DoS Attacks on Remote State Estimation With Asymmetric Information

Kemi Ding, Xiaoqiang Ren, Daniel E. Quevedo, *Senior Member, IEEE*, Subhrakanti Dey, and Ling Shi

***Abstract*—In this paper, we consider remote state estimation in an adversarial environment. A sensor forwards local state estimates to a remote estimator over a vulnerable network, which may be congested by an intelligent denial-of-service attacker. It is assumed that the acknowledgment information from the remote estimator to the sensor is hidden from the attacker, which, thus, leads to asymmetric information between the sensor and attacker. Considering the infinite-time goals of the two agents and their asymmetric information structure, we model the conflicting nature between the sensor and the attacker by a stochastic Bayesian game. Solutions for this game under two different structures of public information history are investigated, that is, the open-loop structure (in which players cannot observe their opponents' play) and the closed-loop one (in which players can observe the play causally). For the open-loop history case, the original game problem is transformed into a static Bayesian game. We provide the unique mixed-strategy equilibrium explicitly for this game, and analyze the sensor's advantages brought by the extra information. When it comes to the closed-loop case, the dynamic nature of history structure introduces additional difficulties solving the original problem. Thus, to derive *stationary* optimal power schemes for each agent, we convert the original game into a continuous-state stochastic game and discuss the existence of optimal transmission/jamming power strategies. Furthermore, an algorithm based on multiagent reinforcement learning is proposed to find such strategies, and numerical examples are provided to illustrate the developed results.**

***Index Terms*—Asymmetric information, cyber-physical systems, network security, state estimation.**

I. Introduction

CYBER-PHYSICAL systems (CPSs) are systems, which combine dynamic physical processes, sensors and actuators, communication networks, and software [1]. Due to great performance improvements (e.g., stability and robustness) provided by CPSs, they are largely viewed as the next generation of engineering systems. Their applications range from nation-wide smart grids to medium-size transportation and water supply systems, and to small-smart meter and wearable medical devices.

The communication networks provide efficiency for physical systems, but at the same time introduce technical challenges, in particular, security issues, into the system design. Since CPSs are closely related to many safety-critical infrastructures, any intentional attacks on even a single component of a CPS may lead to severe economic losses. Two typical classes of cyber-attacks on CPSs are shown in [2]: 1) deception (integrity) attacks; and 2) denial-of-service (DoS) attacks. The deception attacks focus on deteriorating the system performance by stealthily manipulating the transmitted data packets, whereas DoS attacks compromise the availability of resources by jamming the communication channels. For instance, in the world's first power outage incident caused by a cyber-attack on Ukraine's power system, an exotic virus brought down the information flow from the physical process to the remote management system [3].

In this paper, we focus on remote state estimation (SE) under DoS attacks, which are much easier to implement and are more likely to be encountered in CPSs. Many existing works on DoS attacks rely heavily on quantitative analysis of only one side, such as [4] and [5], which investigated how to launch DoS attacks wisely under power-expenditure constraints from the attacker's perspective. In practice, the sensor should take rational actions (transmission strategy) to avoid jamming attacks, while the attacker will try to recognize these actions, and modify its attack pattern accordingly. Defensive/offensive-scheme designs become complicated when an interaction between the two agents is taken into account. The game-theoretic approach has been successfully applied to capture the strategic interactions between an attacker and a sensor [6]–[9]. The survey conducted by Agah *et al.* [6] studied a cooperative game and proposed a new method for clustering sensors to provide a more reliable communication. A novel discussion on the leader–follower game was presented by Langbort *et al.* [7], which investigated the one-step control problem over a vulnerable communication network under jamming attacks. The latest work by Zhu and Bacşar [8] considered a cross-layer system in which the robust control problem was solved by a zero-sum differential game and a security policy with no power constraints was developed via a zero-sum stochastic game. Li *et al.* [9] considered a scenario in which a sensor sends data for remote estimation via a communication network and the attacker, aiming to degrade

Manuscript received September 15, 2017; revised May 7, 2018 and May 12, 2018; accepted August 6, 2018. Date of publication August 24, 2018; date of current version May 28, 2019. The work of K. Ding, X. Ren, and L. Shi was supported by a Hong Kong RGC theme-based project T23-701/14N. The work of S. Dey was supported by Swedish Research Council Project under Grant 2017-04053. Recommended by Associate Editor L. Xie. (*Corresponding author: Xiaoqiang Ren.*)

K. Ding, X. Ren, and L. Shi are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong (e-mail: kdingaa@connect.ust.hk; zjurenxq@gmail.com; eesling@ust.hk).

D. E. Quevedo is with the Department of Electrical Engineering, Paderborn University, Paderborn D-33098, Germany (e-mail: dquevedo@ieee.org).

S. Dey is with the Department of Engineering Science, Uppsala University, Uppsala SE-751 21, Sweden (e-mail: subhra.dey@signal.uu.se).

Digital Object Identifier 10.1109/TCNS.2018.2867157

the estimation quality, jams the channel in an energy-efficient manner. By employing a two-player zero-sum stochastic game, the authors introduced *stationary* defensive/offensive strategies for each agent (i.e., transmission power and jamming power schemes) with perfect feedback information. That is, the remote estimator will inform the sensor of the packet-loss information via sending back a short acknowledgment (ACK) frame immediately, and the attacker can obtain this online information with intelligent eavesdropping technologies. A major weakness of this approach, however, is that in many practical application scenarios the ACKs may not be accessible to the attacker. For example, peer-to-peer communication uses commutative encryptions to support a secure validation of the ACKs [10]; moreover, current media-access-control protocols and wireless sensor network hardware provide technical support to encrypt full-duplex communication and protect the feedback channel well [11]. Hence, often the sensor knows well about the game state, whereas the attacker, without the ACK's knowledge, only has partial information. Difficulties arise when an attempt is made to implement the (defensive or offensive) policy considering this *asymmetric* information set for the sensor and the attacker. In this paper, we aim to investigate this asymmetric sensor–attacker game on remote estimation, and study the transmission/interference power strategy for the sensor/attacker at *equilibrium*. Compared with previous works, the main contributions of this paper are summarized as follows.

1) Few studies have investigated the impact of *asymmetric information structures* on DoS security issues under the game-theoretic framework. Taking the goals of the sensor and the attacker into consideration, the strategic interaction between them is formulated as a stochastic Bayesian game with asymmetric information (see Problem 1).

2) Considering the solution of the stochastic Bayesian game, this paper begins with the case under a simple open-loop structure of public information history (which excludes all past actions of players). By transforming the original game into a one-stage (static) sensor–attacker game (see Problem 2), we show that the equilibrium (optimal) strategies for the two agents are unique, with the strategy for the sensor having a simple threshold structure. We also show how the sensor can benefit from the online information contained in ACKs.

3) For the closed-loop history case, the original security problem is difficult to solve as the amount of historical information (namely, the past actions of both players) increases with time. We convert the problem into a belief-based continuous-state Markov game with complete information and develop *belief-based* rational strategies for both agents. We prove the existence of the *stationary* equilibrium for the derived Markov game, and also provide a modified Q-learning algorithm to obtain the energy-efficient optimal strategies for the sensor and the attacker.

The remaining paper is organized as follows. Section II contains mathematical models of the system, giving special attention to the asymmetric information structure between the sensor and the attacker. Section III demonstrates the framework of the sensor–attacker game played over an infinite-time horizon. Section IV shows the main theoretical results for the game under two different history structures. Section V provides multiagent Q-learning algorithm to obtain the rational strategies. Sections VI and VII present some examples and concluding remarks, respectively. The Appendix presents the proofs of theorems.

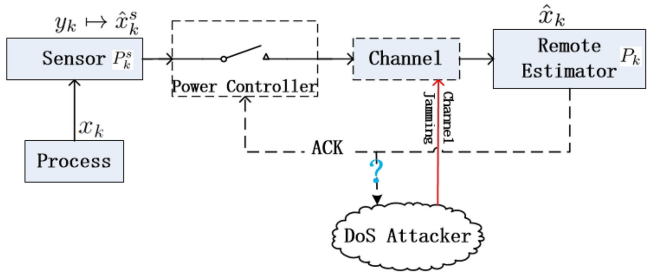


Fig. 1. System model.

Notations: $\mathbb{R}^n$ is the n-dimensional Euclidean space. $\mathbb{S}_+^n$ (or $\mathbb{S}_{++}^n$) is the set of n-by-n positive semidefinite matrices (or positive definite matrices). Let $\mathbb{N}$ denote the set of natural numbers. When $X \in \mathbb{S}_+^n$ (or $X \in \mathbb{S}_{++}^n$), we write $X \geq 0$ (or $X > 0$). For functions h and g, $h \circ g$ is defined as the function composition $h(g(\cdot))$. $\mathbb{E}[\cdot]$ is the expectation of a random variable (r.v.), $\Delta(\cdot)$ refers to the probability measure space over a set, and $\mathtt{Pr}(\cdot)$ refers to the probability. $\mathtt{Tr}(\cdot)$ and $\rho(\cdot)$ denote the trace and the spectral radius of a matrix, respectively. The superscripts $\top$ and $\star$ stand for the matrix transposition and the optimal solution, respectively, while the superscripts s (or subscript 1) and a (or subscript 2) denote the sensor and the attacker, respectively. Moreover, the superscript o represents observation. The capital Θ is the set of types and the small form θ_1 represents the sensor's type variable. Moreover, the term θ is a specific value of the sensor's type. Here, y_0^k stands for the sequence $(y_0, \ldots, y_k)$ and the sequence of actions over time k are defined similarly. Furthermore, $\mathbb{1}(\cdot)$ is the indicator function and the Dirac delta function is

$$\delta_{kj} = \begin{cases} 1, & \text{if } k = j \\ 0, & \text{others.} \end{cases}$$

II. Problem Formulation

Fig. 1 shows the system model, in which the state information of the process is sent to the remote estimator in the presence of a DoS attacker. In this section, essential components of the overall system structure will be introduced in detail.

A. Local Kalman Filter

Consider the following linear time-invariant system:

$$x_{k+1} = Ax_k + w_k \tag{1}$$

$$y_k = Cx_k + v_k \tag{2}$$

where the state vector of the system at time k is denoted by $x_k \in \mathbb{R}^{n_x}$, the noisy measurement obtained by the sensor is $y_k \in \mathbb{R}^{m_y}$, and $w_k \in \mathbb{R}^{n_x}$ and $v_k \in \mathbb{R}^{m_y}$ represent zero-mean independent identically distributed Gaussian random noise with $\mathbb{E}[w_k w_j^\top] = \delta_{kj} Q$ $(Q \geq 0)$, $\mathbb{E}[v_k v_j^\top] = \delta_{kj} R$ $(R > 0)$, and $\mathbb{E}[w_k v_j^\top] = 0\ \forall j, k$. The initial state x_0 is a zero-mean Gaussian random vector with covariance $\Sigma_0 \geq 0$, which is uncorrelated with w_k and v_k. To avoid trivial problems, we assume the system is unstable, that is, $\rho(A) > 1$. The time-invariant pair (A, C) is assumed to be detectable and $(A, \sqrt{Q})$ is stabilizable.

With the advanced smart sensors [12], the estimation/control performance of the current system can be highly improved. The smart sensors are equipped with memory and embedded systems-on-chips, which enable them to query for historical information and execute some simple recursive algorithms on the collected data. With storage and computing abilities, the sensor in Fig. 1 is able to process the collected measurements y_0^k by running a Kalman filter, instead of transmitting them directly, and then estimate the process state x_k locally, denoted by $\hat{x}_k^s$. This minimum mean-squared error (MMSE) estimate of the process state is given by

$$\hat{x}_k^s = \mathbb{E}[x_k | y_0^k]$$

with its corresponding estimation-error covariance

$$P_k^s \triangleq \mathbb{E}[(x_k - \hat{x}_k^s)(x_k - \hat{x}_k^s)^\top].$$

These terms are computed via the Kalman filter [13]. To simplify notations, we define the Lyapunov and Riccati operators h and $\tilde{g} : \mathbb{S}_+^n \to \mathbb{S}_+^n$ as

$$h(X) \triangleq AXA^\top + Q$$

$$g(X) \triangleq X - XC^\top [CXC^\top + R]^{-1} CX.$$

Owing to the stabilizability and detectability assumptions, the estimation-error covariance P_k^s converges exponentially fast to a unique fixed point $\overline{P}$ of $h \circ g$ [13]. For simplicity, we ignore the transient periods and assume that the Kalman filter at the sensor has entered the steady state, i.e., we assume that

$$P_k^s = \overline{P}, \ \forall k \geq 1. \tag{3}$$

According to [14], the steady-state error covariance $\overline{P}$ has the following property.

