Attack Allocation on Remote State Estimation in Multi-Systems: Structural Results and Asymptotic Solution

Xiaoqiang Ren [a], Junfeng Wu [b], Subhrakanti Dey [c], Ling Shi [a]

[a] *Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong.*

[b] *ACCESS Linnaeus Center, School of Electrical Engineering, Royal Institute of Technology, Stockholm, Sweden.*

[c] *Signals and Systems Division, Department of Engineering Sciences, Uppsala University, Uppsala, Sweden*

Abstract

This paper considers optimal attack attention allocation on remote state estimation in multi-systems. Suppose there are `M` independent systems, each of which has a remote sensor monitoring the system and sending its local estimates to a fusion center over a packet-dropping channel. An attacker may generate noises to exacerbate the communication channels between sensors and the fusion center. Due to capacity limitation, at each time the attacker can exacerbate at most `N` of the `M` channels. The goal of the attacker side is to seek an optimal policy maximizing the estimation error at the fusion center. The problem is formulated as a Markov decision process (MDP) problem, and the existence of an optimal deterministic and stationary policy is proved. We further show that the optimal policy has a threshold structure, by which the computational complexity is reduced significantly. Based on the threshold structure, a myopic policy is proposed for homogeneous models and its optimality is established. To overcome the curse of dimensionality of MDP algorithms for general heterogeneous models, we further provide an asymptotically (as `M` and `N` go to infinity) optimal solution, which is easy to compute and implement. Numerical examples are given to illustrate the main results.

Key words: Attack; state estimation; Kalman filtering; structural results; Markov decision process; multi-armed bandit

1 Introduction

Motivations and backgrounds. Cyber-physical systems, integrating information technology infrastructures with physical processes, are ubiquitous and usually critical in modern societies. Examples include sensor networks, power grids, water and gas supply systems, transportation systems, water pollution monitoring systems. The use of open communication networks, though enabling more efficient design and flexible implementation, makes cyber-physical systems more vulnerable to attacks Teixeira et al. [2015], Pasqualetti et al. [2015]. Illustrative examples are Iran's nuclear centrifuges accident Farwell and Rohozinski [2011] and western Ukraine blackout BBC [2016].

Many research works on attackers' possible behaviors for cyber-physical systems have been done recently. Generally speaking, attacks can be classified as either denial of service (DoS) attacks or deception attacks Amin et al. [2009]. DoS attacks, comprising availability of data, are most likely threats Byres and Lowe [2004] due to their easy implementation. DoS attacks in networked control systems are studied in Amin et al. [2009]. Optimal offline DoS attack on remote state estimation over a finite horizon for a single sensor system is investigated in Zhang et al. [2015]. An interactive decision of sending data by sensor and jamming channel by an attacker for remote state estimation in a zero-sum game setting is studied in Li et al. [2015], and a similar setting is investigated for a control system in Gupta et al. [2010]. Optimal DoS attacks were also studied in the context of detection Ren et al. [2014a]. Deception attacks, comprising integrity of data, are more subtle. Various types of deception attacks have been studied, for example, replay attacks Mo and Sinopoli [2009], stealthy deception attacks Guo et al. [2016] and covert attacks Teixeira et al. [2012].

Related works and contributions. In this paper, we consider the DoS attacks. Each sensor monitors a (different)

Email addresses: `xren@connect.ust.hk` (Xiaoqiang Ren), `junfengw@kth.se` (Junfeng Wu), `subhrakanti.dey@signal.uu.se` (Subhrakanti Dey), `eesling@ust.hk` (Ling Shi).

system and sends its estimates to a fusion center over a packet-dropping channel. An attacker is present and is capable of attack a certain number of channels at each time. When a channel is under attack, the packet arrival rate decreases. The problem is to study the optimal attack policy to maximize the averaged estimation error at the fusion center. A threshold structure of optimal policies is proved. The related works are Mo et al. [2012], Ren et al. [2014b], Leong et al. [2015], which study the structure of sensor scheduling policy. Our work differs from these works as follows. First, our work focuses on multi-systems, while a single sensor scenario is studied in aforementioned three papers. Second, we use a fundamentally different methodology. Specifically, both Mo et al. [2012] and Ren et al. [2014b] proved the structure results by analyzing the stationary probability distribution of states, which, however, works only in very special and simple cases (e.g., a single sensor case). On the contrary, we resort to the MDP theory, a more general and powerful tool. Although an MDP approach was also adopted in Leong et al. [2015], the methods used to prove either the existence of optimal stationary and deterministic policy or the threshold structure are significantly different due to the different problem models (multi-systems versus single sensor system, different cost/reward structures [1]). Lastly, we provide an asymptotically optimal policy, which is rather easy to compute and implement.

