Secure Networked Control Systems

Henrik Sandberg,[1] Vijay Gupta,[2] and Karl H. Johansson[1]

[1]Electrical Engineering and Computer Science and Digital Futures, KTH Royal Institute of Technology, Stockholm, Sweden; email: {hsan,kallej}@kth.se
[2]Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA; email: vgupta2@nd.edu

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–22

https://doi.org/10.1146/((please add article doi))

Keywords

Attack space, false data injection attack, replay attack, denial of service attack, networked control systems, cyber-physical security

Abstract

Cyber-vulnerabilities are being exploited in a growing number of control systems. As many of them form the backbone of critical infrastructure systems and these systems are becoming more automated and interconnected, it is of utmost importance to develop methods allowing system designers and operators to do risk analysis and to develop mitigation strategies. Over the last decade, great advances have been made in the control systems community to better understand cyber-threats and what can be their potential impact. This article provides an overview of recent literature on secure networked control systems. Motivated by recent cyberattacks in the power grid, connected road vehicles, and process industry, a system model is introduced covering many of the existing research studies on control systems vulnerabilities. An attack space is introduced illustrating how adversarial resources are allocated in some common attacks. The main part of the article describes three types of attacks: false data injection, replay, and denial of service attacks. Representative models and mathematical formulations of these attacks are given together with some proposed mitigation strategies. The focus is on linear discrete-time plant models but various extensions are presented in an outlook section, where also some interesting research problems for future work are mentioned.

1. INTRODUCTION

Security in networked control systems refers to the study of attack algorithms by which an adversary can degrade the performance of a control system by utilizing a cyber-channel over which sensor measurements or control signals are being transmitted in a control loop, and the corresponding detection and mitigation mechanisms that can be employed to prevent performance degradation in the presence of such attacks. Attacks on control systems are notably different from attacks in which the goal of the adversary is degradation of the cyber-network itself, e.g., through distributed denial of service attacks on Internet routers (20, 39). They are also different from faults, which may also degrade the performance of a control system, since in the traditional fault diagnosis literature, a basic assumption is that the fault is happening without active participation by an adversary (19). Finally, cyberattacks are differentiated from physical attacks, as the latter are carried out by an adversary who does not utilize a cyber-channel to carry out the attack.

Although many cyberattacks on industrial control systems have continuously been reported for quite some time, the area was catapulted into public awareness through a few high profile attacks, such as the Stuxnet computer worm attacking industrial sites in Iran in 2010 (25) and the cyberattack on the Ukrainian power grid in 2015 (57). The latter attack consisted of interfering with the supervisory control and data acquisition (SCADA) system, remotely switching substations off and disabling infrastructure components to interrupt the proper functioning of the power network. More generally, such cyberattacks have been noted or demonstrated against a variety of systems including infrastructure systems, automobiles, manufacturing industries, oil refineries, smart homes, among many others. As more control systems become *networked*, meaning that sensor and controller signals are transmitted over wireless or wired networks, the possibilities available to the attacker, and hence the importance of security in control systems, are only increasing.

The security problem in networked control systems is essentially a *game* between the attacker and defender. This can be interpreted formally in the context of game theory, or in an informal sense as referring to a competition between these two entities: The attacker seeks to stay one step ahead of the defender in terms of designing an attack that is sophisticated enough to beat the detection and mitigation mechanisms employed by the defender. Similarly, the defender seeks to estimate the capabilities or sophistication of the attacker and deploy mechanisms to detect or mitigate the attack. This point of view has three immediate consequences. The first is that technically it makes sense to define and identify a spectrum of capabilities or resources available to the attacker and the defender in order to calculate what the outcome for the control loop will be in a variety of possible scenarios. The second is that algorithms that can identify the capabilities of an attacker or a defender can be very useful in practice. The final consequence is that the interaction between defender and attacker needs to be ideally captured by a dynamic or iterative model, in which the defender and attacker have the ability to learn from previous interactions and to improve their strategies or resource allocations.

There are several important technologies available and commonly used to defend industrial control systems against cyberattacks. How to engineer secure systems in general is a large and broad cross-disciplinary topic with specific solutions for certain application domains (4). Cryptography is a critical component and is about securing information systems in general against adversarial attacks by encrypting data for secret communication, deploying protocols for secure key exchange, allowing proposer user authentication etc. (23). Different from cryptography, there are also many specialized defense mechanisms, such as

so called honeypots aimed at luring an adversary into the system and then monitoring and analyzing its behavior to create counter actions (37). Obfuscation has often played a key role in traditional control system security in that the knowledge of dedicated computer systems and communication protocols for industrial control systems have not been widely available. The trend of building modern control systems based on off-the-shelf components and standard technologies has reduced the role of obfuscation and made these systems more vulnerable to adversarial organizations and individuals. It is a mistake for a defender to rely only on the lack of knowledge by an attacker about the system or the strategy employed by the defender. Kerckhoff's Principle (40) on military ciphers from 1883 and Shannon's Maxim "the enemy knows the system" from 1948 can be interpreted as saying that the reliance on secrets in a system design should be kept to a minimum and the defender should implement security algorithms under the assumption that the algorithms are public and thus known to the adversary. This principle leads to stronger systems in the long run and is also adopted for the algorithms discussed in this article.

Our focus is on security for networked control systems and particularly on attack and mitigation techniques targeting the requirements and nature of such systems. A key requirement is the need for real-time response as the systems are monitoring and controlling physical processes, which thus impose a time scale that needs to be followed for detection and mitigation of any attacks. It also meant that solutions that rely on collecting and analyzing large amounts of data need to be evaluated for the latency that they impose. A related requirement is the need for the closed-loop control system to be continuously and safely operating, since breaking the loop might deteriorate the performance of the controlled physical process or even render it unstable. While one solution for mitigating an attack on an IT system may simply be to restart a device or a server, interrupting a physical process such as an infrastructure system or an autonomous car is in most practical situations impossible. In some applications, such as the power grid or the transport infrastructure, legacy devices may also be present that need to be considered while designing secure systems.

Confidentiality, integrity, and availability (often abbreviated CIA) are fundamental requirements for computer security (4). The CIA paradigm can be applied also to the security of networked control systems while keeping in mind the differences above. In the present context, confidentiality refers to authorized access to information. In other words, sensor or control information should be encrypted and otherwise authenticated. Authentication attacks seek to impersonate authorized users or otherwise defeat the measures in place. If networked control systems have legacy components, the problem is even more difficult. Similarly, integrity refers to the property that the data being transmitted are not replaced by malicious data by the adversary. Such data can lead to incorrect control updates and thus the state of the process may evolve in a manner that does not satisfy constraints such as safety or stability. Most of the attacks that we consider will be data integrity attacks. The physics of the process imposes some constraints on what the attacker can transmit; however, there is still considerable freedom available that can be exploited. Finally, availability refers to the property that the control system components are always available. Denial of service (DoS) or jamming attacks can lead to sensor or state data not being available to the controller, thus challenging the decisions-making process and possibly degrading the system performance to unacceptable levels.

The outline of this article is as follows. Section 2 presents three motivating examples and describes how they relate to the methodologies presented later. The system and attacker models used throughout the article are introduced in Section 3. Three types of cyberattacks

Figure 1

An example of a substation. Substations are transforming electric voltage levels as part of the power grid, often between the transmission and distribution networks. They typically have several sensing and actuation devices and have been shown to be vulnerable to cyberattacks. (Photo courtesy by ETA+ through Unsplash.)

that have been studied extensively are then discussed together with some proposed mitigation mechanisms: false data injection (FDI) attack (Section 4), replay attack (Section 5), and DoS attack (Section 6). The paper is concluded with an outlook on extensions and future work in Section 7.