Proposition 1: For $0 \leq t_1 < t_2$, the following inequality holds:

$$\mathtt{Tr}[\overline{P}] \leq \mathtt{Tr}[h^{t_1}(\overline{P})] < \mathtt{Tr}[h^{t_2}(\overline{P})]. \tag{4}$$

B. Communication Channel

Since most sensor nodes use onboard batteries, which are difficult to replace or recharge, the energy for sensing, computation, and transmission is restricted for sensor nodes. Hence, as depicted in Fig. 1, the sensor is required to decide the transmission-energy level at which to send the obtained estimates $\hat{x}_k^s$ to the remote estimator. At the same time, by emitting a signal to interfere with the channel, the attacker is capable of sabotaging the delivery of $\hat{x}_k^s$ and thus degrading the estimation quality. Like the sensor, the attacker has a limited energy budget and has to determine the jamming energy at each time. We denote the available transmission power set with M power levels as $\mathbb{E}^s = \{e_1^s, \dots e_M^s\}$ and attack the power set with L levels as $\mathbb{E}^a = \{e_1^a, \dots, e_L^a\}$, in which $e_i^s, 1 \leq i \leq M$ and $e_j^a, 1 \leq j \leq L$ represent the ith transmission power level and the jth jamming power level, respectively. Let $\alpha_{1,k} \in \mathbb{E}^s$ and $\alpha_{2,k} \in \mathbb{E}^a$ denote the transmission power of the sensor and the interference power of the attacker at time k, respectively. Note that $\alpha_{1,k} = 0$ and $\alpha_{2,k} = 0$ represent that the sensor does not transmit data packet $\hat{x}_k^s$ (i.e., the sensor is inactive) and that no DoS attack is launched, respectively. The transmission (or jamming) schedule of the sensor (or attacker) over the infinite-horizon is denoted by $(\alpha_{1,0}, \alpha_{1,1}, \alpha_{1,2}, \dots)$ [or $(\alpha_{2,0}, \alpha_{2,1}, \alpha_{2,2}, \dots)$].

We assume the channel between the sensor and the estimator is memoryless, and that it has independent additive white Gaussian noises. The transmitted data packet may arrive at the remote estimator with unknown errors due to channel noise, signal fading, multipath effects, etc. Some channel-coding methods can detect these errors, and the incorrect packets will be dropped before uploading to the estimator (see [15]). We introduce the packet-error-rate (PER), which is monotonically increasing with the signal-to-noise-ratio (SNR) for any modulation scheme, to measure the packet losses. To quantify the packet loss under DoS attacks, we adopt the signal-to-interference-plus-noise-ratio (SINR) [15] rather than the SNR

$$\mathrm{SINR}_k = \frac{\alpha_{1,k}}{\alpha_{2,k} + n_0}, \qquad \mathrm{PER}_k = \hat{f}(\mathrm{SINR}_k) \tag{5}$$

in which n_0 is the power of the additive white channel noise and $\hat{f}(\cdot)$ is a nonincreasing function. Without loss of generality, the channel gain is taken to be unity, and therefore, the received SINR can be defined based on the transmission powers instead of the actual received power. Notice that here we use a general function $\hat{f}(\cdot)$ to describe PER_k, which has various forms that correspond to different modulation modes. Interested readers are referred to [15].

Under this scenario (the erasure channel), the arrival of the packet can be characterized by a binary random process, denoted by η_k. Let $\eta_k = 0$ represents the packet loss, and $\eta_k = 1$ otherwise. When given the action of two agents ($\alpha_{1,k}$ and $\alpha_{2,k}$), the packet-arrival probability is defined as follows:

$$\mathtt{Pr}(\eta_k = 1) = q(\alpha_{1,k}, \alpha_{2,k}) \triangleq 1 - PER_k. \tag{6}$$

In this paper, we consider a communication-feedback mechanism between the estimator and the sensor, as shown in Fig. 1. The remote estimator will inform the sensor of the packet-loss information η_k via sending back a short ACK frame immediately, i.e., before instant $k+1$. This mechanism is an essential part of Internet protocols (e.g., the TCP/IP protocol). Since the sensor has a comprehensive understanding of the communication dynamics based on the collected ACKs, it can develop an effective transmission schedule to improve the system performance; i.e., $\alpha_{1,k}$ may depend on the previous ACK sequence η_1^{k-1}. Here, the ACK is assumed to be reliably received by the sensor. We shall assume that the attacker has no access to the ACKs. Thus, the jamming scheme $\alpha_{2,k}$ adopted by the attacker depends only on its previous observations: the initial transmission status, its historical jamming power $\alpha_{2,0}^{k-1}$, and the historical transmission energy of the sensor $\alpha_{1,0}^{k-1}$. The sensor can access the channel-state information using pilot-aided channel-estimation techniques [15]. As for the attacker, it will adopt a full-duplex technology to simultaneously generate interference and monitor the channel. Specifically, it treats the packet transmission of the sensor as an unknown deterministic signal and adopts energy-detection technologies to estimate the transmission power [16], [17]. More details about a practical hardware implementation of the full-duplex attack can be found in [18] and [19]. Thus, the sensor and the attacker can monitor the transmission power and jamming power at each time after packet transmission. The estimation error of the transmission energy is not taken into account in this paper.

In conclusion, the decision-making information set of the attacker is different from that of the sensor. The distinction

between them arises from the availability of packet ACKs, that is, the information structure for the sensor is TCPlike, while for the attacker, it is UDPlike.

C. Remote Estimation

Based on the received data packets, the remote estimator generates the MMSE estimate of the process x_k, denoted by $\hat{x}_k$, with corresponding error covariance P_k. The estimate $\hat{x}_k$ [14] is obtained by

$$\hat{x}_k = \eta_k \hat{x}_k^s + (1-\eta_k) A \hat{x}_{k-1}. \tag{7}$$

Consequently, the error covariance P_k at time k is

$$\begin{aligned} P_k &\triangleq \mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^\top] \\ &= \begin{cases} \overline{P}, & \eta_k = 1 \\ h(P_{k-1}), & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

where $\overline{P}$ stands for the steady-state error covariance shown in (3).

Without loss of generality, we assume that the initial packet $\hat{x}_0^s$ is known by the estimator, and hence, $P_0 = \overline{P}$. From (8), it follows that at a given time k, P_k can only take values in the finite set $\{\overline{P}, h(\overline{P}), h^2(\overline{P}), \ldots, h^k(\overline{P})\}$.

D. Problem of Interest

Given a budget of transmission and congestion power, the strategy designs of the sensor and the attacker are coupled. In general, the task of the sensor is to make sure that its end user (i.e., the remote estimator) is sufficiently informed of the process, without wasting energy; as for the attacker, it intends to disrupt the reliable communication between the sensor and the user, also without expending more energy than required. With opposite goals, the decision-making procedures of the sensor and the attacker are interactively linked. The difficulty of designing the power schemes for the two agents arises from their distinct information structures. In the following sections, we will investigate the decision-making procedures for the defending sensor and the malicious attacker with asymmetric information. In Section III, an asymmetric-information game played over time will be developed to model this interactive process over an infinite-time horizon, and its solutions are demonstrated in Section IV.

III. Game Model

In this section, the problem of scheduling energy-efficient actions in an infinite-time scenario, i.e., how to decide the transmission energy (or interference power) for the sensor (or the attacker) is modeled within a game-theoretic framework.

A. Preliminaries

First, we define an r.v. $\tau_k \in \mathbb{Z}$ as the holding time[1]

$$\tau_k \triangleq k - \max_{0 \le l \le k} \{l : \eta_l = 1\} \tag{9}$$

[1]In the remaining paper, we will omit the subscript of τ_k when the underlying time index k is obvious from the context; when it is ambiguous, the subscript will be indicated.

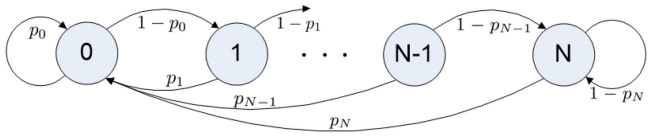


Fig. 2. Markov chain transition process of holding time τ_k.

which represents the intervals between the present moment k and the most recent time when the data packet has been successfully received by the estimator. As mentioned before, we shall assume that the estimator receives the packet $\hat{x}_0^s$ at the beginning of the transmission; i.e., $\eta_0 = 1$. Based on (8), it is easy to obtain the relationship between the holding time and the estimation-error covariance at the remote estimator P_k,

$$P_k = h^{\tau_k}(\overline{P}), \tag{10}$$

and the iteration of the holding time

$$\tau_k = \begin{cases} 0, & \text{if } \eta_k = 1 \\ \tau_{k-1} + 1, & \text{otherwise.} \end{cases} \tag{11}$$

Owing to the communication feedback, the sensor will obtain the online information η_0^{k-1} at the end of the $(k-1)$th time interval and then infer the remote error covariance P_{k-1} from (10) before deciding on the transmission power $\alpha_{1,k}$ for time k. Different from the sensor, the attacker has no access to the online information η_0^{k-1} or the actual P_{k-1}, while it will make a guess about the error covariance P_{k-1} (or τ_{k-1}) based on its observations.

Obviously, τ_k may take values from $\mathbb{Z}_k \triangleq \{0, 1, 2, \ldots, k\}$. As $k \to \infty$, $\mathbb{Z}_k$ will be countably infinite. Notice that the current state τ_k depends only on the last state τ_{k-1} and the r.v. η_k from (11). Hence, the sequence of random states τ_k forms a Markov chain, and the transition process is depicted in Fig. 2. With the powers of the two agents $\alpha_{1,k}$ and $\alpha_{2,k}$, the transition can be described by a simple transition probability matrix

$$\mathrm{T}_{\{\alpha_{1,k},\alpha_{2,k}\}} = \begin{pmatrix} t & 1-t & & & \\ t & & 1-t & & \\ t & & & 1-t & \\ \vdots & & & & \ddots \end{pmatrix} \tag{12}$$

where the entry $\mathrm{T}(i,j)$ represents the transition probability from the state $\tau_k = i$ to $\tau_{k+1} = j$. Notice that the probability $t = q(\alpha_{1,k}, \alpha_{2,k})$ is given by (6) and the other default entries are 0.

To simplify the problem, we truncate $\mathbb{Z}_k$ and consider a finite set, that is, $\mathbb{Z} \triangleq \cup_{k \ge 1} \mathbb{Z}_k = \{0, \ldots, N\}$. The final state N represents all the states $\tau_k \ge N$. The effect of the truncation operator is analyzed when the nontruncated Markov chain is bounded, i.e., $\sum_{k=0}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)] < \infty$. The effect is negligible when N is large enough, which is formalized in the following lemma. Moreover, the truncation effect on the game solution is explained directly in Remark 1.

Lemma 1: If the sensor's transmission and attacker's jamming strategies are such that the nontruncated Markov chain is bounded, i.e., $\sum_{k=0}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)] < \infty$, then there holds that as $N \to \infty$, $\mathrm{D}(N) \to 0$ with $\mathrm{D}(N) \triangleq \sum_{k=0}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k) - \mathbb{E}(P_k|\mathbb{Z})]$ being the performance gap.

Proof: See Appendix A. ■

B. Game Description

A unique feature of the problem is that the two agents take actions simultaneously based on different information sets. This leads to a Bayesian-game structure that captures the asymmetrical-information setting between the sensor and the attacker. Additionally, considering the dynamic nature of the underlying physical process, we then model the strategic interaction process as a stochastic Bayesian game, which can be viewed as a combination of the Bayesian game and the stochastic game. The stochastic Bayesian game, denoted by $\mathcal{G}^B$, is characterized by a sextuplet $(\mathcal{I}^B, \mathcal{A}^B, \mathcal{S}, \Theta, \Delta(\Theta), \mathcal{R}^B)$. Each item is elaborated as follows.

