the total load and the load change as percentage of the total load. The load change consists of step changes at all the load buses at $t = 140$ s. The step change is realistic given increasing adoption of demand response and distributed renewable energy sources that can trigger sudden changes in load. From the table, for each tested total load setting, the AGC's stability is unaffected by the change. This result is consistent with our discussions in Section III-C2. Moreover, with less total load, the system can tolerate longer delays, which is consistent with the results in Fig. 6.

2) AGC's safety: We impose the following two safety requirements. First, the grid frequency deviation must be within $[-0.5\,\text{Hz}, 0.5\,\text{Hz}]$. In real systems, if the deviation exceeds this safe range, disruptive remedial actions such as load shedding will be automatically initiated to protect the grid from infrastructural damage [13]. Second, the power flows must be within capacities of the transmission lines. Otherwise, the lines will trip due to overheating. In our simulations, we adopt the default line capacities of the 37-bus system.

TABLE II
TIME TO CROSS THE SAFE RANGE VS. DELAY AND LOAD CHANGE.

Load change (MW)	Delay (s) 0	1	2	3
-80	105.45	105.45	105.7	105.8
-40	∞	∞	∞	276.1
0	∞	∞	∞	944.3
40	148.6	148.6	148.6	148.6
80	146.1	146.1	146.1	146.1

*The time values are in seconds; ∞ means the system is safe.

AGC's safety depends on load change: The total load is 800 MW. Table II shows the time from the launch of the delay attack to the breach of the safety requirement under different delays and load changes. The symbol ∞ means that the safety limits are never crossed, i.e., the system is safe. From the table, the AGC's safety is affected by the load change, which is consistent with our discussion in Section III-C2. For instance, when the load change is 5% of the total load (i.e., 40 MW), the system will be unsafe, regardless of the delay. When the load change is small, the system will be safe if the delay is also small. Thus, the load change and delay jointly affect the safety.

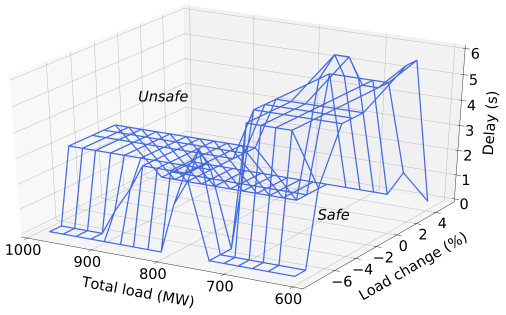
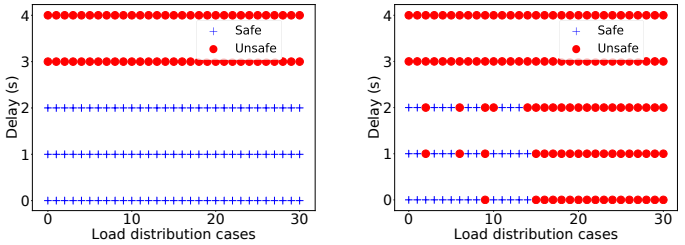


Fig. 7. The minimum delay leading to unsafety vs. total load and load change.

AGC's safety depends on total load: Fig. 7 shows the minimum delays that lead to unsafety under different total loads and load changes. Each grid point represents such a minimum delay obtained by running a set of simulations under different delays. Note that, to simplify the illustration, we relax the transmission line capacities to infinite, such that the load distribution does not affect the safety. The next set of experiments will show the impact of the load distribution on the safety under finite line capacities. In Fig. 7, the surface formed by the grid points that represent the obtained minimum delays leading to unsafety divides the space into safe and unsafe regions, which are below and above the surface, respectively. The result shows that the total load, the load change, and the delay jointly affect the AGC's safety.

AGC's safety depends on load distribution: We fix the total load at 800 MW and distribute it among the load buses randomly. Fig. 8(a) and Fig. 8(b) show the classification of the AGC's safety given different delays in 30 cases of the load distributions, when the line capacities are set to be infinite and finite, respectively. Although the line capacities are finite in practice, we present the infinite case to help understand the affecting factors of the AGC's safety. Under infinite line capacities, the AGC's safety depends on the frequency deviation only. The deviation depends on the total load, rather than the load distribution. Thus, in Fig. 8(a), the safety is independent of the load distribution. In contrast, since power flows depend on the load distribution, under finite line capacities, the load distribution will affect the AGC's safety. In Fig. 8(b), for a given delay, the system may be safe or unsafe depending on the load distribution.

3) Summary: The above experiments show that the AGC's stability depends on the total load and the delay, while its safety additionally depends on the load change and the load distribution. This observation is mostly consistent with that for the barebone control system in Section III-C, except that the AGC's stability depends on the total load, a statistics of the



(a) (b)

Fig. 8. AGC's safety under different load distributions. (a) Infinite line capacity; (b) Finite line capacity.

system state. This can be explained from the fact that AGC is a nonlinear system, although its control-theoretic analysis is often based on a linearization at the system's current condition as characterized by the total load [13]. Thus, the AGC's stability condition is also affected by the total load. However, this minor deviation will not impede the application of the tandem stability-safety assessment, since the scalar total load will not lead to a state explosion problem.

C. Stability-Safety Assessment for AGC under Delay Attack

This section applies the proposed tandem stability-safety assessment to AGC. From Fig. 6, since the AGC's stability has a clear cut boundary in the two-dimensional space formed by the total load and the delay, it can be classified quickly at run time based on the boundary *a priori* obtained through extensive offline transient simulations. We call this classification approach *boundary-based stability classification*. Specifically, if the system's current operating point (i.e., total load and delay) is below the boundary, such as that shown in Fig. 6, the system is stable; otherwise, it is unstable. This classification avoids running a time-consuming online transient simulation based on the system's current operating point. In particular, due to the limited dimension of the stability space (i.e., two), we can achieve any granularity in enumerating operating points within any specified range. As a result, the boundary-based approach achieves perfect classification accuracy asymptotically as the enumerating granularity goes to zero.