2. MOTIVATING EXAMPLES

Cyber-physical security vulnerabilities are present in a wide range of control applications. Some of them have already been exploited over the last few years in various attacks. In this section, we will describe three examples to illustrate this trend and to motivate models and defense strategies introduced in later sections.

2.1. Power Grid

Towards the end of 2015, the operation of part of the power grid in a region of Western Ukraine was interrupted through a sophisticated cyberattack impacting about 225,000 customers. The attack was probably the first one of such a scale on a power system and has been widely reported in the media (57) and described in detail by security corporations and Government organizations (12). By spear-phishing efforts prior to the actual attack, adversaries were able to get Virtual Private Network (VPN) credentials to the SCADA network of the grid. It allowed them on the day of the attack to gain remote access to several substations (similar to the one shown in Figure 1) thereby controlling circuit breakers. Opening such breakers took down the power delivery for several hours for many customers. As part of the attack, the adversaries remotely installed malicious firmware on field devices at the substations, making it impossible for the remote operators to automatically restore their operation, rather requiring extensive manual work.

Figure 2

A vehicle platoon of three trucks. The control of the distance between the trucks are based on vehicle states being wirelessly communicated between the vehicles. (Photo courtesy by Scania.)

Many security vulnerabilities of the power grid has been discussed over the last decade. The grid is a large-scale networked control system properly regulating the frequency and voltage, by ensuring that the supply equals the demand. Sensors, actuators, and control algorithms are distributed over the network and over large geographic areas. All the cyberattacks and mitigation strategies discussed in this article are relevant for the grid. As an example, consider the state estimator for the transmission grid generating estimates for the voltage phasor magnitudes and angles, which are extensively used by control and optimization algorithms in the grid control center. It was shown more than ten years ago how FDI attacks could fool the state estimator (31) and that this vulnerability can be systematically evaluated (47). FDI attacks for networked control systems are discussed in detail in Section 4. DoS attacks, which are presented in Section 6, have also been studied for load frequency control (30).

2.2. Connected and Automated Vehicles

In 2015, Miller and Valasek demonstrated that it was possible to take over the essential functionalities of a 2014 Jeep Cherokee remotely, while a journalist at Wired was driving the vehicle on a highway (22). At a conference and in an extended report (32), they detailed the vulnerabilities in the system and the procedure they went through to take control of the car, basically providing a recipe for how to perform cyberattacks on a large number of car models. The attack surface was the radio interface in the car supporting both Wi-Fi and cellular communication. It gave indirect access to the Controller Area Network (CAN) connecting several Electronic Control Units (ECUs) handling steering, transmission, breaking and many other critical functionalities. By reprogramming an embedded microcontroller with new firmware, it became possible to send commands to an ECU shutting down the engine remotely through the cellular network. The demonstrated remote attack could hence be done from any location, did not require any modification to the vehicle, and showed that hundreds of thousands of vehicles were vulnerable.

Future intelligent transport systems with connected and automated vehicles will have many more radio interfaces than today's vehicles and thus many more potential attack surfaces for remote intruders than the ones described above. An example of an emerging

Figure 3

A paper mill. Such a plant consists of thousands of control loops, which in the future might be connected to the cloud to enable basically unlimited computing power for data analytics to be used for instance in advanced predictive maintenance. (Photo courtesy by Iggesund Paperboard.)

technology is heavy-duty vehicle platooning, as illustrated in Figure 2. Such road trains provide more energy-efficient freight transport under almost driver-less operations, but requires wireless communication among the vehicles to regulate the inter-vehicle distances tightly (9). Vehicle platoons and automated vehicles in general will be supported by an advanced wireless sensing and communication infrastructure to enable the vehicles to have beyond-human situational awareness for smooth interactions between all kind of road users and traffic controllers. Such infrastructure needs to be resilient to a large range of potential cyberattacks, such as DoS attacks overloading the communication networks and servers, which are further discussed in Section 6.

2.3. Industrial Process Automation

The Stuxnet cyberattack on industrial sites in Iran 2010 was the first large scale attack on a SCADA system directly exploiting the nature of the targeted control system (25). It consisted of a number of phases and used specific vulnerabilities including four zero-day exploits, which are software flaws unknown to the target software vendor. The Stuxnet worm entered the system via a USB stick and then spread to multiple computers, while checking if each machine was part of the targeted industrial control system or not. For those machines part of a control system, such as the control of centrifuges in a uranium-enrichment plant, their programmable logic controllers (PLCs) were compromised in a sophisticated way. The compromised PLC first silently recorded plant information such as sensing and actuation data for a while. Then it used these data to generate control commands destabilizing and destroying the process; meanwhile, the PLC provided supervisory controllers and operators with false data, giving them the impression that the status of the process was normal. Such a cyberattack is nowadays called a replay attack and will be further discussed in Section 5.

Wireless sensors and networks are not commonly used in today's process industry con-

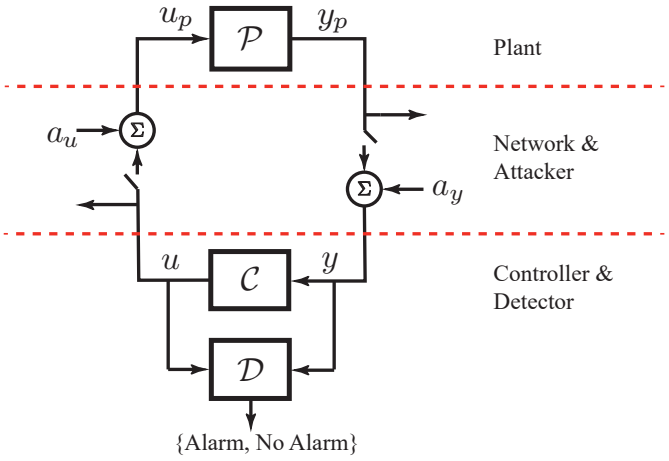

Figure 4

Block diagram of networked control system. The Network & Attacker layer illustrates three modes of attack: denial of service (switches), eavesdropping (outgoing arrows), and false data injection (a_u, a_y).

trol loops, as in the industrial plant in Figure 3. With the current trend in embedded and cloud computing, wireless and cellular communication, and IoT technology, however, drastic changes can be expected in the architecture and operation of industrial automation systems (1). Existing and new embedded devices will collect data online and feed them into the cloud for data analytics, predictive maintenance, operation optimization etc. Such systems will provide huge advantages in performance, resource efficiency, and flexibility, but several new attack surfaces could be established if this technological development is not done properly. Industrial processes could be open to all the cyberattacks discussed in this article, including the replay attack described above.

3. SYSTEM AND ATTACK MODELS

The networked control systems we consider are schematically illustrated by the block diagram in Figure 4. The physical layer consists of the plant $\mathcal{P}$ and also includes devices such as sensors and actuators. Examples of relevant plants were given in Section 2. The cyber-layer consists of a communication network, for example, a SCADA system or a field network, together with the controller $\mathcal{C}$ and a possible detector $\mathcal{D}$. The controller and detector could either be centralized and located at a control center, or be distributed, in PLCs for example. In the following, we will introduce an abstract modeling framework for system-theoretic security analysis. Note that for practical security considerations implementation details, such as the choice of hardware and protocols, are also essential, see (4) for a comprehensive overview.