In summary, the main contributions of this paper are as follows.

(1) The problem of attack on remote state estimation in multi-systems is studied by an MDP formulation. The existence of a deterministic and stationary optimal policy is proved, which means that standard MDP algorithms (e.g., value iteration algorithm) can be utilized to compute the optimal policy. Moreover, a threshold structure of optimal policy is proved, by exploiting which a specialized algorithm may be developed to reduce the computational complexity. By the threshold structure, a myopic policy is proposed and its optimality is established for homogeneous models. The myopic policy is such that the expected reward at the next time is maximized.
(2) To overcome the curse of dimensionality of MDP algorithms for general heterogeneous models, we provide an asymptotically optimal index-based policy using the multi-armed bandit theory. Since the indices are computed based on each system *solely*, they are quite easy to compute. The index-based policy is implemented just by comparing these indices. What is more, our numerical examples show that this asymptotically optimal policy works quite well even when the number of total systems is small.

[1] See the details in Footnote 8.

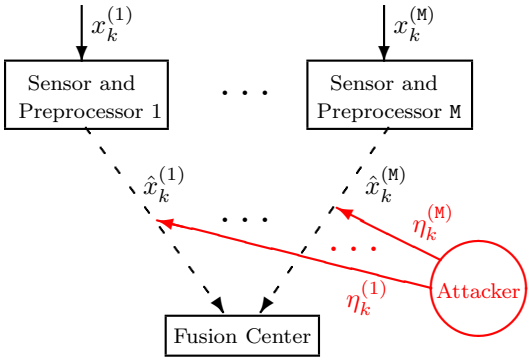

Fig. 1. Remote state estimation with an attacker.

The remainder of this paper is organized as follows. In Section 2, the mathematical formulation of the considered problem is given. The main results, including the MDP formulation, existence of a stationary and deterministic optimal policy, threshold structure of the optimal policy and the asymptotically optimal index-based policy, are provided in Section 3. Numerical examples are given in Section 4 to illustrate the main results, after which we conclude the paper in Section 5. All the proofs are presented in Appendices.

Notation: $\mathbb{R}$ ($\mathbb{R}_+$) is the set of real (nonnegative) numbers and $\mathbb{N}$ the set of nonnegative integer numbers. $\mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$) is the set of n by n real positive semi-definite (definite) matrices. For a matrix X, we use $\mathrm{Tr}(X), X^\top$ and $|X|$ to denote its trace, transpose and spectral radius, respectively. We write $X \succeq 0$ ($X \succ 0$) if $X \in \mathbb{S}_+^n$ ($X \in \mathbb{S}_{++}^n$). For a vector x, denote its i-th element as $x_{[i]}$. We use $\circ$ to denote function composition, i.e., for two functions f and g, $(f \circ g)(x) = f(g(x))$, and $g^i(x) \triangleq \underbrace{g \circ g \circ \cdots \circ g}_{i \text{ times}}(x)$ with $g^0(x) \triangleq x$. Let $\times$ denote Cartesian product. For a set $\mathbb{A}$, define the indicator function as $\mathbf{1}_{\mathbb{A}}(x) = 1$, if $x \in \mathbb{A}$; 0 otherwise. Let $\boldsymbol{Pr}(\cdot)(\boldsymbol{Pr}(\cdot|\cdot))$ be the (conditional) probability. For $x \in \mathbb{R}$, denote by $\lfloor x \rfloor$ the largest integer less than or equal to x. Let $\boldsymbol{E}[\cdot]$ be the expectation of a random variable.

2 Problem Formulation

2.1 Remote Estimation with Packet-dropping Channels

There are totally $\mathtt{M}$ independent discrete-time (i.e., sampled) linear time-invariant systems and $\mathtt{M}$ sensors. The i-th sensor monitors the i-th system (Fig. 1):

$$x_{k+1}^{(i)} = A_i x_k^{(i)} + \omega_k^{(i)}, \tag{1a}$$
$$y_k^{(i)} = C_i x_k^{(i)} + \upsilon_k^{(i)}, \tag{1b}$$

where $x_k^{(i)} \in \mathbb{R}^{n_i}$ is the system state vector and $y_k^{(i)} \in \mathbb{R}^{m_i}$ is the observation vector. The noises $\omega_k^{(i)}$ and $\upsilon_k^{(i)}$ are i.i.d. white Gaussian random variables with zero mean and covariance $Q_i \succeq 0, R_i \succ 0$, respectively. The initial state $x_0^{(i)}$ is a zero-mean Gaussian random variable that is uncorrelated with $\omega_k^{(i)}$ and $\upsilon_k^{(i)}$. It is assumed that the systems at different sensors are independent of each other. To avoid trivial problems, we assume the systems are unstable, i.e., $|A_i| > 1, \forall i = 1, \ldots, \mathtt{M}$. The pair (C_i, A_i) is assumed to be detectable and $(A_i, Q_i^{1/2})$ stabilizable.

