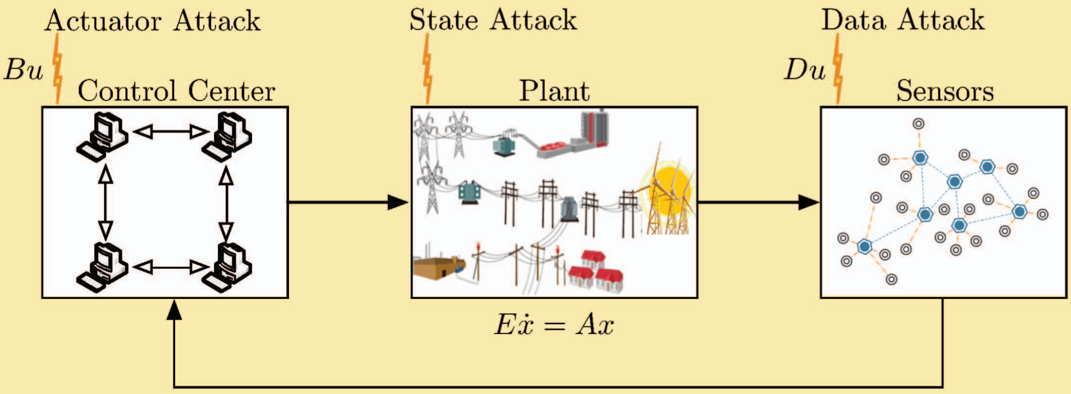
# Control-Theoretic Methods for Cyberphysical Security

## GEOMETRIC PRINCIPLES FOR OPTIMAL CROSS-LAYER RESILIENT CONTROL SYSTEMS



FABIO PASQUALETTI, FLORIAN DÖRFLER,
and FRANCESCO BULLO

*Cyberphysical systems* integrate physical processes, computational resources, and communication capabilities. Cyberphysical systems have permeated modern society, becoming prevalent in many domains, including energy production, health care, and telecommunications. Examples of cyberphysical systems include sensor networks, industrial automation systems, and critical infrastructures such as transportation networks, power generation and distribution networks, water and gas distribution networks, and advanced manufacturing systems. The integration of cybertechnologies with physical processes increases system efficiencies and, at the same time, introduces vulnerabilities that undermine the reliability of critical infrastructures. As recently highlighted by the Maroochy water breach in March 2000 [1], multiple recent power blackouts in Brazil [2], the SQL Slammer worm attack on the Davis-Besse nuclear plant in January 2003 [3], the StuxNet computer worm in June 2010 [4], and various industrial security incidents [5],

*Digital Object Identifier 10.1109/MCS.2014.2364725*
*Date of publication: 19 January 2015*

cyberphysical systems are prone to failures and attacks on their physical infrastructure and cyberattacks on their data management and communication layer [6], [7].

Concerns about the security of systems are not new, as the numerous manuscripts on systems fault detection, isolation, and recovery testify [8], [9]. The literature on fault-tolerant control considers mainly generic or accidental faults. Cyberphysical systems, however, suffer from specific vulnerabilities that do not affect classical systems, and for which appropriate detection and identification techniques need to be developed. For instance, the reliance of cyberphysical systems on communication networks and standard communication protocols to transmit measurements and control packets increases the possibility of intentional and unforeseen attacks against physical plants. On the other hand, information security methods alone can only guarantee secure communication and code execution but may be insufficient for systems comprising physical processes. In fact, security methods such as authentication, access control, and message integrity do not exploit the compatibility of the measurements and data with the underlying physical process and control architecture and are ineffective, for instance, against zero-day attacks [10] or insider attacks performed by entities with authorized access to the control platform, actuators, and sensors [1]. A holistic approach is necessary to protect cyberphysical systems, where information security mechanisms are complemented with system-theoretic monitors and security methods.

The StuxNet attack is a concrete example of cyberattack targeting a physical plant [4]. In June 2010, a carefully designed computer worm infected certain control systems of a nuclear-enrichment plant in Iran. The worm, which spread through standard USB devices, managed to corrupt the centrifuges' measurements to indicate a regular operation and, at the same time, to modify the centrifuges' actuation signals to force them to spin out of control. This cyberattack breached the implemented cyberprotection schemes, altered both the measurement and actuation signals, and caused instabilities and damage to the physical plant. The StuxNet examples illustrates the unique vulnerabilities of cyberphysical systems and motivates the need for a holistic approach combining cyber and physical protection methods to ensure cyberphysical security.

The analysis of vulnerabilities of cyberphysical systems to external attacks has received increasing attention in recent years. The general approach has been to study the effect of specific attacks against particular systems. For instance, in [11], *deception* and *denial of service* attacks against a networked control system are defined, and, for the latter ones, a countermeasure based on semidefinite programming is proposed. Deception attacks refer to the possibility of compromising the integrity of control packets or measurements and are carried out by altering the behavior of sensors and actuators. Denial of service attacks, instead, compromise the availability of resources by, for instance, jamming the communication channel. In [12], *false data* injection attacks against static state estimators are introduced. False data injection attacks are specific deception attacks in the context of static estimators. It is shown that undetectable false data injection attacks can be designed even when the attacker has limited resources. In a similar fashion, *stealthy deception attacks* against the supervisory control and data acquisition system are studied, among others, in [13]. Stealth attacks against legacy systems and possible remedial schemes are considered in [14]–[16]. In [17] and [18], the effect of *replay attacks* on a control system is discussed. Replay attacks are carried out by hijacking the sensors, recording the readings for a certain time, and repeating such readings while injecting an exogenous signal into the system. It is shown that these attacks can be detected by injecting a random signal unknown to the attacker into the system. In [19] the effect of *covert attacks* against control systems is investigated. Specifically, a parameterized decoupling structure allows a covert agent to alter the behavior of the physical plant while remaining undetected from the original controller. In [20], a resilient control problem is studied, in which control packets transmitted over a network are corrupted by a human adversary. A receding-horizon Stackelberg control law is proposed to stabilize the control system despite the attack. Recently, the problem of estimating the state of a linear system with corrupted measurements has been studied [21]. More precisely, the maximum number of tolerable faulty sensors is characterized, and a decoding algorithm is proposed to detect corrupted measurements. Finally, security issues of specific cyberphysical systems have received considerable attention, such as power networks [22]–[27], linear networks with misbehaving components [28]–[30], and water networks [31]–[33].

This article provides a self-contained presentation of recent control-theoretic approaches to cyberphysical security. The unified modeling framework for cyberphysical systems and attacks proposed in [34] is adopted, where cyberphysical systems under attack are modeled as descriptor systems subject to unknown inputs altering the state and the measurements. With respect to [34], this article provides a tutorial and self-contained presentation of the necessary background material, detailed modeling sections, and additional examples of attacks against power systems and water networks. The framework presented here is sufficiently general to include the previously described attack scenarios, yet it allows for a rigorous study of the detectability and identifiability of attacks, for a comprehensive analysis of the effects of attacks on the system, and for the design of monitors and attack-remedy schemes. The article starts with the models of cyberphysical systems, monitors, and attacks. For these models, the detectability and identifiability of attacks are defined as well as fundamental detection and identification limitations from system- and graph-theoretic perspectives. The article concludes by discussing the monitor design problem and a case study on coordinated attacks against power networks.

## Power Network Model and Attack Example

Future power networks will be equipped with a sophisticated coordination infrastructure to control the volatile physical dynamics due to renewable energy sources and deregulation of energy markets. The cyberphysical security of the future "smart grid" has been identified as an issue of primary concern [6], [26], and it has recently attracted the interest of the control and power systems communities; see [13], [22]–[27], [37], and [47].

The small-signal version of the classic structure-preserving power network model is adopted to describe the dynamics of a power network. The interested reader is referred to [37] and [47] for a detailed derivation from the full nonlinear structure-preserving power network model. Consider a connected power network consisting of $n$ generators $\{g_1, \ldots, g_n\}$ and $m$ load buses $\{b_{n+1}, \ldots, b_{n+m}\}$. The interconnection structure of the power network is encoded by a connected susceptance-weighted graph $G$. The vertices of $G$ are the generators $g_i$ and the buses $b_i$. The edges of $G$ are the transmission lines $\{b_i, b_j\}$ and the connections $\{g_i, b_i\}$ weighted by their susceptance values. The Laplacian associated with the susceptance-weighted graph is the symmetric susceptance matrix $\mathcal{L} \in \mathbb{R}^{(n+m)\times(n+m)}$ defined by

$$\mathcal{L} = \begin{bmatrix} L_{gg} & L_{gl} \\ L_{lg} & L_{ll} \end{bmatrix}, \tag{S1}$$

where generators and load buses have been labeled so that the first $n$ rows of $\mathcal{L}$ are associated with the generators and the last $m$ rows of $\mathcal{L}$ correspond to the load buses. The dynamic model of the power network is

$$\begin{bmatrix} I & 0 & 0 \\ 0 & M_g & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\delta} \\ \dot{\omega} \\ \dot{\theta} \end{bmatrix} = -\begin{bmatrix} 0 & -I & 0 \\ L_{gg} & D_g & L_{gl} \\ L_{lg} & 0 & L_{ll} \end{bmatrix} \begin{bmatrix} \delta \\ \omega \\ \theta \end{bmatrix} + \begin{bmatrix} 0 \\ P_\omega \\ P_\theta \end{bmatrix}, \tag{S2}$$

where $\delta: \mathbb{R} \to \mathbb{R}^n$ and $\omega: \mathbb{R} \to \mathbb{R}^n$ denote the generator rotor angles and frequencies, and $\theta: \mathbb{R} \to \mathbb{R}^m$ is the voltage angles at the buses. The matrices $M_g$ and $D_g$ are diagonal matrices of the generator inertial and damping coefficients, and the inputs $P_\omega: \mathbb{R} \to \mathbb{R}^n$ and $P_\theta: \mathbb{R} \to \mathbb{R}^m$ are due to *known* changes in the mechanical input power to the generators or real power demand at the loads.

Consider the power network in Figure S1 subject to a load altering attack [25] at the buses $b_4$ and $b_5$. Due to this attack, the angles $\theta_4$ and $\theta_5$ are altered by the attack signals $u_1: \mathbb{R} \to \mathbb{R}$ and $u_2: \mathbb{R} \to \mathbb{R}$, respectively. Suppose that a monitor measures directly the state variables of the first generator as $y_1 = \delta_1$ and $y_2 = \omega_1$. Let the system matrices be as in (S1) and (S2) with $M_g = \text{blkdiag}(.125, .034, .016)$, $D_g = \text{blkdiag}(.125, .068, .048)$, and in the box at the bottom of the page.

Let $U_1$ and $U_2$ be Laplace transforms of the attack signals $u_1$ and $u_2$ and let

$$\begin{bmatrix} U_1(s) \\ U_2(s) \end{bmatrix} = \underbrace{\begin{bmatrix} \dfrac{-1.024s^4 - 5.121s^3 - 10.34s^2 - 9.584s - 3.531}{s^4 + 5s^3 + 9.865s^2 + 9.173s + 3.531} \\ 1 \end{bmatrix}}_{\mathcal{N}(s)} \bar{U}(s),$$

for *some arbitrary* nonzero signal $\bar{U}(s)$. The attack signals $u_1$ and $u_2$ are carefully chosen by the attacker to avoid detection by a monitor measuring the variables $\delta_1$ and $\omega_1$. In fact, it can be verified that $\mathcal{N}$ coincides with the null space of the transfer matrix between the attack at the buses $b_4$ and $b_5$ and the measurements $\delta_1$, $\omega_1$. From the analysis in the section "Fundamental Attack Detection and Identification Limitations," the attack is undetectable by the monitor because it does not affect the measurements. Figure S2 shows the frequency of the network generators for a

$$\mathcal{L} = \begin{bmatrix} .058 & 0 & 0 & -.058 & 0 & 0 & 0 & 0 & 0 \\ 0 & .063 & 0 & 0 & -.063 & 0 & 0 & 0 & 0 \\ 0 & 0 & .059 & 0 & 0 & -.059 & 0 & 0 & 0 \\ -.058 & 0 & 0 & .235 & 0 & 0 & -.085 & -.092 & 0 \\ 0 & -.063 & 0 & 0 & .296 & 0 & -.161 & 0 & -.072 \\ 0 & 0 & -.059 & 0 & 0 & .330 & 0 & -.170 & -.101 \\ 0 & 0 & 0 & -.085 & -.161 & 0 & .246 & 0 & 0 \\ 0 & 0 & 0 & -.092 & 0 & -.170 & 0 & .262 & 0 \\ 0 & 0 & 0 & 0 & -.072 & -.101 & 0 & 0 & .173 \end{bmatrix}.$$

## MODELS OF CYBERPHYSICAL SYSTEMS, MONITORS, AND ATTACKS

Cyberphysical systems are ubiquitous in various domains, including power networks, water distribution networks, sensor networks, dynamic Leontief models of multisector economies, mixed gas-electricity networks, and large-scale industrial control systems. In this section, cyberphysical systems under attack are modeled as linear time-invariant descriptor systems subject to unknown inputs. This modeling framework is very general and includes most of the existing cyberphysical models, attacks, and faults. In fact, as shown in "Power Network Model and Attack Example" for power networks and in "Water Network Model and Attack Example" for water distribution networks, important real-world cyberphysical systems contain conserved physical quantities, leading to differential-algebraic system descriptions. Additionally, most attack and fault scenarios can be modeled by additive inputs affecting the state and the measurements; see "Stealth, Replay, Covert, and Injection Attacks."

