A Tutorial Introduction to Security and Privacy for Cyber-Physical Systems

Michelle S. Chong, Henrik Sandberg, André M.H. Teixeira

***Abstract*— This tutorial provides a high-level introduction to novel control-theoretic approaches for the security and privacy of cyber-physical systems (CPS). It takes a risk-based approach to the problem and develops a model framework that allows us to introduce and relate many of the recent contributions to the area. In particular, we explore the concept of risk in the context of CPS under cyber-attacks, paying special attention to the characterization of attack scenarios and to the interpretation of impact and likelihood for CPS. The risk management framework is then used to give an overview of and map different contributions in the area to three core parts of the framework: attack scenario description, quantification of impact and likelihood, and mitigation strategies. The overview is by no means complete, but it illustrates the breadth of the problems considered and the control-theoretic solutions proposed so far.**

I. Introduction

Cyber-physical systems (CPS) represent a class of networked control systems with vast and promising applications, such as smart cities, distributed sensing and control based on Internet-of-Things (IoT) devices, or ground-breaking transportation systems based on fleets of cooperative and autonomous vehicles. CPS also include more traditional large-scale control infrastructures found in the process control and power industries. These systems all provide outstanding functionalities and positively influence our life and society. However, such positive outcomes may be hindered by novel threats to the safety of CPS, such as malicious cyber-attacks that may negatively affect the physical domain. Indeed, the past years have witnessed some dramatic cyber-attacks against CPS, with significant media coverage. Several malwares dedicated to attacks against CPS have been discovered, with names such as Stuxnet, Black Energy, Industoyer, and Triton[1].

For these reasons, there has been a surge in the interest for security and privacy in control systems during the past ten years. The IEEE Control Systems Magazine special issue [1] gave an overview of some of the early work in the area. Four years have passed since the special issue, and in this tutorial introduction we aim to also introduce some of the more recent work. However, before turning to the CPS security and privacy problems, we should remind ourselves of the basic security properties analyzed in IT systems.

Information is a key asset in knowledge-driven societies, which require a reliable and continuous availability of data and services. Redundant and fault-tolerant architectures are thus required to build IT systems resilient to faults and disturbances [2]. Additionally, IT systems must also be defended against malicious adversaries whose aim is in disrupting or gaining access to the information flow.

Three fundamental properties of information and services in IT systems are mentioned in the computer security literature [3] using the acronym CIA: *confidentiality, integrity*, and *availability*. Confidentiality concerns the concealment of data, ensuring it remains known to the authorized parties alone. Integrity relates to the trustworthiness of data, meaning there is no unauthorized change to the information between the source and destination. Availability considers the timely access to information or system functionalities.

These three properties can be violated through disclosure, deception, and denial-of-service attacks, respectively. While in IT systems, the impact of such cyber-attacks remains in the cyber-realm, they may cause dire consequences to the physical system in networked control systems. This will be further explored in our tutorial.

Outline: The paper is structured as follows. In Section II, an overview of the risk management framework is given, focusing on some of its core elements: attack scenario description, risk analysis, and risk treatment. This framework serves to contextualize the overall formulation of cyber-security problems, as well as different analysis design questions to assess and improve cyber-security. The following three sections look deeper into each core element. Classical cyber-attack scenarios for CPS are succinctly described in Section III, along with the respective adversarial capabilities. Section IV discusses recent work addressing the evaluation of risk (i.e., the impact and likelihood of each attack scenario). Recent approaches to reduce risk of specific cyber-attacks are summarized in Section V. The paper concludes with final remarks and future research directions in Section VI.

II. Risk Management Framework

The risk management framework [3]–[5] is a common methodology to enhance a system's cyber-security. The main objective of risk management is to identify, assess, and minimize the risk of threats.

M. Chong and H. Sandberg are with the Division of Decision and Control Systems at the KTH Royal Institute of Technology, Stockholm, Sweden. {mchong, hsan}@kth.se

A. Teixeira is with the Department of Engineering Sciences at the Uppsala University, Uppsala, Sweden. andre.teixeira@angstrom.uu.se

This work was supported in part by the Swedish Research Council (grants 2016-00861 and 2018-04396), the Swedish Civil Contingencies Agency through the CERCES project, and the Swedish Energy Agency through the ERA-Net project LarGo!.

[1]See, for instance, https://securingtomorrow.mcafee.com/other-blogs/mcafee-labs/triton-malware-spearheads-latest-generation-of-attacks-on-industrial-systems/

Since risk may vary over time, with the appearance of new threat scenarios and the aging of the system, risk must be continuously managed to ensure security. Such a requirement leads to the cyclic execution of the risk management process, which includes, among others, two core stages: risk analysis and risk mitigation.

A. The Concept of Risk

The classical notion of risk is defined as follows [6]. Consider a given set of threat scenarios, the corresponding impact to the system, and the likelihood of such scenarios. The risk of the system corresponding to the set of threat scenarios is denoted as the set of triplets $\textit{Risk} \triangleq \{(\textit{Scenario}, \textit{Impact}, \textit{Likelihood})\}$.

Although somewhat implicit in the definition, risk is not a property of a system. Instead, it will depend on the specific attack scenarios under which the system is examined. For instance, different scenarios could be considered depending on elements such as: what devices are compromised, how knowledgeable about the system the adversary is, what type of adversary is attacking the system, what security mechanisms are in place, among others. For each specific attack scenario, the impact of the attack and the likelihood of the attack being successfully implemented can be assessed.

In information security risk management [5], *impact* typically relates to the damage that the attack can have on the information (cyber) system itself, such as the disclosure of sensitive information, or denying service from critical functionalities. In CPS however, attacks on the digital components have damage that extends beyond the cyber-realm, affecting also the physical side of the system. Therefore, one of the new aspects of security of CPS is related to the assessment of the impact of cyber-attacks on physical processes, which is further examined in Section IV.

In contrast to impact, likelihood (of an attack scenario) is an elusive concept in cyber-security risk management, and so is its assessment. In safety risk management, which deals with risk against natural phenomena, e.g. earthquakes [7], likelihood is understood as the posterior probability of a given consequence event occurring as a result of the physical phenomena under study.

In the case of security risk management [8], [9], the phenomenon of interest is not a cause of nature, but rather the act of a malicious and intelligent adversary. As such, there is no objective form of computing the prior probability of such an event, which renders the notion of likelihood somewhat void of meaning. One can, however, look at likelihood of cyber-attacks without the prior probabilities, and instead use only the conditional probabilities. In this way, the term *likelihood* then captures not probabilities of attacks occurring, per se, but rather captures the likelihood of attacks already in progress of being successful. Still, due to the lack of historical data regarding cyber-attacks, and their intentional and rational nature, it is unfeasible to assess likelihood based on probabilities in general. To circumvent this issue, proxies for likelihood are often used, such as the complexity of the cyber-attack itself, the level of knowledge and resources required, and the amount of devices that must be corrupted [10].