1) Player: $\mathcal{I}^B = \{1, 2\}$ is the set of players, in which $i = 1$ represents the sensor and $i = 2$ stands for the attacker. The two players are assumed to be rational players,[2] i.e., each of them makes the best choice in terms of their own benefits among all actions available to them. Each player knows that the other player is rational. Furthermore, they also know that their opponent knows that they know this, and so on, ad infinitum.

2) Action: $\mathcal{A}^B = \mathcal{A}_1^B \times \mathcal{A}_2^B$ where $\mathcal{A}_i^B, i = 1, 2$ is the action (or pure-strategy) space for player i. Note that the sensor/attacker selects the transmission/attack power from the corresponding power set. Hence, $\mathcal{A}_1^B = \mathbb{E}^s$ and $\mathcal{A}_2^B = \mathbb{E}^a$. Moreover, the mixed action for each player, denoted by $\mathbf{A}_i^B \in \Delta(\mathcal{A}_i^B), i = 1, 2$, is a probability distribution over the pure action space $\mathcal{A}_i^B$.

3) State: $\mathcal{S} = \{0, 1, \ldots, N\}$ is the state space, with element $s_k \in \mathcal{S}$ representing the state of the game at time step k. We view the Markov state τ_k as the state of the game s_k with the transition matrix shown in (12). As presented in (10), the state is closely related to the estimation-error covariance.

4) Type: $\Theta = \{\Theta_i, i \in \mathcal{I}\}$ illustrates the set of types for each player. The type is the private information for each player that is relevant to his decision making. In this game, the state τ of the Markov chain is known by the sensor only, and hence, it is regarded as the (finite) type of the sensor, denoted by $\theta_1 \in \mathbb{Z}$. Suppose that the sensor is sure that the attacker cannot access the well-protected ACKs or obtain the exact state τ (i.e., the type of the sensor θ_1). Hence, the information collected by the attacker is visible to the sensor completely. In other words, the type of the attacker is known by the sensor, and θ_2 is a singleton.

5) Belief: $\mathbf{B}_k \in \Delta(\Theta_1)$ is a joint probability distribution over the sensor's type Θ_1 assigning θ_1 with probability (w.p.) $\mathbf{B}_k(\theta_1)$ at time k. The initial belief is $\mathbf{B}_0 = \mathbf{b}_0$, which is assumed to be the *common knowledge*[3] shared by all players. Recall that the type θ_1 is known by the sensor only, and the attacker will build a belief of the sensor's type based on Bayes' rule.

6) Reward: $\mathcal{R}^B = \{r_i^B, i \in \mathcal{I}\}$ is the one-stage reward set, and r_i^B represents the reward function for player i. Rewards are computed by each player via taking expectations over types under its own conditional beliefs about opponents' types, thus, $r_i^B : \Theta_i \times \Delta(\Theta_{-i}) \times \mathcal{A}^B \to \mathbb{R}$. As Θ_2 is a singleton, the reward functions for each player can be simplified, namely, $r_1^B : \Theta_1 \times \mathcal{A}^B \to \mathbb{R}$ and $r_2^B : \Delta(\Theta_1) \times \mathcal{A}^B \to \mathbb{R}$.

As discussed previously, the sensor focuses on improving the estimation accuracy, and the attacker hopes to degrade the system performance. By quantifying the benefit of each player as the trace of the expected estimation-error covariance and taking the energy cost into account, we can obtain that

$$\begin{aligned} &r_1^B(\theta_{1,k}, \alpha_{1,k}, \alpha_{2,k}) \\ &= -q(\alpha_{1,k}, \alpha_{2,k})\mathtt{Tr}[\overline{P}] \\ &\quad - (1 - q(\alpha_{1,k}, \alpha_{2,k}))\mathtt{Tr}[h^{\theta_1+1}(\overline{P})] - \delta_1 \alpha_{1,k} \end{aligned} \tag{13}$$

and

$$r_2^B(\mathbf{B}_k, \alpha_{1,k}, \alpha_{2,k}) = \sum_{\theta_1 \in \mathbb{Z}} \mathbf{B}_k(\theta_1) \tilde{r}_2^B(\theta_1, \alpha_{1,k}, \alpha_{2,k}) \tag{14}$$

with

$$\begin{aligned} \tilde{r}_2^B(\theta_1, \alpha_{1,k}, \alpha_{2,k}) &\triangleq q(\alpha_{1,k}, \alpha_{2,k})\mathtt{Tr}[\overline{P}] \\ &\quad + (1 - q(\alpha_{1,k}, \alpha_{2,k}))\mathtt{Tr}[h^{\theta_1+1}(\overline{P})] - \delta_2 \alpha_{2,k}, \end{aligned} \tag{15}$$

in which $\delta_1 \geq 0$ and $\delta_2 \geq 0$ represent the proportions of the energy term in the reward functions. The sensor's reward function depends on its type, whereas the attacker's reward function is an expectation developed, based on the player's belief. This demonstrates the difference between the Nash equilibrium (NE) and the Bayesian NE (BNE).

In general, the information available to all players throughout the play is described by the public information history. Here, we denote $\mathcal{H}$ as the set of all histories and h_k as the public information history at time step k. The game can be carried out under two different history structures [20] as follows.

1) *Open-loop structure.* Both players have the knowledge of the *prior* distribution $\mathbf{B}_0 = \mathbf{b}_0$ and the time k, i.e., $h_k^o = \{\mathbf{b}_0, k\}$.

2) *Closed-loop structure.* Players can observe the actions of their opponents. This history concerns the *prior* distribution $\mathbf{b}_0$ and the history of all players' actions, i.e., $h_k^c = \{\mathbf{b}_0, \alpha_{1,1}, \alpha_{2,1}, \ldots, \alpha_{1,k}, \alpha_{2,k}\}$.

Recall that the private information of player i is contained in Θ_i. Here, we use $\pi_i : \mathcal{H} \times \Theta_i \to \Delta(\mathcal{A}_i^B)$ to represent the strategy (or decision rule) for player i, and $\pi_i(\alpha_{i,k}|h_k, \theta_{i,k})$ is the probability with which the specific action $\alpha_{i,k}$ is played by player i based on public information history h_k and the current type information $\theta_{i,k}$. The scenarios that the game is played under these two history structures are analyzed in details in the next section.

The game is played as follows:

1) In state s_k, the players' types are determined, and each player is informed only about its own type.

2) Based on the history h_k and its own type $\theta_{i,k}$, each player i chooses an action $\alpha_{i,k}$ according to a randomized strategy π_i.

3) Each player receives an immediate reward r_i^B, and the game moves to a new state $s_{k+1} \in \mathcal{S}$ with a transition probability $\mathrm{T}_{\{\alpha_{1,k}, \alpha_{2,k}\}}(s_k, s_{k+1})$.

In an abuse of notation, we use the same notation $r_i^B(\cdot)$ to denote, at time k, the sensor's reward to the randomized strategy

[2]The rationality assumption is a necessary condition for our subsequent equilibrium analysis.

[3]In game theory, the common knowledge is stated in an informal way, that is, every player knows them, knows that all the others know them, and so on.

π_1 when the attacker uses π_2, which is given by

$$r_1^B(\pi_1,\pi_2|h_k,\theta_{1,k}) = \sum_{\alpha_1^B\in\mathcal{A}_1^B}\sum_{\alpha_2^B\in\mathcal{A}_2^B}\Big\{\pi_1(\alpha_1^B|h_k,\theta_{1,k}) \times \pi_2(\alpha_2^B|h_k,\theta_{2,k})r_1^B(\theta_{1,k},\alpha_1^B,\alpha_2^B)\Big\}$$

in which $\theta_{2,k}\in\emptyset,\forall k\geq 0$. Similarly, we can obtain the attacker's reward to the pair of mixed strategies $\{\pi_1,\pi_2\}$, denoted by $r_2^B(\pi_1,\pi_2|h_k)$.

As for this game, from the long-term point of view, the infinite-time discounted payoff of the ith player under the strategies π_1 and π_2 is

$$\begin{aligned}\mathcal{J}_1(\mathbf{b}_0,\pi_1,\pi_2) &= \sum_{k=0}^{\infty}\delta^k r_1^B(\pi_1,\pi_2|h_k,\theta_{1,k})\\ \mathcal{J}_2(\mathbf{b}_0,\pi_1,\pi_2) &= \sum_{k=0}^{\infty}\delta^k r_2^B(\pi_1,\pi_2|h_k)\end{aligned}\tag{16}$$

where $\mathbf{b}_0$ is the initial *prior* belief distribution, and the parameter $\delta\in[0,1)$ stands for the discount factor. Each player's objective is to maximize its own payoff function. Here, we consider the discounted sum of rewards since we put more weight on the rewards obtained in the early periods. The sensor–attacker security problem is summarized as follows.

Problem 1: Find a pair of strategies $(\pi_1^\star,\pi_2^\star)$ such that for any π_1 and π_2 the following hold:

$$\begin{aligned}\mathcal{J}_1(\mathbf{b}_0,\pi_1^\star,\pi_2^\star)&\geq\mathcal{J}_1(\mathbf{b}_0,\pi_1,\pi_2^\star)\\ \mathcal{J}_2(\mathbf{b}_0,\pi_1^\star,\pi_2^\star)&\geq\mathcal{J}_2(\mathbf{b}_0,\pi_1^\star,\pi_2).\end{aligned}\tag{17}$$

Remark 1: This pair of strategies $(\pi_1^\star,\pi_2^\star)$ is called an NE of the truncated game (the truncation operation is discussed in Section III-A). Notice that if a pair of policies leads to an unbounded discounted payoff for each player, then the game is dominated by the attacker since it can adopt the corresponding jamming scheme to obtain an unbounded expected-error covariance. To avoid such trivial cases, in this paper, we focus on the pairs of policies under which the discounted payoff functions are bounded. Notice that the bounded payoff implies the bounded sum of discounted estimation-error covariance, i.e., $\sum_{k=0}^{\infty}\delta^k\mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)]<\infty$. Next, we show that the truncation effect on the game equilibrium is negligible when N is large enough. Let $\mathcal{J}_i(\mathbf{b}_0,\pi_1,\pi_2|\mathbb{Z}_k)$ and $\mathcal{J}_i(\mathbf{b}_0,\pi_1,\pi_2|\mathbb{Z})$ denote the discounted payoff of the ith player under the strategies (π_1,π_2) for the nontruncated and truncated games, respectively. From Lemma 1 and (17), we have for $i=1,2$

$$\begin{aligned}&\mathcal{J}_i(\mathbf{b}_0,\pi_i^\star,\pi_{-i}^\star|\mathbb{Z}_k)-\mathcal{J}_i(\mathbf{b}_0,\pi_i^\star,\pi_{-i}|\mathbb{Z}_k)\\ &=\lim_{N\to+\infty}\mathcal{J}_i(\mathbf{b}_0,\pi_i^\star,\pi_{-i}^\star|\mathbb{Z})-\mathcal{J}_i(\mathbf{b}_0,\pi_i^\star,\pi_{-i}|\mathbb{Z})\geq 0.\end{aligned}$$

Thereby, we conclude that $(\pi_1^\star,\pi_2^\star)$ is also an NE of the nontruncated game when the truncation parameter N goes to infinity.

IV. Main Results

In this section, we present some results of Problem 1 under two different history structures h_k^o and h_k^c.

A. Open-Loop Structure

As mentioned previously, the information available to a player at time k is the initial prior distribution $\mathbf{B}_0=\mathbf{b}_0$ and time k. Note that the players' strategies are determined by the public information history h_k^o, of which the total amount remains constant, and $\mathbf{B}_k=\mathbf{b}_0,\forall k\geq 0$. Accordingly, the strategies for the sensor and the attacker are reduced to $\pi_1^B:\Delta(\Theta)\times\Theta_1\to\Delta(\mathcal{A}_1^B)$ and $\pi_2^B:\Delta(\Theta)\to\Delta(\mathcal{A}_2^B)$ with $\Pi_i^B, i=1,2$ denoting the set of strategies for player i. Thus, we can simplify the multistage game (see Problem 1) to the following one-stage problem, which can be solved easily.