### Model of Cyberphysical Systems and Attacks

The following linear time-invariant descriptor system is considered

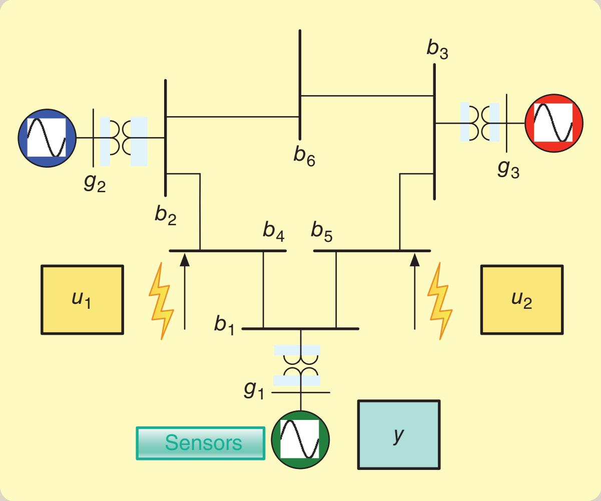


**FIGURE S1** The Western Electricity Coordinating Council (WECC) power system with three generators and six buses. The attacker modifies the power injection at the buses $b_4$ and $b_5$ via a load-altering attack. The monitor measures the rotor angle and frequency of the generator $g_1$. For some attack inputs $u_1$ and $u_2$, the attacker compromises the generators $g_2$ and $g_3$ while remaining undetected to the monitor.

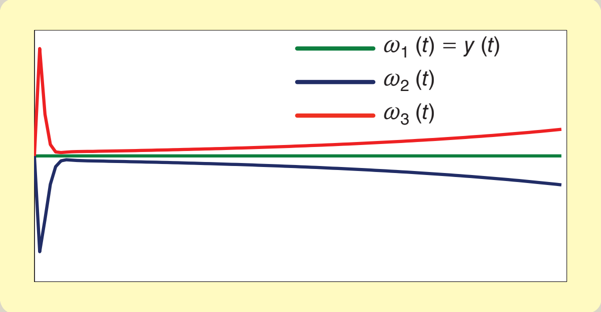


**FIGURE S2** The effect of the attack on the generators' frequencies discussed in "Power Network Model and Attack Example." Notice that generators $g_2$ and $g_3$ are driven unstable by the attack, while generator $g_1$ is not affected by the attack.

specific choice of $\bar{U}$. Notice that the second and the third generators are driven unstable by the attack input, yet the first generator does not deviate from the nominal operating condition.

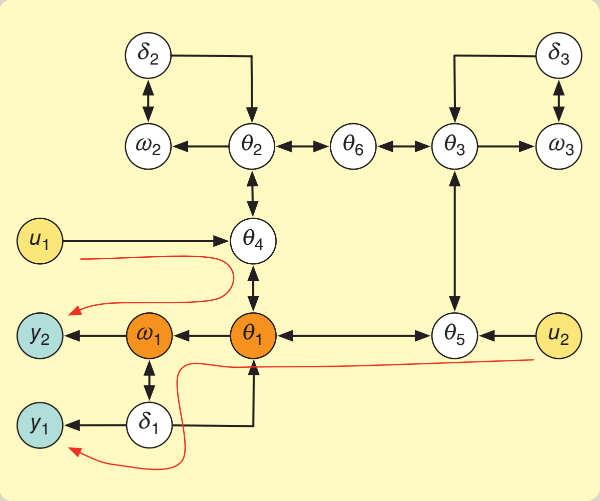


**FIGURE S3** The digraph associated with the network in Figure S1 (the self-loops of the vertices $\{\delta_1, \delta_2, \delta_3\}, \{\omega_1, \omega_2, \omega_3\}$, and $\{\theta_1, \ldots, \theta_6\}$ are not drawn). Inputs $u_1$ and $u_2$ affect the buses $b_4$ and $b_5$, respectively. The measured variables are the rotor angle and frequency of the first generator. Notice that there are no two mutually disjoint paths from the attack vertices to the output vertices, and, equivalently, there are only linkings of size at most one between the attack vertices and the output vertices. In fact, the vertices $\theta_1$ and $\omega_1$ belong to every path from $\{u_1, u_2\}$ to $\{y_1, y_2\}$. Two sample paths from the attack vertices to the output vertices are depicted in red.

In other words, if the attack signals $u_1$ and $u_2$ are regarded as additional loads, then they are entirely sustained by the second and third generators.

Graph-theoretic analysis methods are now applied to analyze the above load altering attack. The directed graph describing the network in Figure S1 is reported in Figure S3. Notice from Figure S3 that the maximum size of a linking from the attack vertices to the output vertices is one so that, by Theorem 3, the attack configuration admits generically undetectable attacks. In other words, for every choice of numerical values for the network matrices, there exist nonzero attack modes $u_1$ and $u_2$ that are not detectable through the measurements $y_1$ and $y_2$.

□

$$\begin{aligned} E\dot{x} &= Ax + Bu, \\ y &= Cx + Du, \end{aligned} \tag{1}$$

$x\colon \mathbb{R} \to \mathbb{R}^n$ and $y\colon \mathbb{R} \to \mathbb{R}^p$ are the maps describing the evolution of the system state and measurements, respectively, and $E \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$ are constant matrices. In this article, the matrix $E$ is allowed to be singular. The case of nonsingular systems $(E = I)$ is a particular instance of this model. The inputs $Bu$ and $Du$ are unknown signals that describe disturbances affecting the system state and measurements. Besides reflecting the genuine failure of systems components, these disturbances model the effect of attacks against the cyberphysical system (see below for the attack model). Finally, it should be observed that the presence of known inputs affecting system (1) is neglected because they do not affect the results on the detectability and identifiability of unknown input attacks; see [34] for a complete analysis and Figure 1 for an illustration of the setup.

For notational convenience, and without affecting generality, each state and output can be independently compromised by an attacker. Thus, the input matrices

## Water Network Model and Attack Example

Mass transport networks are cyberphysical systems modeled by differential-algebraic equations. Examples include gas transmission and distribution networks [48], large-scale process engineering plants [49], and water networks. The vulnerability of water networks to cyberphysical attacks has been shown in [31] and [19] for the case of open-channel networks [50] and in [1] and [32] for the case of municipal networks.

Following [51] and [52], water networks can be modeled as directed graphs with a vertex set consisting of reservoirs, junctions, and storage tanks, and with edge set given by pipes, pumps, and valves that are used to convey water from source points to consumers. The key variables are the pressure head $h_i$ at each network node $i$ and the flows $Q_{ij}$ from node $i$ to $j$. The hydraulic model governing the network dynamics includes constant reservoir heads, flow balance equations at junctions and tanks, and pressure difference equations along all edges:

$$\begin{aligned}
\text{reservoir } i: h_i &= h_i^{\text{reservoir}} = \text{constant},\\
\text{junction } i: d_i &= \sum\nolimits_{j\to i} Q_{ji} - \sum\nolimits_{i\to k} Q_{ik},\\
\text{tank } i: A_i\dot{h}_i &= \sum\nolimits_{j\to i} Q_{ji} - \sum\nolimits_{i\to k} Q_{ik},\\
\text{pipe } (i,j): Q_{ij} &= Q_{ij}(h_i - h_j),\\
\text{pump } (i,j): h_j - h_i &= +\Delta h_{ij}^{\text{pump}} = \text{constant},\\
\text{valve } (i,j): h_j - h_i &= -\Delta h_{ij}^{\text{valve}} = \text{constant}.
\end{aligned} \tag{S3}$$

Here, $d_i$ is the demand at junction $i$, $A_i$ is the (constant) cross-sectional area of storage tank $i$, and the notation "$j \to i$" denotes the set of nodes $j$ connected to node $i$. The flow $Q_{ij}$ depends on the pressure drop $h_i - h_j$ along pipe according to the Hazen-Williams equation $Q_{ij}(h_i - h_j)| = g_{ij} | h_i - h_j |^{1/1.85-1} \cdot (h_i - h_j)$, where $g_{ij} > 0$ is the pipe loss coefficient [51].

Consider the water supply network EPANET 3 [53] linearized at steady state with nonzero pressure drops. The topology of the water network and the attack locations are illustrated in Figure S4. For notational convenience, let $x_1, x_2, x_3$, and $x_4$ denote, respectively, the pressure at the reservoir $R_2$, at the reservoir $R_1$ and at the tanks $T_1, T_2$ and $T_3$, at the junction $P_2$, and at the remaining junctions, respectively. The descriptor model for the EPANET 3 network is

$$\begin{bmatrix} \dot{x}_1(t) \\ M\dot{x}_2(t) \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & A_{24} \\ A_{31} & 0 & A_{33} & A_{34} \\ 0 & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix},$$
$$y = [C_1 \;\; C_2 \;\; C_3 \;\; C_4]x,$$

where the pattern of zeros is due to the network interconnection structure, and $M = \operatorname{diag}(1, A_1, A_2, A_3)$ corresponds to the dynamics of the reservoir $R_1$ and the tanks $T_1, T_2$, and $T_3$.

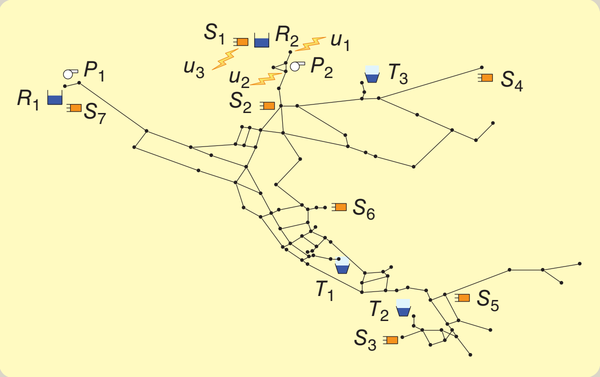


FIGURE S4 The structure of the EPANET water supply network model 3, which features three tanks $(T_1, T_2, T_3,)$, two reservoirs $(R_1, R_2)$, two pumps $(P_1, P_2)$, 96 junctions, and 119 pipes. Seven pressure sensors $(S_1, \ldots, S_7)$ have been installed to monitor the network. A cyberphysical attack to steal water from the reservoir $R_2$ is reported. Notice that the cyberphysical attack features two state attacks $(u_1, u_2)$ and one output attack $(u_3)$.

The following attack is considered where the attacker's intention is to steal water from the reservoir $R_2$. To remain undetected from the sensors measurements, the attacker simultaneously corrupts the measurements at sensor $S_1$ and modifies the pressure at pump $P_2$. Formally, the attack matrices are $B = [B_1 \; B_2 \; 0]$ and $D = [0 \; 0 \; D_1]$, with

$$B_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \text{and } D_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Let the attack input be

$$u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix},$$

where $u_1 = -x_1$ (the attacker physically subtracts water from $R_2$), $u_2 = -A_{31}x_1$, and $u_3 = -x_1$. The dynamics of the EPANET 3 network under attack are

$$\begin{bmatrix} \dot{x}_1 \\ M\dot{x}_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$
$$y = [0 \;\; C_2 \;\; C_3 \;\; C_4]x.$$

Observe that the attacker subtracts water from $R_2$ while remaining undetected from the sensors measurements. In fact, the effect of the attack does not affect the measurements $y$ because the pressure change at the reservoir $R_2$ does not appear in the measurements $y$.

Some comments are in order. First, the attack can be implemented with knowledge of the submatrix $A_{31}$ only, without knowing the whole network structure and initial state. Second, the effectiveness of the proposed attack strategy is independent of the sensors measuring the variables $x_3$ and $x_4$. On the other hand, if additional sensors are used to measure the flow between the reservoir $R_2$ and the pump $P_2$, then the attacker would need to corrupt these measurements as well to remain undetected. Third and finally, due to the reliance on networks to control actuators in cyberphysical systems, the attack $u_2$ on the pump $P_2$ could be generated by a cyberattack as for the case of power grids [25].

$$B = [I \;\; 0], \quad \text{and} \quad D = [0 \;\; I],$$

are partitioned into identity and zero matrices of appropriate dimensions and, accordingly,

$$u = \begin{bmatrix} u_x \\ u_y \end{bmatrix}.$$

The *attack* $(Bu, Du) = (u_x, u_y)$ can be classified as a *state attack* $(Bu, 0)$, affecting the system dynamics, or as an *output attack* $(0, Du)$, corrupting directly the measurements vector.

The attack signal $u: \mathbb{R} \to \mathbb{R}^{n+p}$ depends on the attack strategy. In the presence of $k \in \{1, \ldots, n+p\}$ attackers (likewise $k$ attacked variables) indexed by the attack set $K \subseteq \{1, \ldots, n+p\}$, only the entries $K$ of $u$ are nonzero over time. To underline this sparsity relation, let $u_K$ denote the *attack mode*, that is the subvector of $u$ indexed by $K$. Accordingly, the pair $(B_K, D_K)$ denotes the *attack signature*, where $B_K$ and $D_K$ are the submatrices of $B$ and $D$ with columns in $K$. Thus, $Bu = B_K u_K$, and $Du = D_K u_K$. Because the matrix $E$ can be singular, the following assumptions are made on system (1):

A1) The pair $(E, A)$ is regular, that is, the determinant $\det(sE - A)$ does not vanish identically.