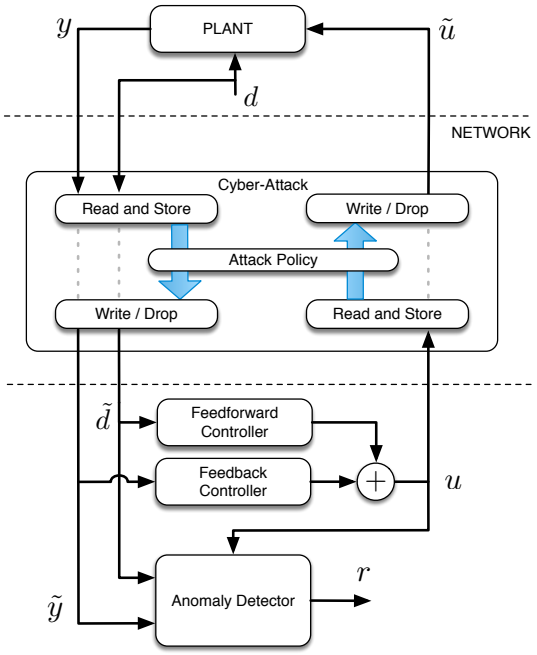

Fig. 1. Illustration of a networked control system architecture subject to cyber-attacks.

The three elements composing risk are further discussed in the following subsections.

B. Attack Scenario Description

A typical architecture of CPS is depicted in Fig. 1. At the top, we have a physical process or plant, which is controlled and monitored through digital computers over a possibly unsecured communication network. Measurement of the plant $y[k] \in \mathbb{R}^m$ and of possible disturbances $d[k] \in \mathbb{R}^v$ are transmitted through the network to digital computers, on which controller and anomaly detector algorithms have been implemented. These algorithms compute the control signal $u[k] \in \mathbb{R}^p$ used to steer the physical plant, as well as alarm variables, namely the residual $r[k] \in \mathbb{R}^s$, that are evaluated to determine whether anomalies have been detected, or not, within the overall system. The control signal $u[k]$ is transmitted to the actuators over the possibly unsecured network. Given that the network may be unsecured, the measurement and control signals may be manipulated along the communication channels, as depicted in Fig. 1.

In general, data exchanged through the communication network can be eavesdropped by the adversary, as well as compromised before being sent to the destination. A wide-class of attacks are described by attack policies of the form

$$\begin{aligned} \tilde{y}[k] &= \phi_y\left(Y_{[0,k]}, U_{[0,k]}, D_{[0,k]}\right) \\ \tilde{u}[k] &= \phi_u\left(Y_{[0,k]}, U_{[0,k]}, D_{[0,k]}\right) \\ \tilde{d}[k] &= \phi_d\left(Y_{[0,k]}, U_{[0,k]}, D_{[0,k]}\right), \end{aligned} \tag{1}$$

where, given a vector variable $x[j] \in \mathbb{R}^n$, $X_{[0,k]} \triangleq$

$\{x[0], x[1] \cdots, x[k]\}$ denotes the set of all samples of $x[j]$ from time 0 to time k. Specific instances of attack policies will be revisited in Section III, including those where the adversary wishes to damage the plant as much as possible while remaining undetected, and therefore delaying the triggering of any sort of pro-active mitigation scheme.

C. Risk Analysis

Risk analysis identifies threats and assesses the respective likelihood and impact on the system. Threat scenarios may be identified based on historical and empirical data of cyber-attacks, expert knowledge, and known vulnerabilities in the system [5]. The report [11] provides a good example of power system related threat scenarios identified from expert knowledge. The likelihood of a given threat depends on the components compromised by the adversary in a given attack scenario and their respective vulnerability. Quantitative methods can be used to identify the minimal set of components that need to be compromised for each attack scenario [12], [13], while the vulnerability of each compromised components is obtained by qualitative means such as expert knowledge and historical and empirical data [12]. The potential impact of a threat may be assessed by qualitative and quantitative methods, for instance, by modeling the system and simulating the attack scenarios [14].

The risk of different threat scenarios may be summarized in a two-dimensional risk matrix [5], where each dimension corresponds to the likelihood and impact of threats, respectively. The objective of the risk management framework is to reduce the overall risk of attacks, which typically requires the selection of critical attacks (with high impact and high likelihood) for deploying targeted risk mitigation strategies.

D. Risk Mitigation

Actions minimizing the risk of threats are determined within the risk mitigation step. The different actions can be classified as prevention, detection, and treatment. A brief overview is provided next, while specific risk mitigation solutions for CPS are further described in Section V.

1) Prevention: Prevention aims at decreasing the likelihood of attacks by reducing the vulnerability of the system components, e.g., by encrypting the communication channels [15] and using firewalls. For instance, with respect to disclosure attacks violating confidentiality, encryption of the communication link corresponds to a preventive action.

Through the risk management framework, prevention can be more efficiently deployed, by taking into account the available security resources and the most critical scenarios with higher risk (i.e., higher impact and likelihood). Examples of recent work addressing the rational allocation of security are given in Section V-I.

2) Detection: Detection is an approach in which the system is continuously monitored for anomalies caused by adversarial actions. Anomaly detection mechanisms (see Fig. 1) typically involve the processing of measurement and input data, for instance by fault detection algorithms [16], to produce the so-called residual sequences $R_{[j,k]}$, which are then evaluated by detector schemes. Traditional detectors considered are χ^2-detectors (see, for instance, [17]), CUSUM detectors (see, for instance, [18]), or more generic 2-norm measures of residual sequences (see, for instance, [19], [20]).

3) Treatment: Once an anomaly or attack is detected, treatment actions may be taken to disrupt and neutralize the attack. The attack may be neutralized by replacing the compromised components or using redundant components. In the case of the denial-of-service attacks violating availability, one could have a treatment scheme where the data are re-sent using a different path from source to destination, thus avoiding the compromised links [21].

As a concluding remark regarding the risk management framework, it must be highlighted that risk is a variable that evolves over time. Novel attack scenarios appear more often than not, and impact and likelihood of attacks also varies with the aging of the system components.

Therefore, the effectiveness of the defensive actions and the evolution of risk over time must be evaluated regularly using the risk management framework. For instance, in the case of deception attacks, the attacker may find novel attack strategies that bypass the current detection mechanisms. This particular scenario is explored in Section IV-A, and novel risk mitigation schemes are described in Section V.

III. Attack Scenarios

In the following, we succinctly revisit some of the key attack scenarios considered in the literature, examining each attack in terms of its attack policy and in terms of the resources required of the adversary for a successful attack: CPS model knowledge, disclosure resources, and disruption resources. To provide a wider overview, a broader but still incomplete sample of common attack scenarios and recent work tackling them is presented in Fig. 2, where scenarios are placed along three axes, each corresponding to a particular resource requirement. In relation to Fig. 1, 'Disclosure resources' corresponds to the number of channels the attacker can 'Read and Store' data from. 'Disruption resources' corresponds to the number of channels in which the attacker can 'Write/Drop' data. Finally, 'CPS model knowledge' corresponds to the extent the attacker has access to models of the plant, controllers, and anomaly detector schemes. An advanced attack could very well consist of multiple stages/scenarios, and evolve in the diagram. For instance, a first stage could consist of learning of CPS models by means of eavesdropping attacks, while a second stage could be to implement undetectable attacks based on the models learnt.