In the following sections, we suppose that the plant $\mathcal{P}$ can be modeled as a linear discrete-time system

$$\mathcal{P}: \begin{cases} x(k+1) = Ax(k) + Bu(k) + B_a a(k) + w(k), & x(0) = x_{\mathrm{i}} \\ y(k) = Cx(k) + D_a a(k) + v(k), \end{cases} \quad 1.$$

for times $k \geq 0$, with state $x(k) \in \mathbb{R}^n$, communicated control input $u(k) \in \mathbb{R}^m$, and received measurement $y(k) \in \mathbb{R}^p$. We assume that the plant is subject to process disturbance $w(k) \in \mathbb{R}^n$ and measurement noise $v(k) \in \mathbb{R}^p$. The signal $a(k) \in \mathbb{R}^q$ models attacks and is introduced in Section 4. Depending on the attack scenario considered, we will also specify models of the controller $\mathcal{C}$ and detector $\mathcal{D}$.

We consider malicious attacks carried out through the network communication system as indicated in Figure 4. As explained in Section 1, the process operator desires CIA: confidentiality, integrity, and availability. Conversely, the attacker seeks to violate one or several of these properties. For example, the attacker may inject, or manipulate, the data content in the packets between the controller, sensors, and actuators. This violates integrity, and is modeled by an additive signal $a = (a_u, a_y)$ injected into the control loop in Figure 4. Such FDI attacks are the focus of Section 4. An attacker may also eavesdrop on communication, and thus violate confidentiality. This is modeled by outgoing arrows in the network layer of Figure 4. For example, the attacker may record sensor measurements over a period of time to later inject and fool the controller about the plant state. Possibly this can be coordinated with harmful false-data injection in the actuator channel. Such attacks are called replay attacks and are the focus of Section 5. The actuator or sensor channels may also be taken out (or rendered unavailable) by the attacker by overwhelming the network with communication packets. This is modeled in Figure 4 by switches, which show that such attacks effectively render the system open loop. DoS attacks are the focus of Section 6.

Figure 5 shows a three-dimensional attack space serving to illustrate the relative resources required to implement attacks against networked control systems. 'Disclosure resources' measure the possibilities the attacker has to read the channels u and y in Figure 4. This can vary in a network depending on the attacker's access to links and physical devices, and also on whether the operator has encrypted some channels. 'Disruption resources' measure the possibilities to overwrite or inject new u and y packets. This depends on factors similar to those affecting 'Disclosure resources', but also on whether the operator authenticates some links. Finally, 'Model knowledge' measures how much the attacker knows about the system and models of $\mathcal{P}$, $\mathcal{C}$, and $\mathcal{D}$ in use by the designer and the operator. An attacker can use such knowledge to mask data attacks. For example, FDI attacks can be made undetectable or stealthy. Replay attacks do not require such knowledge, but instead exploit the ability of the adversary to read, record, and write data packets. Finally, DoS attacks only exploit the ability of the adversary to send large numbers of data packets to specific target devices and thereby overload the device or the communication channel. More advanced multistage attacks could very well move in the attack space after the attack has been commenced. For example, one could envisage an attacker that first passively eavesdrop channels to learn the models in use and later stages a stealthy FDI attack by exploiting the learned models.

4. FDI ATTACK

In this section, the physical plant $\mathcal{P}$ is targeted by FDI attacks, modeled by $a \neq 0$ in equation 1. FDI attacks generally originate from interception and corruption of communication channels to and from the plant, see Figure 4. In such scenarios, we split the attack into the components directly affecting the control signal u and measurement y as $a = (a_u, a_y)$. In this case, $B_a = \begin{bmatrix} B & 0 \end{bmatrix}$ and $D_a = \begin{bmatrix} 0 & I \end{bmatrix}$. We define three types of FDI attacks: *Sen-*

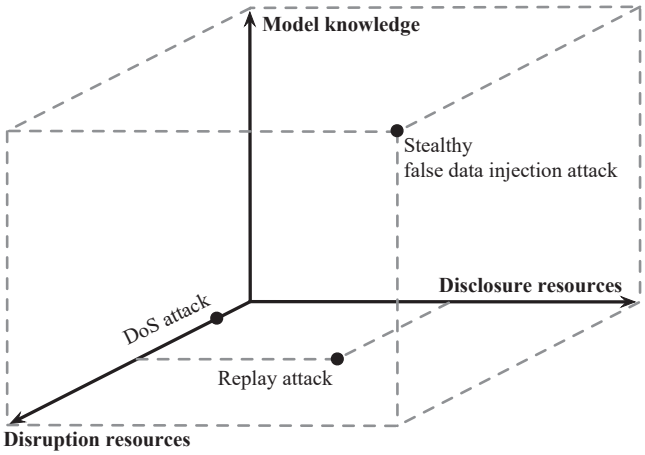

Figure 5

Attack space illustrating the adversarial resources required to conduct attacks. 'Disclosure resources' quantify the amount of channels the attacker can read, whereas 'Disruption resources' quantify channel writing ability. 'Model knowledge' refers to the ability of the adversary to models of system components $\mathcal{P}$, $\mathcal{C}$, and $\mathcal{D}$.

sor attacks ($a_y \neq 0$, $a_u = 0$), *actuator attacks* ($a_u = 0$, $a_y \neq 0$), and *coordinated attacks* ($a_y \neq 0$, $a_u \neq 0$). Under FDI attacks, the state, control, and sensor signals will generally be affected as compared to the un-attacked scenario. The reason for the difference in plant state under actuator attack is obvious. For sensor attacks, the plant state may be affected when operating in a closed loop in which a feedback controller reacts to the manipulated signal y and applies the affected control u, and thus perturbs the state as compared to the un-attacked scenario. We assume, however, that the disturbances and noise, w and v, are unaffected by FDI attacks.

4.1. Attack Detection

A fundamental problem for the system operator is to determine whether the plant is subject to an FDI attack, or not. That is, if $a \neq 0$ or $a \equiv 0$ in equation 1. Using fault diagnosis (15) terminology, this is a *detection* problem. If an attack is detected, one may proceed to the *isolation* problem to determine the attack type and affected channels. Finally, the *identification* problem, which is the most complex problem, concerns the reconstruction of the exact attack sequence $a_0^k := (a(0), a(1), \ldots, a(k))$.

Assuming a statistical setup where probability distributions of $w_0^k := (w(0), w(1), \ldots, w(k))$, $v_0^k := (v(0), v(1), \ldots, v(k))$, and x_i are known, and with given data (u_0^k, y_0^k), the detection problem can be understood as testing between the multiple hypothesis

$$\mathcal{H}_0 : \ a \equiv 0 \qquad \text{(No FDI attack)},$$
$$\mathcal{H}_{k_0} : \ a_0^{k_0-1} \equiv 0, \ a_{k_0}^k \neq 0, \qquad \text{(FDI attack starts at time } k_0 > 0).$$

The hypothesis test should decide whether there is sufficient evidence to reject the null

hypothesis $\mathcal{H}_0$ in favor of one of the alternative hypothesis $\mathcal{H}_{>0}$. A difficulty here is for the operator to characterize and model the alternative hypothesis $\mathcal{H}_{>0}$ since the capabilities and goals of the attacker are generally unknown, although risk assessment may give valuable insights. Assuming deterministic, but unknown, attack signals a, we may use recursive generalized likelihood-ratio tests on online data (7, 55) to generate a detection alarm as soon as possible after the actual attack has started. For optimality properties and relations to the so-called CUmulative SUM (CUSUM) tests, see (7, 26). The books (7, 15) explain how, and when, it is possible to proceed to isolate and identify attacks following a detection alarm. In general, this follow-up analysis is best done in an offline mode using all available data.