In contrast, AGC's safety additionally depends on the load distribution vector, which has exponential complexity with respect to the number of load buses that is often tens to hundreds. To avoid the exponential complexity, we use a Monte Carlo method to randomly sample the operating points in a discretized state space and generate extensive offline simulation results with determined safety labels to train an ELM [15] to characterize the AGC's safety. The ELM is a single hidden layer feedforward neural network with a training algorithm much faster than conventional gradient-based learning algorithms. At run time, the trained ELM classifies the AGC's safety based on the current operating point (i.e., total load, load change, load distribution, and delay). We summarize the algorithm for safety assessment in Fig. 9. In Section VII, we will compare the performance of the ELM with a baseline approach that also uses the training data to classify safety.

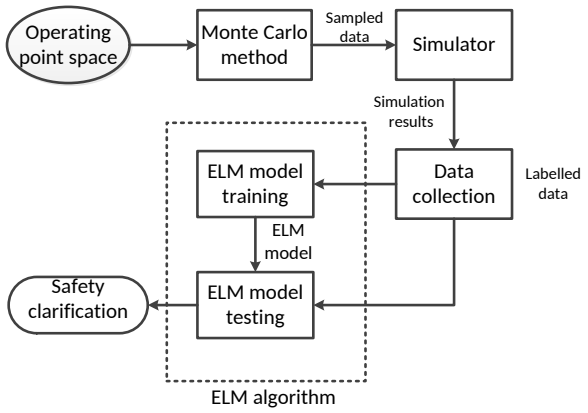
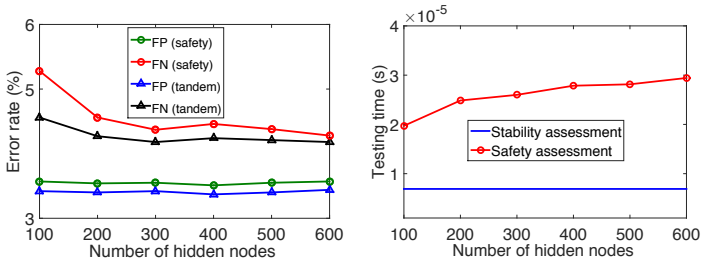


Fig. 9. Algorithm overview for safety assessment.

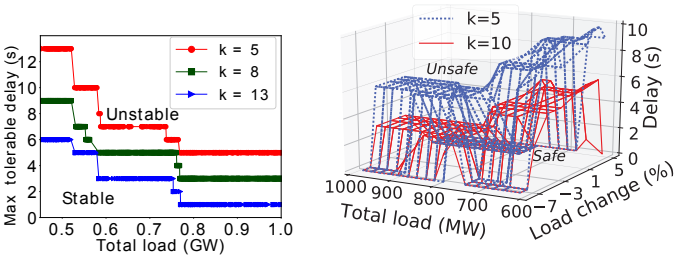


(a) (b)

Fig. 10. FP, FN rates, and testing time versus the number of hidden nodes in ELM. (a) FP and FN rates. (b) Testing time.

We present the following numeric results to show the effectiveness of the ELM-based safety assessment. The training and testing data sets consist of 11,000 and 7,000 operating points and their safety labels, respectively. We use the false positive (FP) and false negative (FN) rates as the accuracy metrics, which are the percentages of safe (resp., unsafe) cases that are wrongly classified to be unsafe (resp., safe). The green and red curves in Fig. 10(a) show the ELM's FP and FN rates versus the number of hidden nodes in the ELM. The two rates are generally below 5%. In Section VII-C, we will discuss how to deal with the FPs and FNs. When the number of hidden nodes is 300, both the two rates reach their knee points. Thus, 300 is a satisfactory setting, since using more hidden nodes does not improve the accuracy much, but it increases the testing time as shown by the red curve in Fig. 10(b). Under the setting of 300, the testing time is around 0.03 ms only on an Intel i7 2.2GHz CPU. This time is short compared with the time horizon of a power grid's fault clearing (e.g., 200 ms for lightning strike overcurrent clearing). The testing time can be further reduced significantly by using hardware acceleration.

Lastly, we show the benefits of the tandem stability-safety assessment. First, as the boundary-based stability classification gives asymptotically perfect accuracy, it helps reduce FNs of the ELM-based safety classification. The blue and black curves in Fig. 10(a) show the FP and FN rates of the tandem stability-safety assessment. FN rate is reduced by up to 1%. Second, the blue curve in Fig. 10(b) shows the testing time of the boundary-based stability classification, which is 11 microseconds only, 3 times shorter than that of the ELM's



(a) Stability boundaries (b) Safety boundaries

Fig. 11. System stability and safety boundaries under different k settings, where k is the gain of the PI controller to produce a reference signal for the plant's generator as introduced in Section V-A.

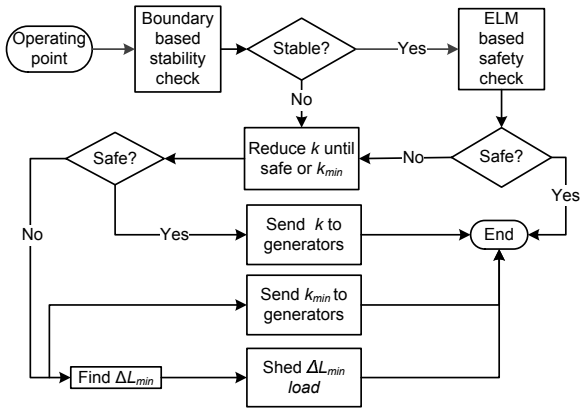


Fig. 12. Two-tier delay attack impact mitigation.

testing time with 300 hidden nodes. Thus, under the tandem approach, any instability will be detected by the fast stability classification, which improves the timeliness of the needed mitigation (cf. Section VI). In Section VII-C, we will evaluate the impact of an FP and describe an approach to further reduce the FN rate.

VI. Mitigating Impact of Attack against AGC

This section presents an approach to mitigating the delay attack impact on AGC. As the total load is an important determining factor for both stability and safety, a feasible approach is to shed load to restore safety. However, clearly, load shedding will affect customers adversely, sometimes severely. Hence, it should be avoided if possible. This section proposes a two-tier approach that firstly tunes the AGC gain as a first-line defense, and resorts to shedding load only when the gain tuning is insufficient. This section studies the impact of the gain on the AGC's stability and safety first in Section VI-A. Then, it presents the two-tier approach in Section VI-B.