A. Eavesdropping Attacks

Eavesdropping attacks aim at violating the confidentiality of the data, and thus disclose private information of the system [41]. Such attacks are of high importance in privacy-sensitive applications, for instance when personal health data is collected and used in medical studies. Additionally, eavesdropping attacks could also be the first step in a more damaging and disruptive attack (for instance, replay attacks).

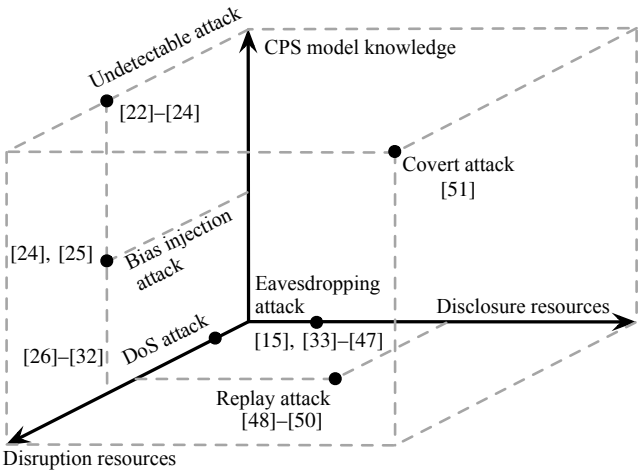

Fig. 2. Illustration of common attack scenarios. Each axis indicates the relative amount of a certain resource required to carry out an attack.

CPS model knowledge: the adversary does not have a model of the plant.

Disclosure resources: the adversary is able to read some or all of the transmitted data, and store it for later processing, which yields $Y_{[0,k]} = \{y[0], \cdots, y[k]\}$, $D_{[0,k]} = \{d[0], \cdots, d[k]\}$, and $U_{[0,k]} = \{u[0], \cdots, u[k]\}$

Disruption resources: the attacker is not able to corrupt the integrity and availability of the data being transmitted through the network.

Attack policy: in such an attack scenario, the attack policies do not affect the transmitted signals, which leads to:

$$\begin{aligned} \tilde{y}[k] &= y[k] \\ \tilde{u}[k] &= u[k] \\ \tilde{d}[k] &= d[k]. \end{aligned} \tag{2}$$

B. Open-loop False-data Injection Attacks

In the case of open-loop false-data injection attacks, the adversary aims at corrupting the integrity of the transmitted data to affect the system without being detected, a scenario that has been considered in [22], for instance. Such an attack scenario has been extensively covered in the literature, and the reader is referred to Section IV for further references.

CPS model knowledge: the adversary is considered to have a full model of the cyber-physical system.

Disclosure resources: the adversary does not record any of the transmitted data for later processing, leading to $Y_{[0,k]} = U_{[0,k]} = D_{[0,k]} = \emptyset$. However, for the adversary to be able to inject malicious data correctly, disclosure of the current transmitted data may be needed.

Disruption resources: the attacker is able to corrupt the integrity of specific data being transmitted through the network.

Attack policy: in such an attack scenario, the attack policies $\phi_y(\emptyset)$, $\phi_d(\emptyset)$ and $\phi_u(\emptyset)$ are computed offline based solely on the known plant models. The real-time implementation the attack, however, may require that the adversary can read the current data, so that the data corruption can be successfully implemented as follows:

$$\begin{aligned} \tilde{y}[k] &= y[k] + \phi_y(\emptyset) = y[k] + a_y[k] \\ \tilde{u}[k] &= u[k] + \phi_u(\emptyset) = u[k] + a_u[k] \\ \tilde{d}[k] &= d[k] + \phi_d(\emptyset) = d[k] + a_d[k]. \end{aligned} \tag{3}$$

C. Replay Attacks

In replay attacks, the adversary records measurement data into a buffer, and later replays the recorded measurements while tampering with the system through other channels, such as actuators or physical disturbances. Due to the replaying of old measurement data, the tampering goes unnoticed by anomaly detectors. Replay attacks have been first considered in [48], where their undetectability was analyzed, and an additive watermarking scheme was proposed to detect the attacks. Further developments in the analysis and detection of replay attacks may be found in the references discussed in Section V-D.

CPS model knowledge: the adversary does not have a model of the cyber-physical system.

Disclosure resources: the adversary is able to break confidentiality of all the measurement data, and records all of the transmitted measurement data up to a time T, leading to $Y_{[0,k]} = Y_{[0,T]} = \{y[0], \cdots, y[T]\}$, $D_{[0,k]} = D_{[0,T]} = \{d[0], \cdots, d[T]\}$ for $k \geq T$, and $U_{[0,k]} = \emptyset$.

Disruption resources: the attacker is able to replace measurement data with previously recorded data (a form of integrity violation). Additionally, the adversary is also able to freely disturb the physical plant directly, e.g., through the actuators.

Attack policy: in such an attack scenario, the attack policies become

$$\begin{aligned} \tilde{y}[k] &= \phi_y(Y_{[0,T]}) = y[k-T] \\ \tilde{d}[k] &= \phi_d(D_{[0,T]}) = d[k-T] \\ \tilde{u}[k] &= \phi_u(\emptyset) = a_u[k], \end{aligned} \tag{4}$$

where $a_u[k]$ is a free signal computed offline.

D. Denial-of-Service Attacks

In Denial-of-Service (DoS) attacks, the adversary aims at dropping transmitted data packets so that the performance of the closed-loop system is deteriorated [26], [27], perhaps even resulting in instability.

CPS model knowledge: the adversary does not have a model of the cyber-physical system.

Disclosure resources: the adversary does not need to break confidentiality of the transmitted data, leading to $Y_{[0,k]} = D_{[0,k]} = U_{[0,k]} = \emptyset$.

Disruption resources: the attacker is able to drop transmitted data packets and thus block them from reaching the intended destination (a pure data availability violation). Such a mechanism is similar to packet drops commonly analyzed within the research area of control over communication networks.

Attack policy: in DoS attacks, the attack policies become

$$\begin{aligned}\tilde{y}[k] &= \gamma_y[k]y[k] + (1-\gamma_y[k])\emptyset \\ \tilde{d}[k] &= \gamma_d[k]d[k] + (1-\gamma_d[k])\emptyset \\ \tilde{u}[k] &= \gamma_u[k]u[k] + (1-\gamma_u[k])\emptyset,\end{aligned} \tag{5}$$

where $\tilde{x}[k] = \emptyset$ is used to denote the absence of the data $\tilde{x}[k]$ at the receiver, and $\gamma_x[k] \in \{0, 1\}$ is a binary variable determined by the corresponding attack policy $\phi_x(\emptyset)$.

IV. Quantifying Impact and Likelihood

Having characterized common types of attacks, quantitative tools for their risk analysis are now discussed in more detail. In Section IV-A, we discuss work that identifies non-trivial attack scenarios, in Section IV-B some tools for likelihood assessment are presented, and in Section IV-C tools for impact assessment are presented.