Problem 2: Find a pair of strategies $(\pi_1^{B,\star},\pi_2^{B,\star})\in\Pi_1^B\times\Pi_2^B$ to maximize the one-stage reward function for each player with $\mathbf{B}_0=\mathbf{b}_0$, i.e.,

$$\begin{aligned}&\max_{\pi_1,\pi_2} r_1^B(\pi_1,\pi_2|\mathbf{b}_0,\theta_1),\ \forall\theta_1\in\Theta\\ &\max_{\pi_1,\pi_2} r_2^B(\pi_1,\pi_2|\mathbf{b}_0).\end{aligned}$$

Next, we define the solution for Problem 2, called BNE.

Definition 1 (BNE): In this attacker–sensor one-stage game with a finite number of types for the sensor and a *prior* distribution $\mathbf{B}_0$, the strategy profile $(\pi_1^{B,\star},\pi_2^{B,\star})$ for the players is a BNE if no player can benefit from changing strategies, while the other keeps its own unchanged, *i.e.*, for any type of the sensor

$$\begin{aligned}r_1^\star(\mathbf{B}_0,\theta_1)&\triangleq r_1^B(\pi_1^{B,\star},\pi_2^{B,\star}|\mathbf{B}_0,\theta_1)\\ &\geq r_1^B(\pi_1^B,\pi_2^{B,\star}|\mathbf{B}_0,\theta_1),\forall\pi_1^B\in\Pi_1^B\end{aligned}$$

and

$$\begin{aligned}r_2^\star(\mathbf{B}_0)&\triangleq r_2^B(\pi_1^{B,\star},\pi_2^{B,\star}|\mathbf{B}_0)\\ &\geq r_2^B(\pi_1^{B,\star},\pi_2^B|\mathbf{B}_0),\forall\pi_2^B\in\Pi_2^B.\end{aligned}$$

The attacker could use its belief $\mathbf{B}_0(\theta_1)$ to compute the expected benefit of each action choice and, thus, find its optimal response. By [21], the existence of a BNE with a mixed strategy for Problem 2 is an immediate consequence as the type space Θ is finite. The uniqueness and structure of a BNE with a mixed-strategy form is analyzed in Theorem 1. To simplify the problem, in this section, we suppose that the pure action set of the sensor is $\mathcal{A}_1^B=\{0,e_1\}$, in which the action $\alpha_1^B=0$ means no packet is sent. Similarly, the attacker also has two possible actions —to attack or not to attack, that is, $\mathcal{A}_2^B=\{0,e_2\}$. This scenario is representative, (see [22] and [5]); due to the limited available communication energy, the sensor/attacker has to decide to send/jam or not. For notational convenience, we define the packet-arrival rate in different cases as $\lambda_1\triangleq q(\alpha_2=e_1,\alpha_2=e_2)$ and $\lambda_2\triangleq q(\alpha_1=e_1,\alpha_2=0)$, in which we have $\lambda_1<\lambda_2$.

Theorem 1: Consider the attacker–sensor static game Problem 2 with the aforementioned action sets. If there exists an $m\in\mathbb{N}$ that satisfies the inequality $\sum_{\theta=m+1}^{n}\mathbf{B}(\theta)f(\theta)\leq\delta_2 e_2\leq\sum_{\theta=m}^{n}\mathbf{B}(\theta)f(\theta)$, then the mixed-strategy BNE is unique. In particular, in the mixed-strategy BNE, the sensor transmits the packets w.p. $q_{1,\theta}^\star$, which has the following thresh-

old structure:

$$q_{1,\theta}^{\star}=\begin{cases}0, & \text{if } \theta<m\\ \frac{\delta_2 e_2-\sum_{\theta=m+1}^{n}\mathbf{B}(\theta)f(\theta)}{\mathbf{B}(m)f(m)}, & \text{if } \theta=m\\ 1, & \text{otherwise}\end{cases} \tag{18}$$

where $f(\theta)=(\lambda_2-\lambda_1)\mathtt{Tr}[h^{\theta+1}(\overline{P})-\overline{P}]$. Moreover, the attacker jams the channel w.p. $q_2^{\star}=\frac{\delta_1 e_1-\frac{\lambda_2}{\lambda_2-\lambda_1}f(m)}{f(m)}$.

Proof: See Appendix B. ■

The Bayesian game is the partially observable counterpart of the normal-form game. The one-stage sensor–attacker game with symmetric information can be regarded as a normal-form game. Comparing the sensor's optimal reward in the asymmetric game with that in the normal-form game, we can see that the sensor will earn extra benefits by protecting the ACK information from the attacker.

Theorem 2: Denote the optimal reward for the sensor in the normal-form game (obtained by the NE) by $r_1^{\mathrm{NE}}(\theta)$, and that of the Bayesian game $\mathcal{G}^B$ by $r_1^{\mathrm{BNE}}(\mathbf{B}_0,\theta)$. We then have for any initial belief $\mathbf{B}_0\in\Delta(\Theta)$

$$r_1^{\mathrm{NE}}(\theta)\le r_1^{\mathrm{BNE}}(\mathbf{B}_0,\theta),\ \forall\theta\in\mathbb{Z}.$$

Proof: See Appendix C. ■

B. Closed-Loop Structure

In the previously discussed open-loop case, the attacker makes decisions based on an *a priori* guess (i.e., initial belief $\mathbf{B}_0$) at the sensor's type. In contrast to this, the guesses in the closed-loop case are made based on historical (observed) actions up to time k applied both by the attacker and the sensor (i.e., $\mathbf{B}_k\neq\mathbf{B}_0$ for $k>0$). Obviously, these dynamic guesses allow the attacker to make better decisions in the future. Clearly, the sensor may notice this and take actions with the consideration of the attacker's guess. Therefore, both players design their optimal power strategies (i.e., the BNE) with the history information taken into account. Unfortunately, the total information amount increases with k for each player (i.e., h_k increases with time). However, not all information accumulated up to time k turns out to be relevant for the decision-making process. Motivated by this, next, we attempt to circumvent the complexities induced by an increasing information set, and investigate easy-to-implement strategies (i.e., *stationary* strategies as shown below) for both players.

This game between the sensor and the attacker is played with the property that at each stage k, two players simultaneously select actions that will be revealed at the end of the stage k. Moreover, the sensor can directly obtain the underlying online information τ_k[4] from the observed ACKs, whereas the attacker cannot. Hence, as illustrated in Fig. 3, the planning problem for the sensor is akin to a Markov decision process (MDP), and that of the attacker is like a partially observed MDP (POMDP). To overcome difficulties arising from the absence of online information, the conventional treatment to POMDP consists in taking the internally generated belief as the state of the new MDP, since

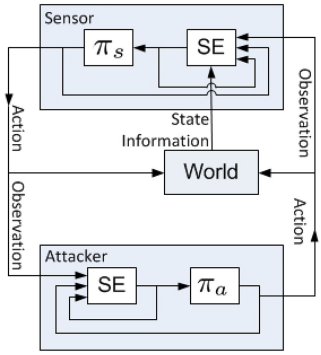


Fig. 3. Dynamic game with asymmetric information. At each stage, the belief is updated by the SE device.

the belief is sufficient statistic (i.e., satisfies the Markov property [23]). Inspired by the idea of the "belief-based" MDP, we construct a stochastic game with complete information, which is composed of five tuples: $\mathcal{G}^S\triangleq(\mathcal{I},\mathcal{B},\mathcal{A},\mathbf{Q},r)$[5]; details are demonstrated as follows.

1) Player: $\mathcal{I}=\{0,1,\ldots,N+1\}$, where each player type is treated as a separate player. Specifically, $i=N+1$ represents the attacker, and the other $i\in\{0,\ldots,N\}$ stands for the respective "type player" of the sensor. In effect, when the type of the sensor at stage k is $\tau_k=m$, the sensor will adopt the strategy of the mth type player correspondingly. The general idea is to view the different types of sensor as different "individuals," and one of them is selected by nature to "appear" when the game is played. The different "individuals" can make their strategies at the "interim" stage (i.e., after knowing their type), which is equivalent to a single sensor making *ex ante* decisions before learning its type.

2) Belief State Space: $\mathcal{B}=\Delta(\mathcal{S})$ stands for the continuous belief state space defined on $\mathcal{S}$ with $\mathcal{S}=[0,\ldots,N]$. Let $\mathbf{B}_k(m)$ denote the probability that $s_k\triangleq\tau_k=m$. The distinction between the belief state and the original state s_k is that the former is a public knowledge for all players $\mathcal{I}$, whereas the latter is only observed by the type players. Let $\mathcal{B}$ be endowed with the topology of weak convergence, then it is a Polish space (i.e., complete and separable metric space) [24].

3) Action: $\mathcal{A}=\prod_{i\in\mathcal{I}}\mathcal{A}_i$ is the joint action set. The action set for the attacker $\mathcal{A}_{N+1}=\Delta(\mathcal{A}_2^B)$ and each type player share the same action set: $\mathcal{A}_i=\Delta(\mathcal{A}_1^B)\ \forall i\in\{0,\ldots,N\}$. We let $\mathbf{A}_{i,k}\in\mathcal{A}_i$ denote the action for the ith player at stage k: $\mathbf{A}_{i,k}(\alpha_i)$ represents the probability of the pure action α_i taken by the ith player at stage k. Notice that unlike in games with fully observable states, the distributions of actions that are not taken affect the evolution of belief state, we thus consider such an action space. As for belief state space, we let $\mathcal{A}_i$ be endowed with the topology of weak convergence. Let $\mathbf{a}=\{\mathbf{a}_0,\ldots,\mathbf{a}_{N+1}\}$ be the extended joint action. We then define the metric for action space $\mathcal{A}$ as $d(\mathbf{a},\mathbf{a}')=\max_{i\in\mathcal{I}}\{d_P(\mathbf{a}_i,\mathbf{a}_i')\}$, where $d_P(\cdot,\cdot)$ is the Prokhorov metric [24] that induces the weak convergence topology for $\mathcal{A}_i$.

4) Transition Probability: Denote the extended joint action at time k by $\mathbf{A}_k=\{\mathbf{A}_{0,k},\ldots,\mathbf{A}_{N+1,k}\}$. The transition

[4]In the open-loop game, we use θ_1 to represent the type of the sensor. In the closed-loop case, to distinguish from the open-loop game, let τ_k denotes that of the sensor at stage k.

[5]In the remaining paper, we will omit the superscript of S without ambiguity.

function $\mathbf{Q}:\mathcal{B}\times\mathcal{A}\times\mathcal{B}\to[0,1]$ is defined as $\mathbf{Q}(\mathbf{b}'|\mathbf{b},\mathbf{a})\triangleq \mathtt{Pr}(\mathbf{B}_{k+1}=\mathbf{b}'|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})\ \forall k$, i.e., $\mathbf{Q}$ gives the probability of state at the next time conditioned on the current belief state and the joint action.

The update of the belief consists of two parts: *correction* and *prediction*.

Correction: At stage k, based on the actual transmission energy $\alpha^o_{1,k}$ adopted by the sensor, the attacker will correct its *a priori* belief $\mathbf{B}_k$ using Bayes' rule. Let $\mathbf{B}^+_k$ denote the corrected (posterior) belief, which is computed by the following: equation (19–21) as shown at the bottom of this page.

Note that in (19), $\mathbf{a}_m(\alpha^o_1), m\in\{0,\dots,N\}$ is the probability that the mth type player takes action α^o_1.

Prediction: Then, the attacker will predict the belief $\mathbf{B}_{k+1}$ based on the posterior belief $\mathbf{B}^+_k$ and the observed joint action $\alpha^o_k\triangleq\{\alpha^o_{1,k},\alpha^o_{2,k}\}$

$$\mathbf{B}_{k+1}\triangleq\varphi_2(\mathbf{B}^+_k,\alpha^o_k)=(\mathbf{B}^+_k)^\top \mathrm{T}_{\{\alpha^o_{1,k},\alpha^o_{2,k}\}} \tag{22}$$

in which $\alpha^o_{2,k}$ is the actual jamming energy adopted by the attacker, and $\mathrm{T}_{\{\alpha^o_1,\alpha^o_2\}}$ is the transition matrix defined in (12).