A. Impact of AGC Gain on Stability and Safety

As discussed in Section V-A, each power plant applies a PI controller with a gain of k to the received ACE to produce a reference signal for the plant's generator. We conduct simulations based on the 37-bus system model to investigate the impact of k on the AGC's stability and safety. The curves and surfaces in Figs. 11(a) and (b) show the stability and safety boundaries, respectively, under different settings of k. By reducing k, we can expand the stable and safe regions. However, from control theory, a smaller k will result in slower convergence when there is a load change. Hence, we have a trade-off between (i) AGC's tolerance to the delay in terms of stability and safety, and (ii) AGC's convergence speed in response to a load change. As AGC generally also needs to meet some required convergence speed, there exists in practice a minimum allowable setting for k [13], which is denoted as $k_{\min}$. Multiple ELMs are trained to characterize the safety boundaries under different settings of k. This *ELM bank* will be used in Section VI-B to find a k to restore safety where needed.

B. Two-Tier Delay Attack Impact Mitigation

Fig. 12 illustrates the integrated stability-safety assessment and attack impact mitigation. When a system is classified unstable or unsafe, the two-tier mitigation is activated. No mitigation is needed only when the system is classified safe. The two-tier mitigation works as follows. First, within the range from $k_{\min}$ to the current setting of k, we search for the maximum setting of k that can restore safety using the ELM bank discussed in Section VI-A. If such a k setting is found, it is piggybacked onto the next ACE signal that will be sent to generators. Otherwise, load shedding should be applied. We use the ELM bank to find the minimum amount of load that needs to be shed to restore safety under the setting $k_{\min}$. This minimum amount is denoted by $\Delta L_{\min}$. The grid operator sheds $\Delta L_{\min}$ load and piggybacks the $k_{\min}$ to the next ACE signal that will be sent to generators. The shedding amount can be shared among load buses equally or using existing scheduling algorithms addressing other grid operation optimization objectives and constraints [37]. Once a generator receives the new AGC gain, it updates its setting accordingly.

In our mitigation approach, we use the classical frequency restoration techniques to mitigate the impact of the delay attack. This is different from the techniques like time stamping or watchdog, where they are used for the attack detection as mentioned in Section III-B but not for mitigating the impact of delay attack before it is removed from the system. There can be other techniques that is capable of managing the delays in the control system, e.g., software-defined networking (SDN). However, managing the delays using SDN is a non-trivial task. We have a recent work in [38] to study this problem.

VII. Performance Evaluation

This section evaluates several key aspects of our attack impact assessment and mitigation designed for the AGC of the 37-bus system shown in Fig. V-A.

A. Effectiveness of ELM-Based Safety Classification

We compare the proposed ELM-based approach with a data-driven baseline approach. Specifically, the baseline finds a system operating point within the ELM's training data that has the smallest Euclidean distance to the system's current operating point, and yields the found operating point's safety label. Fig. 13 shows the classification error rates of our ELM-based and the baseline approaches under different settings of

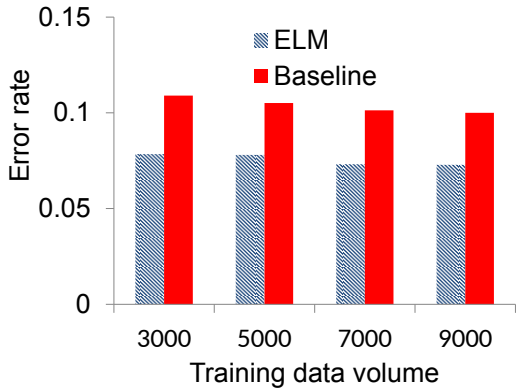


Fig. 13. Comparison between ELM-based and data-driven baseline approaches.

training data volume. Consistent with intuition, the error rate decreases with the volume of training data. The ELM-based approach gives lower error rates. Moreover, the running time for the ELM-based approach is up to 6,000 times shorter than that of the baseline approach.

B. Effectiveness of Attack Impact Mitigation

We conduct two simulations to show the effectiveness of our two-tier attack mitigation. The system's total load is 1000 MW. The initial setting for k is 10. The safety requirement for the grid frequency deviation is $[-0.5\,\text{Hz}, 0.5\,\text{Hz}]$. The attacker delays the ACE signal by 4 s from $t = 120$ s. The attack impact assessment classifies the system safe until a step load change is introduced at $t = 140$ s. In Fig. 14(a), the load change is 5% of the total load. At this moment, the system is classified unsafe. The red curve in Fig. 14(a) shows the system's trajectory if no mitigation is applied. It confirms the assessment result. The mitigation approach starts searching for a k setting to regain safety. By decreasing k from 10 to $k_{\min} = 5$, the system is classified safe under the attack. The thick green curve in Fig. 14(a) shows the system's trajectory after the new setting $k = 5$ is applied. We can see that the system becomes safe after the mitigation. In Fig. 14(b), the load change is 8% of the total load. Because of the increased load change, tuning k to $k_{\min} = 5$ is insufficient and shedding 10% of load is needed to restore safety. The thick green curve in Fig. 14(b) shows the system's trajectory after load shedding and reconfiguring k. The system is safe after the mitigation. The effects of different mitigation approaches on the customers are different. In Fig. 14(b), as tuning k to $k_{\min}$ still cannot mitigate the attack impact, we have to shed some of the customer loads, which results in lower utility to the owners. In Fig. 14(a), as the mitigation is achieved by adjusting the AGC parameters only, no customers will be affected.

C. False Positives and Negatives in Safety Classification

While the ML deals with the state explosion problem, it results in FPs and FNs. An FP will trigger the attack mitigation. Fig. 15(a) shows the system's trajectory after the mitigation wrongly triggered by a safety classification FP caused by a load change that is 0.5% of the total load, where the ACE signal is delayed by 2 s from $t = 120$ s. As the

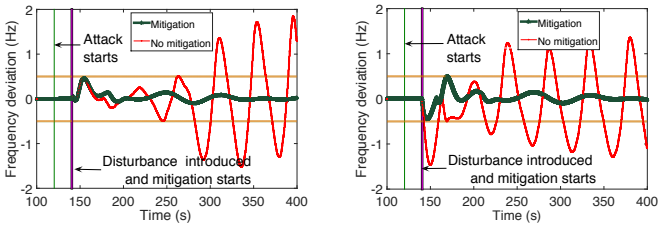


Fig. 14. Attack impact mitigation examples. (a) Tuning k only; (b) Tuning k and shedding load.

mitigation applies a small adjustment only (i.e., decrease k from 10 to 8), the frequency deviation has a slightly longer settling time. Moreover, Fig. 15(b) shows another scenario of the system's trajectory after the mitigation wrongly triggered by a safety classification FP caused by a load change that is 0.5% of the total load, where the ACE signal is delayed by 5 s from $t = 120$ s. The mitigation sheds 8% of the total load after decreasing k from 10 to 5; the frequency deviation can even have a shorter settling time. This is because the mitigation speeds up the system to diminish the small fluctuations due to the delay. Therefore, as FPs mostly occur for marginally safe operating conditions, the triggered mitigation is generally of small strength. The weak mitigation can lead to a slight settling time increase. Sometimes, it can even help decrease the settling time, which mitigates the concern for FPs.