A. Undetectable and Stealth Attacks

In [22], [23], the notion of *undetectable attacks* is defined for linear systems. The idea behind the definition is simple, yet powerful: A false-data injection attack (Section III-B) is undetectable if the resulting input to the anomaly detector block in Fig. 1 is identical to a signal the system could produce without the attacker present. The motivation behind the concept is that if the received signal could be the result of a naturally occurring state, then the operator may have no reason to expect an attack. The possibility of such undetectable attacks is equivalent to the existence of invariant zeros in the system. For this reason, these attacks are sometimes also called *zero-dynamics attacks*. Especially dangerous destabilizing undetectable attacks exist when the plant has unstable zeros (using $a_u[k]$ in (3)) or unstable poles (using $a_y[k]$ in (3)). It should be noted that an anomaly detector could also check if the received signal is likely or not (not only feasible), and in principle also could raise an alarm when an 'undetectable' attack is staged. Hence, an attacker who would like to remain hidden typically would like to include a model of the anomaly detector while designing attacks. This leads to the concept of *stealth attacks*, which occurred around the same time as the undetectable attacks.

In [18], [52]–[54], attacks against process and power systems controlled and monitored using SCADA systems were considered. It was natural to characterize the classes of attacks that would not raise alarms in the installed anomaly detectors, and these were called stealth attacks. Note that these attacks may result in 'un-natural' inputs to the anomaly detector block in Fig. 1, but that the detector output still does not trigger an alarm by design of the attack. Hence, the set of stealth attacks is larger than the set of undetectable attacks, but requires a stronger attacker with (at least partial) knowledge of the detector. Stealthy attacks have been studied in deterministic [20] and in stochastic [19], [55] settings. All attack scenarios of Section III, such as replay and DoS attacks, can be studied in the context of stealth attacks [56].

Many applications and extensions of the undetectable and stealth attacks have been published. Some examples are listed next. Coordinated undetectable attacks on the plant input and output are called *covert attacks* [51]. In *bias attacks* [24], [25], a bias is added to certain signals in a stealthy manner. Extensions for nonlinear uncertain systems [57], and for sampled-data systems [58] are also available.

B. Security Indices

The term 'security index' was first introduced for a linear dynamical system in steady state in [13] to quantify the vulnerability of sensor i against attacks modelled as an additive signal to the sensor measurements. The computation of this security index α_i for sensor i is also provided in [59] for a power network. In other words, security index in the context of [13], [59] is a static index. This concept was then generalized to dynamical systems in [60]. In contrast, the 'security index' in [61] is defined for the whole dynamical system. Indeed, the security index in [61] is the maximum value over all the security indices α_i in [13], [59].

The security indices are tools for estimating the complexity of certain undetectable attacks, which serves as a proxy for their likelihood. A large index means the attacker is required to have large resources in terms of disruption resources and model knowledge in order to conduct the corresponding attack. Hence, such attacks may be deemed less likely compared to attacks with small indices.

C. Worst-Case Stealth Attacks

In Section IV-A, we have discussed attacks that are hard, or even impossible, to detect, and in Section IV-B security indices have been introduced to quantify the difficulty of implementing such attacks. The next natural problem to consider is the quantification of the worst possible physical impact of these attacks. Several works have addressed this problem by applying tools from the optimal control theory. In [17], [19], the reachable set of the system state and estimation error under stealthy data injection attacks are characterized and bounded. In [18], several sensor attack schemes that maximize impact subject to a no-alarm condition in a CUSUM detector are characterized. In [20], the worst-case 2-norm impact of stealthy attacks is characterized in terms of generalized eigenvalue analysis, and worst-case infinity-norm impact is characterized in terms of linear programs. These works also highlight how the results can be used to compare the severity of various attack scenarios, and as such are useful in risk assessment. Further developments and applications of these ideas can be found in [56], [62]–[64], for instance.

V. Mitigation Strategies

In the previous sections, recent literature in the security of CPS have been revisited under the lens of a risk management framework. Special attention was paid to the three components of risk, by means of a description of different attack scenarios, and the quantification of the impact and likelihood of attacks with detectability constraints.

As a natural continuation, the present section reviews several approaches that have been proposed to mitigate the risk of attacks, by means of minimizing their impact

and/or likelihood. We label each class of approaches with 'Prevention', 'Detection', or 'Treatment' as introduced in Section II-D.

A. Tuning of Detector Thresholds [Detection]

The system architecture considered in Fig. 1 includes anomaly detectors which raise alarms if the received signals are sufficiently far away from nominal trajectories. For instance, in the case of χ^2-detectors, alarms would be triggered if the energy of the residual sequence $R_{[j,k]}$ exceeds a given threshold $\delta > 0$, i.e., $\|R_{[j,k]}\|_2^2 > \delta$. Due to the presence of natural disturbances, noise, and modeling errors, the thresholds of these detectors always need to be tuned to avoid raising false alarms constantly. However, thresholds which are too high will expose the system to stealth attacks, see Section IV-A. A common tool from detection theory to tune detectors is the ROC curve (Receiver Operating Characteristic curve, see, for instance, [65]), which plots the true positive detection rate against the false positive rate. In [66], it is argued, however, that the ROC curve is not suitable for tuning detectors against stealth attacks. This is because stealth attacks by definition have zero true positive detection rate, and hence the trade-off illustrated by the ROC curve is not helpful. Rather it is pertinent to map the thresholds to the worst-case physical impact, using methods from Section IV-C. A very low detector threshold typically yields a small worst-case impact, but gives a high false-alarm rate. Hence, [66] advocates that the worst-case impact should be plotted against the mean time between false alarms, to aid security-aware tuning of detectors. Following this recommendation, several works have developed techniques for explicitly computing such curves. See, [67], [68], for instance.

B. Secure State Estimation [Detection, Treatment]

The case for secure state estimation of CPS have concentrated on the scenario where n_a out of m sensors, where $n_a < m$ can be arbitrarily compromised by the adversary. The attack is modeled as an additive signal to the measurement, as shown in (3). A necessary and sufficient condition for estimating the states of linear dynamical systems under sensor attacks have been derived for both continuous [69] and discrete-time [70], which is combinatorial in nature. The redundancy of the state estimates obtained through multiple state observers is exploited to achieve exact reconstruction of the states, despite the presence of sensor attacks as long as less than half of the sensors have been compromised. The required number of state observers in both [69], [70] is $\binom{m}{m-n_a} + \binom{m}{m-2n_a}$, which becomes computationally infeasible for a large number of sensors m. This complexity issue has been addressed using various approaches to limit the search space, including satisfiability modulo theory [71], l_0 minimization [72], set covering [73], set partitioning [74], and an adaptive switching mechanism [75].