A2) The initial condition $x(0) \in \mathbb{R}^n$ is consistent, that is, the relation $(Ax(0) + Bu(0)) \in \operatorname{Im}(E)$ holds.

A3) The input signal $u$ is smooth.

The regularity assumption A1) ensures the existence of a unique solution $x$ to (1). Assumptions A2) and A3) simplify the technical presentation in this article since they guarantee smoothness of the state trajectory $x$ and the measurements $y$; see [35] and [34] for further details.

### Model of Monitors

A monitor is a device to detect and identify attacks in a cyberphysical system. A general class of monitors is considered that has knowledge of the system dynamics and measurements, that is, the monitor knows the system matrices $E, A, C$, and it has access to the measurements $y$ at all times. No additional constraints are imposed on monitors.

An example of a monitor is the *bad data detector* [36]. The bad data detector takes as inputs the matrix $C$ and the measurements $y$ and detects an attack whenever there is no physical state that satisfies the measurement equation $y = Cx$. In other words, the bad data detector detects an attack whenever the residual

$$r = y - CC^{\dagger} y \tag{2}$$

is nonzero, where $C^{\dagger}$ denotes the Moore-Penrose pseudoinverse of the matrix $C$. Observe that the bad data detector detects only attacks of the form $(0, Du)$ with $Du \notin \operatorname{Im}(C)$. Other examples of monitors can be found in [13], [24], and [17] and in the section "Design of Attack Detection and Identification Monitors."

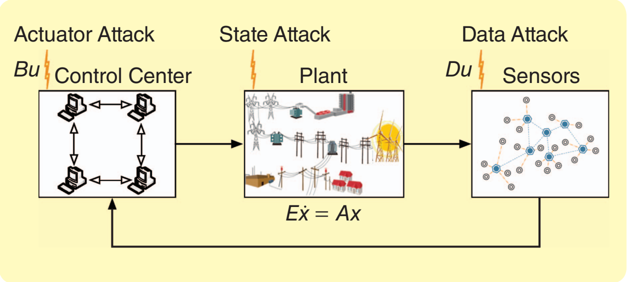


**FIGURE 1** Cyberphysical systems integrate physical and cyberlayers and are prone to attacks on all components. The dynamics of cyberphysical systems can be represented as $E\dot{x} = Ax$, and attacks as unknown inputs $(Bu, Du)$. State and actuator attacks are modeled by an input $Bu$ that directly affects the system dynamics. Output (or data) attacks corrupt the system measurements and are modeled by the input $Du$. State, actuator, and data attacks can be implemented by cyber or physical tampering with the system components. See the section "Models of Cyberphysical Systems, Monitors, and Attacks" for specific examples of cyber and physical attacks.

### Model of Attackers

This article considers colluding omniscient attackers with the ability to alter the cyberphysical dynamics through exogenous inputs. In particular, the attack $(Bu, Du)$ in (1) is designed based on knowledge of the system matrices $E, A, C$ and the full state $x$ at all times. Additionally, attackers have unlimited computation capabilities, and their objective is to disrupt the physical state or the measurements while avoiding detection.

For a power network (see "Power Network Model and Attack Example"), attacks and faults modeled by additive inputs include the following:

- » A change in the mechanical power input to generator $i$ is described by the attack signature $(B_i, 0)$ and an arbitrary attack mode $u_{n+i}$. This attack can originate from a genuine loss of generation or load, a malicious attack via the governor control to disrupt the system functionality [22], or an Internet-based load-altering attack [25].
- » A line outage occurring on the line $\{r, s\}$ is modeled by the signature $([B_r \; B_s], [0 \; 0])$ and an arbitrary attack mode $[u_r \; u_s]^{\top}$; see [37].
- » The failure of sensor $i$, or the corruption of the $i$th measurement by an attacker is captured by the signature $(0, D_{2n+m+i})$ and a nonzero mode $u_{2n+m+i}$; see [12], [13], [21], and [27] for examples of sensor attacks.

Likewise, for a water network (see "Water Network Model and Attack Example**"),** faults modeled by additive inputs include leakages, sudden changes of demand, and failures of pumps and sensors. Possible cyberphysical attacks include compromising the flow and pressure measurements to divert flow and attacks on the hydraulic control architecture (pumps and valves). These attacks are modeled similarly to the power network attacks above.

## Stealth, Replay, Covert, and Injection Attacks

Several attacks against cyberphysical systems have recently been identified and analyzed. These attacks are particular instances of the general framework introduced in the section "Models of Cyberphysical Systems, Monitors, and Attacks."

### STEALTH ATTACK [12], [23]

In a stealth attack, the attacker modifies some sensors readings by physically tampering with the individual meters or by getting access to some communication channels. Following the notation in the section "Models of Cyberphysical Systems, Monitors, and Attacks," stealth attacks are modeled by the exogenous input $(0, Du)$, with $\operatorname{Im}(D) \subseteq \operatorname{Im}(C)$ and $u$ an arbitrary signal. Notice that stealth attacks modify only the measurements equation so that the system dynamics become

$$E\dot{x} = Ax,$$
$$y = Cx + Du.$$

See Figure S5 for a block diagram representation of a stealth attack.

Stealth attacks have three important features. First, they can be formulated without knowing or tampering with the system dynamics. Second, they are undetectable by bad data detectors (see the section "Models of Cyberphysical Systems, Monitors, and Attacks"). To see this, notice that the residual of the bad data detector $r = y - CC^{\dagger}y$ is identically zero for a stealth attack, because $\operatorname{Im}(D) \subseteq \operatorname{Im}(C)$ (the measurements $y$ are compatible with the measurement matrix $C$). Third and finally, stealth attacks may be detectable by the monitor (6) because such a monitor verifies the compatibility of the measurements with the system dynamics, and not only with the measurements equation. See [24] for a discussion of the detectability of attacks via static and dynamic monitors.

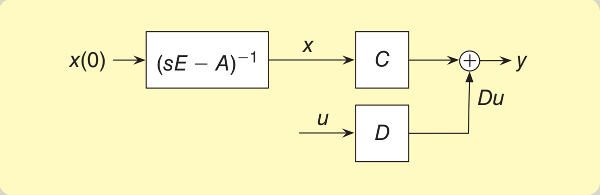


FIGURE S5 A block diagram of a stealth attack. The attacker corrupts the measurements $y$ with the signal $Du \in \operatorname{Im}(C)$. Notice that stealth attacks can be carried out by tampering with the sensors measurements and without knowing the system dynamics. In fact, stealth attacks do not alter the system dynamics.

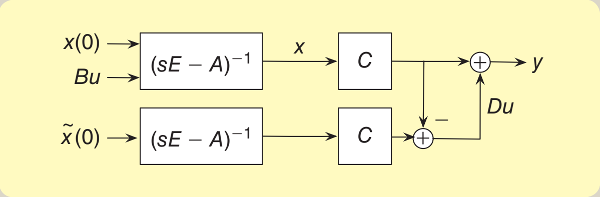


FIGURE S6 A block diagram of a replay attack. The attacker corrupts the system dynamics and the measurements. In particular, the attacker resets the measurements to reflect a prerecorded nominal operating condition $\tilde{x}(0)$ and to hide the effect of the state attack on the system dynamics. Replay attacks can be carried out with access to all sensors, and without knowing the system dynamics.

### REPLAY ATTACK [17]

In a replay attack, the attacker performs three main actions. First, the system output corresponding to a nominal operating condition is recorded. Second, the sensor measurements are modified to replicate previously recorded measurements corresponding to a nominal operating condition. Third, a control signal is injected to disrupt the system functionality. Replay attacks can be modeled by the input $(Bu, -Cx + C\tilde{x})$, where $x$ and $\tilde{x}$ are the state trajectories of the system under attack and without the attack, respectively. In other words, $x$ and $\tilde{x}$ satisfy the differential equations

$$E\dot{x} = Ax + Bu,$$
$$E\dot{\tilde{x}} = A\tilde{x},$$

## FUNDAMENTAL ATTACK DETECTION AND IDENTIFICATION LIMITATIONS

In this section, system- and graph-theoretic conditions for the detectability and identifiability of attacks are presented. These conditions are fundamental, in the sense that they hold independently of the monitoring device.

### *System-Theoretic Conditions*

As discussed in the section "Models of Cyberphysical Systems, Monitors, and Attacks," monitors exploit only the system dynamics and measurements to reveal attacks. Consequently, an attack is undetectable if the measurements due to the attack are compatible with the measurements without the attack, that is, they coincide with the measurements due to some nominal operating condition. On the other hand, if the measurements due to the attack are not compatible with the system dynamics and measurements without attacks, then the attack can be detected. The following definitions summarize this discussion, where $y(x_0, u, t)$ denotes the system measurements at time $t$ due to the attack $u$ and initial state $x_0$.

### Definition 1 (Undetectable Attack)

For the descriptor system (1) with initial state $x_0$, the attack $(B_K u_K, D_K u_K)$ is undetectable if $y(x_0, u_K, t) = y(x_1, 0, t)$ for some initial state $x_1 \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$.

This means that if the measurements are compatible with the physical process, then the effect of an undetectable attack cannot be distinguished from regular system operation. A more general concern than detectability is

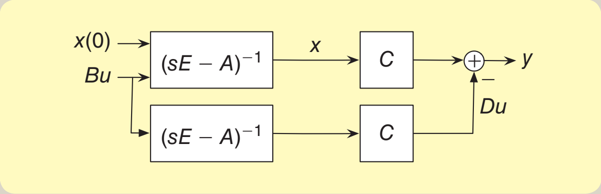


FIGURE S7 A block diagram of a covert attack. The attacker corrupts the system dynamics and the measurements. In particular, the attacker modifies the measurements to cancel out the effect of its attack on the system dynamics. Covert attacks are closed-loop replay attacks, and they require access to some sensors and knowledge of the system dynamics to be implemented.

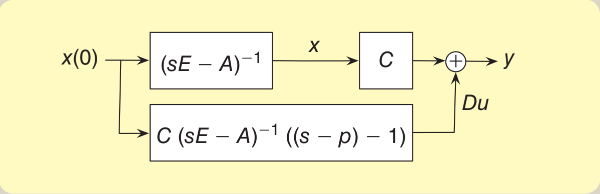


FIGURE S8 A block diagram of a dynamic false-data injection attack. The attacker corrupts the system dynamics and measurements to render the unstable mode $p$ unobservable from the measurements. Dynamic false-data injection attacks require access to some sensors and knowledge of the system dynamics to be implemented.

and $C\tilde{x}$ are the measurements corresponding to the system without attack. The system dynamics with replay attack are

$$\begin{aligned} E\dot{x} &= Ax + Bu, \\ y &= C\tilde{x}. \end{aligned}$$

See Figure S6 for a block diagram representation of a replay attack. Notice that replay attacks can be carried out without knowing the system dynamics, provided that the attacker has access to all sensors. The reader is referred to [17] for a method to reveal replay attacks.

### COVERT ATTACKS [19]

Covert attacks are closed-loop replay attacks, where the attacker modifies the system measurements to cancel out only the effect of its attack on the system dynamics. In particular, the covert attack input is $(Bu, -C\tilde{x})$, where $\tilde{x}$ is the state trajectory due to the attack input. The system dynamics with covert attacks are

$$\begin{aligned} E\dot{x} &= Ax + Bu, \\ y &= C(x - \tilde{x}), \end{aligned}$$

where $\tilde{x}$ satisfies

$$E\dot{\tilde{x}} = A\tilde{x} + Bu, \quad \text{with } \tilde{x}(0) = 0.$$

See Figure S7 for a block diagram representation of a covert attack. Notice that covert attacks require the attacker to know the exact system dynamics and to hijack some sensors (only those sensors affected by the state attack). On the other hand, covert attacks are undetectable by static and dynamic monitors [24] and by the active method proposed in [17]. In fact, covert attacks excite only the zero dynamics of the attack/measurements dynamical system, and they are therefore undetectable as discussed in the section “Fundamental Attack Detection and Identification Limitations.”

### DYNAMIC FALSE-DATA INJECTION ATTACKS [54]

Dynamic false-data injection attacks can be formulated against systems with unstable modes, and they aim to modify the system measurements to make some unstable modes unobservable. The attack input for a dynamic false-data injection attack is $(0, -C\tilde{x})$, where

$$E\dot{\tilde{x}} = A\tilde{x},$$

and $\tilde{x}(0)$ is the projection of the system state $x(0)$ along the eigenvector of an unstable mode. Dynamic false-data injection attacks are illustrated in Figure S8. As for the case of covert attacks, dynamic false-data injection attacks are undetectable as in Definition 1. □

identifiability of attacks, that is, the possibility for a monitor to distinguish between two distinct sets of attackers. Recall that attackers can independently compromise any state variable or measurement.

### Definition 2 (Unidentifiable Attack)

For the descriptor system (1) with initial state $x_0$, the attack $(B_K u_K, D_K u_K)$ is unidentifiable if $y(x_0, u_K, t) = y(x_1, u_R, t)$ for some initial state $x_1 \in \mathbb{R}^n$, attack $(B_R u_R, D_R u_R)$ with $|R| \leq |K|$ and $R \neq K$, and for all $t \in \mathbb{R}_{\geq 0}$.

Thus, an attack $K$ is not identifiable if it cannot be distinguished from another attack $R$ corrupting equally many or fewer variables $|R| \leq |K|$. Here, the attack set $K$ is compared only with other attack sets $R$ with $|R| \leq |K|$ because sufficiently large attack sets can always be designed to be unidentifiable, for instance, by corrupting sufficiently many sensors.