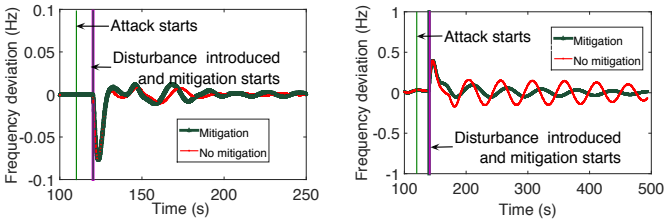


Fig. 15. Mitigation is wrongly triggered. (a) By a safety classification false positive and k is tuned; (b) By a safety classification false positive and load shedding is conducted.

In contrast, the system may become unsafe due to FNs. We discuss a sliding window approach as illustrated in Fig. 16(a) to reduce the FNs. In this approach, the load change is defined as the difference between the current load and the load in the

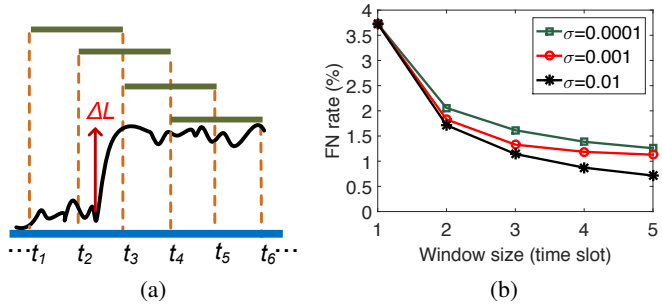
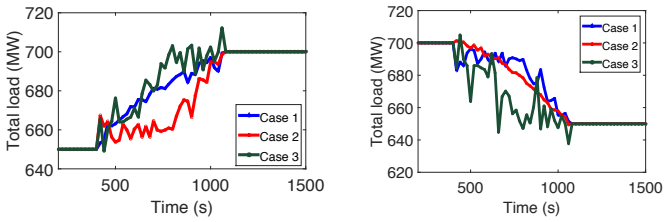
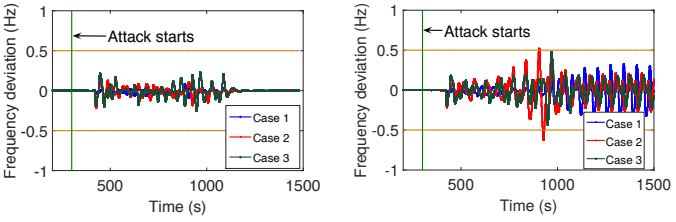


Fig. 16. (a) Sliding window approach. The time window is set as two time slots and the step load change will be assessed twice at $t = t_3$ and $t = t_4$. (b) FN rate vs. window size. The window size is increased from 1 to 5 and the standard deviations of three different random load fluctuations are illustrated.



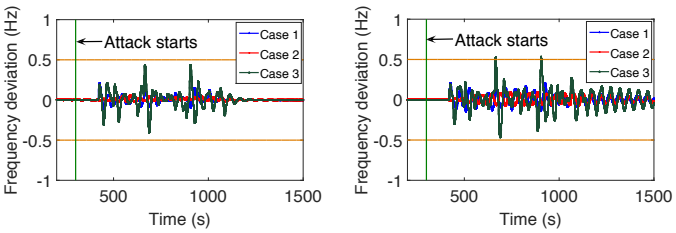
(a) (b)

Fig. 17. Trajectories of load changes. (a) Trajectories of load increasing. The load starts to change from $t = 450$ s and lasts 600 s. The total load increases by 50 MW in all three randomly generated trajectories. (b) The load starts to change from $t = 450$ s and lasts 600 s. The total load decreases by 50 MW in all three randomly generated trajectories.



(a) (b)

Fig. 18. System output for different load increasing trajectories in Fig. 17(a) under different delay. (a) Delay by 3 s (b) Delay by 6 s.



(a) (b)

Fig. 19. System output for different load decreasing trajectories in Fig. 17(b) under different delay. (a) Delay by 2 s (b) Delay by 4 s.

previous time window. As a result, a step load change will be assessed for multiple times. For instance, in Fig. 16(a), the time window is two time slots and the step load change will be assessed twice at $t = t_3$ and $t = t_4$. Due to random temporal fluctuations of the load, the probability that an unsafety can be detected in at least one of the multiple assessments will increase, thus reducing the FN rate. By increasing the window size, a load change will be assessed for more times. Fig. 16(b) shows the FN rate versus the window size under different random load fluctuations' standard deviations (σ). The FN rate decreases with the window size. Thus, this approach can effectively reduce the FN rate. The concern for increased FP rate due to this approach is minor since the FPs cause little impact on the system as illustrated earlier.

D. Impact of Gradual Load Changes

In the above, we have considered the sudden change of the load, i.e., the disturbance. We now discuss a more complicated case where the load varies gradually. In the following, we will show how the different trajectories of the load change affects the system stability and safety using the three-area 37-bus system as shown in Fig. V-A. Instead of mimicking the sudden change of the load, we now randomly generate the load profiles that vary differently within the same period. Fig. 17(a) and Fig. 17(b) show different trajectories of the increasing and decreasing load, respectively. In each figure, we randomly generate three different loads trajectories. Fig. 18 shows the system frequency deviation for increasing load trajectories in Fig. 17 under 3 s and 6 s delays. Figs. 18(a) and (b) show that the system stability is not affected by the load trajectories. This result is consistent with our analysis. Specifically, in Fig. 18(a), even under the 3 s delay, the system can still converge to the nominal value. In Fig. 18(b), although the system suffers small oscillations around the nominal value due to the delay, the system is still stable for different load trajectories. For the system safety, it can be relevant to the load trajectories, i.e., the frequency deviation increases with the change of load. For example, at around $t = 900$ s both Case 2 and Case 3 have large load changes, and the corresponding frequency deviations are also high. For Case 2, it can even violate our safety range, i.e., violate the lower bound. We can obtain similar results in Fig. 19 too.