C. Privacy-preservation by Noise Injection [Prevention]

Privacy in CPS means sharing only the necessary information between subsystems in order to achieve either estimation or control objectives, while preventing eavesdroppers from obtaining sensitive data. The information shared is most often the aggregate to preserve the privacy of the individual subsystems. However, aggregation is generally not sufficient, and often the transmitted data is also intentionally corrupted by some mechanism. One popular mechanism, deriving from the database literature, is differential privacy. To attain differential privacy of a pre-defined level, the data holder typically adds a noise signal of sufficient variance to the released data to make estimation of the source data hard. In many cases, the optimal noise distribution is Laplacian [76]. A comprehensive tutorial paper [37] defines differential privacy in the context of systems and control and summarizes the various privacy-aware algorithms that have been devised for estimation [35], distributed control [34] and distributed convex optimization [39]. Differential privacy is one way of quantifying privacy, but there are other notions that also rely on random perturbations. Examples include perturbations adapted to the average consensus protocol [33], [38] or estimation problems [41], minimizing directed information in control loops [40], and minimizing Fisher information in estimation [47]. Game theory is also a viable approach to the privacy problem [36].

D. Watermarking and Moving Target Defense [Detection, Prevention]

Traditionally, watermarking is used in audio and image processing, where information is embedded in a carrier signal which is then used to verify the authenticity of the data. In the case of CPS security, watermarking or more commonly known in the literature as *physical* watermarking in this context, serves to authenticate that the CPS is operating normally by detecting irregularities in the measurable signals of the CPS when a watermark is embedded in any of the accessible signals.

In [49], this is achieved for a discrete-time linear dynamical system by designing a watermark signal which is superimposed on the control input that is optimal under the linear quadratic Gaussian (LQG) framework. In [50], a multiplicative watermarking signal is introduced in the sensor output of a discrete-time linear dynamical system, which is then removed on the controller side using a bank of filters. The authors showed that this watermarking scheme does not affect the closed-loop performance of the system in the absence of replay attacks.

However, watermarking is vulnerable against an adversary who has knowledge of the system model. One mechanism to address this shortfall is known as moving target defense, which was introduced in [77]. The idea is to obfuscate the adversary's knowledge of the system by varying the system's parameters over time (thereby acting as a moving target) to counter possible system identification which can be carried out by the adversary. A related idea involving the defender creating model uncertainty for the attacker was also proposed in [78]. Different mechanisms for moving target defense are further analyzed in [79], which include switching between different modes; an extended system interconnected

with the plant which do not affect system performance; and introducing random nonlinearities to the output.

E. Coding and Encryption Strategies [Prevention]

A plant which is controlled over a communication channel runs the risk of being exposed to adversaries who have access to the channel, who can then mount an attack based on the information gathered between the plant and estimator/controller. As such, coding and encryption methods are employed to protect the flow of information over a network. On one hand, we have coding schemes which are generally publicly known and any interception of data may expose private information of the CPS to attackers. On the other, encryption schemes are private by means of secret keys and hence adds a layer of security.

A common setup using encryption methods is where the information flow on the plant side is encrypted [15], [42]–[46]. Homomorphic encryption is a class of encryption which allows for computation on the encrypted data. This enables the implementation of controllers directly on encrypted data. *Semi* or *partially*-homomorphic encryption allows only a subset of computations to be performed on the encrypted data. The Pallier method [80] is one such example, which enables the summation of non-encrypted data through the multiplication of encrypted data.

Stability and performance guarantees for stabilizing linear dynamical systems under semi-homomorphic encryption with static and dynamic controllers are provided by [15], [42], [43], [45] and [81], respectively. In [46], the authors provide sufficient conditions to stabilize a nonlinear dynamical system, in the form of the encryption parameters, which recovers the results in [45] when linear controllers are considered.

In contrast to the aforementioned encryption techniques, coding schemes do not use secret keys. If data is intercepted without any errors, it may be decoded by the eavesdropper. Hence, smart design of the coding scheme either at the plant or estimator/controller side only (one-way coding) [82]–[84] or on both ends (two-way coding) [85] aim to mitigate attacks. In comparison to the homomorphic encryption techniques discussed above, these coding schemes encode the transmitted signal, which is then decoded when received by the controller. Hence, this does not involve designing a controller that acts upon the encoded signal, and often offers a solution that is low in computation.

F. Countering DoS Attacks [Treatment]

As described in Section III-D, DoS attacks refers to the malicious disruption of information flow over the communication medium. Current state-of-the-art strategies in countering DoS attacks include game-theoretic approaches [28]–[30], optimal control [26] and event-triggered control [27], [31], [32], [86], to name a few.

In the game-theoretic works of [28], [29], the interaction between the adversary and the controller is formulated as a zero-sum dynamic game. These works aim to derive an optimal strategy for the adversary to cause a DoS attack, in order to maximize the impact on control performance. The authors of [30] however, consider the state estimation problem, where the interplay between the sensor and adversary is modeled as a stochastic Bayesian game. Here, the optimal DoS attack strategy is devised to degrade estimation performance. The same objective is also studied in [26] using tools from optimal control theory, where safety specifications are also taken into consideration.

The works in [27], [31], [32], [86] provide explicit characterization of the adversary's frequency and duration of implementing the DoS attacks in order to adversely affect stabilization [27], [31], [86] or consensus [32] of the dynamical system. Deterministic guarantees are provided in [27], [32], [86] and stochastic ones are given in [31].

G. Distributed Algorithms [Detection, Treatment]

The networked nature of CPS often means that there are nodes which only have partial access to the entire system, such that efficient mitigation of adversarial attacks cannot be performed using regular centralized techniques. Hence, distributed algorithms are attractive in this regard.

A control theoretic perspective of characterizing the vulnerability of a CPS towards sensor and actuator attacks was studied in [87] for a linear consensus network. Further, distributed algorithms were also devised for attack detection and identification. On the other hand, distributed algorithms under adversarial attacks have also been developed for state estimation [25], [88], consensus [89]–[92], and optimization [93].

H. Methods Related to Robust Statistics [Detection, Treatment]

For the problem of state estimation in the presence of possibly attacked sensors, several works have employed methods inspired by robust statistics. In essence, the idea is to design filters which provide estimates that are insensitive to large fractions of faulty or attacked sensors ('outliers'), but still are accurate. For example, the least-squares method is highly sensitive to outliers, whereas a median-based filter can tolerate up to a fraction $1/2$ corrupted sensors at the expense of general estimation quality. The problem considered is very similar to that addressed in Section V-B, but the starting point here is often probabilistic, and the effect of the attack is generally not guaranteed to be completely eliminated.

In [94], the *least trimmed squares* method is applied to reduce the influence of sensor attacks on estimating the states in a power system. In [95], a more general convex-optimization-based framework for robust state estimation is proposed. A similar state estimation problem with a binary state is considered in [96], and an optimal threshold-based estimator is characterized. An extension is presented in [97], where also a fundamental trade-off between the estimation performance and tolerance to attacks is identified. In [90]–[93], local filters that discard extreme values are used to robustify distributed algorithms in the presence of adversarial nodes (see also Section V-G).

I. Rational Security Allocation [Prevention]

When facing a difficult control problem, one should always try to change the process to make the problem easy. This insight could also be applied to the CPS security problem, and the risk assessment in Section II could serve as a guide. A simple example is plants with unstable zeros, which allow for unbounded undetectable attacks on the input (see Section IV-A). The addition, or movement, of a sensor can remove such an unstable zero, however. Hence, if the risk assessment concludes that an undetectable attack against the actuator is a high-risk scenario, moving a sensor could block the scenario. This may be easier than implementing other security mitigation strategies.