Following Definition 1, an attack set is undetectable if it can result in undetectable attacks. Likewise, an attack set is unidentifiable if it can result in unidentifiable attacks. Observe that, due to the linearity of (1), the detectability condition in Definition 1 can be equivalently rewritten: for the descriptor system (1) with initial state $x_0$, the attack $(B_K u_K, D_K u_K)$ is undetectable if and only if $y(x_2, u_K, t) = 0$ for some initial state $x_2 \in \mathbb{R}^n$ (namely $x_2 = x_0 - x_1$ for some $x_1 \in \mathbb{R}^n$) and for all $t \in \mathbb{R}_{\geq 0}$. The relation $y(x_2, u_K, t) = 0$ can be satisfied at all times if and only if the attack $u_K$ excites only the *zero dynamics* of the input/output dynamical system; see “Invariant Zeros and Zero Dynamics” and [38], [35], and [39]. Thanks to this interpretation and the notion of *invariant zeros*, undetectable attack sets can be algebraically characterized as follows.

### Theorem 1 (Detectability of Cyberphysical Attacks)

For the descriptor system (1) and an attack set $K$, the following statements are equivalent:

i) The attack set $K$ is undetectable.

ii) There exist $s \in \mathbb{C}$, $g \in \mathbb{C}^{|K|}$, and $x \in \mathbb{C}^n$, with $x \neq 0$, such that

$$\begin{aligned}(sE - A)x - B_K g &= 0,\\ Cx + D_K g &= 0.\end{aligned}$$

In other words, the existence of undetectable attacks for the system $(E, A, B_K, C, D_K)$ is equivalent to the existence of invariant zeros for the same attack/measurements system. On the other hand, undetectable attacks exist only if the cardinality of the attack set is sufficiently large. To see this, let $\mathrm{supp}(x)$ be the set of nonzero components of the vector $x$, and let the zero norm of $x$, that is the number of nonzero components of $x$, be denoted as $\|x\|_0 = |\mathrm{supp}(x)|$. Observe that condition ii) of Theorem 1 can be satisfied if and only if the cardinality of the attack set satisfies $|K| \geq \|(sE - A)x\|_0 + \|Cx\|_0$ for some vector $x$. The choice of the vector $x$ determines the cardinality of the attack set. This makes it, therefore, a suitable optimization variable for the design of undetectable attacks with smallest cardinality.

The ability to modify the system dynamics by state feedback may improve detectability of attacks. To see this, consider the attack $(B_K, D_K)$ and the state feedback matrix $B$ with $\mathrm{Im}(B) \not\subseteq \mathrm{Im}(B_K)$. Theorem 1 ensures that the detectability of attacks in the closed-loop system is determined by the invariant zeros of $(E, A + BF, B_K, C, D_K)$, which may differ from the invariant zeros of the open-loop system $(E, A, B_K, C, D_K)$. Thus, attacks designed to excite only the zero dynamics of the system $(E, A, B_K, C, D_K)$, and hence undetectable in the open-loop system, may excite detectable dynamics in the closed-loop system $(E, A + BF, B_K, C, D_K)$. It should be observed that i) the choice of feedback matrix $F$ is arbitrary, provided that $F$ alters the invariant zeros of the system, is unknown to the attacker, and compatible with predefined control objectives; and ii) the condition $\mathrm{Im}(B) \not\subseteq \mathrm{Im}(B_K)$ is necessary because the invariant zeros of the open-loop system $(E, A, B_K, C, D_K)$, which are exploited by the attack $(B_K, D_K)$, cannot be changed by static state feedback through the attack matrix $B_K$ [40, Chapter 3].

Analogously to the detectability condition, the identifiability condition in Definition 2 can be equivalently rewritten: for the descriptor system (1) with initial state $x_0$, the attack $(B_K u_K, D_K u_K)$ is unidentifiable if and only if $y(x_2, u_K - u_R, t) = 0$ for some initial state $x_2 \in \mathbb{R}^n$, for some attack $(B_R u_R, D_R u_R)$ with $|R| \leq |K|$ and $R \neq K$, and for all $t \in \mathbb{R}_{\geq 0}$. The following result gives an algebraic characterization of identifiability.

### Theorem 2 (Identifiability of Cyberphysical Attacks)

For the descriptor system (1) and an attack set $K$, the following statements are equivalent:

i) The attack set $K$ is unidentifiable.

## Invariant Zeros and Zero Dynamics

The concept of a zero of a dynamical system plays an important role in several control problems, and it refers to the possibility of having a nonzero state trajectory while the system output is identically zero. To be specific, consider the system $(E, A, B, C, D)$, where $E \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $D \in \mathbb{R}^{p\times m}$. Assume that

$$\mathrm{Rank}\left(\begin{bmatrix} B \\ D \end{bmatrix}\right) = m,$$

where $\mathrm{Rank}(\cdot)$ denotes the rank of a matrix. Define the Rosenbrock matrix associated with the system $(E, A, B, C, D)$ as

$$P(s) = \begin{bmatrix} sE - A & -B \\ C & D \end{bmatrix}.$$

The invariant zeros of $(E, A, B, C, D)$ are the complex values $s \in \mathbb{C}$ satisfying

$$\mathrm{Rank}(P(s)) < n + m.$$

Let $z$ be an invariant zero, and let $x_0$ and $u_0$ be such that

$$\begin{aligned}(sE - A)x_0 - Bu_0 &= 0,\\ Cx_0 + Du_0 &= 0.\end{aligned}$$

The vectors $x_0$ and $u_0$ are referred to as the *state-zero direction* and the *input-zero direction*, and they can be used to excite the system $(E, A, B, C, D)$ so that the state trajectory is nonzero while the output is identically zero. To see this, let $E = I$, and let the system initial state $x(0)$ and input $u$ be $x_0$ and $t \to e^{zt}u_0$, respectively. Notice that the state trajectory $x$ is $t \to e^{zt}x_0$ because it is the unique solution to the differential equation $\dot{x} = Ax + Bu$. Finally, observe that the system output is identically zero at all times $t \in \mathbb{R}_{\geq 0}$ because

$$y(t) = Ce^{zt}x_0 + De^{zt}u_0 = e^{zt}(Cx_0 + Du_0) = 0.$$

The state trajectory $x$ is called *zero dynamics*. Given the relationship between zero dynamics and invariant zeros, it can be shown that a system exhibits zero dynamics if and only if it features invariant zeros. Notice that the number of invariant zeros can be infinite. A system with a finite number of invariant zeros is called *left invertible*, and it satisfies $y(0, u_1, t) \neq y(0, u_2, t)$ for some times $t \in \mathbb{R}$ and for all inputs $u_1$ and $u_2$. Likewise, a system that fails to be left invertible can be characterized in the Laplace domain by a rank-deficient transfer matrix, as illustrated in the example in "Power Network Model and Attack Example."

□

## Graph Theory and Generic Properties

The *graph-theoretic* approach to the control of dynamical systems tries to express system properties based on the features of an associated graph [42], [41]. A directed graph $G = (\mathcal{V}_G, \varepsilon_G)$ consists of a set of vertices $\mathcal{V}_G$ and a set of directed edges $\varepsilon_G \subseteq \mathcal{V}_G \times \mathcal{V}_G$. An edge $(v,w) \in \varepsilon_G$ is directed from vertex $v$ to vertex $w$. A subgraph of a graph $G = (\mathcal{V}_G, \varepsilon_G)$ is a graph $H = (\mathcal{V}_H, \varepsilon_H)$ such that $\mathcal{V}_H \subseteq \mathcal{V}_G$ and $\varepsilon_H \subseteq \varepsilon_G$. A graph is undirected if $(v,w) \in \varepsilon_G$ implies that $(w,v) \in \varepsilon_G$, and in this case $\{v,w\} \in \varepsilon_G$. A path in $G$ is a subgraph $P = (\{v_1, \ldots, v_{k+1}\}, \{e_1, \ldots, e_k\})$ such that $v_i \neq v_j$ for all $i \neq j$, and $e_i = (v_i, v_{i+1})$ for each $i \in \{1, \ldots, k\}$. A set of $\ell$ mutually disjoint paths between two sets of vertices $S_1$ and $S_2$ is called a *linking* of size $\ell$ from $S_1$ to $S_2$ (see Figure S3).

For a system $(E,A,B,C,D)$, let $d$ be the number of its nonzero entries. A system $(\bar{E},\bar{A},\bar{B},\bar{C},\bar{D})$ is compatible with $(E,A,B,C,D)$ if the two systems have the same pattern of nonzero entries. By collecting the nonzero parameters into a vector, every system compatible with $(E,A,B,C,D)$ can be represented by a point in the Euclidean space $\mathbb{R}^d$. A property that can be asserted on a dynamical system is called *generic* if, informally, it holds for *almost all* compatible systems. To be more precise, a property is generic if and only if the set of compatible systems satisfying such property forms a dense subset of the parameter space. For instance, controllability, observability, and left invertibility of a dynamical system are generic properties with respect to the parameter space $\mathbb{R}^d$ [41]. The interested reader is referred to [42] and [39] for a comprehensive discussion of structured systems and generic properties.

□

ii) There exists an attack set $R$, with $|R| \leq |K|$ and $R \neq K$, $s \in \mathbb{C}$, $g_K \in \mathbb{C}^{|K|}$, $g_R \in \mathbb{C}^{|R|}$, and $x \in \mathbb{C}^n$, with $x \neq 0$, such that

$$\begin{aligned}(sE - A)x - B_K g_K - B_R g_R &= 0,\\ Cx + D_K g_K + D_R g_R &= 0.\end{aligned}$$

Condition ii) in Theorem 2 can be written by collecting the input matrices

$$\begin{aligned}(sE - A)x - [B_K \;\; B_R]\begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0,\\ Cx + [D_K \;\; D_R]\begin{bmatrix} g_K \\ g_R \end{bmatrix} &= 0. \qquad (3)\end{aligned}$$

From (3) and Theorem 1 the existence of unidentifiable attack sets of cardinality $k$ is equivalent to the existence of undetectable attack sets of cardinality $2k$, that is, to the existence of invariant zeros for the system $(E,\ A,\ B_{\bar{K}},\ C,\ D_{\bar{K}})$ with $|\bar{K}| \leq 2k$.

### Graph-Theoretic Conditions

In this section, graph-theoretic conditions for the detectability of attacks are described. The reader is referred to "Graph Theory and Generic Properties" for the background information on the graph theory and algebraic geometry used in this section.

For the system $(E,A,B,C,D)$, construct the directed *attack/state/output graph* $\mathcal{G}_{\text{aso}} = (\mathcal{V}_{\text{aso}}, \mathcal{E}_{\text{aso}})$ by defining the vertex set as

$$\mathcal{V}_{\text{aso}} = \mathcal{U}_{\text{aso}} \cup \mathcal{X}_{\text{aso}} \cup \mathcal{Y}_{\text{aso}},$$

where $\mathcal{U}_{\text{aso}} = \{u_1, \ldots, u_m\}$ is the set of attack vertices, $\mathcal{X}_{\text{aso}} = \{x_1, \ldots, x_n\}$ is the set of state vertices, and $\mathcal{X}_{\text{aso}} = \{y_1, \ldots, y_p\}$ is the set of output vertices, and the edge set is

$$\mathcal{E}_{\text{aso}} = \mathcal{E}_E \cup \mathcal{E}_A \cup \mathcal{E}_B \cup \mathcal{E}_C \cup \mathcal{E}_D,$$

where

$$\begin{aligned}&\mathcal{E}_E = \{(x_j, x_i) : E_{ij} \neq 0\},\ \mathcal{E}_A = \{(x_j, x_i) : A_{ij} \neq 0\},\\ &\mathcal{E}_B = \{(u_j, x_i) : B_{ij} \neq 0\},\ \mathcal{E}_C = \{(x_j, y_i) : C_{ij} \neq 0\},\\ &\mathcal{E}_D = \{(u_j, y_i) : D_{ij} \neq 0\}.\end{aligned}$$

Various properties of the dynamical system $(E,A,B,C,D)$ can be expressed as properties of its associated graph $\mathcal{G}_{\text{aso}}$ [41], [42].

The dynamical system $(\bar{E},\bar{A},\bar{B},\bar{C},\bar{D})$ with attack/state/output graph $\bar{\mathcal{G}}_{\text{aso}} = (\bar{\mathcal{V}}_{\text{aso}}, \bar{\mathcal{E}}_{\text{aso}})$ is *compatible* with $(E,A,B,C,D)$ if $\bar{\mathcal{G}}_{\text{aso}}$ is a subgraph of $\mathcal{G}_{\text{aso}}$ with $\bar{\mathcal{V}}_{\text{aso}} = \mathcal{V}_{\text{aso}}$ and $\bar{\mathcal{E}}_{\text{aso}} \subseteq \mathcal{E}_{\text{aso}}$. In other words, the system $(\bar{E},\bar{A},\bar{B},\bar{C},\bar{D})$ is compatible with the system $(E,A,B,C,D)$ if the matrices $\bar{E},\bar{A},\bar{B},\bar{C}$, and $\bar{D}$ can be obtained from the matrices $E,A,B,C$, and $D$ by changing only their nonzero entries. A system property is *generic* if it holds for *almost all* compatible systems. Many system properties turn out to be generic and hence robust to uncertainties in the system parameters.