If the alternative hypothesis $\mathcal{H}_{>0}$ are not well characterized, so-called non-parametric CUSUM tests (21) can be run online in order to quickly detect certain deviations from the null hypothesis $\mathcal{H}_0$. How to tune such detectors to minimize the physical damage in the presence of malicious FDI attacks is discussed in (21). An example of a non-parametric CUSUM test to detect attacks is the following: First compute residual sequence r_0^k using a state estimator

$$\begin{aligned} \hat{x}(k+1) &= A\hat{x}(k) + Bu(k) + Kr(k), \quad \hat{x}(0) = \mathbb{E}x_{\mathrm{i}} \\ r(k) &= y(k) - C\hat{x}(k). \end{aligned} \tag{2.}$$

A common choice is to use a Kalman gain K tuned under the null hypothesis $\mathcal{H}_0$, in which case the residual is the innovation sequence with desirable i.i.d. statistical properties. Then, compute the CUSUM statistic as

$$S(k) = \max\{S(k-1) + |r(k)|^2 - \delta, 0\}, \quad S(0) = 0,$$

where δ is a tunable forgetting parameter chosen such that $\mathbb{E}|r(k)|^2 - \delta < 0$ under the null hypothesis $\mathcal{H}_0$. The test is now

$$S(k) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_{>0}}{\gtrless}} \tau,$$

where τ is a tunable alarm threshold chosen to trade-off the false-alarm and correct detection rates. An alarm at time k_{a} supports that an abrupt change has occurred in the system at time $k_0 \leq k_{\mathrm{a}}$, and we can estimate the change time as $k_0 \approx k_{\mathrm{a}} - N_{k_{\mathrm{a}}} + 1$, where $N_k = N_{k-1}\mathbf{1}_{\{S(k-1)>0\}} + 1$, see (7). Here N_k counts the number of time steps since $S(k)$ was set to zero last, and $\mathbf{1}$ is the indicator function.

The CUSUM statistic is an example of a so-called *stateful* test, which takes the history of the residual into account. Popular alternatives are *stateless* tests, such as the χ^2 test (15, 21), where only the current residual is monitored,

$$|r(k)|^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_{>0}}{\gtrless}} \tau. \tag{3.}$$

Generally such tests are easier to theoretically analyze, but may perform worse in practice (higher false alarm rate, easier to bypass, etc.) (21). Windowed χ^2 tests (34) is a possible stateful extension of the regular χ^2 test, with better performance.

The tests and schemes mentioned above were originally designed for detection of faults with no malicious intent. More recent anomaly detection schemes and mitigation tools are discussed in Section 4.3. As discussed in the following section, particular detection tests can, however, often be bypassed by a resourceful attacker.

4.2. Undetectable Attack

Design of attack detection algorithms is clearly very important and challenging even under strong assumptions on the attacker. A related line of work has focused on characterizing types of attacks that are particularly difficult to detect for *any* detection test. Such attacks are called *undetectable* or *stealthy*. The possibility of these attacks highlights inherent weaknesses in the system itself. As part of a risk management process, knowledge of such attacks and their impact can be used to guide the allocation of new security resources (33). Examples of security resources are redundant sensors and redundant actuators, and deployment of encryption or authentication mechanisms on selected communication channels.

We review some classes of undetectable and stealthy attacks next.

Deterministic Undetectable Attack A deterministic approach to characterizing attacks that are hard to detect leads to the concept of *undetectable attacks* (38). Informally, an attack a is undetectable if it does not perturb the inputs and outputs available to the controller and detector. More formally, $a \neq 0$ is undetectable if there exists initial states $x_\mathrm{i} = x_\mathrm{i}^a$ such that

$$y(k; x_\mathrm{i}, u, 0) = y(k; x_\mathrm{i}^a, u, a), \quad k \geq 0, \qquad 4.$$

for at least some input u[1]. Hence, there exists (at least) two different possibilities for the same input and output sequences. One possibility is that the system is not under attack, and another is that there is an FDI attack (albeit with a different initial condition). Typically the operator does not exactly know what the initial state is, and hence cannot determine if there is an attack, or not. The attack is thus undetectable.

Under certain situations (generically when there are more attack signals than outputs), the condition given in equation 4 may hold with $x_\mathrm{i} = \tilde{x}_\mathrm{i}$. In this case, even if the operator knows the initial state, undetectable attacks exist. Such attacks are *perfectly undetectable attacks* (or shorter, *perfect attacks*, see (53)).

For linear systems, the condition in equation 4 translates into the existence of zero dynamics (transmission zeros) (38), and these attacks are also sometimes referred to as *zero-dynamics attacks* (49).

Deterministic Stealthy Attack Another deterministic approach leads to the concept of *stealthy attacks* (49). Whereas undetectable attacks are designed to be independent of possible fault or anomaly detection mechanisms (Section 4.1), the stealthy attacks take them explicitly into account. For simplicity, consider a constraint derived from equation 3:

$$|r(k; y, x_\mathrm{i}, u, a)|^2 \leq \tau + \epsilon, \quad k \geq 0, \qquad 5.$$

where $r(\cdot; y, x_\mathrm{i}, u, a)$ emphasizes the dependence of the residual on the system signals and attack under $\mathcal{H}_{>0}$. A stealthy attack $a \neq 0$ against the detector given in equation 3 satisfies the constraint from equation 5 for some threshold $\epsilon \geq 0$ chosen by the attacker. By choosing $\epsilon = 0$, the attack will not generate an alarm from this particular detector. Using $\epsilon > 0$ increases the risk for an alarm, but may also increase the impact of the attack. In a similar manner, stealthy attacks may be defined for other detectors.

[1] $y(k; x_\mathrm{i}, u, a)$ denotes the output of equation 1 at time k when the input is u and attack is a. For simplicity, we leave out w, v in the deterministic scenario here.

Intuitively, the class of deterministic stealthy attacks is larger than the class of undetectable attacks. Indeed, undetectable attacks are often stealthy attacks. However, it is possible to design detectors that trigger alarms for dangerous, or unlikely, actual outputs y. In general, it is not possible to conclude that undetectable attacks are always stealthy, or vice versa.