Many industrial control systems are of a large scale which include hundreds of sensors and actuators, and are often originally built without cyber-security in mind. Securing these systems under a budget constraint often leads to formidable combinatorial optimization problems unless special structures can be exploited. In the literature, there are several approaches in treating the rational security allocation problem. Distinguishing features include identifying the measures needed to block high-risk scenarios, and determining the method of measurement for security.

Optimal allocation of new sensors, actuators, or leaders (in multi-agent systems) is one possible measure to increase security. Although inherently being a combinatorial optimization problem, it can sometimes be efficiently (albeit suboptimally) solved [98]–[100]. Although these works are not always motivated by security concerns, the increase of component redundancy often improves the security as measured by security indices (Section IV-B). Other security measures that can be allocated include authentication, encryption, or coding of critical channels, physical protection of especially important actuators and sensors [10], [21], [101], multirate sampling in certain sensors [58], and saturation settings on actuators [102], [103]. By securing only a few well-selected components, one can generally block large classes of attack scenarios [10]. For instance, if a few protected sensors can always be trusted, classical methods from the field of fault-tolerant control [16], such as 'virtual sensing', can be used in place of other more computationally expensive security mitigation strategies [18], [104], [105]. Game theory is another approach to optimally allocate and configure CPS defense components, see, for instance [106]–[108].

VI. Concluding Remarks

In this paper, we have provided a risk-based introduction to CPS security and privacy. The authors believe that a careful risk analysis singling out the most relevant attack scenarios is a good starting point for the design of secure and resilient CPS. The reason for this is that many security mitigation strategies are fragile to changes in the attacker or defender models. For instance, some schemes can completely eliminate attacks that affect up to half of the sensors, but cannot guarantee anything when the attacker is able to control more sensors than that. Similarly, by securing a single actuator or sensor it may be possible to completely block certain classes of attacks. See Section V-I for further examples.

We have also provided an overview of recent work in the area of CPS security and privacy. The overview is by no means complete, but has hopefully illustrated the breadth of the problems considered and the proposed solutions so far. It is interesting to note that the field has been in rapid development, and almost all references are no more than 10 years old. But indeed there are still many promising research directions to pursue. We discuss a few of them next.

Machine learning and artificial intelligence will undoubtedly have a huge impact on security and privacy in CPS. On one hand, machine learning has clear applications in the data-driven detection and diagnosis of anomalies and attacks. On the other hand, applications of machine learning to control systems are becoming increasingly relevant, but also raise privacy issues, as well as concerns about its vulnerability to data poisoning attacks. Furthermore, from a control systems perspective, such methods typically come with few formal guarantees, which poses substantial challenges to safety-critical applications. This aspect (among others) is certainly worth future investigation.

Investigating possible fundamental trade-offs between properties discussed in the paper is another interesting problem. One example is the trade-off between privacy and utility, where deliberate injection of noise increases the level of privacy at the expense of system performance. Another example is possible trade-offs between safety and security. For example, many CPS are safety critical and security solutions which are too rigid could interfere with physical safety procedures.

The areas of fault detection and fault-tolerant control are well established, and many tools developed there have inspired the work discussed in this paper. Nevertheless, there are certainly more connections to be made, for instance, with regards to the diagnosis of root-causes of faults and attacks. A key difference between faults and attacks is that the latter are caused by an intelligent agent with an incentive. In contrast, faults are generally random, which simplifies the risk analysis.

The works discussed in this paper have often used differential or difference equations for modeling the CPS. Discrete-event systems is another relevant modeling framework to be considered. In fact, these models may be more appropriate for modeling certain aspects, such as operational procedures in control centers and low-level safety functionality. These models may also be more amenable for the application of formal methods and verification.

We conclude this tutorial on security and privacy for CPS from a control theoretic perspective by noting the rich proliferation of results which provide provable guarantees through understanding the dynamics of the CPS considered. Nonetheless, its importance and the emergence of new approaches discussed above ensures continual progress in this exciting area which has captured the attention of our community.

References

[1] H. Sandberg, S. Amin, and K. H. Johansson, "Cyberphysical security in networked control systems: An introduction to the issue," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 20–23, Feb 2015.

[2] I. Koren and C. M. Krishna, *Fault-tolerant systems*. Morgan Kaufmann, 2010.

[3] M. Bishop, *Computer Security: Art and Science*. Addison-Wesley Professional, 2002.

[4] U.S. DHS, "Risk management fundamentals," U.S. Department of Homeland Security, Apr. 2011, last accessed: 10 Sep. 2014. [Online]. Available: www.dhs.gov/xlibrary/assets/rma-risk-management-fundamentals.pdf

[5] NIST, "Special publication 800-30: Guide for conducting risk assessments," National Institute of Standards and Technology, Sep. 2012, last accessed: 10 Sep. 2014. [Online]. Available: csrc.nist.gov/publications/nistpubs/800-30-rev1/sp800_30_r1.pdf

[6] S. Kaplan and B. J. Garrick, "On the quantitative definition of risk," *Risk Analysis*, vol. 1, no. 1, pp. 11–27, 1981.

[7] Y. Y. Bayraktarli, J.-P. Ulfkjaer, U. Yazgan, and M. H. Faber, "On the Application of Bayesian Probabilistic Networks for Earthquake Risk Management," *Ninth Int. Conf. Struct. Saf. Reliab. (ICOSSAR 05)*, pp. 3505–3512, 2005.

[8] Y. Cherdantseva, P. Burnap, A. Blyth, P. Eden, K. Jones, H. Soulsby, and K. Stoddart, "A review of cyber security risk assessment methods for SCADA systems," *Comput. Secur.*, vol. 56, pp. 1–27, feb 2016.

[9] S. Chockalingam, W. Pieters, A. Teixeira, and P. van Gelder, "Bayesian network models in cyber security: A systematic review," in *Secur. IT Syst. Nord. 2017. Lect. Notes Comput. Sci.*, H. Lipmaa, A. Mitrokotsa, and R. Matulevičius, Eds. Springer, Cham, 2017, pp. 105–122.

[10] J. Milosevic, 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*, 2018, accepted.

[11] (2014, Jun.) Electric sector failure scenarios and impact analyses. NESCOR, Electric Power Research Institute. Last accessed: 10 Sep. 2014. [Online]. Available: www.smartgrid.epri.com/doc/NESCOR%20failure%20scenarios%2006-30-14a.pdf

[12] T. Sommestad, M. Ekstedt, and H. Holm, "The cyber security modeling language: A tool for assessing the vulnerability of enterprise system architectures," *IEEE Systems Journal*, vol. 7, no. 3, pp. 363–373, 2013.

[13] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First Workshop on Secure Control Systems (SCS), Stockholm, 2010*, 2010.

[14] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber-physical system security for the electric power grid," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 210–224, 2012.

[15] F. Farokhi, I. Shames, and N. Batterham, "Secure and private control using semi-homomorphic encryption," *Control Engineering Practice*, vol. 67, pp. 13–20, 2017.

[16] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*, 3rd ed. Springer Publishing Company, Incorporated, 2015.