From these observations, we can see that the system stability is independent of the load change, which is the same as our observation in Section V-B1. For safety, as the load trajectory keeps changing, the safety can also be affected, which is also the same as our summary in Section V-B2.

VIII. Stability-Safety Assessment for PPC

In this section, similar to the assessment in AGC, we first present extensive simulations to understand PPC's stability and safety under the delay attack. Then we apply the stability-safety assessment approach to a PPC system. In the end, the mitigation approach is used to mitigate the attack impact.

A. Stability and Safety in PPC under the Delay Attack

We use a PPC model in ThermoPower [39], an open-source library based on Modelica, to under simulate the PPC system. Note that Modelica is an object-oriented complex physical system modeling language [40]. The signal flow graph of the system is shown in Fig. 20. The controlled power plant admits two inputs, the power control signal and the void fraction control signal. The void fraction is also known as porosity, which is an important parameter characterizing two-phase fluid flow, especially gas-liquid flow. The two control signals are determined respectively by two PID controllers. The power controller's feedback signal is corrupted by additive zero-mean Gaussian noises acting as disturbances to the system. The adversary delays the power controller's output signal.

Similar to the discussion in Section V-B, we consider total load and disturbance as the factors that can affect the system stability and safety. In the PPC system, the power setpoint corresponds to the total load in the connected power grid. The noise is introduced to represent the disturbance in the system, while we use load changes to represent the disturbance in discussing AGC. Different from the AGC, since the transmission and distribution systems are transparent to the

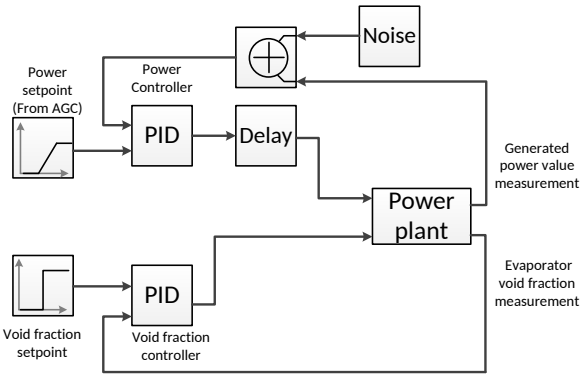
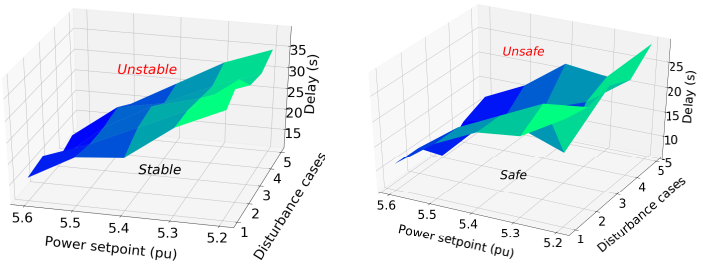


Fig. 20. A PPC system. The system has two inputs, the power control signal and the void fraction control signal.



(a) (b)

Fig. 21. The minimum delay leading to instability (a) and unsafety (b) vs. load setpoint and disturbance. For each load setpoint, we consider 5 different disturbance cases.

power plant, the load distribution is not a factor related to the assessment.

Fig. 21 shows the minimum delays that lead to unstable (a) and unsafe (b) under different total loads (i.e., power setpoint in per unit (pu)) and disturbance cases. Each grid point represents such a minimum delay obtained by running a set of simulations under different delays. In Fig. 21, the surface formed by the grid points that represent the obtained minimum delays leading to instability (a) or unsafety (b) divides the space into unstable and stable regions in (a) or unsafe and safe regions in (b), which are above and below the surface, respectively.

1) PPC's stability: The stability is assessed by checking the system's convergence, i.e., whether the system state signal, e.g., gas flow pressure, converges after the attack is launched. In the worst case, the system can halt automatically due to the system state divergence.

PPC's stability depends on the total load: The result in Fig. 21(a) shows that, when the total load is increased, the PPC has the trend to be more unstable, i.e., shorter delay can cause the system unstable. For example, for the disturbance case 1, when we increase the total load, i.e., the power setpoint, from 5.2 to 5.6, the minimum delay to make PPC unstable decreases from 35 s to 11 s, which means that a shorter delay can make the system unstable.

PPC's stability depends on the disturbance: In Fig. 21(a), the result also shows that, under different disturbances, the minimum delay leading to system unstable is quite different even under the same power setpoint. For example, when power setpoint is 5.6, the minimum delay to make PPC unstable can change from 12 s to 17 s. This is different from the case in AGC where the disturbance does not affect the stability.

2) PPC's safety: The safety is assessed by imposing the safety requirement, i.e., in our case, we require the gas flow pressure deviation must be within $[-0.02, 0.02]\times$ 1e5 Pa. Note that, this range is defined by our observations in thousands of simulations that gas flow pressure signal is mostly in this range under our system settings. As we discussed the definition for safety in Section III-A, the safety range can be other specified ranges based on the system setting and the operator's requirements.

PPC's safety depends on the total load: The result in Fig. 21(b) shows that, generally, when the total load is increased, the PPC generally has the trend to be unsafe, i.e., shorter delay can cause the system unstable. But this is not always true. For example, for the disturbance case 3, when we increase the total load, i.e., the power setpoint, from 5.2 to 5.3, the minimum delay to make PPC unsafe is deceased from 27 s to 12 s. But, if we keep increasing the power setpoint from 5.3 to 5.5, the minimum delay increases to 15 s first when the power setpoint is 5.4 and goes down to 14 s when power setpoint is 5.5, which means the safety boundary is not strictly decreasing when we increase the load. This is consistent with the case in AGC as introduced in Section V-B2.

PPC's safety depends on the disturbance: In Fig. 21(b), the result also shows that, under different disturbances, the minimum delay leading to system unsafe is quite different even under the same power setpoint. For example, when power setpoint is 5.3, the minimum delay to make PPC unstable can change from 12 s to 25 s. This is also similar to the case in AGC that the disturbance affects the safety.