In summary, the belief state $\mathbf{B}_k$ transitions deterministically given the public observations, and we can obtain transition probabilities $\mathbf{Q}$ as (20), in which

$$\mathtt{Pr}(\alpha^o_k=\alpha^o|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})=\sum_{i=1}^{N+1} b(i)\mathbf{a}_i(\alpha^o_1)\mathbf{a}_{N+1}(\alpha^o_2).$$

The element of the initial belief state is $\mathbf{b}_0(s_0)=\mathbb{1}_{\{m_0\}}(s_0)\ \forall s_0\in\mathcal{S}$, in which m_0 is the initial state known by each player.

5) Reward: The one-stage reward function for each player $r_i:\mathcal{B}\times\mathcal{A}\to\mathbb{R}$ is given as follows. For $i\le N$, we have

$$\begin{aligned} &r_i(\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})\\ &\triangleq\sum_{\alpha_1\in\mathcal{A}_1^B}\sum_{\alpha_2\in\mathcal{A}_2^B}\mathbf{a}_i(\alpha_1)\mathbf{a}_{N+1}(\alpha_2)r_1^B(i,\alpha_1,\alpha_2). \end{aligned}\tag{23}$$

For the attacker

$$\begin{aligned} &r_{N+1}(\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})\\ &\triangleq\sum_{\alpha_1\in\mathcal{A}_1^B}\sum_{\alpha_2\in\mathcal{A}_2^B}\sum_{i=0}^{N}\mathbf{a}_i(\alpha_1)\mathbf{a}_{N+1}(\alpha_2)\mathbf{b}(i)\tilde{r}_2^B(i,\alpha_1,\alpha_2). \end{aligned}\tag{24}$$

Recall that $r_1^B(\cdot)$ and $\tilde{r}_2^B(\cdot)$ are given in (13) and (15).

Notice that we limit our attention to easy-to-implement *stationary strategies*, which are defined as time-independent mappings from the belief state space into the players' actions, i.e., $\pi:\mathcal{B}\to\mathcal{A}$. We denote by $\pi(\mathbf{b})_{[\alpha_i]}$ the probability given to energy choice $\alpha_i\in\mathcal{A}_1^B$ (or $\alpha_i\in\mathcal{A}_2^B$) by the ith player when the joint strategy π is adopted and the current state is $\mathbf{b}\in\mathcal{B}$.

As for this game, from the long-term viewpoint, the infinite-time discounted payoff of the ith player under the joint *stationary* strategy π is

$$\mathcal{J}_i(\mathbf{b}_0,\pi)=\sum_{k=0}^{\infty}\delta^k r_i(\mathbf{B}_k=\mathbf{b},\pi(\mathbf{b})),\ i\in\{0,\dots,N+1\}$$

in which the parameter $\delta\in[0,1)$ stands for the discount factor. Each player's objective is to maximize its own payoff function. Here, we consider the discounted sum of rewards since we put more weight on the rewards obtained in the early periods.

The belief-based stochastic game at any given time every player knows the belief state. Consequently, the dynamic sensor-attacker game with asymmetric information structure is converted into a stochastic game $\mathcal{G}^S$ with a continuous state space (akin to how POMDPs are converted into continuous belief-space MDPs). Hence, the strategy-design problem for the two agents is equivalent to finding the NE of the stochastic game. Many preliminary works have investigated the existence of stationary equilibria in stochastic games with a finite number of states and actions [21]. However, when extending noncooperative stochastic games to the case where the state and the actions

$$\begin{aligned} \mathbf{B}^+_k(m^+)&\triangleq\varphi_1(\mathbf{B}_k,\mathbf{A}_k,\alpha^o_{1,k})\triangleq\mathtt{Pr}(s_k=m^+|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a},\alpha^o_{1,k}=\alpha^o_1)\\ &=\frac{\mathtt{Pr}(s_k=m^+,\alpha^o_{1,k}=\alpha^o_1|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})}{\sum_{m=0}^{N}\mathtt{Pr}(s_k=m,\alpha^o_{1,k}=\alpha^o_1|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})}\\ &=\frac{\mathtt{Pr}(\alpha^o_{1,k}=\alpha^o_1|s_k=m^+,\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})\mathtt{Pr}(s_k=m^+|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})}{\sum_{m=0}^{N}\mathtt{Pr}(\alpha^o_{1,k}=\alpha^o_1|s_k=m,\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})\mathtt{Pr}(s_k=m|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a})}\\ &=\frac{\mathbf{a}_{m^+}(\alpha^o_1)\mathbf{b}(m^+)}{\sum_{m=0}^{N}\mathbf{a}_m(\alpha^o_1)\mathbf{b}(m)} \end{aligned}\tag{19}$$

$$\mathbf{Q}(\mathbf{b}'|\mathbf{b},\mathbf{a})=\begin{cases}\mathtt{Pr}(\alpha^o_k=\alpha^o|\mathbf{B}_k=\mathbf{b},\mathbf{A}_k=\mathbf{a}), & \text{if } \mathbf{b}'=\varphi_2(\varphi_1(\mathbf{b},\mathbf{a},\alpha^o_1),\alpha^o)\ \text{ for some } \alpha^o\in\mathcal{A}_1^B\times\mathcal{A}_2^B\\ 0, & \text{otherwise}\end{cases}\tag{20}$$

$$\widehat{\mathbf{Q}}(\mathbf{b}'|\mathbf{b},\mathbf{a},\alpha)=\begin{cases}\mathtt{Pr}(\alpha^o_k=\alpha^o|\mathbf{B}_k=\mathbf{b},\alpha_k=\alpha), & \text{if } \mathbf{b}'=\varphi_2(\varphi_1(\mathbf{b},\mathbf{a},\alpha^o_1),\alpha^o)\ \text{ for some } \alpha^o\in\mathcal{A}_1^B\times\mathcal{A}_2^B\\ 0, & \text{otherwise}\end{cases}\tag{21}$$

are in uncountable (e.g., continuous) sets, the existence of stationary equilibria is much harder to access.

Next, we prove the existence of a *stationary* equilibrium for this game. Recall that, an NE is a probability distribution over actions for each player, from which no agent is motivated to deviate unilaterally. We also concentrate on the class of *stationary* policy, and let $\pi^\star \triangleq \{\pi_0^\star, \ldots, \pi_{N+1}^\star\}$ denote the joint *stationary* equilibrium. Hence, at each time k, the players obtain the current belief state $\mathbf{B}_k = \mathbf{b}$ and choose their (transmission/jamming) energies independently. That is to say, the energy choice α_i for the ith player is sampled as a mode of play w.p. $\pi^\star(\mathbf{b})_{[\alpha_i]}$. At a *stationary* NE, no player adopting a meta-strategy (denoted by a transition $\psi(\pi_i^\star)$) can improve its expected payoff, if the others are assumed to play according to the equilibrium policy (denoted by $\pi_{-i}^\star$). Hence, we have the following definition:

Definition 2 (Stationary NE): For this stochastic game $\mathcal{G}^S$, a policy $\pi_i^\star, \forall i \in \mathcal{I}$ is a *stationary* NE if

$$\mathcal{J}_i(\mathbf{B}_0, [\pi_i^\star, \pi_{-i}^\star]) \geq \mathcal{J}_i(\mathbf{B}_0, [\psi(\pi_i^\star), \pi_{-i}^\star]),\ \forall i \in \mathcal{I},\ \forall \mathbf{B}_0 \in \mathcal{B}. \tag{25}$$

The corresponding optimal game value for each player is denoted by $\mathcal{J}_i^\star$. ■

To investigate the existence of stationary equilibrium *strategies* for the players in the continuous-state stochastic game, we have the following results.

Theorem 3: The game $\mathcal{G}^S$ has a *stationary* NE.

Proof: See Appendix D. ■

Even though the existence of a *stationary* NE is proved, it is difficult to build practically a lookup table about the pairs of continuous state and optimal strategy (namely, *stationary* NE). Therefore, we consider the practical implementation in the next section.

V. Practical Implementation

In the previous section, we have shown how Problem 1 with the open-loop history can be simplified into a static Bayesian game and also demonstrated its solution explicitly. When it comes to the closed-loop case, We have proved that the formulated stochastic game $\mathcal{G}^S$ has a *stationary* NE based on the results in [25]. Notice that both the state space and the action space are probability measure spaces, the abstractness of which renders the analysis and implementation of NE quite challenging. In this section, we provide a practical implementation to find the *stationary* NE for each player.

First, we need a method for sampling from the continuous belief state space. In light of the approaches used to solve POMDP problems, we can discretize the state space at the first step. In order to improve the accuracy of discretization, many existing works have used particle filters to represent beliefs over continuous state spaces [26]. Moreover, taking efficient computation into account, exponential family principal components analysis (E-PCA) was proposed to reduce dimensionality of the state space by taking advantage of its sparsity [27]. Here, pursuing the conciseness and tractability of the implementation, we shall discretize the state space with a regular grid, for details see Section VI.

Consequently, this continuous stochastic game can be approximated by discretizing the belief state space. Note that the finite states ensure the existence of the *stationary* NE for the discretized game, which can be directly derived from [28]. Next, we present an algorithm to find the NE of the discretized stochastic game, based on the multiagent Q-learning method [29]. There are many traditional algorithmic techniques for solving stochastic games, such as *value iteration* and *strategy improvement* [30], quadratic programming, etc. These algorithms assume that the environment model parameters about the state transition and the reward function are known; however, such perfect environmental knowledge is not available in many real applications. The proposed multiagent Q-learning method (also called model-free learning) can overcome these limitations.

First, we derive an analog of Bellman's theorem [30] via the Markov property, which is viewed as the set-valued backup operator for the learning algorithm. Given a joint *stationary* equilibrium $\pi^\star$, the expected payoff value for each player for all $\mathbf{b} \in \mathcal{B}$ has the following recursive property:

$$\begin{aligned}\mathcal{J}_i^\star(\mathbf{b}) &\triangleq \mathcal{J}_i(\mathbf{b}, \pi^\star) \\ &= \mathbf{Nash}_i\{Q_0^\star(\mathbf{b}, \mathbf{a}), \ldots, Q_{N+1}^\star(\mathbf{b}, \mathbf{a})\} \end{aligned} \tag{26}$$

$$Q_i^\star(\mathbf{b}, \mathbf{a}) = r_i(\mathbf{b}, \mathbf{a}) + \delta \sum_{\mathbf{b}' \in \mathcal{B}} \mathbf{Q}(\mathbf{b}'|\mathbf{b}, \mathbf{a}) \mathcal{J}_i^\star(\mathbf{b}') \tag{27}$$

in which $Q^\star(\mathbf{b}, \mathbf{a})$ represents the expected cumulative discounted reward of action $\mathbf{a}$ taken in state $\mathbf{b}$ and then obeying the optimal policy $\pi^\star$ afterwards. Note that the Q-value for the ith player is defined over states $\mathbf{b}$ and joint action pairs $\mathbf{a}$. Moreover, $r_i(\mathbf{b}, \mathbf{a})$ is shown in (23) and (24). The discretized state space is also represented by $\mathcal{B}$ without ambiguity. The notation $\mathbf{Nash}_i$ describes the operation that finds the NE point (that is, $\pi^\star(\mathbf{b}) = \arg\max_{\mathbf{a}_i \in \mathcal{A}_i} Q_i^\star(\mathbf{b}, [\mathbf{a}_i, \pi_{-i}^\star(\mathbf{b})])$) and provides the corresponding optimal game value for the ith player. Note that the number of $Q^\star(\mathbf{b}, \mathbf{a})$ is uncountable since $\mathbf{a} \in \mathcal{A}$. To overcome this disadvantage, we abuse the notation $Q_i^\star(\cdot)$ and define Q-value for the ith player over states $\mathbf{b}$ and joint energy action pairs $\alpha_k = \alpha$, denoted by $Q_i^\star(\mathbf{b}, \alpha)$. Note that $\alpha \triangleq \{\alpha_0, \ldots, \alpha_{N+1}\} \in \mathcal{A}_1^B \times \cdots \times \mathcal{A}_2^B$ and $\mathbf{a}(\alpha) = \prod_{i=0}^{N+1} \mathbf{a}_i(\alpha_i)$ since each player takes independent actions. Consequently, (26) is equal to

$$\mathcal{J}_i^\star(\mathbf{b}) = \mathbf{Nash}_i\{Q_0^\star(\mathbf{b}, \alpha), \ldots, Q_{N+1}^\star(\mathbf{b}, \alpha)\}. \tag{28}$$

Note that $Q_i^\star(\mathbf{b}, \mathbf{a}) = \sum_\alpha \mathbf{a}(\alpha) Q_i^\star(\mathbf{b}, \alpha)$. From (27), we construct that

$$Q_i^\star(\mathbf{b}, \alpha) = \widehat{r}_i(\mathbf{b}, \alpha) + \delta \sum_{\mathbf{b}' \in \mathcal{B}} \widehat{\mathbf{Q}}(\mathbf{b}'|\mathbf{b}, \alpha, \mathbf{a}) \mathcal{J}_i^\star(\mathbf{b}')$$

in which

$$\widehat{r}_i(\mathbf{b}, \alpha) = \begin{cases} r_1^B(i, \alpha_i, \alpha_{N+1}), & \text{if } i \leq N, \\ \sum_{i=0}^{N+1} \mathbf{b}(i) \tilde{r}_2^B(i, \alpha_i, \alpha_{N+1}), & \text{otherwise.} \end{cases}$$

and $\widehat{\mathbf{Q}}(\mathbf{b}'|\mathbf{b}, \alpha, \mathbf{a})$ is shown in (21).