Recall from Definition 1 that an attack $u$ is undetectable if $y(x_0, u, t) = y(x_1, 0, t)$ at all times $t$ for some initial states $x_0$ and $x_1$. As a particular case, if the system initial state is known, an attack $u$ is undetectable if $y(x_0, u, t) = y(x_0, 0, t)$ for some initial state $x_0$. This attack undetectability condition is equivalent to the system $(E,A,B,C,D)$ failing to be left invertible; see "Invariant Zeros and Zero Dynamics."

### Theorem 3: (Generically Undetectable Attack)

Let $\mathcal{G}_{\text{aso}}$ be the attack/state/output graph associated with the descriptor system (1) and attack set $K$. Assume that the system initial state is known and that the determinant $\det(sE - A) \neq 0$ for some values of $s \in \mathbb{C}$. The following statements are equivalent (see Figure S3 for an example of linking):

i) The attack set $K$ is generically undetectable.

ii) The graph $\mathcal{G}_{\text{aso}}$ contains no linking of size $|K|$ from $\mathcal{U}_{\text{aso}}$ to $\mathcal{Y}_{\text{aso}}$.

Theorem 3 shows that if the attack/state/output graph is sufficiently connected and the system initial state is known, then there are no undetectable attacks for almost all compatible systems, that is, for almost every choice of the numerical entries of the system matrices. Conversely, if a system admits a generically undetectable attack set, then every compatible system admits an undetectable attack set. See "Power Network Model and Attack Example" for an illustrative example of this result.

If the system initial state is unknown, then an undetectable attack $u$ is characterized by the existence of a pair of initial conditions $x_0$ and $x_1$ such that $y(x_0,u,t) = y(x_1,0,t)$ or, equivalently, by the existence of invariant zeros for the given cyberphysical system. It is now shown that, provided that a cyberphysical system is left invertible, its invariant zeros can be computed by simply looking at an associated nonsingular state space system. Let the state vector $x$ of the descriptor system (1) be partitioned as $[\xi_1^\top \ \xi_2^\top]^\top$, where $\xi_1$ corresponds to the dynamic variables. Let the network matrices $E,A,B,C$, and $D$ be partitioned accordingly, and assume that the descriptor system (1) is given in *semiexplicit* form, that is $E = \mathrm{blkdiag}(E_{11},0)$ and $E_{11}$ is nonsingular, where $\mathrm{blkdiag}(M_1,\dots,M_n)$ is the block-diagonal matrix with diagonal blocks $M_1,\dots,M_n$. In fact, many cyberphysical systems, such as power and mass-transportation networks, are readily given in semiexplicit form. In this case, the descriptor system (1) is

$$\begin{aligned} E_{11}\dot{\xi}_1 &= A_{11}\xi_1 + A_{12}\xi_2 + B_1 u, \\ 0 &= A_{21}\xi_1 + A_{22}\xi_2 + B_2 u, \\ y &= C_1\xi_1 + C_2\xi_2 + Du. \end{aligned} \tag{4}$$

Consider now the associated nonsingular state-space system that is obtained by regarding $\xi_2$ as an external input and the algebraic constraint as an output

$$\begin{aligned} \dot{\xi}_1 &= E_{11}^{-1}A_{11}\xi_1 + E_{11}^{-1}A_{12}\xi_2 + E_{11}^{-1}B_1 u, \\ \tilde{y} &= \begin{bmatrix} A_{21} \\ C_1 \end{bmatrix}\xi_1 + \begin{bmatrix} A_{22} & B_2 \\ C_2 & D \end{bmatrix}\begin{bmatrix} \xi_2 \\ u \end{bmatrix}. \end{aligned} \tag{5}$$

Under the assumption of left invertibility of system (4), the invariant zeros of systems (4) and (5) coincide. Because system (5) is nonsingular, graph-theoretic results in control can be used to investigate the presence of generically undetectable attacks in singular cyberphysical systems. For instance, from [41, Theorem 4], system (4) admits generically undetectable attacks if i) the system initial state is unknown, ii) the number of attack vertices equals the number of output vertices, iii) the system is left invertible, and iv) in the graph $\mathcal{G}_{\mathrm{aso}}$ the vertices $\mathcal{X}_{\mathrm{aso}}$ are not contained in some linking of size $|K|$ from $\mathcal{U}_{\mathrm{aso}}$ to $\mathcal{Y}_{\mathrm{aso}}$.

## DESIGN OF ATTACK DETECTION AND IDENTIFICATION MONITORS

In the previous sections, fundamental limitations and conditions characterizing attack detectability and identifiability by monitors are derived. In this section, the complementary problem of designing monitors to detect and identify attacks is addressed. Monitors can be designed in different ways, depending on the knowledge of the system dynamics, the available measurements, and the communication constraints. For the considered setup, monitors are designed by leveraging and extending fault detection and isolation techniques; see "Geometric Control Theory and Its Application to Fault Detection and Isolation" and [8]. This article focuses on the design of centralized monitors with access to all measurements $y$ and with detailed knowledge of the system matrices $(E,A,C)$. We refer to [34], [43], and [44] for extensions to distributed monitors with local knowledge of the system dynamics, access to locally available measurements, and subject to communication constraints.

The design of an attack detection monitor is first considered. The design consists of a continuous-time residual filter with the system measurements $y: \mathbb{R}_{\geq 0} \to \mathbb{R}^p$ as input and outputs the residual signal $r: \mathbb{R}_{\geq 0} \to \mathbb{R}^p$. Consider the modified Luenberger observer

$$\begin{aligned} E\dot{w} &= (A + GC)w - Gy, \\ r &= Cw - y, \end{aligned} \tag{6}$$

where the output injection matrix $G \in \mathbb{R}^{n\times p}$ is selected so that the pair $(E, A+GC)$ is regular and Hurwitz, that is, its finite spectrum $\sigma(E,A) = \{\lambda : \lambda \in \mathbb{C}, [\lambda] < \infty, \det(\lambda E - A) = 0\}$ lies in the open left-half plane. If the system initial state $x(0)$ is known and the filter (6) is initialized with $w(0) = x(0)$, then an analysis of the filter error dynamics $w - x$ yields that the residual $r$ is identically zero if and only if the attack $(B_K u_K, D_K u_K)$ is either identically zero (no attack) or undetectable. Thus, the proposed filter (6) is a *complete* monitor, that is, it detects every detectable attack. Some initial results based on system design and reconfiguration for the detection of undetectable attacks are discussed in [14]–[16].

### Theorem 4 (Complete Attack Detection Monitor)

Consider the descriptor system (1) and assume that the attack set $K$ is detectable and the initial state $x(0) \in \mathbb{R}^n$ is known. Consider the *attack detection filter* (6), where $w(0) = x(0)$ and $G \in \mathbb{R}^{n\times p}$ is such that the pair $(E, A+GC)$ is regular and Hurwitz. Then, $r(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if $u_K(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$.

Several comments are in order. First, if the initial state $x(0)$ is not available, then an arbitrary initial state $w(0) \in \mathbb{R}^n$ can be chosen and the filter (6) has an asymptotic performance: the filter error $w - x$ converges asymptotically, and the residual $r$ (in the absence of attacks) becomes zero only in the limit as time goes to infinity. Second, if the filter (6) is implemented only over a finite and nontrivial interval of time, then the residual $r$ being zero in this interval is equivalent to the attack signal $u_K$ being zero for this interval. Third, the filter (6) can be implemented using locally available information and distributed computation; see [44] for details. Fourth, the dynamics and the measurements of (1) may be affected by modeling uncertainties and noise with known statistics. In a practical implementation, the output injection matrix $G$ should be chosen to optimize the sensitivity of the residual $r$ to attacks versus the effect of noise or to optimize

## Geometric Control Theory and Its Application to Fault Detection and Isolation

The *geometric approach* to the control of dynamical systems aims to develop a set of tools and techniques based on geometric notions and operations, such as subspaces, sets, linear transformation, direct sum, and orthogonalization, to analyze and control dynamical systems. The geometric approach has been developed over the last decades, and it has found applicability in many classic control problems. The interested reader is referred to [39], [38], and [55] for a comprehensive treatment of the geometric approach to control of linear dynamical systems. Some basic concepts and applications of the geometric approach are now reviewed.

Consider the system

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \end{aligned}$$

where $A, B$, and $C$ are constant matrices of appropriate dimensions. A subspace $\mathcal{V} \subseteq \mathbb{R}^n$ is an $(A, \mathrm{Im}(B))$-controlled invariant subspace [38, Chapter 4] if

$$A\mathcal{V} \subseteq \mathcal{V} + \mathrm{Im}(B),$$

or, equivalently, if there exists a matrix $F$ such that

$$(A + BF)\mathcal{V} \subseteq \mathcal{V}.$$

The notion of a controlled invariant subspace refers to the possibility of confining the state trajectory of the system $(A, B, C)$ within a subspace. Specifically, a subspace $\mathcal{V} \subseteq \mathbb{R}^{n \times n}$ is an $(A, \mathrm{Im}(B))$ -controlled invariant if, for every initial state $x_0 \in \mathcal{V}$, there exists a control input $u$ such that the state $x \in \mathcal{V}$ at all times $t \in \mathbb{R}_{\geq 0}$. For instance, the controllability subspace $\mathrm{Im}([B\ AB \cdots A^{n-1}B])$ is an $(A, \mathrm{Im}(B))$-controlled invariant subspace. The set of controlled invariant subspaces contained in a subspace $\mathcal{E} \subseteq \mathbb{R}^{n \times n}$ admits a supremum $\mathcal{V}^*$, that is, there exists an $(A, \mathrm{Im}(B))$-controlled invariant subspace satisfying $\mathcal{V} \subseteq \mathcal{V}^* \subseteq \mathcal{E}$, for any $(A, \mathrm{Im}(B))$-controlled invariant subspace $\mathcal{V}$. If $\mathcal{E} = \mathrm{Ker}(C)$, then the subspace $\mathcal{V}^*$ contains all the state trajectories driven by the input $u$ and resulting in the output $y$ being identically zero.

Controlled invariant subspaces are dual to conditioned invariant subspaces. A subspace $\mathcal{S} \subseteq \mathbb{R}^n$ is an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace [38, Chapter 4] if

$$A(\mathcal{S} \cap \mathrm{Ker}(C)) \subseteq \mathcal{S},$$

or, equivalently, if there exists a matrix $G$ such that

$$(A + GC)\mathcal{S} \subseteq \mathcal{S}.$$

Conditioned invariant subspaces arise in the context of state estimation. Specifically, the subspace $\mathcal{S}$ is an $(A, C)$-conditioned invariant if it is possible to estimate the trajectory $x \setminus \mathcal{S}$ by processing the initial condition $x_0 \setminus \mathcal{S}$, the input $u$, and the measurements $y$ through an observer [55, Chapter 5]. For instance, the unobservability subspace $\mathrm{Ker}([C^\top\ A^\top C^\top \cdots (A^{n-1})^\top C^\top]^\top)$ is an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace. The set of conditioned invariant subspaces containing $\mathcal{E} \subseteq \mathbb{R}^{n \times n}$ admits an infimum $\mathcal{S}^*$, that is, an $(A, \mathrm{Ker}(C))$-conditioned invariant subspace satisfying $\mathcal{E} \subseteq \mathcal{S}^* \subseteq \mathcal{S}$, for any $(A, \mathrm{Ker}(C))$-conditioned invariant subspace $\mathcal{S}$. If $\mathcal{E} = \mathrm{Im}(B)$, then the subspace $\mathcal{S}^*$ defines the largest subspace of the state space that can be estimated in the presence of an unknown input signal $u$. Controlled and conditioned invariant subspaces can be extended to systems with direct feedthrough matrix, and to singular systems; see [38] and [56]. Several problems, including disturbance decoupling, noninteracting control, fault detection and isolation, and state estimation in the presence of unknown inputs, have been addressed and solved in the geometric framework.

the transient behavior of the filter. Statistical hypothesis testing techniques [9] are subsequently used to analyze the residual $r$ for sufficiently large but finite horizons. Notice that attacks hiding in the transient dynamics or aligned with the noise statistics may remain undetected.

In contrast to the attack detection problem, the attack identification problem is inherently combinatorial and computationally hard. If the cardinality of the attack set is known, the identification of the attack set $K$ requires a combinatorial procedure because, a priori, $K$ is one of the $\binom{n+p}{|K|}$ possible attack sets. The following attack identification procedure consists of designing a residual filter for a candidate attack set to determine whether the candidate set coincides with the actual attack set.