Stochastic Stealthy Attack The previous classes of attacks have taken a deterministic approach. There are various ways and assumptions in the literature to handle uncertainty in noise, disturbance, and initial state, see (38, 49). If probabilistic models of the uncertain variables are available, they can be used to define *stochastic stealthy attacks* (5, 6, 24). Let $p_0(y_0^k)$ and $p_a(y_0^k)$ be the probability densities of the output sequence in closed loop under no attack ($\mathcal{H}_0$) and under attack ($\mathcal{H}_{>0}$), respectively. A (possibly stochastic) attack a is then ϵ-stealthy if

$$\limsup_{k\to\infty} \frac{1}{k+1} D_{\mathrm{KL}}(p_a(y_0^k) \,\|\, p_0(y_0^k)) \leq \epsilon, \tag{6.}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler divergence between two probability densities. Intuitively, if $\epsilon > 0$ is small, it requires a very long output sequence to be able to correctly distinguish between the two hypothesis with high probability. In fact, as is shown in (6), Stein's lemma directly yields that for any $\delta \in (0, 1)$ there exists no detector with detection probability $p^{\mathrm{D}}(k)$ (deciding $\mathcal{H}_{>0}$ when $\mathcal{H}_{>0}$ is true) satisfying $0 < 1 - p^D(k) < \delta$ and simultaneously a false alarm probability $p^{\mathrm{F}}(k)$ (deciding $\mathcal{H}_{>0}$ when $\mathcal{H}_0$ is true) decaying faster than rate ϵ:

$$\limsup_{k\to\infty} -\frac{1}{k+1} \ln p^{\mathrm{F}}(k) > \epsilon,$$

illustrating the difficulty for an operator to detect such an attack.

Impact of Stealthy Attack Malicious adversaries often have specific attack objectives. In practice, a resourceful system operator can stop many attacks soon after they are detected. It becomes important then to evaluate the possible impact of stealthy attacks. Indeed, if a stealthy attack cannot cause much damage, it may not be of the highest priority in a risk management process (33). A system operator may tune his detection system to balance the impact of possible attacks and the cost for false alarms, which is a central idea in (21). We illustrate such trade-offs in a simple example next.

For detectors in stochastic systems, the worst mean delay for detection, T_{D}, is asymptotically bounded by the mean time between false alarms, T_{F}, as

$$T_{\mathrm{D}} \geq \frac{\ln T_{\mathrm{F}}}{\epsilon}, \quad T_{\mathrm{F}} \to \infty,$$

where ϵ is a bound on the Kullback-Leibler divergence between the attacked and un-attacked scenario as expressed in equation 6. We point the reader to (26) for a precise statement and definitions of the involved quantities. Since operators typically require very long times between false alarms, this bound is of practical relevance. Assume next that the damage an attack can incur at every time step it is undetected is approximately quadratic in that attack, $D(a) \approx \frac{1}{2} a^T D_{aa} a$ where D_{aa} is a positive semi-definite matrix and a is a constant (bias) attack vector. If the attack magnitude is small, we can furthermore approximate the Kullback-Leibler divergence as $D_{\mathrm{KL}}(p_a(y)\|p_0(y)) \approx \frac{1}{2} a^T I_{aa} a$, where I_{aa} is the Fisher information matrix. Hence, the total impact of an optimal attack $a^\star$, before it is detectable

and stoppable on average, can be bounded by a generalized Rayleigh quotient,

$$\ln T_{\mathrm{F}} \frac{a^T D_{aa} a}{a^T I_{aa} a} \leq \ln T_{\mathrm{F}} \frac{(a^\star)^T D_{aa} a^\star}{(a^\star)^T I_{aa} a^\star} = \ln T_{\mathrm{F}} \cdot \lambda_{\max}(I_{aa}^{-1} D_{aa}) \leq T_{\mathrm{D}} D(a^\star). \qquad 7.$$

Here $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue and we assume that the Fisher information matrix is positive definite (which corresponds to the condition that the attack cannot be undetectable)[2]. Note that under the quadratic approximation mentioned above, the total impact is independent on the stealthiness parameter ϵ.

The system operator can use expressions such as those provided in equation 7 to assess whether a stealthy attack is *serious*, or not. This depends on the visibility of the attack (I_{aa}), how the attack affects the plant (D_{aa}), and the threshold τ of the detector ($\tau \mapsto T_{\mathrm{F}}(\tau)$). Assuming that there is also a cost proportional to the frequency of false alarms, C/T_{F} (cf. (51)), the overall cost

$$\ln T_{\mathrm{F}}(\tau) \cdot \lambda_{\max}(I_{aa}^{-1} D_{aa}) + \frac{C}{T_{\mathrm{F}}(\tau)}$$

can be minimized by choosing an optimal detection threshold $\tau^\star$ such that

$$T_{\mathrm{F}}(\tau^\star) = \frac{C}{\lambda_{\max}(I_{aa}^{-1} D_{aa})}.$$

This expression optimally balances false alarm and attack impact costs.

4.3. Mitigation

Several mitigation strategies against FDI attacks have been proposed in the literature. Fault diagnosis and change detection methods together with threshold tuning (7, 15, 21) have been discussed in Sections 4.1–4.2. In (38), fundamental limitations and distributed implementations of FDI attack detectors are derived. Such fundamental limitations can be exploited in preventive security resource allocation (33). Secure estimators (11, 17) exploit sensor redundancy to always reconstruct a correct state estimate that can be used for resilient control. Data-based anomaly detection schemes based on machine learning methodology is currently also an active research area (see, for example, (2, 27)). Approaches for increasing the opportunities for detecting stealthy FDI attacks include moving target defense (54) and multiplicative watermarking (48).

5. REPLAY ATTACK

In this section, we consider replay attacks, which proceed by replaying or transmitting a delayed version of the true sensor measurements, while changing the control input to degrade the control performance. Since the received measurements correspond to a measurement sequence that is *possible*, it can be very hard to detect via statistical means that an attacker has altered the measurements in this way. It is, thus, considered to be a different class of attacks than FDI attacks considered in the previous section, even though the control signal

[2]The first upper bound in equation 7 is attained by choosing the optimal attack $a^\star$ as the eigenvector of $I_{aa}^{-1} D_{aa}$ corresponding to $\lambda_{\max}$. By normalizing such that $(a^\star)^T I_{aa} a^\star = 2\epsilon$ the desired level of stealthiness is achieved.

is still corrupted by an adversary. As discussed in Section 2, the Stuxnet attack is believed to have been a replay attack.

A formal definition of a replay attack was considered in (35, 34). In this version, the attacker is able to collect real-time sensor measurements being transmitted from the sensor to the controller, and has the capability to change the true sensor measurements to such values collected previously. This could be done, for instance, by knowing the cryptographic keys, or in simpler systems that do not employ cryptography or time-stamping, by simply retransmitting the sensor measurements collected earlier. After the data collection phase, the attacker implements a different control action than the one transmitted by the controller. Thus, a replay attack can be implemented starting at time $k = 0$ for T steps by the attacker making the substitution

$$y_a(k) = y(k - T), \qquad 0 \leq k \leq T - 1, \tag{8.}$$

and simultaneously injecting an external input signal $u_a(k)$ for $0 \leq k \leq T - 1$. This attack is best applied when the system is at a steady state since the measurements will then correspond to a system state that is well-regulated. In this case, the attack is difficult to detect with any statistical test on the sequence of received measurements. Depending on the actuation capability available to the attacker, the true system state can degrade rapidly.

5.1. Attack Detection

The detection methodology for a replay attack relies on the basic idea of the controller purposefully and unpredictably moving the system away from a steady-state value. This variation would not show up in the measurements that the attacker substitutes and hence the attack can be detected. The price to pay for such variations is that even in the absence of an attack, the plant state moves away from the desired value. Further, there is a tradeoff between the amount of control energy expended in such variations and the ease with which an attack can be detected.