Note that the operation $\mathbf{Nash}_i$ in (28) is similar to the definition of the NE in a one-stage game, except that the value $Q_i^\star(\mathbf{b}, \alpha)$ is unknown. A learning process, called a Q-learning algorithm, can be developed to approximate the Q-value through repeated play. The updated equation of the Q-value is developed

Algorithm 1: Nash Q-Learning Algorithm.

Input: Belief state space $\mathcal{B}$, action space $\mathcal{A}$, packet loss function $\hat{f}(\cdot)$, discount δ.
Output: Q-value $Q^\star(\cdot,\cdot)$, NE $\pi^\star$.
Initialization:
$k=0$ and set the initial state s_0 and the belief state $\mathbf{b}_0 \in \mathcal{B}$
Initialize the Q-value $\mathcal{Q}_i(\mathbf{b},\alpha,k)$ for all states $\mathbf{b}$ and arbitrary joint energy actions α, where $i \in \mathcal{I}$
While $||\mathcal{Q}_i(\cdot,\cdot,k+1)-\mathcal{Q}_i(\cdot,\cdot,k)|| < \varepsilon$
For each state $\mathbf{b}$, find an NE (i.e., optimal mixedg strategies) $\pi^\star$ based on (29)
Randomly select the energy actions α based on the optimal mixed strategy profiles
Observe α^o and compute the next state $\mathbf{b}'$ and update the Q-value for each player according to (30)
Update the state: $\mathbf{B}_{k+1}=b'$ and decay the learning rate κ_k
$k := k+1$
End

based on the iteration in (26)

$$\mathcal{J}_i(\mathbf{b},k)=\mathbf{Nash}_i\{Q_0(\mathbf{b},\alpha,k),\ldots,Q^\star_{N+1}(\mathbf{b},\alpha,k)\} \tag{29}$$

$$Q_i(\mathbf{b},\alpha,k+1)=(1-\kappa_k)Q_i(\mathbf{b},\alpha,k) + \kappa_k[\widehat{r}_i(\mathbf{b},\alpha)+\delta\mathcal{J}_i(\mathbf{b}',k)] \tag{30}$$

in which κ_k is the learning rate. With an arbitrary guess at the beginning, the Q-value for each player is updated via using the current reward to amend the historical Q-value. In addition, one of the remaining difficulties in learning NE policies $\pi^\star$ stems from the fact that in general-sum multiplayer one-stage games, multiple Nash equilibria may exist and the equilibrium selection process is out of the scope of this paper. Without considering the efficiency of Nash equilibria, in Section VI, we adopt an equilibrium selection mechanism as an example to learn an NE for each learning stage.

A summarized version of the Nash Q-learning algorithm is given in Algorithm 1. Note that $||\cdot||$ is the induced norm and ε represents the accuracy condition.

As required for the convergence in the general multiagent learning algorithm, the following two conditions should be satisfied.

1) Every state $\mathbf{b}\in\mathcal{B}$ and every joint energy action $\alpha \in \mathcal{A}_1^B\times\cdots\times\mathcal{A}_2^B$ is visited infinitely often during the learning process.
2) The learning rate κ_k satisfies: $\kappa_k\in[0,1), \sum_{k=0}^{\infty}\kappa_k=\infty$, and $\sum_{k=0}^{\infty}\kappa_k^2<\infty$; $\kappa_k(\mathbf{b},\alpha)\neq 0$ if $(\mathbf{b},\alpha)$ is the state-action pair visited at stage k.

These two assumptions are easy to satisfy. Condition 1 can be satisfied using a large number of iterations and samplings. Condition 2 is about the decaying of the learning rate: The first term restricts its convergence, whereas the second term states that the players only update the Q-values that correspond to the current state-action pair. To satisfy it, the learning rate is designed as a nonzero decreasing function of time t and the current state-action pair. The specific representation is provided in the simulation part (see Section VI).

TABLE I
SUMMARY FOR PARAMETERS

Parameters for dynamic system		Channel and discount				Weight	
R	$\mathtt{Tr}[P]$	c	L	n_0	δ	δ_1	δ_2
0.3	1.40	1	2	0.1	0.96	1.5	1

Remark 2: The Nash Q-learning algorithm provably converges to the NE for the general-sum stochastic game if either every stage game during learning has a globally optimal strategy or a saddle point. However, such conditions are not necessary [29]. As shown in many experiments [31], we find the consistent convergence of this algorithm despite violating the condition. With respect to the implementation, we test this algorithm on the stochastic game $\mathcal{G}^S$ under different tuples of parameters, and all the results show the empirical convergence of the Q-value. If needed, we can adopt a new approach but with a high computational complexity to find all *stationary* equilibria of this game [32].

Remark 3: Aiming at learning an equilibrium policy of a stochastic game, we adopt the equilibrium-based multiagent reinforcement learning algorithm. The key idea is to compute an NE of the one-stage game for each belief state and then sample the actions for the players to update the expected payoff functions. Also, we can consider this scenario, in which each player learns by itself (namely, in self-play) and adapts to the others' behavior with the best response. Many pieces of work have modified the reinforcement learning algorithm, for example, via using a variable learning rate [33], to tackle this problem.

Remark 4: In this paper, we assume that the sensor and the attacker have abundant computation abilities and develop the algorithm finding out the optimal strategies. If the computation costs matters, we can adopt the concept of the ϵ-equilibrium to achieve a tradeoff between the system performance and computation budgets by adjusting ϵ. We refer the readers to [34] for more details about computational complexity of ϵ-equilibrium.

VI. EXAMPLES

In this section, we will illustrate the results developed, using some examples. We consider a high-dimensional dynamic system with parameters

$$A=\begin{pmatrix}1 & 0.5\\ 0 & 1.15\end{pmatrix}, C=\begin{pmatrix}0.8 & 0.8\end{pmatrix}, Q=\begin{pmatrix}0.5 & 0\\ 0 & 0.5\end{pmatrix}$$

and other parameters are shown in Table I. Suppose that the communication channel between the sensor and the estimator is wireless fast-fading channel with Gaussian noise n_0; the general form of the $\hat{f}$-function in (5) is $\hat{f}(x)=cx^{-L}$, where c and L are constants dependent on channel characteristics.

In the examples, the energy level set of the sensor is $\mathcal{A}_1^B=\{0.5,\ 0.6\}$ and that of the attacker is $\mathcal{A}_2^B=\{0.1,\ 0.2\}$. Hence, the maximum packet-dropout rate for one transmission is 0.36. As proposed in Section III-A, the index $\mathrm{D}(N)$ is used to measure the performance loss caused by the finite-state approximation. Since $\mathrm{D}(4)<10^{-1}$, to reduce the computation, we impose the restriction that the states set $\mathbb{Z}$ is finite and $N=4$. The learning

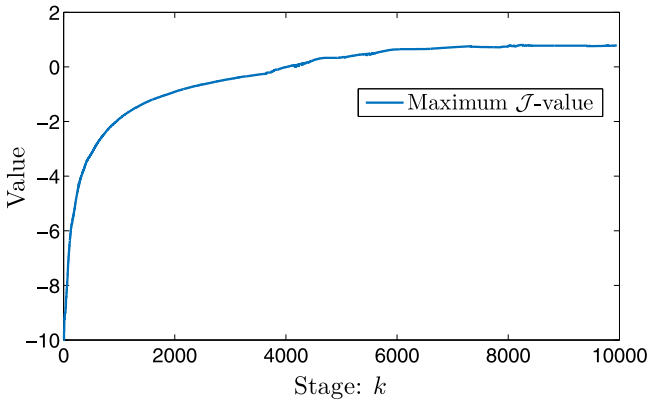


Fig. 4. Converged maximum J-value.

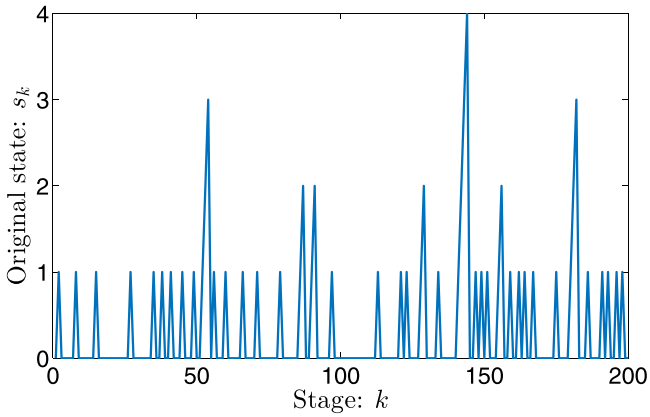


Fig. 5. Iteration of original state s_k.

rate is $\kappa_k = 10/[15 + \text{count}(\mathbf{b}, \alpha)]$, where count$(\cdot)$ is a function to calculate the occurrence of the state-action pair $(\mathbf{b}, \alpha)$ from stage 0 to k. Before the learning process, we discretize the belief state space with resolution rate 0.05, and the amount of discrete belief states is 9267. We employ the algorithm under 10 000 iterations and obtain the following results.

1) *Result 1:* After 10 000 learning stages, we can see that for each player, the procedure converges to an expected payoff value $\mathcal{J}_i^\star(\cdot)$ for each state and a Q-value $Q_i^\star(\cdot,\cdot)$ for each state-action pair. To describe the convergent result for whole states, we take the attacker as an example and adopt a useful statistic, that is, $\max_{\mathbf{b}\in\mathcal{B}} \mathcal{J}_6(\mathbf{b})$. The results are depicted in Fig. 4. Extensive numerical simulations show that the number of iterations needed to converge for our algorithm is around 6000.
2) *Result 2:* In the first 200 stages, the iteration of the original state s_k is shown in Fig. 5, and the corresponding belief state $\mathbf{B}_k$ is depicted in Fig. 6. Note that since the learning process starts with little information, the belief (or probability distribution) cannot capture the fluctuation of state s_k. For example, when k is around 50, $s_k = 3$ but the probability $\texttt{Pr}(s_k = 3)$ shown in Fig. 6 is close to 0.1. But, after a sufficiently large number of steps, sufficient information is available for the attacker to develop an accurate guess about state value. For instance, the probability $\texttt{Pr}(s_k = 3)$ is relatively high when k is about 180 according to the original state $s_k = 3$.
3) *Result 3:* As mentioned previously, Algorithm 1 develops a lookup table of the pairs of discrete belief state and optimal strategies for each of the sensor and the attacker. Taking belief state $\mathbf{b} = [0.75,\ 0.25,\ 0,\ 0,\ 0]$ (which occurs most frequently based on statistical results) as an example, one entry of the lookup table is shown in Table II. If the discretized belief state developed by the sensor/attacker based on their observations is $\mathbf{b} = [0.75\ 0.25\ 0\ 0\ 0]$, the sensor and the attacker will play according to Table II. Specifically, as the sensor has the interim status of the game (i.e., $s_k = i$), it will execute the ith *type agent*'s optimal strategy in Table II. Whereas, the attacker will select the fifth *type agent*'s

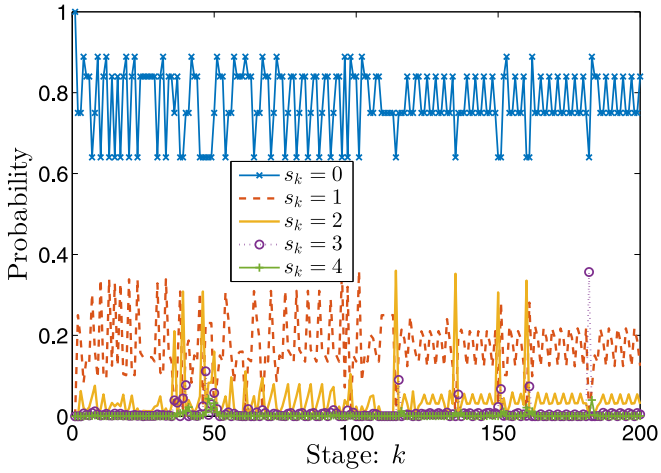


Fig. 6. Iteration of belief state $\mathbf{B}_k$.