For simplicity, only the case of nonsingular systems $(E = I)$ in the absence of output attacks $D_K = 0$ is considered; see [34] for a more general treatment. Identification monitor design is similar to the design of residual generators (S4) in fault detection and isolation, and it relies on the notion of *conditioned invariant* subspaces from geometric control theory; see "Geometric Control Theory and Its Application to Fault Detection and Isolation." Define the subspace $\mathcal{S}_K$ to be the smallest $(A, \mathrm{Ker}(C))$-conditioned invariant subspace containing $\mathrm{Im}(B_K)$, where $\mathrm{Ker}(C)$ and $\mathrm{Im}(B_K)$ denote the null space of $C$ and the range space of $B_K$, respectively. Let $J_K \in \mathbb{R}^{p \times n}$ be an output injection matrix rendering this subspace invariant, that is,

$$(A + J_K C)\mathcal{S}_K \subseteq \mathcal{S}_K.$$

Consider the orthonormal matrix $T_K = [W_K\ P_K] \in \mathbb{R}^{n \times n}$, where $W_K$ is a basis of $\mathcal{S}_K$ and $P_K$ is a basis of the quotient space $\mathbb{R}^n \setminus \mathcal{S}_K$. In the coordinates $[\xi_1, \xi_2] = [W_K x, P_K x]$ and with the output injection $J_K$, system (1) is

$$\begin{aligned} \begin{bmatrix} \dot{\xi}_1 \\ \dot{\xi}_2 \end{bmatrix} &= \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & \hat{A}_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \hat{B}_K \\ 0 \end{bmatrix} u_K, \\ y(t) &= [\hat{C}_1\ \ \hat{C}_2] \begin{bmatrix} \xi_1(t) \\ \xi_2(t) \end{bmatrix}, \end{aligned} \tag{7}$$

where $\hat{A} = T_K^\top (A + J_K C) T_K, \hat{B}_K = T_K^\top B_K$, and $\hat{C} = CT_K$. Hence, the effect of the input $u_K$ is contained in the "contaminated" $\xi_1$-dynamics, and the $\xi_2$-dynamics are "secure." The measurement equation $y = \hat{C}\xi$ can be projected on the image of $\hat{C}_1$ and its orthogonal complement as

$$\begin{bmatrix} \tilde{y} \\ \bar{y} \end{bmatrix} = \begin{bmatrix} \hat{C}_1 & \hat{C}_1 \hat{C}_1^\dagger \hat{C}_2 \\ 0 & (I - \hat{C}_1 \hat{C}_1^\dagger)\hat{C}_2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \tag{8}$$

where $\bar{y} = (I - \hat{C}_1 \hat{C}_1^\dagger) y$ is the secure component of the output unaffected by $\xi_1$. Hence, a residual filter for the secure $\xi_2$-dynamics can be designed using the secure output $\bar{y}$.

### Theorem 5 (Complete Attack Identification Monitor for the Attack Set $K$)

Consider the descriptor system (1) with attack set $K$ in the coordinates (7). Assume that the attack set is identifiable and the network initial state $x(0)$ is known. Consider the *attack identification filter for the attack signature* $(B_R, D_R)$ with $|R| = |K|$

$$\begin{aligned} \dot{w} &= (\hat{A}_{22} + G(I - \hat{C}_1 \hat{C}_1^\dagger)\hat{C}_2) w - G\bar{y}, \\ r_R &= (I - \hat{C}_1 \hat{C}_1^\dagger)\hat{C}_2 w - \bar{y}, \end{aligned} \tag{9}$$

where $w(0) = \xi_2(0)$ and $G$ is such that $\hat{A}_{22} + G(I - \hat{C}_1 \hat{C}_1^\dagger)\hat{C}_2$ is Hurwitz and $\bar{y}$ is the secure output defined in (8). The residual satisfies $r_R(t) = 0$ at all times $t \in \mathbb{R}_{\geq 0}$ if and only if $R$ coincides with the attack set, that is, if and only if $R = K$.

Theorem 5 implies that the attack set $K$ can be identified by constructing $\binom{n+p}{|K|}$ residual filters (9), one for each distinct attack set of cardinality $|K|$; see "An Example of Monitor Design." In [34] it is shown that this nonpolynomial complexity is inherent to the attack identification problem, which is generally NP-hard. As a remark, for output attacks, an efficient (yet incomplete) approach is to reformulate the attack identification problem as a convex optimization problem using heuristic convex relaxations [21]; see "An Example of Monitor Design."

## An Example of Monitor Design

Consider the undirected consensus network $G$ in Figure S9. Notice that $G$ has connectivity three, because there exist three vertex-disjoint paths between any two vertices. Let the network evolve according to the continuous-time nonsingular system

$$\dot{x} = Ax,$$

where $x: \mathbb{R}_{\geq 0} \to \mathbb{R}^8$ contains the agents states and $A$ is the Laplacian matrix of $G$ [57]

$$A = \begin{bmatrix} -3 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & -3 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -3 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & -3 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & -3 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & -3 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & -3 \end{bmatrix}.$$

Assume that each node measures its own and its neighboring states. In particular, let the attack and measurements sets be $\{7\}$ and $\{1,2,4\}$, respectively, and define

$$B_K^T = [0\,0\,0\,0\,0\,0\,1\,0], \; C = \begin{bmatrix} 1\,0\,0\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0\,0 \end{bmatrix}.$$

From the analysis in the section "Fundamental Attack Detection and Identification Limitations" and [34], [29], and [28], every set of at most two attackers is detectable (three attackers are detectable if the system initial state is known), and any set of at most one attacker is identifiable. To see this, i) construct the attack/state/output graph associated with $G$ as in Figure S10, and ii) notice that there exists a linking of size 2 from any set of two nodes to the output vertices (see Figure S10).

A procedure to identify the attacker from measurements follows from the analysis in the section "Design of Attack Detection and Identification Monitors." Assuming state attacks only, design eight residual generators, where the $i$th residual generator is made insensitive to any input entering at node $i$ and sensitive to all other inputs. In other words, due to the identifiability of the attack set, the output of the $i$th residual generator is identically zero if and only if the attacker compromises the $i$th node.

**RESIDUAL FILTER FOR NODE *i***

1) Compute $\mathcal{S}_i$, the smallest $(A, \mathrm{Ker}(C))$-conditioned invariant subspace containing $\mathrm{Im}(B_i)$, and its injection matrix $J_i$, and define the *conditioned filter*

$$\begin{aligned} \dot{w}_i &= (A + J_i C) w - J_i y, \\ z_i &= C w_i. \end{aligned}$$

2) Define the orthonormal change of coordinates $T_i = [W_i \; P_i]$, where $\mathrm{Im}(W_i) = \mathcal{S}_i$ and the coordinates $w = T\xi$. In the new coordinates the conditioned filter is

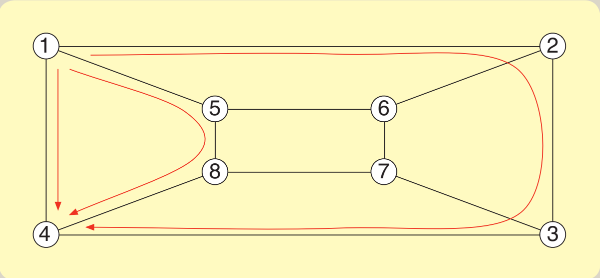


FIGURE S9 An undirected consensus network with eight nodes. The network has connectivity three, because there exist three vertex-disjoint paths between any two nodes. Three vertex-disjoint paths between nodes 1 and 4 are shown in red.

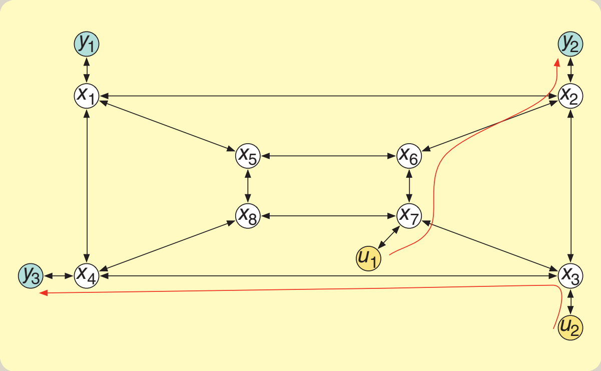


**FIGURE S10** The attack/state/output graph associated with the consensus network in Figure S9, with attack set $\{3,7\}$ and measurements $\{1,2,4\}$ (see the section "Graph-Theoretic Conditions"). Because there exists a linking of size two from any two attack nodes to the output vertices, every set of two attackers is detectable, and every set of one attacker is identifiable (see the section "Fundamental Attack Detection and Identification Limitations" and [34], [29], [28]).

$$
\begin{aligned}
\dot{\xi}_i &= \underbrace{T_i^T(A+J_iC)T_i}_{\hat{A}_i}\xi_i - \underbrace{T_i^TJ}_{\hat{B}_i}y,\\
z_i &= \underbrace{CT_i}_{\hat{C}_i}\xi_i.
\end{aligned}
$$

3) Partition $\hat{C}_i$ as $\hat{C}_i = [\hat{C}_i^1 \ \hat{C}_i^2]$, where $\hat{C}_i^1$ has as many columns as the dimension of $\mathcal{S}_i$. Let $\bar{C}_i = (I - \hat{C}_i^1\hat{C}_i^{1\dagger})$ and $\tilde{C}_i = \bar{C}_i\hat{C}_i$.
4) Define the residual filter as

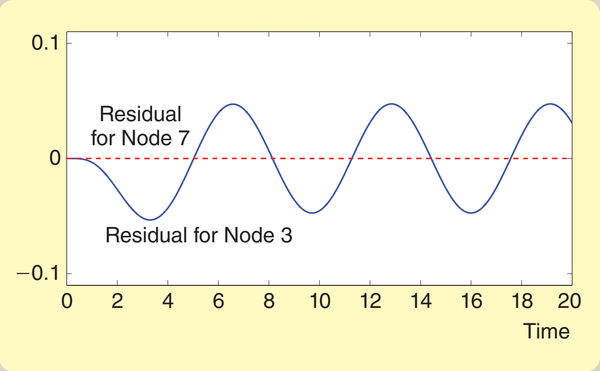


**FIGURE S11** The output of the residual filters designed in "An Example of Monitor Design." Because the only zero residual is associated with node 7, the attacker has compromised node 7 in the network in Figure S10.

$$
\begin{aligned}
\dot{\xi}_i &= \underbrace{(\hat{A}_i + G_i\tilde{C}_i)}_{F_i}\xi_i - \hat{B}_iy - G_i\tilde{C}_iy,\\
\hat{z}_i &= \tilde{C}_i\xi_i - \bar{C}_iy,
\end{aligned}
$$

where $G_i$ is such that $\hat{A}_i + G_i\tilde{C}_i$ is stable.

By leveraging the geometric routines developed in [38], the residual filters for node 3 and node 7 are computed as in the equations shown at the bottom of the page.

Figure S11 shows the residuals computed by the above filters when the attack input is a sinusoidal wave. Notice that the residual associated with node 7 is identically zero, while the residual associated with node 3 is nonzero (as well as the residuals associated with other network nodes). Thus, the attacker is detected and identified by the set of residual filters.

$$
F_3 = \begin{bmatrix}
-3 & 1.7321 & -1.1547 & 0 & 0 & 0 & -1.1547 & 1.1547\\
1.7321 & -3 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & -1.5 & 0 & 0 & -1 & 0 & 0\\
0 & 0 & 0.2041 & -1.8750 & 0.6495 & 0 & 0.4082 & 0.8165\\
0 & 0 & -0.3536 & 0.6495 & -2.6250 & 0 & 0.7071 & 0\\
0 & 0 & -0.5 & 0 & 0 & -3 & -1 & 1\\
0 & 0 & 0 & -0.3062 & 1.2374 & -1 & -3 & 0\\
0 & 0 & 0 & 0.9186 & -0.8839 & 1 & 0 & -3
\end{bmatrix},\quad
B_3 = \begin{bmatrix}
0 & 0 & 0\\
0 & 0 & 0\\
1.5 & -1 & -1\\
-0.2041 & 0.9186 & -0.9186\\
0.3536 & 0.5303 & -0.5303\\
0.5 & 0 & 0\\
0 & -1.2500 & -0.7500\\
0 & 0.7500 & 1.2500
\end{bmatrix},
$$

$$
\tilde{C}_3 = \begin{bmatrix}
0 & 0 & -1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & -0.6124 & -0.3536 & 0 & 0 & 0\\
0 & 0 & 0 & 0.6124 & 0.3536 & 0 & 0 & 0
\end{bmatrix},\quad
\bar{C}_3 = \begin{bmatrix}
-1 & 0 & 0\\
0 & -0.5 & 0.5\\
0 & 0.5 & -0.5
\end{bmatrix},
$$

and

$$
F_7 = \begin{bmatrix}
-3 & 2 & 0 & 1.7321 & 0 & 0 & 0 & 0\\
2 & -3 & -1.7321 & 0 & 0 & 0 & 0 & 0\\
0 & -1.7321 & -3 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & -1.5 & 0.6124 & 0 & 1.0607 & 0\\
0 & 0 & 0 & 0 & -1.8750 & 0.5 & -0.6495 & 0.8660\\
0 & 0 & 0 & 0 & 0.5 & -3 & 0.8660 & 0\\
0 & 0 & 0 & 0 & -0.6495 & 0.8660 & -2.6250 & -0.5\\
0 & 0 & 0 & 0 & 0.4330 & 0 & -0.2500 & -3
\end{bmatrix},\quad
B_7 = \begin{bmatrix}
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
-1.5 & 1.5 & 1.5\\
0 & 0.9186 & -0.9186\\
0 & 0 & 0\\
0 & -0.5303 & 0.5303\\
0 & -0.3536 & 0.3536
\end{bmatrix},
$$

$$
\tilde{C}_7 = \begin{bmatrix}
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & -0.6124 & 0 & 0.3536 & 0\\
0 & 0 & 0 & 0 & 0.6124 & 0 & -0.3536 & 0
\end{bmatrix},\quad
\bar{C}_7 = \begin{bmatrix}
-1 & 0 & 0\\
0 & -0.5 & 0.5\\
0 & 0.5 & -0.5
\end{bmatrix}.
$$

## COORDINATED ATTACKS IN POWER NETWORKS

This section considers a network of utility companies that compete in the production of electrical energy. In particular, the case is considered where a group of utility companies form a coalition to compromise the functionality of their business rivals through a coordinated and destabilizing attack. A similar power network scenario is studied in [22].