To consider a simple example, consider a system of the form in equation 1 with $a(k) \equiv 0$ and zero-mean white Gaussian noises w and v with covariances Σ_w and Σ_v, respectively. Assume that the nominal controller is an LQG controller that minimizes the standard quadratic cost

$$J = \lim_{T\to\infty} \mathbb{E}\frac{1}{T}\left(\sum_{k=0}^{T-1}\left(x^T(k)Qx(k) + u^T(k)Ru(k)\right)\right). \tag{9.}$$

In steady state, the controller first computes the minimum mean squared error estimate $\hat{x}(k|k)$ of the state $x(k)$ given measurements $y(0), \ldots, y(k)$ and then calculates the control input

$$u(k) = u^\star(k) = -\left(B^T SB + R\right)^{-1} B^T SA\hat{x}(k|k), \tag{10.}$$

where S satisfies the Riccati equation

$$S = A^T SA + Q - A^T SB - \left(B^T SB + R\right)^{-1} B^T SA. \tag{11.}$$

If the covariance of the steady-state estimation error for the estimate $\hat{x}(k|k-1)$ is given by P, then the cost achieved by the LQG controller is given by

$$J = \text{trace}\,(S\Sigma_w) + \text{trace}\left(\left(A^T SA + Q - S\right)\left(P - PC^T\left(CPC^T + R\right)^{-1} CP\right)\right). \tag{12.}$$

One could imagine that if an attacker changes the measurements, it might show up in the statistics of the innovation sequence of the MMSE estimates being calculated. We could propose checking the mean and variance of the sequence $y(k) - C\hat{x}(k|k-1)$ to see if an attack is present. It is easy to see that, in the absence of an attack, this sequence is an i.i.d. sequence of Gaussian variables with mean zero and variance $CPC^T + \Sigma_v$. Thus, for instance, for a window size of detection N, a χ^2 detector could be used to check for mean and variance of this sequence being calculated. However, for a replay attack, if the measurements being replayed were collected when the plant was at steady state, these statistics do not change and such a detector would not be able to identify an attack in progress.

5.2. Mitigation

In order to detect a replay attack, we redesign the controller as

$$u(k) = u^\star(k) + \Delta u(k), \tag{13.}$$

where $\Delta u(k)$ is drawn from an i.i.d. Gaussian sequence with mean 0 and covariance Σ_u in a manner that is independent of the inputs $u^\star(k)$ that are calculated as in equation 10. These variables can be viewed as an authentication signal, and have also been viewed in the literature as a physical watermarking signal since the detector expects to see a signature of this signal in the future time steps. Clearly, the controller is no longer LQG optimal and the choice of the authentication signal above is rather ad hoc. However, for this sequence, the tradeoff between the ease of detection of the attack (say by a χ^2 detector) and the degradation in control performance can be easily characterized. Specifically, under some technical conditions, we can calculate the LQG performance now as

$$\bar{J} = J + \text{trace}\left(\left(R + B^T S B\right)\Sigma_u\right). \tag{14.}$$

We can characterize the expectation of a χ^2 detector as follows. First, we note that the following holds:

$$\lim_{k\to\infty} \mathbb{E}\left[(y(k) - C\hat{x}(k|k-1)^T (CPC(T+R)^{-1}(y(k) - C\hat{x}(k|k-1)\right] = \begin{cases} p & \text{if no attack} \\ p + 2\text{trace}(C^T(CPC(T+R)^{-1}C\mathcal{U} & \text{if attack,} \end{cases} \tag{15.}$$

where p is the dimension of $y(k)$, P is the solution of the Riccati equation corresponding to the estimation problem, and $\mathcal{U}$ is the solution of a Lyapunov equation (see more details in (35)). This result can then be used to show that for a χ^2 detector that considers a window of T measurements, its steady state expectation is increased from pT without attack to $pT + 2\text{trace}(C^T(CPC(T+R)^{-1}C\mathcal{U}T$.

Since $\mathcal{U}$ is an increasing function of Σ_u, there is a trade-off between the ease of detection of the attack, and the loss in LQG performance without an attack. In a SISO system, there is only one way to insert the random signal Δu, and only one way to observe it. Thus, to achieve a certain detection rate, a certain performance loss has to be accepted. In MIMO systems, however, we have more degrees of freedom. The random signal can be optimized such that the detection requirements are met while minimizing the effect on controller performance. Two ways in which the optimization problem can be posed and solved are provided in (34).

The formulation discussed above can be extended in various ways. For instance, (36) removed both the assumptions of the authentication (or watermarking) signal $\Delta u(k)$ being chosen to be an i.i.d. sequence, as well as the χ^2 detector being employed to detect if an attack is in progress. That work considered instead the case when the authentication signal is designed as the output of a linear dynamical system expressed in a state space form. Further, instead of the χ^2 detector, an optimal Neyman-Pearson detector was identified to decide among the two hypotheses of an attack being present or not. Together, these assumptions allow consideration of an adversary employing more intelligent attack strategies. The performance of this detector can be characterized by an appropriate Kullback-Leibler divergence between the probability density functions characterizing the output sequence under the two rival hypotheses. An optimization problem for designing the authentication signal can once again be posed.

The paper (42), besides presenting a nice overview of the area of replay attacks, also extends these results in many directions, such as considering systems with arbitrary delay, partially observed systems, and non-Gaussian systems. The latter, in particular, remains a significant open problem since the assumption of linear systems driven by Gaussian noises provides a significant advantage in designing and analyzing the statistical tests for detecting an attack. We would also like to mention the paper (18), which considered a multiplicative watermarking algorithm that utilized a watermark removing function to prevent any sacrifice of control performance.

6. DoS ATTACK

A DoS attack is in general a cyberattack in which a malicious actor aims to render a computer, machine or system unavailable for its user. This is often done by flooding a computer network or server with data traffic making them unable to serve their user or even crash. In this section, we discuss DoS attacks for networked control systems. Such attacks interrupts temporarily or indefinitely the feedback control loop and can thereby have drastic consequences for the system operation. Obviously, if the loop of an unstable plant is opened, the attack leads to the overall system becoming unstable. It is natural to model DoS attacks as switches introduced into the networked control system, where an open switch indicates a control loop under attack. We present such a model next followed by some mitigation strategies proposed in the literature. A more extensive overview of DoS attacks in control systems is given in (10).

Consider the networked control system in Figure 4 again with the plant dynamics given in equation 1, but with no additive injected signal: $a(k) \equiv 0$. Suppose that the closed-loop system is exposed to DoS attacks in the communication of the control commands and sensor measurements. Such attacks are represented by the switches in Figure 4, and in the model by introducing the multiplicative relations

$$u_p(k) = \gamma(k)u(k), \qquad y(k) = \delta(k)y_p(k),$$

where u_p is the control applied to the plant and y_p the plant output. The binary variables $\gamma(k)$ and $\delta(k)$ represent DoS attacks, such that $\gamma(k) = 0$ if the control command communication is interrupted by an attack and $\gamma(k) = 1$ otherwise, and $\delta(k)$ defined in a similar way. In this model γ and δ are decision variables of the adversary. The attack space in Figure 5 shows that a DoS attack can be performed using only disruption resources, while in a more sophisticated scenario the adversary could use also other available information

and resources.