TABLE II
OPTIMAL STRATEGIES

Player	Strategy	Transmission Power 0.5	Transmission Power 0.6	Jamming Power 0.1	Jamming Power 0.2
Sensor	0	0.61	0.39	N/A	N/A
	1	0.83	0.17	N/A	N/A
	2	0.06	0.94	N/A	N/A
	3	0.55	0.45	N/A	N/A
	4	0	1	N/A	N/A
Attacker	5	N/A	N/A	1	0

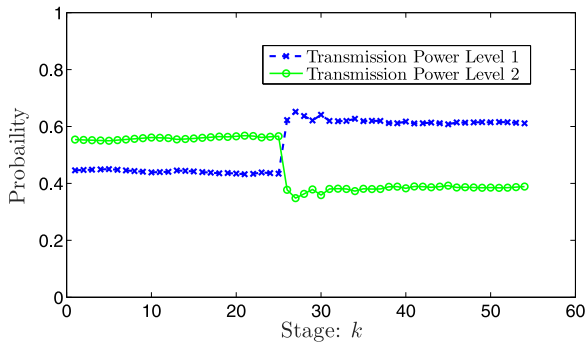


Fig. 7. Iteration of transmission-power strategy for player $i' = 0$ in the last 200 learning stages.

optimal strategy. Moreover, in the last 200 stages, the iteration of the transmission-power scheme for player $i = 0$ (namely, when the original state $s_k = 0$, the sensor's strategy), as demonstrated in Fig. 7, also converges.

VII. Conclusion

This paper has discussed a CPS security issue, where a hostile agent can launch DoS attacks to jam the communication process between a sensor observing a correlated process and a remote estimator. Successfully received data is acknowledged to the sensor, which is perfectly protected from the attacker. The purpose of the attacker is to deteriorate the estimation performance. The interaction between the sensor and the attacker, which has no feedback information from the remote estimator to the sensor, was characterized by a dynamic game under asymmetric information structure. To obtain the optimal strategies for each agent, this game was converted into a continuous-state symmetric-information one and solved by the multiagent reinforcement learning method. Being limited to the simple case of the game between one sensor and one attacker, more research involving multisensors or multiattackers can be further investigated. For instance, how to design an efficient collaboration among a sensor team to avoid DoS attacks. We can also investigate cases where ACKs are randomly observed by both parties.

Appendix A
Proof of Lemma 1

The key point is that for any $k < N$, we have $\mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)] = \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z})]$. Then, $\mathrm{D}(N) \triangleq \sum_{k=0}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k) - \mathbb{E}(P_k|\mathbb{Z})] \leq \sum_{k=N}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)] \to 0$, as $n \to \infty$, which is due to the fact that $\sum_{k=1}^{\infty} \mathtt{Tr}[\mathbb{E}(P_k|\mathbb{Z}_k)] < \infty$. Moreover, based on Proposition 1, we have $\mathrm{D}(N) \geq 0$. Hence, $\lim_{N\to\infty} \mathrm{D}(N) = 0$. ■

Appendix B
Proof of Theorem 1

As proposed by Harsanyi [21], we model this Bayesian game via introducing a *prior* move of nature that determines the type of the attacker.

To interpret the mixed strategies, the sensor in the θth type will transmit the data packet w.p. $q_{1,\theta}$, and the attacker jams the channel w.p. q_2. The reward functions under different actions for each player are given as follows:

$$\begin{aligned} r_1^B(\theta)|_S =& -q_2\mathtt{Tr}[\lambda_1\overline{P} + (1-\lambda_1)h^{\theta+1}(\overline{P})] \\ & - (1-q_2)\mathtt{Tr}[\lambda_2\overline{P} + (1-\lambda_2)h^{\theta+1}(\overline{P})]\} - \delta_1 e_1 \\ r_1^B(\theta)|_{NS} =& -\mathtt{Tr}[h^{\theta+1}(\overline{P})] \end{aligned}$$

in which $r_1^B(\theta)|_S$ represents the reward for sending the packets for the sensor if its type is θ. For the attacker, we have

$$r_2^B|_A = \sum_{\theta\in\mathbb{Z}} \mathbf{B}(\theta)\tilde{f}(\lambda_1,\theta),\ r_2^B|_{NA} = \sum_{\theta\in\mathbb{Z}} \mathbf{B}(\theta)\tilde{f}(\lambda_2,\theta)$$

where $\tilde{f}(\lambda,\theta) = q_{1,\theta}\mathtt{Tr}[\lambda\overline{P} + (1-\lambda)h^{\theta+1}(\overline{P})] + (1-q_{1,\theta})\mathtt{Tr}[h^{\theta+1}(\overline{P})]$.

Notice that a mixed-strategy BNE $(q_{1,\theta}^\star, q_2^\star)$ satisfies the conditions in Definition 1. That is, under $(q_{1,\theta}^\star, q_2^\star)$ the following equations must hold:

$$r_1^B(\theta)|_S = r_1^B(\theta)|_{NS}, \forall\theta \geq 0; \exists\theta \geq 0 : r_2^B(\theta)|_A = r_2^B(\theta)|_{NA}.$$

It is convenient to introduce

$$\begin{aligned} g(q_2,\theta) &\triangleq r_1^B(\theta)|_S - r_1^B(\theta)|_{NS} \\ &= [\lambda_1 q_2 + (1-q_2)\lambda_2]\mathtt{Tr}[h^{\theta+1}(\overline{P}) - \overline{P}] - \delta_1 e_1. \end{aligned}$$

Based on Proposition 1, we can obtain that $\frac{\partial g(q_2,\theta)}{\partial\theta} > 0$ and $\frac{\partial g(q_2,\theta)}{\partial q_2} < 0$. Hence, we can conclude that if $\min_{q_2,\theta} g(q_2,\theta) = g(1,0) > 0$, then there exists a dominant strategy for the sensor $(q_{1,\theta}^\star = 1, \forall\theta)$; otherwise, there exist several pairs $(\tilde{q}_2, m)$ s.t. $g(\tilde{q}_2, m) = 0$ and $q_2^\star = \tilde{q}_2$. The possible value of m is in a finite set, denoted by $\{m_1, m_2, \ldots\}$ (in ascending order), and its corresponding probability values are $\{\tilde{q}_{2,1}, \tilde{q}_{2,2}, \ldots\}$. Obviously, $\tilde{q}_{2,i}$ is monotonically increasing on i.

If $q_2^\star = \tilde{q}_{2,i}$, then the optimal strategy for the sensor is as follows. When $\theta < m_j$, then $g(q_2^\star,\theta) < 0$ and the best choice for the sensor is not sending the packet (i.e., $q_{1,\theta}^\star = 0$); moreover, we have $g(q_2^\star,\theta) > 0$ when $\theta > m_i$, and the sensor is suggested to be active (i.e., $q_{1,\theta}^\star = 1$). Next, we will interpret the computation of $q_{1,m_j}^\star$ and prove that there exists a unique mixed-strategy BNE. We note that $\frac{df(\theta)}{d\theta} > 0$, and obtain that

$$\begin{aligned} r_2^B(\theta)|_A - r_2^B(\theta)|_{NA} &= \sum_{\theta\in\mathbb{Z}} \mathbf{B}(\theta)q_{1,\theta}f(\theta) \\ &= \mathbf{B}(m_j)q_{1,m_j}f(\theta_1) \\ &\quad + \sum_{\theta>m_j} \mathbf{B}(\theta)f(\theta) - \delta_2 e_2. \end{aligned}$$

Therefore, it holds that if $\delta_2 e_2 \in \Sigma_i = [\sum_{\theta=m_j+1}^{n} \mathbf{B}(\theta)f(\theta), \sum_{\theta=m_j}^{n} \mathbf{B}(\theta)f(\theta)]$, then there exists an optimal $q_{1,m_i}^\star$ s.t. $r_2^B(\theta)|_A - r_2^B(\theta)|_{NA} = 0$. Note that the intervals $\Sigma_i\ \forall i$ are disjoined, hence, $\delta_2 e_2$ resides in a single interval Σ_i. Consequently, $q_{1,m_j}^\star$ is unique. ■

Appendix C
Proof of Theorem 2

As discussed previously, the optimal reward for the sensor in game $\mathcal{G}^B$ is obtained by the BNE

$$r_1^{\mathrm{BNE}}(\theta) = \begin{cases} r_1^B(\theta)|_{NS} = -\mathtt{Tr}[h^{\theta+1}(\overline{P})], & \text{if } \theta \leq m \\ r_1^B(\theta)|_S, & \text{otherwise.} \end{cases}$$

Note that the attacker will learn the type information in the normal-form game and adopt type-contingent strategies [i.e., $q_2(\theta)$]. Via analyzing the property of the function $g(q_2,\theta)$, we obtain that $\exists\theta_1 < m$ s.t. $g(q_2,\theta_1) \leq 0,\ \forall q_2 \in [0,1]$, and $\exists\theta_2 > m$ s.t. $g(q_2,\theta_2) \geq 0,\ \forall q_2 \in [0,1]$. Then, we have as follows.

1) if $\theta \leq \theta_1$, then $g(q_2,\theta) < 0$ and the sensor will choose not to transmit in the NE. Hence, $r_1^{\mathrm{NE}}(\theta) = r_1^B(\theta)|_{NS} = r_1^{\mathrm{BNE}}(\theta)$.
2) if $\theta \geq \theta_2$, then $g(q_2,\theta) > 0$ and the sensor will decide to send in the NE. Hence, $r_1^{\mathrm{NE}}(\theta) = r_1^B(\theta)|_S = r_1^{\mathrm{BNE}}(\theta)$.

3) if $\theta \in (\theta_1, m]$, then there exists $q_2^\star(\theta) \in [0,1]$ such that $g(q_2^\star(\theta), \theta) = 0$ [that is, $r_1^B(q_2^\star(\theta), \theta)|_S = r_1^B(\theta)|_{NS}$]. Hence, $r_1^{\text{NE}}(\theta) = r_1^B(\theta)|_{NS} = r_1^{\text{BNE}}(\theta)$.
4) if $\theta \in (m, \theta_2)$, there also exists such $q_2^\star(\theta) \in [0,1]$. Hence, $r_1^{\text{NE}}(\theta) = r_1^B(\theta)|_{NS} < r_1^B(\theta)|_S = r_1^{\text{BNE}}(\theta)$. ■

Appendix D
Proof of Theorem 3

Using [25], to prove Theorem 3, it is sufficient to verify the following conditions.

C1 (state space) $\mathcal{B}$ is a compact metric space.
C2 (action space) $\mathcal{A}_i$ is a compact metric space for every $i \in \mathcal{I}$.
C3 (reward functions) $r_i(\cdot,\cdot)$ is continuous on $\mathcal{B} \times \mathcal{A}$ for every $i \in \mathcal{I}$.
C4 (transition probability) $\mathbf{Q}$ is weakly continuous on $\mathcal{B} \times \mathcal{A}$, i.e., if $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, then $\mathbf{Q}(\cdot|\mathbf{b}^n, \mathbf{a}^n)$ converges weakly[6] to $\mathbf{Q}(\cdot|b, \mathbf{a})$.