[17] Y. Mo and B. Sinopoli, "Integrity attacks on cyber-physical systems," in *Proceedings of the 1st International Conference on High Confidence Networked Systems*, ser. HiCoNS '12. New York, NY, USA: ACM, 2012, pp. 47–54.

[18] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: Risk assessment, detection, and response," in *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '11. New York, NY, USA: ACM, 2011, pp. 355–366.

[19] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *49th IEEE Conference on Decision and Control (CDC)*, Dec 2010, pp. 5967–5972.

[20] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Quantifying cyber-security for networked control systems," in *Control of Cyber-Physical Systems*, ser. Lecture Notes in Control and Information Sciences, D. C. Tarraf, Ed. Springer International Publishing, 2013, vol. 449, pp. 123–142.

[21] 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, July 2012.

[22] 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.

[23] F. Pasqualetti, F. Dörfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 110–127, Feb 2015.

[24] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. 1, pp. 135–148, 2015.

[25] M. Deghat, V. Ugrinovskii, I. Shames, and C. Langbort, "Detection and mitigation of biasing attacks on distributed estimation networks," *Automatica*, vol. 99, pp. 369–381, 2019.

[26] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *International Workshop on Hybrid Systems: Computation and Control*. Springer, 2009, pp. 31–45.

[27] 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.

[28] V. Ugrinovskii and C. Langbort, "Controller–jammer game models of denial of service in control systems operating over packet-dropping links," *Automatica*, vol. 84, pp. 128–141, 2017.

[29] A. Gupta, C. Langbort, and T. Başar, "Dynamic games with asymmetric information and resource constrained players with applications to security of cyberphysical systems," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 71–81, 2017.

[30] K. Ding, X. Ren, D. E. Quevedo, S. Dey, and L. Shi, "Dos attacks on remote state estimation with asymmetric information," *IEEE Transactions on Control of Network Systems*, 2018.

[31] A. Cetinkaya, H. Ishii, and T. Hayakawa, "Networked control under random and malicious packet losses," *IEEE Transactions on Automatic Control*, vol. 62, no. 5, pp. 2434–2449, 2017.

[32] 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.

[33] N. E. Manitara and C. N. Hadjicostis, "Privacy-preserving asymptotic average consensus," in *2013 European Control Conference (ECC)*, July 2013, pp. 760–765.

[34] Z. Huang, Y. Wang, S. Mitra, and G. E. Dullerud, "On the cost of differential privacy in distributed control systems," in *Proceedings of the 3rd international conference on High confidence networked systems*. ACM, 2014, pp. 105–114.

[35] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 341–354, 2014.

[36] E. Akyol, C. Langbort, and T. Basar, "Privacy constrained information processing," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 4511–4516.

[37] J. Cortés, G. E. Dullerud, S. Han, J. Le Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, 2016, pp. 4252–4272.

[38] Y. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, Feb 2017.

[39] S. Han, U. Topcu, and G. J. Pappas, "Differentially private distributed constrained optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 1, pp. 50–64, 2017.

[40] T. Tanaka, M. Skoglund, H. Sandberg, and K. H. Johansson, "Directed information and privacy loss in cloud-based control," in *2017 American Control Conference (ACC)*, May 2017, pp. 1666–1672.

[41] A. S. Leong, A. Redder, D. E. Quevedo, and S. Dey, "On the use of artificial noise for secure state estimation in the presence of eavesdroppers," in *2018 European Control Conference (ECC)*, June 2018, pp. 325–330.

[42] A. B. Alexandru, M. Morari, and G. J. Pappas, "Cloud-based MPC with Encrypted Data," *arXiv preprint arXiv:1803.09891*, 2018.

[43] M. S. Darup, A. Redder, I. Shames, F. Farokhi, and D. Quevedo, "Towards encrypted mpc for linear constrained systems," *IEEE Control Systems Letters*, vol. 2, no. 2, pp. 195–200, 2018.

[44] Y. Lu and M. Zhu, "Privacy preserving distributed optimization using homomorphic encryption," *Automatica*, vol. 96, pp. 314 – 325, 2018.

[45] J. Kim, C. Lee, H. Shim, J. H. Cheon, A. Kim, M. Kim, and Y. Song, "Encrypting controller using fully homomorphic encryption for security of cyber-physical systems," *IFAC-PapersOnLine*, vol. 49, no. 22, pp. 175–180, 2016.

[46] Y. Lin, F. Farokhi, I. Shames, and D. Nešić, "Secure control of nonlinear systems using semi-homomorphic encryption," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 5002–5007.

[47] F. Farokhi and H. Sandberg, "Ensuring privacy with constrained additive noise by minimizing fisher information," *Automatica*, vol. 99, pp. 275 – 288, 2019.

[48] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2009, pp. 911–918.

[49] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems*, vol. 35, no. 1, pp. 93–109, 2015.

[50] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7363–7368, 2017.

[51] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 90 – 95, 2011, 18th IFAC World Congress.

[52] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, ser. CCS '09. New York, NY, USA: ACM, 2009, pp. 21–32.

[53] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *2010 First IEEE International Conference on Smart Grid Communications*, Oct 2010, pp. 214–219.

[54] S. Amin, X. Litrico, S. Sastry, and A. M. Bayen, "Cyber security of water SCADA systems-Part I: analysis and experimentation of stealthy deception attacks," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 5, pp. 1963–1970, Sep. 2013.

[55] 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.

[56] J. Milosevic, D. Umsonst, H. Sandberg, and K. H. Johansson, "Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector," in *2018 European Control Conference (ECC)*, June 2018, pp. 331–337.

[57] G. Park, C. Lee, and H. Shim, "On stealthiness of zero-dynamics attacks against uncertain nonlinear systems: A case study with quadruple-tank process," in *Proceedings of the 23rd International Symposium on Mathematical Theory of Networks and Systems*, 2018.

[58] M. Naghnaeian, N. Hirzallah, and P. G. Voulgaris, "Dual rate control for security in cyber-physical systems," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 1415–1420.

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

[60] H. Sandberg and A. M. Teixeira, "From control system security indices to attack identifiability," in *Science of Security for Cyber-Physical Systems Workshop (SOSCYPS)*. IEEE, 2016, pp. 1–6.

[61] Z. Tang, M. Kuijper, M. S. Chong, I. Mareels, and C. Leckie, "Linear system security-detection and correction of adversarial sensor attacks in the noise-free case," *Automatica*, vol. 101, pp. 53–59, 2019.

[62] A. Teixeira, H. Sandberg, and K. H. Johansson, "Strategic stealthy attacks: The output-to-output ℓ_2-gain," in *2015 54th IEEE Conference on Decision and Control (CDC)*, Dec 2015, pp. 2582–2587.

[63] C. M. Ahmed, C. Murguia, and J. Ruths, "Model-based attack detection scheme for smart water distribution networks," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ser. ASIA CCS '17. New York, NY, USA: ACM, 2017, pp. 101–113.

[64] C. Murguia, I. Shames, J. Ruths, and D. Nesic, "Security metrics of networked control systems under sensor attacks (extended preprint)," 2018, arXiv:1809.01808.