The interplay between the decision by controller and the one by the adversary leads to different analysis and design problems. One relevant case is when the controller is given, for instance, as a static feedback law $u(k) = -Ky(k)$. The closed-loop system then becomes a switched system, which without noise ($w(k) = v(k) \equiv 0$) and with scalar controls and measurements, can be written as

$$x(k+1) = (A - BK\gamma(k)\delta(k))x(k). \qquad 16.$$

Such systems can be analyzed using Lyapunov methods based on input-to-state stability (45) as was done in (14), which derived conditions on how long and how often DoS attacks can happen without rendering the closed-loop system unstable. In this work, the adversary is represented by deterministic sequences $\gamma(k)$ and $\delta(k)$, $k = 0, 1, \ldots$. These results make it possible to reason about worst-case situations and can give an indication about what combination of open- and closed-loop dynamics together with DoS attack patterns are especially undesirable. Uncertainty in the model and communication can also be conveniently included in the analysis. Nonlinear and continuous-time plant models can be considered as well (13).

Networked control systems under DoS attacks have been considered also when the plant is exposed to stochastic uncertainties and the attack signal is randomly generated. Amin et al. (3) considered an LQG setting, where w and v in equation 1 are assumed to be i.i.d. Gaussian noise. The adversary is supposed to let the sequences γ and δ be Bernoulli distributed with fixed success probabilities $\bar{\gamma}$ and $\bar{\delta}$. Under such a model, it is shown that the optimal controller subject to power constraints can be derived by solving a semidefinite program. A slightly more realistic attack model is also considered, where the adversary has to decide how a finite budget of drop outs should be utilized, denoted as block attacks. For an LQG controlled system, the worst possible (optimal for the adversary) DoS attack was derived in (58) and it was shown under general assumptions to have such a block structure. If the drop outs are due to a jamming attack of a wireless communication channel, it is possible to model the influence of the jamming signal on the signal-to-interference-plus-noise ratio (SINR). Such an attack is studied for a remote state estimation problem in (28), where it is shown that under a power-constrained setup, a game can be formulated between the transmitter and the attacker and that it admits a pure-strategy Nash equilibrium. Another modification of the memoryless Bernoulli process attack model above is to extend it to a hidden Markov model. Befekadu et al. (8) considered this problem and came up with a risk-sensitive control formulation, which is shown to have a nice separation principle that allows the optimal control to be recursively computed.

6.1. Mitigation

Several mitigation strategies for networked control systems under DoS attacks have been proposed in the literature. Some strategies are based on the analysis provided by the attacks described above, such as making sure that sufficiently many sensor or control packets are transmitted (3, 14) or are transmitted with sufficiently high power (28), to weaken the influence of the malicious intent. For a networked control system, it is natural to utilize the inherent resilience provided by the network, such as multi-path routing and data authentication. By allowing, for instance, certain sensor data packets to take multiple paths or to randomly change path, the overall security can be enhanced considerably (52). A

more sophisticated defense setup is considered in (41), where a two-level defense mechanism with an intrusion detection system and continuous authentication is studied. The intrusion detection system is responsible for detecting the attacker and network anomalies. Since traditional authentication is insufficient to prevent identity theft attacks, continuous authentication based on the characterization of user behavior has emerged. The authors show how the operator can combine these defense mechanisms to make the system more secure even under advanced learning-based attacks.

7. OUTLOOK

To conclude, we present a brief outlook and discuss some aspects of secure networked control systems not covered in the previous sections. To simplify the exposure, we focused in this article on a discrete-time linear system setup. For some of the results discussed on FDI, replay, and DoS attacks, there exist extensions to continuous-time as well as nonlinear systems. However, many questions remain open in that setting, particularly in the stochastic setting.

Another important aspect covered only to a limited extent in the literature is the robustness to system model uncertainties and to limited knowledge about the adversary, cf. Section 3 on system and attack models. For example, the problems in which the system parameters are unknown are becoming increasingly important. In practice, both attack and defense strategies based on probing and learning the system and user responses from available data seem to be more and more common (50). Some recent works exist in which the attacks above are combined with learning algorithms. For example, the watermark design problem mentioned as a mitigation strategy for the replay attack in Section 5 was extended in this direction in (29), which presents an online learning algorithm to simultaneously infer the parameters of the system and generate the watermark signal as well as the optimal detector. Interestingly, the watermark signal asymptotically converge to a signal that does not satisfy the persistent excitation condition. The authors show that by controlling the convergence rate, the system parameters can still be guaranteed to converge to the true parameters almost surely. This work could expand watermarking in a number of directions, including consideration of nonlinear systems, through appropriate learning techniques. It is also interesting to consider other attack scenarios, where a mitigation strategy could be based on a joint online system identification and defense algorithm design problems.

Machine learning and data analytics are widely applied to cyber-security problems (16, 56). Some of these studies are also relevant for networked control systems. Methods not relying on a known system model can be advantageous in many situations. One aspect where they are particularly relevant even if the system model is known is when the model is non-linear or the noises are not Gaussian, since the resulting statistics of the signals make tools depending on Kalman filters or hypothesis testing difficult to apply. Signal-based methods include some mitigation strategies based on classical detection schemes, as discussed in Sections 4–5. Causality measures can be useful in detecting anomalies in a networked control system without accurate knowledge of the model or statistics. Transfer entropy is one such measure from physics indicating how much a signal can improve the prediction capability of another signal. By detecting changes in the transfer entropy, it was shown in (44) that FDI, replay, and DoS attacks can be mitigated in a data-driven fashion without relying on a model of the underlying system.

Another assumption so far was that a stand-alone process was considered. There is a

rich literature on secure multi-agent control systems, studying a variety of problems from the simple consensus protocol under cyberattacks to more general multi-agent dynamics, see (46), for example. As a further illustration of these results, consider the consensus network under DoS attacks in (43). The authors show that consensus can be preserved despite the attack by suitable time-varying control and communication policies. However, for more general distributed control systems with multiple attackers, the problem becomes quite complicated (especially if coordination, detection, and mitigation need to be done in a distributed manner) as issues of signaling become relevant.

Finally, to develop courses and curricula in which security of networked control systems takes an essential role is important for the wider deployment of the protection methods developed over the last couple of decades. Today, security is seldom taught as part of control courses in chemical, civil, electrical, and mechanical engineering, despite the fact that systems from these domains are continuously exposed to cyber vulnerabilities. Curriculum at various levels—undergraduate, graduate, and continuing education—needs to be developed in a holistic manner including both theory and practice.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partially supported by the Swedish Research Council, Swedish Strategic Research Foundation, Swedish Civil Contingencies Agency, Knut and Alice Wallenberg Foundation, the US Army Research Office, and the US Air Force Office of Scientific Research.