3) Summary: The above experiments show that both PPC's stability and safety depend on total load, disturbances and the delay. This observation is still partly consistent with the results for the barebone control system in Section III-C, except that the PPC's stability also depends the total load and the disturbance. Similar to the case in AGC, this can be explained from the fact that PPC is a non-linear system, although the control-theoretic analysis in Section III-C is based on the linearization at the system's current condition as characterized by both the total load and the disturbance. Since both stability and safety depend on the disturbance, which can have exponential complexity with different settings, there are no clear boundaries for both stability and safety. Therefore, similar to the safety case in AGC, we use the ELM-based machine learning approach to classify both stability and safety.

B. Effectiveness of ELM-Based Stability-Safety Classification

Due to the highly complexity in the power plant, the clear stability boundary is not available in PPC. Thus, we use offline simulations together with ML to model the system's stability and safety. Specifically, we use OpenModelica [41], a Modelica-based simulator, to run massive offline simulations

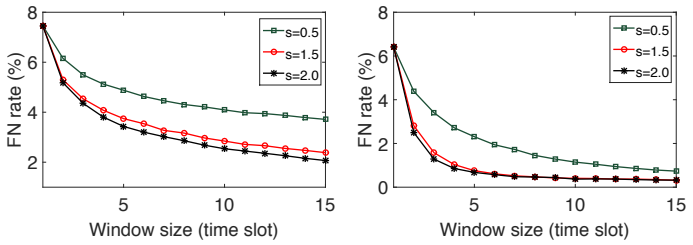


Fig. 22. FN rates of stability (a) and safety (b) assessment vs. window size. The s is the noise generator sampling period. A smaller s value means a larger level of disturbance to the system.

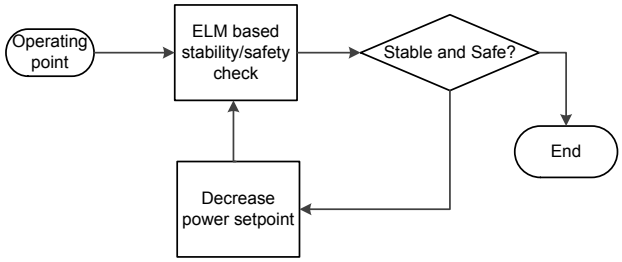


Fig. 23. The diagram of mitigation in PPC.

under a wide range of settings to generate training data. Then, we train the ELMs to model the stability and safety. Inputs to the ELMs include the sampling period and variance of the Gaussian noise as well as the delay. The output is the stability or safety assessment result. We apply the sliding window approach illustrated in Fig. 16 to improve the FN rate. The training and testing data sets consist of 10,000 and 3,500 operating points, respectively.

Figs. 22(a) and 22(b) show the FN rates of the stability and safety assessments, respectively, under various settings of the sliding window size. The different curves are the results under different settings of the Gaussian noise generator's sampling period in seconds, i.e., the s values in the legends. We note that different from the simple control loop in Section III-C the AGC, which are discrete-time systems, the PPC is a hybrid system with discrete-time sensing but continuous-time control and actuation. Thus, the Gaussian noise generation, which belongs to the sensing part, is in discrete time. As a result, the frequency at which we update the noise affects the level of the disturbance to the system. Specifically, a smaller noise sampling period causes a higher disturbance. From Figs. 22(a) and 22(b), similar to the results for the AGC, the FN rate increases with the window size and decreases with the disturbance level.

C. Effectiveness of Attack Impact Mitigation

For attack mitigation, similar to the mitigation approach presented in Section VI, we can build ELMs for a range of PID configurations and then tune the PID configuration. Since there are two PID controllers, we can tune either of them. In our approach, we choose the PID controller with the power setpoint as the input, which corresponds to the total load in the power system, to tune. We use Fig. 23 to illustrate

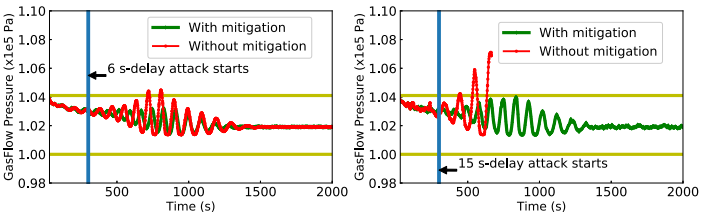


Fig. 24. Attack mitigation examples in PPC. The attack is launched at $t = 300s$ s by delaying the PID control command by 6 s and 15 s, respectively. We consider the effect of applying and without applying mitigation to the system.

the mitigation approach in the PPC system. Once we know the system will be either unstable or unsafe by running the ELM algorithm, we decease the power setpoint, i.e., decrease the total load in the power system. This is motivated by the observations in Fig. 21 that when the system power setpoint is lower, the system is generally more stable and resilient. After that, by applying our optimal load scheduling strategy in [37], the load in the power system can be balanced. In Fig. 24, we show that the attack is launched at $t = 300$ s. The 6 s and 15 s delay attacks are respectively applied to the PPC system. In Fig. 24(a), the 6 s delay attack makes the system unsafe, i.e., the peak point of the fluctuation is larger than the safety threshold defined by our setting. Moreover, in Fig. 24(b), the 15 s delay attack makes the system unstable, i.e., the system is crashed without mitigation under the attack. But if we apply the mitigation approach as shown in Fig. 23, we can greatly shrink the fluctuation when the attack is launched and ensure the system is safe and stable in both figures in Fig. 24.

IX. Conclusion and Future Work

This paper presented an efficient delay attack impact assessment approach that applies a stability classifier and an ML-based safety classifier sequentially. The ML addresses the state explosion problem in the safety classification due to the dependence of the system's safety on the multi-dimensional system state. The tandem stability-safety design improves the accuracy of the unsafety detection and speeds up the overall assessment. We applied our approach to power grid AGC, and developed a two-tier attack impact mitigation that tunes the control gain as a first-line defense and resorts to shedding load only if the gain tuning is insufficient to regain safety. Simulations based on a 37-bus system model verified and illustrated the effectiveness of our assessment and mitigation approaches. We also applied our approach to assess the stability and safety of a PPC system and proposed the mitigation approach. We presented the evaluation results based on Modelica simulations. Although we have evaluated the impact of load changes on safety, considering the trajectories details can make the evaluation of the system safety becomes even more complicated, as it is relevant to the dynamics of the system load. Other advanced machine learning techniques, for example, the recurrent neural network [42], may be needed to capture the relationship among the time series. This is an open issue for future research.