Consider a connected power transmission network with $n$ generators $G_m = \{g_1, \dots, g_n\}$, where the generators' rotor dynamics are modeled by second-order linear swing equations subject to governor control, and the power flows along lines are modeled by the dc approximation; see "Power Network Model and Attack Example." Assume that a subset $K = \{k_1, \dots, k_m\}$ of generators is driven by an additional control action besides the primary frequency control. After elimination of the load bus variables through Kron reduction, the power network dynamics subject to the additional control $u$ at the generators $K$ are

$$\dot{x} = Ax + B_K u_K, \tag{10}$$

where $x = [\theta^\top, \omega^\top]^\top$ contains the generators' rotor angles and frequencies, $A \in \mathbb{R}^{2n \times 2n}$, and $B_K = I_K \in \mathbb{R}^{2n \times m}$, where $I_K = [e_{n+k_1} \cdots e_{n+k_m}]$ and $e_i$ is the $i$th canonical vector in $\mathbb{R}^{2n}$. Consider an attack where the $K$ generators form a coalition, select some sacrificial machines $\bar{K} \subseteq K$, and implement a coordinated control strategy (see below) to destabilize the other generators $G_m \setminus K$, while maintaining satisfactory performance within the group $K \setminus \bar{K}$.

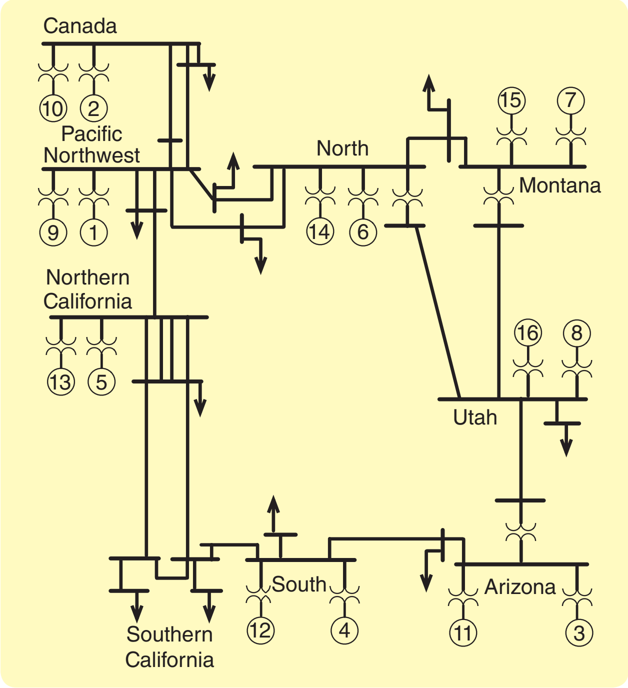


**FIGURE 2** A schematic diagram of the western North American power grid. The grid contains 16 generators, numbered from 1 to 16, that are connected to the grid through transformers. Loads are denoted by arrows and buses by straight lines.

The attack strategy relies on the notion of *controlled invariant* subspace from geometric control theory; see "Geometric Control Theory and Its Application to Fault Detection and Isolation." In particular, the colluding generators inject an attack input that remains undetectable by the generators $K \setminus \bar{K}$, while affecting the generators $G_m \setminus K$. The attack input is of the form

$$u_K = Fx + \bar{B}_K^\dagger v, \tag{11}$$

where the matrix $F$ and $\bar{B}_K$ satisfy the conditions

$$(A + B_K F)\mathcal{V}^* \subseteq \mathcal{V}^* \tag{12}$$

and

$$\bar{B}_K = \text{Basis}(\mathcal{V}^* \cap \text{Im}(B_K)). \tag{13}$$

In (12), $\mathcal{V}^*$ denotes the largest $(A, \text{Im}(B_K))$-controlled invariant subpace contained in $\text{Ker}(C)$, where $C$ is the vector of the frequencies of the generators $K \setminus \bar{K}$. Notice that the subspace $\text{Im}(C)$ identifies the generators $K \setminus \bar{K}$, while $\text{Ker}(C)$ identifies the generators $G_m \setminus K$ and the sacrificial machines $\bar{K}$.

The attack input (11) consists of two components. The open-loop component $\bar{B}_K^\dagger v$ alters the behavior of the sacrificial machines only. In fact, $\text{Im}(\bar{B}_K) \subseteq \mathcal{V}^* \subseteq \text{Ker}(C)$. The input $v: \mathbb{R} \to \mathbb{R}^n$ is an arbitrary signal designed by the attackers to optimize some performance function, such as the effect of the malicious control on the sacrificial machines, the energy of the malicious control, or the information pattern required to implement the malicious control. The closed-loop component $F$ ensures that the generators $K \setminus \bar{K}$ are not affected by network dynamics evolving in the subspace $\mathcal{V}^*$. In fact, dynamics in the subspace $\mathcal{V}^*$ are invariant due to (12) and do not affect the generators $K \setminus \bar{K}$ because $\mathcal{V}^* \subseteq \text{Ker}\, C$. Because the open-loop component of the attack excites only dynamics in $\mathcal{V}^*$ due to (13), the attack (11) does not affect the generators $K \setminus \bar{K}$, while altering the behavior of the sacrificial machines and, consequently, of the generators $G_m \setminus K$. Notice that the attack (11) is undetectable from the measurements taken at the generators $K \setminus \bar{K}$. Let $I_{K \setminus \bar{K}}$ be the matrix obtained by selecting the columns $K \setminus \bar{K}$ from the identity matrix $I$.

### Theorem 6 (Malicious Attacks)

Consider the network-reduced power system model (10) with controlled generators $K$ and sacrificial machines $\bar{K} \subseteq K$. Let $\bar{C} = I_{K \setminus \bar{K}}^\top$, let $\mathcal{V}^*$ be the largest $(A, \text{Im}(B_K))$-controlled invariant subspace contained in $\text{Ker}(\bar{C})$, let the state feedback $F$ satisfy $(A + B_K F)\mathcal{V}^* \subseteq \mathcal{V}^*$, let $\bar{B}_K = \text{Basis}(\mathcal{V}^* \cap \text{Im}(B_K))$, and let $\mathcal{S}^*$ be the smallest $(A, \text{Ker}(\bar{C}))$-conditioned invariant subspace containing $\text{Im}(B_K)$. Then, for every input $v: \mathbb{R}_{\geq 0} \to \mathbb{R}^n$, the attack $u = Fx + \bar{B}_K^\dagger v$ affects the generators $\bar{K} \cup G_m \setminus K$ only.

The attack (11) is very general and, in fact, includes all attacks that can be formulated by the generators $K$ without affecting the generators $K \setminus \bar{K}$, including the strategy proposed in [22]. To illustrate the effectiveness of the attack (11), consider an aggregated model of the Western North American power grid as illustrated in Figure 2. This model is often studied in the context of interarea oscillations [45]. Assume that the generators {1,9} form a coalition and that generator 9 is the sacrificial machine. Following Theorem 6, the malicious attack illustrated in Figure 3, which is of the form $u = Fx + \bar{B}_K^{\dagger} v$, is carried out by the generators {1,9} such that i) generator 1 is not affected by the attack, ii) generator 2 maintains an acceptable working condition even in the presence of the attack, and iii) large frequency oscillations are induced at all other generators $G_m \backslash K$. See Figure 4 for a pictorial representation of the effect of the coordinated attack. As a consequence of the attack, the linear model (10) is driven far away from the operating point, and the corresponding original nonlinear model eventually may lose stability. In a real-world scenario, instability and equipment damage would be prevented by promptly disconnecting the generators $G_m \backslash K$ from the grid [46]. It is worth noting that the methods derived in this article for linearized dynamics are, in fact, robust to model uncertainties and nonlinearities. A related example is presented in [34, Section V.D], where the performance of the presented detection and identification methods is validated for noisy and nonlinear dynamics.

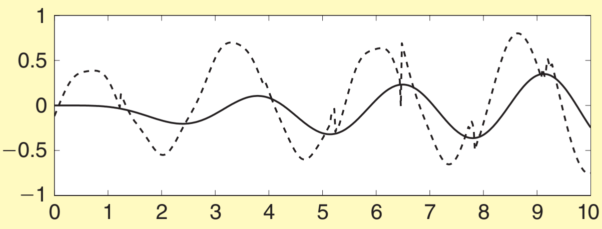


FIGURE 3 The coordinated attack input discussed in the section "Coordinated Attacks in Power Networks." The attack is implemented by modifying the governor control input of generator 1 (solid) and generator 9 (dashed). Both signals are represented as deviations from steady state and normalized by the base power for the linear system (10). All signals are plotted as a function of time measured in seconds.

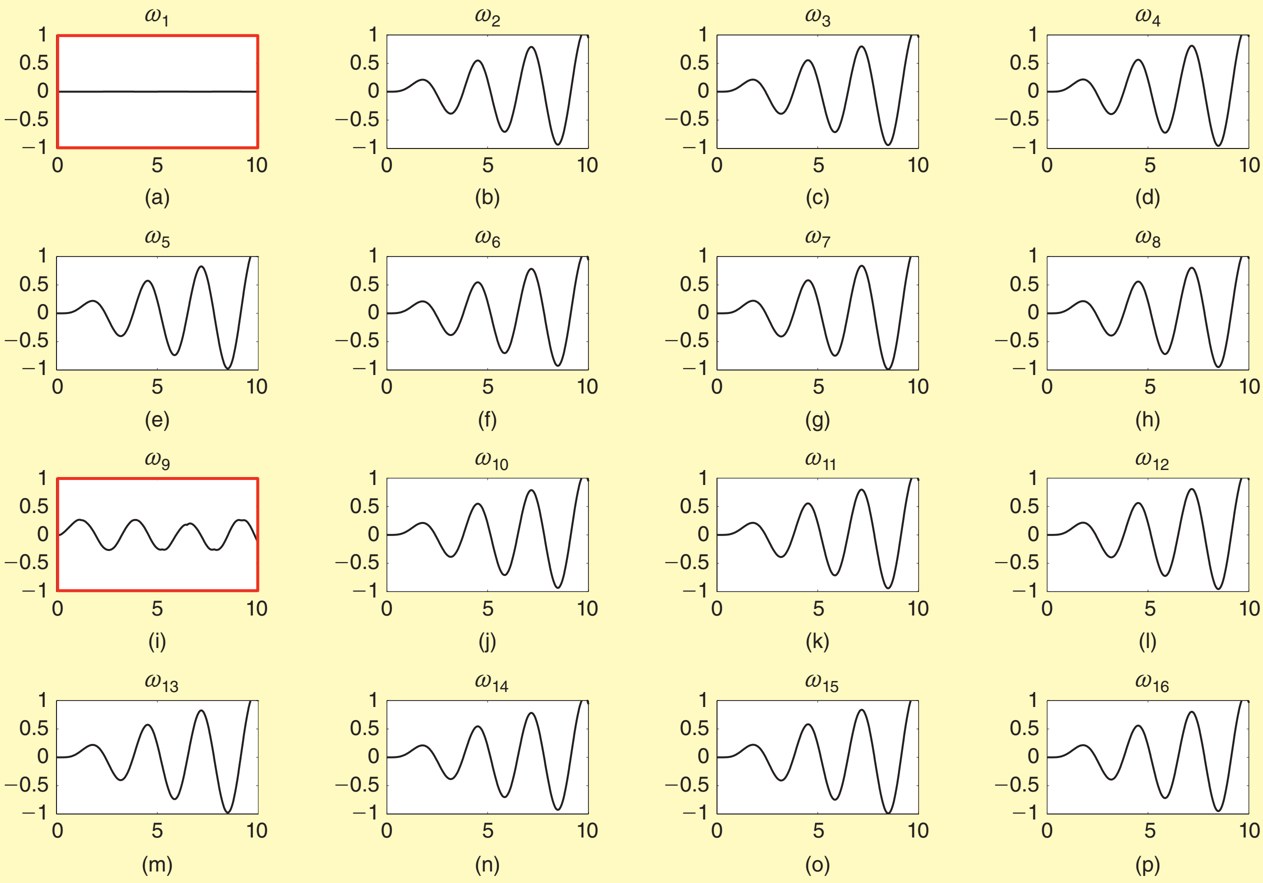


FIGURE 4 The deviations from steady state of the generators' frequencies due to the coordinated attack in the section "Coordinated Attacks in Power Networks." The system model is given by (10), with parameters taken from [45]; see also the section "Coordinated Attacks in Power Networks." All deviations have been normalized so that the unit value indicates a safety limit. All signals are plotted as a function of time (measured in seconds). The attack input is of the form (11), where the input $v$ is chosen such that the infinity norm of $\omega_9$ is minimized, subject to the infinity norm of $\omega_{16}$ being no less than one. The attack input is reported in Figure 3. Notice that i) generator 1 is not affected by the attack, ii) generator 9 maintains satisfactory performance, and iii) the remaining generators are severely affected by the coordinated attack.