LITERATURE CITED

1. A. Ahlen, J. Akerberg, M. Eriksson, A. J. Isaksson, T. Iwaki, K. H. Johansson, S. Knorn, T. Lindh, and H. Sandberg, "Toward wireless control in industrial process automation: A case study at a paper mill," *IEEE Control Systems Magazine*, vol. 39, no. 5, pp. 36–57, 2019.
2. A. A. Al Makdah, V. Katewa, and F. Pasqualetti, "A fundamental performance limitation for adversarial classification," *IEEE Control Systems Letters*, vol. 4, no. 1, pp. 169–174, 2020.
3. S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*, R. Majumdar and P. Tabuada, Eds. Berlin, Heidelberg: Springer, 2009, pp. 31–45.
4. R. Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*, 3rd ed. Wiley, 2020.
5. C.-Z. Bai, V. Gupta, and F. Pasqualetti, "On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6641–6648, 2017.
6. C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, 2017.
7. M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. USA: Prentice-Hall, Inc., 1993.
8. G. K. Befekadu, V. Gupta, and P. J. Antsaklis, "Risk-sensitive control under Markov modulated

Denial-of-Service (DoS) attack strategies," *IEEE Transactions on Automatic Control*, vol. 60, no. 12, pp. 3299–3304, 2015.
9. B. Besselink, V. Turri, S. van de Hoef, K.-Y. Liang, A. Alam, J. Mårtensson, and K. H. Johansson, "Cyber-physical control of road freight transport," *Proceedings of IEEE*, vol. 104, no. 5, pp. 1128–1141, 2016.
10. A. Cetinkaya, H. Ishii, and T. Hayakawa, "An overview on denial-of-service attacks in control systems: Attack models and security analyses," *Entropy*, vol. 21, no. 2, 2019.
11. M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *2015 American Control Conference (ACC)*, 2015, pp. 2439–2444.
12. Cybersecurity and Infrastructure Security Agency, US Department of Homeland Security, "Ics alert: Cyber-attack against ukrainian critical infrastructure," https://us-cert.cisa.gov/ics/alerts/IR-ALERT-H-16-056-01, 2016.
13. C. De Persis and P. Tesi, "Networked control of nonlinear systems under denial-of-service," *Systems & Control Letters*, vol. 96, pp. 124–131, 2016.
14. C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2930–2944, 2015.
15. S. X. Ding, *Model-Based Fault Diagnosis Techniques: Design Schemes, Algorithms, and Tools*, 1st ed. Springer Publishing Company, Incorporated, 2008.
16. S. Dua and X. Du, *Data mining and machine learning in cybersecurity*. CRC press, 2016.
17. 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.
18. R. Ferrari and A. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7363–7368, 2017, iFAC World Congress.
19. Z. Gao, C. Cecati, and S. X. Ding, "A survey of fault diagnosis and fault-tolerant techniques—part i: Fault diagnosis with model-based and signal-based approaches," *IEEE transactions on industrial electronics*, vol. 62, no. 6, pp. 3757–3767, 2015.
20. L. Garber, "Denial-of-service attacks rip the internet," *Computer*, vol. 33, no. 04, pp. 12–17, 2000.
21. J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, and R. Candell, "A survey of physics-based attack detection in cyber-physical systems," *ACM Comput. Surv.*, vol. 51, no. 4, Jul. 2018.
22. A. Greenberg, "Hackers remotely kill a Jeep on the highway——with me in it," *Wired*, Jul 2015.
23. J. Katz and Y. Lindell, *Introduction to Modern Cryptography*. Chapman and Hall, 2007.
24. E. Kung, S. Dey, and L. Shi, "The performance and limitations of ϵ- stealthy attacks on higher order systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 941–947, 2017.
25. D. Kushner, "The real story of stuxnet," *IEEE Spectrum*, Feb 2013.
26. T. L. Lai, "Information bounds and quick detection of parameter changes in stochastic systems," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2917–2929, 1998.
27. D. Li and S. Martínez, "High-confidence attack detection via Wasserstein-Metric computations," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 379–384, 2021.
28. Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "SINR-based DoS attack on remote state estimation: A game-theoretic approach," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 3, pp. 632–642, 2017.
29. H. Liu, Y. Mo, J. Yan, L. Xie, and K. H. Johansson, "An online approach to physical watermark design," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3895–3902, 2020.
30. S. Liu, X. P. Liu, and A. El Saddik, "Denial-of-Service (DoS) attacks on load frequency control in smart grids," in *IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*, 2013, pp. 1–6.
31. Y. Liu, M. K. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *ACM Conference on Computer and Communications Security*, New

York, NY, USA, 2009, p. 21–32.
32. C. Miller and C. Valasek, "Remote exploitation of an unaltered passenger vehicle," in *Black Hat USA Conference*, Las Vegas, NV, USA, 2015, extended report at illmatics.com.
33. J. Milošević, A. Teixeira, T. Tanaka, K. H. Johansson, and H. Sandberg, "Security measure allocation for industrial control systems: Exploiting systematic search techniques and submodularity," *International Journal of Robust and Nonlinear Control*, vol. 30, no. 11, pp. 4278–4302, 2020.
34. Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.
35. Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Allerton Conference on Communication, Control, and Computing*, 2009, pp. 911–918.
36. Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.
37. I. Mokube and M. Adams, "Honeypots: concepts, approaches, and challenges," in *ACM Proceedings of the 45th Annual Southeast Regional Conference*, 2007.
38. F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, Nov 2013.
39. K. Pelechrinis, M. Iliofotou, and S. V. Krishnamurthy, "Denial of service attacks in wireless networks: The case of jammers," *IEEE Communications Surveys & Tutorials*, vol. 13, no. 2, pp. 245–257, 2010.
40. F. A. P. Petitcolas, *Encyclopedia of Cryptography and Security*, H. C. A. van Tilborg and S. Jajodia, Eds. Springer, 2011.
41. S. Sarıtaş, E. Shereen, H. Sandberg, and G. Dán, "Adversarial attacks on continuous authentication security: A dynamic game approach," in *Decision and Game Theory for Security*, T. Alpcan, Y. Vorobeychik, J. S. Baras, and G. Dán, Eds. Springer International Publishing, 2019, pp. 439–458.
42. B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2017.
43. D. Senejohnny, P. Tesi, and C. De Persis, "A jamming-resilient algorithm for self-triggered network coordination," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 981–990, 2018.
44. D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 386–401, 2018.
45. E. D. Sontag, *Input to State Stability: Basic Concepts and Results*. Berlin, Heidelberg: Springer, 2008, pp. 163–220.
46. S. Sundaram and C. N. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.
47. A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry, "Cyber-security analysis of state estimators in electric power systems," in *IEEE Conference on Decision and Control*, Atlanta, GA, USA, 2010.
48. A. Teixeira and R. M. Ferrari, "Detection of sensor data injection attacks with multiplicative watermarking," in *2018 European Control Conference (ECC)*, 2018, pp. 338–343.
49. A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.
50. K. K. Trejo, J. B. Clempner, and A. S. Poznyak, "Adapting strategies to dynamic environments in controllable stackelberg security games," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 5484–5489.
51. D. Umsonst and H. Sandberg, "A game-theoretic approach for choosing a detector tuning under

stealthy sensor data attacks," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 5975–5981.
52. O. Vukovic, K. C. Sou, G. Dan, and H. Sandberg, "Network-aware mitigation of data integrity attacks on power system state estimation," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1108–1118, 2012.
53. S. Weerakkody, X. Liu, S. H. Son, and B. Sinopoli, "A graph-theoretic characterization of perfect attackability for secure design of distributed control systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 60–70, 2017.
54. S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *2015 54th IEEE Conference on Decision and Control (CDC)*, 2015, pp. 5820–5826.
55. A. Willsky and H. Jones, "A generalized likelihood ratio approach to the detection and estimation of jumps in linear systems," *IEEE Transactions on Automatic Control*, vol. 21, no. 1, pp. 108–112, 1976.
56. Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Gao, H. Hou, and C. Wang, "Machine learning and deep learning methods for cybersecurity," *IEEE access*, vol. 6, pp. 35 365–35 381, 2018.
57. K. Zetter, "Inside the cunning, unprecedented hack of Ukraine's power grid," *Wired*, Mar 2016.
58. H. Zhang, P. Cheng, L. Shi, and J. Chen, "Optimal DoS attack scheduling in wireless networked control system," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 3, pp. 843–852, 2016.