REFERENCES

[1] S. Viswanathan, R. Tan, and D. Yau, "Exploiting electrical grid for accurate and secure clock synchronization," *ACM Trans. on Sensor Networks*, vol. 14, no. 2, 2018.

[2] Y. Zhang and V. Paxson, "Detecting stepping stones," in *USENIX Security Symposium*, 2000.

[3] "Hackers infiltrated power grids in U.S., Spain," 2014, https://bit.ly/2E5FHyE.

[4] Y. Zhang and V. Paxson, "Stuxnet worm impact on industrial cyberphysical system security," in *IECON*, 2011.

[5] B. Chen, S. Mashayekh, K. Butler-Purry, and D. Kundur, "Impact of cyber attacks on transient stability of smart grids with voltage support devices," in *IEEE PES General Meeting*, 2013.

[6] X. Cao, P. Cheng, J. Chen, S. Ge, Y. Cheng, and Y. Sun, "Cognitive radio based state estimation in cyber-physical systems," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 489–502, 2014.

[7] A. Farraj, E. Hammad, and D. Kundur, "A cyber-physical control framework for transient stability in smart grids," *IEEE Trans. Smart Grid*, vol. 4, no. 2, pp. 847–855, 2013.

[8] D. Rabadi, R. Tan, D. Yau, and S. Viswanathan, "Taming asymmetric network delays for clock synchronization using power grid voltage," in *ACM ASIACCS*, 2017.

[9] S. Xu and J. Lam, "On equivalence and efficiency of certain stability criteria for time-delay systems," *IEEE Trans. Autom. Control*, vol. 52, no. 1, pp. 905–101, 2007.

[10] C. Zhang, L. Jiang, Q. Wu, Y. He, and M. Wu, "Further results on delay-dependent stability of multi-area load fequency control," *IEEE Trans. Power Systems*, vol. 28, no. 4, pp. 4465–4474, 2013.

[11] S. Sönmez, S. Ayasun, and C. Nwankpa, "An exact method for computing delay margin for stability of load frequency control systems with constant communication delays," *IEEE Trans. Power Systems*, vol. 31, no. 1, pp. 370–377, 2016.

[12] R. Tan, H. Nguyen, and D. Yau, "Collaborative load management with safety assurance in smart grids," *ACM Trans. on CPS*, vol. 1, no. 2, 2017.

[13] P. Kundur, *Power System Stability and Control*. McGraw-Hill, 1994.

[14] *PowerWorld*, 2018, www.powerworld.com.

[15] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1, 2006.

[16] X. Lou, C. Tran, R. Tan, D. Yau, and Z. Kalbarczyk, "Assessing and mitigating impact of time delay attack: A case study for power grid frequency control," in *ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, 2019.

[17] H. D. Chiang, "Study of the existence of energy functions for power systems with losses," *IEEE Trans. Circuits Syst.*, vol. 36, no. 11, 1989.

[18] T. Mikolinnas and B. Wallenberg, "An advanced contingency selection algorithm," *IEEE Trans. PAS.*, vol. 100, no. 2, 1981.

[19] V. Brandwajn, Y. Liu, and M. Lauby, "Pre-screening of single contingencies causing network topology changes," *IEEE Trans. Power Syst.*, vol. 6, no. 1, pp. 30–36, 1991.

[20] R. Fischl, "Application of neural networks to power system security: Technology and trends," in *IEEE World Congr. Comput. Intell.*, 1994.

[21] B. Wang, B. Fang, Y. Wang, H. Liu, and Y. Liu, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2561–2570, 2016.

[22] M. He, J. Zhang, and V. Vittal, "Robust online dynamic security assessment using adaptive ensemble decision-tree learning," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 4089–4098, 2013.

[23] T. S. Sidhu and C. Lan, "Contingency screening for steady-state security analysis by using FFT and artificial neural networks," *IEEE Trans. Power Syst.*, vol. 15, no. 1, pp. 421–426, 2000.

[24] P. Esfahani, M. Vrakopoulou, K. Margellos, J. Lygeros, and G. Andersson, "A robust policy for automatic generation control cyber attack in two area power network," in *IEEE CDC*, 2010.

[25] ——, "Cyber attack in a two-area power system: Impact identification using reachability," in *ACC*, 2010.

[26] R. Tan, H. Nguyen, E. Foo, X. Dong, D. Yau, Z. Kalbarczyk, R. Iyer, and H. Gooi, "Optimal false data injection attack against automatic generation control in power grids," in *ACM/IEEE ICCPS*, 2016.

[27] K. Rahimi, A. Parchure, V. Centeno, and R. Broadwater, "Effect of communication time-delay attacks on the performance of automatic generation control," in *NAPS*, 2015.

[28] J. Wang and C. Peng, "Analysis of time delay attacks against power grid stability," in *ACM CPSR-SG*, 2017.

[29] A. Sargolzaei, K. Yen, M. Abdelghani, S. Sargolzaei, and B. Carbunar, "Resilient design of networked control systems under time delay switch attacks, application in smart grid," *IEEE Access*, vol. 5, 2018.

[30] M. Roozbehani, M. Dahleh, and S. Mitter, "Volatility of power grids under real-time pricing," *IEEE Trans. Power Syst.*, vol. 27, no. 4, 2012.

[31] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T. V. Cutsem, and V. Vittal, "Definition and classification of power system stability," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1387–1401, 2004.

[32] D. Debeljkovic and B. Sreten, "Asymptotic stability analysis of linear time delay systems: delay dependent approach," *Systems, Structure and Control(Scientific monograph)*, pp. 29–60, 2008.

[33] *MATLAB*, 2018, https://www.mathworks.com/.

[34] J. D. Glover, M. S. Sarma, and T. J. Overbye, *Power System Analysis and Design*, 5th ed. Cengage Learning, 2011.

[35] S. Tripathy, T. Bhatti, C. Jha, O. Malik, and G. Hope, "Sampled data automatic generation control analysis with reheat steam turbines and governor dead-band effects," *IEEE Trans. Power apparatus and systems*, no. 5, pp. 1045–1051, 1984.