[65] M. Barreno, A. Cárdenas, and J. D. Tygar, "Optimal ROC curve for a combination of classifiers," in *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 57–64.

[66] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1092–1105.

[67] C. Murguia and J. Ruths, "Characterization of a CUSUM model-based sensor attack detector," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 1303–1309.

[68] D. Umsonst, H. Sandberg, and A. A. Cárdenas, "Security analysis of control system anomaly detectors," in *2017 American Control Conference (ACC)*, May 2017, pp. 5500–5506.

[69] M. Chong, M. Wakaiki, and J. Hespanha, "Observability of linear systems under adversarial attacks," in *Proceedings of the 2015 American Control Conference (ACC)*. IEEE, 2015, pp. 2439–2444.

[70] 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.

[71] Y. Shoukry, M. Chong, M. Wakaiki, P. Nuzzo, A. Sangiovanni-Vincentelli, S. Seshia, J. Hespanha, and P. Tabuada, "SMT-based observer design for cyber-physical systems under sensor attacks," *ACM Transactions on Cyber-Physical Systems*, vol. 2, no. 1, p. 5, 2018.

[72] C. Lee, H. Shim, and Y. Eun, "On redundant observability: From security index to attack detection and resilient state estimation," *IEEE Transactions on Automatic Control*, 2018.

[73] A.-Y. Lu and G.-H. Yang, "Secure switched observers for cyber-physical systems under sparse sensor attacks: a set cover approach," *IEEE Transactions on Automatic Control*, 2019.

[74] L. An and G.-H. Yang, "State estimation under sparse sensor attacks: A constrained set partitioning approach," *IEEE Transactions on Automatic Control*, 2018.

[75] ——, "Secure state estimation against sparse sensor attacks with adaptive switching mechanism," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2596–2603, 2018.

[76] Y. Wang, Z. Huang, S. Mitra, and G. E. Dullerud, "Entropy-minimizing mechanism for differential privacy of discrete-time linear feedback systems," in *53rd IEEE Conference on Decision and Control*, Dec 2014, pp. 2130–2135.

[77] 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)*. IEEE, 2015, pp. 5820–5826.

[78] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2012, pp. 1806–1813.

[79] P. Griffioen, S. Weerakkody, and B. Sinopoli, "A moving target defense for securing cyber-physical systems," *arXiv preprint arXiv:1902.01423*, 2019.

[80] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.

[81] C. Murguia, F. Farokhi, and I. Shames, "Secure and private implementation of dynamic controllers using semi-homomorphic encryption," *arXiv preprint arXiv:1812.04168*, 2018.

[82] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 1, pp. 106–117, 2017.

[83] M. Wiese, T. J. Oechtering, K. H. Johansson, P. Papadimitratos, H. Sandberg, and M. Skoglund, "Secure estimation and zero-error secrecy capacity," *IEEE Transactions on Automatic Control*, 2018.

[84] A. Tsiamis, K. Gatsis, and G. J. Pappas, "An information matrix approach for state secrecy," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 2062–2067.

[85] S. Fang, K. H. Johansson, M. Skoglund, H. Sandberg, and H. Ishii, "Two-way coding in control systems under injection attacks: From attack detection to attack correction," *arXiv preprint arXiv:1901.05420*, 2019.

[86] V. Dolk, P. Tesi, C. D. Persis, and W. Heemels, "Event-triggered control systems under denial-of-service attacks," *IEEE Transactions on Control of Network Systems*, vol. 4, pp. 93 – 105, 2017.

[87] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, Jan 2012.

[88] A. Mitra, W. Abbas, and S. Sundaram, "On the impact of trusted nodes in resilient distributed state estimation of lti systems," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 4547–4552.

[89] 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.

[90] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, April 2013.

[91] S. M. Dibaji and H. Ishii, "Consensus of second-order multi-agent systems in the presence of locally bounded faults," *Systems & Control Letters*, vol. 79, pp. 23 – 29, 2015.

[92] S. M. Dibaji, H. Ishii, and R. Tempo, "Resilient randomized quantized consensus," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2508–2522, Aug 2018.

[93] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Transactions on Automatic Control*, 2018.

[94] Y. Chakhchoukh and H. Ishii, "Robust estimation for enhancing the cyber security of power state estimation," in *2015 IEEE Power Energy Society General Meeting*, July 2015, pp. 1–5.

[95] D. Han, Y. Mo, and L. Xie, "Convex optimization based state estimation against sparse integrity attacks," *IEEE Transactions on Automatic Control*, 2019, to appear.

[96] Y. Mo, J. P. Hespanha, and B. Sinopoli, "Resilient detection in the presence of integrity attacks," *IEEE Transactions on Signal Processing*, vol. 62, no. 1, pp. 31–43, Jan 2014.

[97] X. Ren, J. Yan, and Y. Mo, "Binary hypothesis testing with Byzantine sensors: Fundamental tradeoff between security and efficiency," *IEEE Transactions on Signal Processing*, vol. 66, no. 6, pp. 1454–1468, March 2018.

[98] A. Clark, L. Bushnell, and R. Poovendran, "A supermodular optimization framework for leader selection under link noise in linear multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 283–296, Feb 2014.

[99] T. H. Summers, F. L. Cortesi, and J. Lygeros, "On submodularity and controllability in complex dynamical networks," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 1, pp. 91–101, March 2016.

[100] L. S. Perelman, W. Abbas, X. Koutsoukos, and S. Amin, "Sensor placement for fault location identification in water networks: A minimum test cover approach," *Automatica*, vol. 72, pp. 166 – 176, 2016.

[101] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye, "Detecting false data injection attacks on dc state estimation," in *Preprints of the First Workshop on Secure Control Systems, CPSWEEK*, 2010.

[102] S. H. Kafash, J. Giraldo, C. Murguia, A. A. Cardenas, and J. Ruths, "Constraining attacker capabilities through actuator saturation," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 986–991.

[103] A. Cristofaro, S. Galeani, and M. L. Corradini, "A saturated dynamic input allocation policy for preventing undetectable attacks in cyber-physical systems," in *2018 European Control Conference (ECC)*, June 2018, pp. 845–850.

[104] K. Paridari, N. O'Mahony, A. El-Din Mady, R. Chabukswar, M. Boubekeur, and H. Sandberg, "A framework for attack-resilient industrial control systems: Attack detection and controller reconfiguration," *Proceedings of the IEEE*, vol. 106, no. 1, pp. 113–128, Jan 2018.

[105] A. F. M. Piedrahita, V. Gaur, J. Giraldo, A. A. Cardenas, and S. J. Rueda, "Virtual incident response functions in control systems," *Computer Networks*, vol. 135, pp. 147 – 159, 2018.

[106] T. Alpcan and T. Basar, *Network Security: A Decision and Game-Theoretic Approach*, 1st ed. New York, NY, USA: Cambridge University Press, 2010.

[107] S. Amin, G. A. Schwartz, and S. S. Sastry, "Security of interdependent and identical networked control systems," *Automatica*, vol. 49, no. 1, pp. 186 – 192, 2013.

[108] Q. Zhu and T. Basar, "Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 46–65, Feb 2015.