In the above scenario, assume that each generator monitors its own state variables and that at most two generators may be colluding to disrupt the network. Notice that detectability of the malicious attacks designed in Theorem 6 is guaranteed for each generator affected by the attack. Unfortunately, the colluding generators cannot be identified from the measurements of any single generator. To see this, let $B_K$ be the input matrix associated with any set $K$ of two generators, and let $C_i = e_i^\top$ be the output matrix associated with generator $i$. It can be verified that for every $K$ and $i$ the system $(A, B_K, C_i)$ is right-invertible [38], that is, the output $C_i x$ can be arbitrarily assigned by any coalition of two generators. Thus, the measurements taken by generator $i$ can be generated by any set of two generators so that the colluding generators are not identifiable by generator $i$.

## CONCLUSION

Cyberphysical systems are complex systems integrating physical processes with cyber infrastructures. For security assessment, cyberphysical systems can be conveniently modeled by linear time-invariant descriptor systems, where the algebraic constraints capture the presence of conserved physical quantities in the system. For cyberphysical systems modeled by descriptor systems, attacks can be represented by exogenous inputs that alter the system dynamics and the measurements. With this representation of attacks, it is possible to i) characterize fundamental attack detection and identification limits, ii) analyze the effect of attacks on the system, and iii) design monitors capable of revealing and locating attacks independently of the attack strategy and implementation. This article presented a self-contained discussion of cyberphysical security, including modeling, system-theoretic and graph-theoretic security analyses, monitor design, and illustrative examples.

## ACKNOWLEDGMENT

This work was supported in part by NSF award ECCS 1405330.

## AUTHOR INFORMATION

***Fabio Pasqualetti*** (fabiopas@engr.ucr.edu) is an assistant professor in the Department of Mechanical Engineering at the University of California, Riverside. He received a Ph.D. degree in mechanical engineering from the University of California, Santa Barbara, in 2012, a Laurea Magistrale degree "summa cum laude" (M.Sc. equivalent) in automation engineering from the University of Pisa, Italy, in 2007, and a Laurea degree "summa cum laude" (B.Sc. equivalent) in computer engineering from the University of Pisa, Italy, in 2004. His main research interest is in secure control systems, with application to multiagent networks, distributed computing, and power networks. Other interests include vehicle routing and combinatorial optimization, with application to distributed patrolling and camera surveillance. He is a Member of the IEEE. He can be contacted at Bourns Hall A309, Department of Mechanical Engineering, University of California, Riverside, 900 University Avenue, Riverside, CA 92521 USA.

***Florian Dörfler*** is an assistant professor at the Automatic Control Laboratory at the Swiss Federal Institute of Technology (ETH) Zürich. He received his Ph.D. degree in mechanical engineering from the University of California, Santa Barbara, in 2013, and a Diplom degree in engineering cybernetics from the University of Stuttgart in 2008. From 2013 to 2014 he was an assistant professor at the University of California, Los Angeles. His primary research interests are centered around distributed control, complex networks, and cyberphysical systems with applications to smart power grids, robotic coordination, and social networks. He is a recipient of the 2009 Regents Special International Fellowship, the 2011 Peter J. Frenkel Foundation Fellowship, the 2010 ACC Student Best Paper Award, the 2011 O. Hugo Schuck Best Paper Award, and the 2012–2014 Automatica Best Paper Award. As a coadvisor and a coauthor, he has been a finalist for the ECC 2013 Best Student Paper Award. He is a Member of the IEEE.

***Francesco Bullo*** is a professor in the Mechanical Engineering Department at the University of California, Santa Barbara. He received the Laurea degree ″summa cum laude″ in electrical engineering from the University of Padova, Italy, in 1994, and the Ph.D. degree in control and dynamical systems from the California Institute of Technology in 1999. From 1998 to 2004, he was an assistant professor with the Coordinated Science Laboratory at the University of Illinois at Urbana-Champaign. His main research interest is multiagent networks with application to robotic coordination, distributed computing, and power networks. Other interests include vehicle routing, geometric control, and motion planning problems. He has published more than 200 papers in international journals, books, and refereed conferences. He is the coauthor, with Andrew D. Lewis, of *Geometric Control of Mechanical Systems* (Springer, 2004) and, with Jorge Cortés and Sonia Martínez, of *Distributed Control of Robotic Networks* (Princeton, 2009). His students' papers were finalists for the Best Student Paper Award at the IEEE Conference on Decision and Control (2002, 2005, 2007) and the American Control Conference (2005, 2006, 2010). He is an IEEE Fellow and has served on the editorial boards of *IEEE Transactions on Automatic Control*, *ESAIM: Control, Optimization, and the Calculus of Variations*, and *SIAM Journal of Control and Optimization*.

## REFERENCES

[1] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," in *Proc. Critical Infrastructure Protection*, 2007, vol. 253, pp. 73–82.

[2] J. P. Conti, "The day the samba stopped," *Eng. Technol.*, vol. 5, no. 4, pp. 46–47, Mar. 06–26, 2010.

[3] S. Kuvshinkova, "SQL Slammer worm lessons learned for consideration by the electricity sector," North Amer. Elec. Reliab. Council, Atlanta, GA, Tech. Rep., 2003.

[4] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[5] G. Richards, "Hackers vs slackers," *Eng. Technol.*, vol. 3, no. 19, pp. 40–43, 2008.

[6] A. R. Metke and R. L. Ekl, "Security technology for smart grid networks," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 99–107, 2010.

[7] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Research challenges for the security of control systems," in *Proc. 3rd Conf. Hot Topics Security*, Berkeley, CA, 2008, pp. 6:1–6:6.

[8] M.-A. Massoumnia, G. C. Verghese, and A. S. Willsky, "Failure detection and identification," *IEEE Trans. Autom. Contr.*, vol. 34, no. 3, pp. 316–321, 1989.

[9] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] R. Axelrod and R. Iliev, "Timing of cyber conflict," *Proc. Natl. Acad. Sci.*, vol. 111, no. 4, pp. 1298–1303, 2014.

[11] S. Amin, A. Cárdenas, and S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Proc. Hybrid Systems: Computation Control*, Apr. 2009, vol. 5469, pp. 31–45.

[12] Y. Liu, M. K. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *Proc. ACM Conf. Computer Communications Security*, Chicago, IL, Nov. 2009, pp. 21–32.

[13] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *Proc. IEEE Conf. Decision Control*, Atlanta, GA, Dec. 2010, pp. 5991–5998.

[14] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *Proc. Allerton Conf. Communications, Control Computing*, Oct. 2012, pp. 1806–1813.

[15] S. D. Bopardikar and A. Speranzon, "On analysis and design of stealth-resilient control systems," in *Proc. Int. Symp. Resilient Control Systems*, San Francisco, CA, Aug. 2013, pp. 48–53.

[16] J. Y. Keller and D. Sauter, "Monitoring of stealthy attack in networked control systems," in *Proc. Conf. Control Fault-Tolerant Systems*, Nice, France, Oct. 2013, pp. 462–467.

[17] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. Allerton Conf. Communications, Control Computing*, Monticello, IL, Sept. 2010, pp. 911–918.

[18] Y. Mo, T.-H. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber-physical security of a smart grid infrastructure," *Proc. IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[19] R. Smith, "A decoupled feedback structure for covertly appropriating network control systems," in *Proc. Int. Federation Automatic Control World Congr.*, Milan, Italy, Aug. 2011, pp. 90–95.

[20] M. Zhu and S. Martínez, "Stackelberg-game analysis of correlated attacks in cyber-physical systems," in *Proc. American Control Conf.*, San Francisco, CA, July 2011, pp. 4063–4068.

[21] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Autom. Contr.*, vol. 59, no. 6, pp. 1454–1467, 2014.

[22] C. L. De Marco, J. V. Sariashkar, and F. Alvarado, "The potential for malicious control in a competitive power systems environment," in *Proc. IEEE Int. Conf. Control Applications*, Dearborn, MI, 1996, pp. 462–467.

[23] G. Dan and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Proc. IEEE Int. Conf. Smart Grid Communications*, Gaithersburg, MD, Oct. 2010, pp. 214–219.

[24] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," in *Proc. IEEE Conf. Decision Control European Control Conf.*, Orlando, FL, Dec. 2011, pp. 2195–2201.

[25] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Distributed internet-based load altering attacks against smart power grids," *IEEE Trans. Smart Grid*, vol. 2, no. 4, pp. 667–674, 2011.

[26] S. Sridhar, A. Hahn, and M. Govindarasu, "Cyber–physical system security for the electric power grid," *Proc. IEEE*, vol. 99, no. 1, pp. 1–15, 2012.

[27] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks: Characterizations and countermeasures," in *Proc. IEEE Int. Conf. Smart Grid Communications*, Brussels, Belgium, 2011, pp. 232–237.

[28] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Autom. Contr.*, vol. 56, no. 7, pp. 1495–1508, 2011.

[29] F. Pasqualetti, A. Bicchi, and F. Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Trans. Autom. Contr.*, vol. 57, no. 1, pp. 90–104, 2012.

[30] M. Zhu and S. Martínez, "On distributed convex optimization under inequality and equality constraints," *IEEE Trans. Autom. Contr.*, vol. 57, no. 1, pp. 151–164, 2012.

[31] S. Amin, X. Litrico, S. S. Sastry, and A. M. Bayen, "Stealthy deception attacks on water SCADA systems," in *Proc. Hybrid Systems: Computation Control*, Stockholm, Sweden, Apr. 2010, pp. 161–170.

[32] D. G. Eliades and M. M. Polycarpou, "A fault diagnosis and security framework for water systems," *IEEE Trans. Control Syst. Technol.*, vol. 18, no. 6, pp. 1254–1265, 2010.

[33] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. (2012). A secure control framework for resource-limited adversaries. [Online]. Available: http://arxiv.org/abs/1212.0226

[34] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Autom. Contr.*, vol. 58, no. 11, pp. 2715–2729, 2013.

[35] T. Geerts, "Invariant subspaces and invertibility properties for singular systems: The general case," *Linear Algebra Applicat.*, vol. 183, pp. 61–88, Apr. 1993.

[36] A. Abur and A. G. Exposito, *Power System State Estimation: Theory and Implementation*. Boca Raton, FL: CRC Press, 2004.

[37] E. Scholtz, "Observer-based monitors and distributed wave controllers for electromechanical disturbances in power systems," Ph.D. dissertation, Dept. Electr. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, 2004.

[38] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1991.

[39] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed. Berlin Heidelberg, Germany: Springer-Verlag, 1985.

[40] K. Zhou, J. C. Doyle, and K. Glover, *Robust and Optimal Control*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[41] J. M. Dion, C. Commault, and J. van der Woude, "Generic properties and control of linear structured systems: A survey," *Automatica*, vol. 39, no. 7, pp. 1125–1144, 2003.

[42] K. J. Reinschke, *Multivariable Control: A Graph-Theoretic Approach*. Berlin Heidelberg, Germany: Springer-Verlag, 1988.

[43] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical security via geometric control: Distributed monitoring and malicious attacks," in *Proc. IEEE Conf. Decision Control*, Maui, HI, Dec. 2012, pp. 3418–3425.

[44] F. Dörfler, F. Pasqualetti, and F. Bullo, "Continuous-time distributed observers with discrete communication," *IEEE J. Sel. Topics Signal Processing*, vol. 7, no. 2, pp. 296–304, 2013.

[45] D. J. Trudnowski, J. R. Smith, T. A. Short, and D. A. Pierre, "An application of Prony methods in PSS design for multimachine systems," *IEEE Trans. Power Syst.*, vol. 6, no. 1, pp. 118–126, 1991.

[46] M. A. Hanley, "Frequency instability problems in North American interconnections," Dept. Energy, Tech. Rep. DOE/NETL-2011/1473, June 2011.

[47] F. Pasqualetti, A. Bicchi, and F. Bullo, "A graph-theoretical characterization of power network vulnerabilities," in *Proc. American Control Conf.*, San Francisco, CA, June 2011, pp. 3918–3923.

[48] A. Osiadacz, *Simulation and Analysis of Gas Networks*. Houston, TX: Gulf Publishing Co., 1987.

[49] A. Kumar and P. Daoutidis, *Control of Nonlinear Differential Algebraic Equation Systems*. Boca Raton, FL: CRC Press, 1999.

[50] X. Litrico and V. Fromion, *Modeling and Control of Hydrosystems*. Berlin Heidelberg, Germany: Springer-Verlag, 2009.

[51] J. Burgschweiger, B. Gnädig, and M. C. Steinbach, "Optimization models for operative planning in drinking water networks," *Optim. Eng.*, vol. 10, no. 1, pp. 43–73, 2009.

[52] P. F. Boulos, K. E. Lansey, and B. W. Karney, *Comprehensive Water Distribution Systems Analysis Handbook for Engineers and Planners*. Denver, CO: Amer. Water Works Assoc., 2006.

[53] L. A. Rossman, "EPANET 2, water distribution system modeling software," U.S. Environ. Protection Agency, Water Supply and Water Resources Div., Tech. Rep., 2000.

[54] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *Proc. 1st Workshop Secure Control Systems*, Stockholm, Sweden, Apr. 2010.

[55] H. L. Trentelman, A. Stoorvogel, and M. Hautus, *Control Theory for Linear Systems*. Berlin Heidelberg, Germany: Springer-Verlag, 2001.

[56] F. L. Lewis, "A tutorial on the geometric analysis of linear time-invariant implicit systems," *Automatica*, vol. 28, no. 1, pp. 119–137, 1992.

[57] C. D. Godsil and G. F. Royle, *Algebraic Graph Theory* (Graduate Texts in Mathematics, vol. 207). Berlin Heidelberg, Germany: Springer-Verlag, 2001.