We next verify the conditions one by one.

C1 and **C2**: Since $\mathcal{B}$ and $\mathcal{A}_i$ all are probability measure spaces on a finite set, then by [24, Th. 6.4], they are compact metric spaces.

C3: For probability measures $\mu, \mu^n, n \in \mathbb{N}$, we write $\mu^n \xrightarrow{w} \mu$ if μ^n converges weakly to μ. By the definition of the metric defined for the action space $\mathcal{A}$, one sees that as $\mathbf{a}^n \to \mathbf{a}$, also $\mathbf{a}_i^n \xrightarrow{w} \mathbf{a}_i, \forall i \in \mathcal{I}$. Since either $\mathcal{S}, \mathcal{A}_1^B, \mathcal{A}_2^B$ is a finite set, of which each subset is a continuity set, then by the Portmanteau Theorem [24], one obtains that

$$\text{for } 0 \le i \le N: \ \mathbf{a}_i^n \xrightarrow{w} \mathbf{a}_i \iff \mathbf{a}_i^n(\alpha) \to \mathbf{a}_i(\alpha), \forall \alpha \in \mathcal{A}_1^B$$

$$\mathbf{a}_{N+1}^n \xrightarrow{w} \mathbf{a}_{N+1} \iff \mathbf{a}_{N+1}^n(\alpha) \to \mathbf{a}_{N+1}(\alpha), \forall \alpha \in \mathcal{A}_2^B$$

$$\mathbf{b}^n \xrightarrow{w} \mathbf{b} \iff \mathbf{b}^n(i) \to \mathbf{b}(i), \forall 0 \le i \le N$$

where $\iff$ means equivalence. Furthermore, notice that the functions r_1^B and $\tilde{r}_2^B$ are bounded. Then, the dominated convergence theorem yields that as $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, $r_i(\mathbf{b}^n, \mathbf{a}^n) \to r_i(\mathbf{b}, \mathbf{a})$ for every $i \in \mathcal{I}$. The continuity of reward functions is thus verified.

C4: Notice that given current state $\mathbf{b}$ and action $\mathbf{a}$, the possible values of the next state are finite. Then, again by the Portmanteau Theorem, to verify this condition, it suffices to prove that for any $\alpha^o \in \mathcal{A}_1^B \times \mathcal{A}_2^B$, if $(\mathbf{b}^n, \mathbf{a}^n) \to (\mathbf{b}, \mathbf{a})$, then

$$\varphi_2(\varphi_1(\mathbf{b}^n, \mathbf{a}^n, \alpha^o), \alpha^o) \xrightarrow{w} \varphi_2(\varphi_1(\mathbf{b}, \mathbf{a}, \alpha^o), \alpha^o)$$

$$\mathtt{Pr}(\alpha^o|\mathbf{b}^n, \mathbf{a}^n) \to \mathtt{Pr}(\alpha^o|\mathbf{b}, \mathbf{a}).$$

This can be done using very similar arguments to those used for the previous **C3** verification. ■

References

[1] K.-D. Kim and P. Kumar, "Cyber-physical systems: A perspective at the centennial," *Proc. IEEE*, vol. 100, no. special centennial issue, pp. 1287–1308, May 2012.

[2] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Proc. IEEE 28th Int. Conf. Distrib. Comput. Syst. Workshops*, 2008, pp. 495–500.

[3] Case, Defense Use, "Analysis of the cyber attack on the Ukrainian power grid," Electricity Information Sharing and Analysis Center (E-ISAC), 2016.

[4] H. Zhang, P. Cheng, L. Shi, and J. Chen, "Optimal denial-of-service attack scheduling with energy constraint," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 3023–3028, Nov. 2015.

[5] H. Zhang and W. X. Zheng, "Denial-of-service power dispatch against linear quadratic control via a fading channel," *IEEE Trans. Autom. Control*, vol. 63, no. 9, pp. 3032–3039, Sep. 2018.

[6] A. Agah, S. K. Das, and K. Basu, "A game theory based approach for security in wireless sensor networks," in *Proc. IEEE Conf. Perform., Comput. Commun.*, 2004, pp. 259–263.

[7] C. Langbort and V. Ugrinovskii, "One-shot control over an AVC-like adversarial channel," in *Proc. IEEE Conf. Amer. Control Conf.*, Jun. 2012, pp. 3528–3533.

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

[9] Y. Li, D. E. Quevedo, S. Dey, and L. Shi, "SINR-based DoS attack on remote state estimation: A game-theoretic approach," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 3, pp. 632–642, Sep. 2017.

[10] Y. He, H. Li, X. Cheng, Y. Liu, and L. Sun, "A bitcoin based incentive mechanism for distributed P2P applications," in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl.*, 2017, pp. 457–468.

[11] A. K. Khandani, "Full duplex wireless transmission with channel phase-based encryption," U.S. Patent 9 572 038, Feb. 14, 2017.

[12] P. Hovareshti, V. Gupta, and J. S. Baras, "Sensor scheduling using smart sensors," in *Proc. IEEE 46th Annu. Conf. Decis. Control*, 2007, pp. 494–499.

[13] B. Anderson and J. Moore, *Optimal Filtering*. Upper Saddle River, NJ, USA: Prentice-Hall, 1979.

[14] L. Shi, K. H. Johansson, and L. Qiu, "Time and event-based sensor scheduling for networks with limited communication resources," in *Proc. 18th IFAC World Congr.*, 2011, pp. 13 263–13 268.

[15] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[16] H. Urkowitz, "Energy detection of unknown deterministic signals," *Proc. IEEE*, vol. 55, no. 4, pp. 523–531, Apr. 1967.

[17] F. F. Digham, M.-S. Alouini, and M. K. Simon, "On the energy detection of unknown signals over fading channels," *IEEE Trans. Commun.*, vol. 55, no. 1, pp. 21–24, Jan. 2007.

[18] M. Wilhelm, I. Martinovic, J. B. Schmitt, and V. Lenders, "Short paper: reactive jamming in wireless networks: How realistic is the threat?" in *Proc. 4th ACM Conf. Wireless Netw. Secur.*, 2011, pp. 47–52.

[19] D. Nguyen, C. Sahin, B. Shishkin, N. Kandasamy, and K. R. Dandekar, "A real-time and protocol-aware reactive jamming framework built on software-defined radios," in *Proc. ACM Workshop Softw. Radio Implementation Forum*, 2014, pp. 15–22.

[20] A. Haurie, J. B. Krawczyk, and G. Zaccour, *Games and Dynamic Games*. Singapore: World Scientific, 2012.

[21] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[22] L. Shi, P. Cheng, and J. Chen, "Sensor data scheduling for optimal state estimation with communication energy constraint," *Automatica*, vol. 47, no. 8, pp. 1693–1698, 2011.

[23] E. J. Sondik, "The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs," *Oper. Res.*, vol. 26, no. 2, pp. 282–304, 1978.

[24] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 2013.

[25] M. J. Sobel, "Continuous stochastic games," *J. Appl. Probability*, vol. 10, pp. 597–604, 1973.

[26] D. Silver and J. Veness, "Monte-Carlo planning in large POMDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 2164–2172.

[27] N. Roy and G. J. Gordon, "Exponential family PCA for belief compression in POMDPs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1635–1642.

[28] W. Whitt, "Representation and approximation of noncooperative sequential games," *SIAM J. Control Optim.*, vol. 18, no. 1, pp. 33–48, 1980.

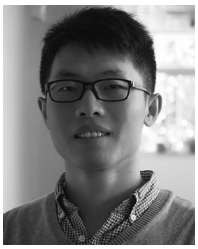
[29] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[6] Interested readers are referred to [24] to see details of weak convergence of probability measures.

[30] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1. Belmont, MA, USA: Athena Scientific, 2005.
[31] A. Greenwald, K. Hall, and R. Serrano, "Correlated Q-learning," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, vol. 3, pp. 242–249.
[32] R. N. Borkovsky, U. Doraszelski, and Y. Kryukov, "A user's guide to solving dynamic stochastic games using the homotopy method," *Oper. Res.*, vol. 58, no. 4-part-2, pp. 1116–1132, 2010.
[33] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artif. Intell.*, vol. 136, no. 2, pp. 215–250, 2002.
[34] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*, vol. 1. Cambridge, MA, USA: Cambridge Univ. Press, 2007.

Kemi Ding received the B.S. degree in electronic and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2014. From September 2016 to December 2016, she was a Visiting Student with the School of Engineering and Applied Sciences in Harvard University, Cambridge, MA, USA. She received the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2018.

She is currently a Postdoctoral Researcher with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ, USA. Her current research interests include cyber-physical system security/privacy, networked state estimation, and wireless sensor networks.



Xiaoqiang Ren received the B.E. degree in control science and engineering, from Zhejiang University, Hangzhou, China, in 2012 and the Ph.D. degree in electronic and computer engineering, from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2016.

He is currently a Postdoctoral Researcher with the Department of Automatic, KTH Royal Institute of Technology, Stockholm, Sweden. Prior to this, he was a Postdoctoral Researcher with the Hong Kong University of Science and Technology, Hong Kong, from September to November 2016, and Nanyang Technological University, Singapore, from December 2016 to February 2018. His research interests include security of cyber-physical systems, sequential decision, and networked estimation and control.

Daniel E. Quevedo (S'97–M'05–SM'14) received Ingeniero Civil Electrónico and M.Sc. degrees from the Universidad Técnica Federico Santa María, Valparasío, Chile, in 2000. He received the Ph.D. degree from the University of Newcastle in Australia, Callaghan, N.S.W., Australia, in 2005. He was supported by a full scholarship from the alumni association during his time at the Universidad Técnica Federico Santa María and received several university-wide prizes upon graduating.

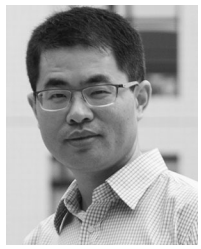
He is the Head of the Chair of Automatic Control (*Regelungs- und Automatisierungstechnik*) with Paderborn University, Paderborn, Germany. His research interest focuses on the control of networked systems and power converters.

Dr. Quevedo was the recipient of the IEEE Conference on Decision and Control Best Student Paper Award in 2003, and was also a finalist in 2002. In 2009, he was awarded a five-year research fellowship from the Australian Research Council. He is an Associate Editor for the IEEE Control Systems Magazine, Editor for the *International Journal of Robust and Nonlinear Control*, and serves as a Chair of the IEEE Control Systems Society *Technical Committee on Networks & Communication Systems*.

Subhrakanti Dey received the B.Tech. and M.Tech. degrees from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 1991 and 1993, respectively, and the Ph.D. degree from Australian National University, Canberra, A.C.T., Australia, in 1996.

He is currently a Professor with the Department of Engineering Sciences in Uppsala University, Uppsala, Sweden. His current research interests include networked control systems, wireless communications and networks, signal processing for sensor networks, and stochastic and adaptive signal processing and control.

Dr. Dey currently serves on the Editorial Board of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS. He was also an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING during 2007–2010 and the IEEE TRANSACTIONS ON AUTOMATIC CONTROL during 2004–2007, and Associate Editor for *Elsevier Systems and Control Letters* during 2003–2013.



Ling Shi received the B.S. degree in electrical and electronic engineering from Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2002, and the Ph.D. degree in control and dynamical systems from California Institute of Technology, Pasadena, CA, USA, in 2008.

He is currently an Associate Professor with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology. His research interests include cyber-physical systems security, networked control systems, sensor scheduling, and event-based state estimation.

Dr. Shi served as an Editorial Board Member for The European Control Conference 2013–2016. He has been serving as a Subject Editor for the *International Journal of Robust and Nonlinear Control* from March 2015, an Associate Editor for the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS from July 2016, and an Associate Editor for the IEEE CONTROL SYSTEMS LETTERS from February 2017. He also served as an Associate Editor for a special issue on secure control of cyber physical systems in the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS in 2015–2017. He serves as the General Chair of the 23rd International Symposium on Mathematical Theory of Networks and Systems (MTNS 2018).