[36] N. Jaleeli, L. S. VanSlyck, D. N. Ewart, L. H. Fink, and A. G. Hoffmann, "Understanding automatic generation control," *IEEE trans. Power Syst.*, vol. 7, no. 3, pp. 1106–1122, 1992.

[37] X. Lou, D. Yau, H. Nguyen, and B. Chen, "Profit-optimal and stability-aware load curtailment in smart grids," *IEEE Trans. Smart Grid*, vol. 4, no. 3, pp. 1411–1420, 2013.

[38] R. Jhaveri, R. Tan, and S. Ramani, "Managing industrial communication delays with software-defined networking," in *IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, 2019.

[39] *ThermoPower*, 2018, https://casella.github.io/ThermoPower/.

[40] *Modelica*, 2018, https://www.modelica.org/.

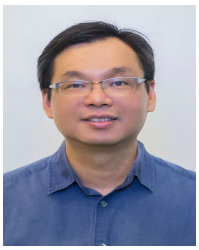
[41] *OpenModelica*, 2018, https://www.openmodelica.org/.

[42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

Xin Lou is a Research Scientist at Advanced Digital Sciences Center (ADSC), a research center established by the University of Illinois at Urbana Champaign (UIUC) in Singapore. He is also a research affiliate at the Coordinated Science Laboratory (CSL) at UIUC. He was a postdoctoral Researcher (2016-2018) at ADSC. He received the Ph.D. (2016) degree in computer science from City University of Hong Kong, Hong Kong SAR and B.E. (2005) degree in telecommunication engineering from Sichuan University, China. His research interests include cyber-physical systems, deep learning for sensing and computing, nonlinear optimization and distributed algorithms.

Cuong Tran is an AI Researcher at VNPT (Vietnam Posts and Telecommunications Group) and a Co-founder of Vimentor (Online Learning and HR Platform), Vietnam. He was a Research Assistant (2016-2018) at Singapore University of Technology and Design. He received the B.C degree (2012) in communication engineering from Hanoi University Of Science and Technology, Hanoi, Vietnam. His interests include knowledge graph, information retrieval, cyber-security, distributed algorithms.

Rui Tan (M'08-SM'18) is an Assistant Professor at School of Computer Science and Engineering, Nanyang Technological University, Singapore. Previously, he was a Research Scientist (2012-2015) and a Senior Research Scientist (2015) at Advanced Digital Sciences Center, a Singapore-based research center of University of Illinois at Urbana-Champaign (UIUC), a Principle Research Affiliate (2012-2015) at Coordinated Science Lab of UIUC, and a postdoctoral Research Associate (2010-2012) at Michigan State University. He received the Ph.D. (2010) degree in computer science from City University of Hong Kong, the B.S. (2004) and M.S. (2007) degrees from Shanghai Jiao Tong University. His research interests include cyberphysical systems, sensor networks, and ubiquitous computing systems. He received the Best Paper Awards from IPSN'17, CPSR-SG'17, Best Paper Runner-Ups from IEEE PerCom'13 and IPSN'14.



David K.Y. Yau received the B.Sc. from the Chinese University of Hong Kong, and M.S. and Ph.D. from the University of Texas at Austin, all in computer science. He has been Professor at Singapore University of Technology and Design since 2013. Since 2010, he has been Distinguished Scientist at the Advanced Digital Sciences Centre, Singapore. He was Associate Professor of Computer Science at Purdue University (West Lafayette). He received an NSF CAREER award. He won Best Paper award in 2017 ACM/IEEE IPSN and 2010 IEEE MFI. His papers in 2008 IEEE MASS, 2013 IEEE PerCom, 2013 IEEE CPSNA, and 2013 ACM BuildSys were Best Paper finalists. His research interests include cyber-physical system and network security/privacy, wireless sensor networks, smart grid IT, and quality of service. He serves as Associate Editor of IEEE Trans. Network Science and Engineering and ACM Trans. Sensor Networks. He was Associate Editor of IEEE Trans. Smart Grid, Special Section on Smart Grid CyberPhysical Security (2017), IEEE/ACM Trans. Networking (2004-09), and Springer Networking Science (2012-2013); Vice General Chair (2006), TPC co-Chair (2007), and TPC Area Chair (2011) of IEEE ICNP; TPC co-Chair (2006) and Steering Committee member (2007-09) of IEEE IWQoS; TPC Track co-Chair of 2012 IEEE ICDCS; and Organizing Committee member of 2014 IEEE SECON.

Zbigniew T. Kalbarczyk (M'95) is a Research Professor at Department of Electrical and Computer Engineering and the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign. Dr. Kalbarczyks research interests are in the area of design and validation of reliable and secure computing systems. His current work explores emerging technologies, such as resource virtualization to provide redundancy and assure system resiliency to accidental errors and malicious attacks. Dr. Kalbarczyks research involves also analysis of data on failures and security attacks in large computing systems, and development of techniques for automated validation and benchmarking of dependable and secure computing systems using formal (e.g., model checking) and experimental methods (e.g., fault/attack injection). He served as a program Chair of Dependable Computing and Communication Symposium (DCCS), a track of the International Conference on Dependable Systems and Networks (DSN) 2007 and Program Co-Chair of Computer Performance and Dependability Symposium, a track of the DSN 2002. He has been an Associate Editor of IEEE Transactions on Dependable and Secure Computing. Dr. Kalbarczyk has published over 130 technical papers and is regularly invited to give tutorials and lectures on issues related to design and assessment of complex computing systems. He is a member of the IEEE, the IEEE Computer Society, and IFIP Working Group 10.4 on Dependable Computing and Fault Tolerance.

Ambarish Banerjee is a third year undergraduate studying computer science and engineering from Indian Institute of Technology Bhubaneswar, India. He will receive his Bachelor of Technology degree in 2021. He worked as an intern (2019) at Advanced Digital Sciences Center (ADSC), a research center established by the University of Illinois at Urbana Champaign (UIUC) in Singapore. His research interests include supervised learning algorithms and their applications in diverse fields like cyber security.

Prakhar Ganesh is currently working as a Research Engineer at Advanced Digital Sciences Center (ADSC), Singapore. He received the B.Tech in Computer Science and Engineering from Indian Institute of Technology (IIT) Delhi, India in 2019. His research interests include Computer Vision and other cross-domain applications of Deep Learning.