Each sensor is assumed to be intelligent in the sense that a Kalman filter is run locally. With the above detectability and stabilizability assumptions, the estimation error covariance associated with each local Kalman filter converges exponentially to a steady state Anderson and Moore [2012]. On the other hand, since the nature of asymptotic behaviors of remote estimation under malicious attacks (which will be elaborated later) over an infinite horizon cost is investigated, without any performance loss, we assume the Kalman filter at each sensor enters into the steady state at $k = 0$. Let the steady state estimation error covariance at sensor i be $\hat{P}^{(i)}$.

At each time k, sensor i sends the output of its local Kalman filter (i.e., the *a posterior* minimum mean square error (MMSE) estimate) $\hat{x}_k^{(i)}$ Anderson and Moore [2012] to a fusion center over a packet-dropping communication channel. Let $\gamma_k^{(i)} \in \{0, 1\}$ denote whether or not the packet is received error-free by the fusion center. If it arrives successfully, $\gamma_k^{(i)} = 1$; $\gamma_k^{(i)} = 0$ otherwise. Again since the asymptotic behavior over an infinite horizon is studied, it is assumed without any performance loss that $\gamma_0^{(i)} = 1, \forall i = 1, \ldots, \mathtt{M}$. Since the sensor sends the local MMSE estimates instead of raw measurements, the MMSE estimate and the associated error covariance at the fusion center (whether or not the attacker introduced later is present) for $k \geq 1$ is:

$$\tilde{x}_k^{(i)} = \begin{cases} \hat{x}_k^{(i)}, & \text{if } \gamma_k^{(i)} = 1, \\ A_i \tilde{x}_{k-1}^{(i)}, & \text{if } \gamma_k^{(i)} = 0, \end{cases}$$

$$\tilde{P}_k^{(i)} = \begin{cases} \hat{P}^{(i)}, & \text{if } \gamma_k^{(i)} = 1, \\ h_i(\tilde{P}_{k-1}^{(i)}), & \text{if } \gamma_k^{(i)} = 0, \end{cases}$$

where functions $h_i, 1 \leq i \leq \mathtt{M}$, are defined as follows:

$$h_i(X) = A_i X A_i^\top + Q_i, \quad \text{for } X \in \mathbb{S}_+^{n_i}.$$

Notice that by the assumption $\gamma_0^{(i)} = 1, \forall i$, the starting point at the fusion center is: $\tilde{x}_0^{(i)} = \hat{x}_0^{(i)}$ and $\tilde{P}_0^{(i)} = \hat{P}^{(i)}$.

2.2 Attack Model

There is an attacker capable of generating noises to exacerbate the communication channels between sensors and the fusion center. Due to capacity limitation, at each time the attacker can only choose at most $\mathtt{N}$ of the $\mathtt{M}$ channels to attack. Let $\eta_k^{(i)} \in \{0, 1\}$ indicate whether or not the i-th channel is under attack: $\eta_k^{(i)} = 1$ if it is; $\eta_k^{(i)} = 0$ otherwise. We make the following assumption about the effects of the attacks on packet dropouts.

Assumption 1 *The packet loss process is memoryless with respect to the considered attacks, i.e., the following equality holds for any $k \geq 1$:*

$$\boldsymbol{Pr}(\gamma_1^{(i)}, \ldots, \gamma_k^{(i)} | \eta_{1:k}^{(i)}) = \prod_{j=1}^{k} \boldsymbol{Pr}(\gamma_j^{(i)} | \eta_j^{(i)}),$$

where $\eta_{1:k}^{(i)} \triangleq (\eta_1^{(i)}, \ldots, \eta_k^{(i)})$. Let $\boldsymbol{Pr}(\gamma_k^{(i)} = 1 | \eta_k^{(i)} = 0) = \epsilon_i$ and $\boldsymbol{Pr}(\gamma_k^{(i)} = 1 | \eta_k^{(i)} = 1) = \underline{\epsilon_i}$. We assume that $0 < \underline{\epsilon_i} < \epsilon_i \leq 1$.

It is assumed that the attacker has the knowledge of system dynamics (i.e., A_i, C_i, Q_i and R_i [2]), has access to the knowledge of $\{\gamma_k^{(i)}\}_{k \in \mathbb{N}}, \forall i = 1, \ldots, \mathtt{M}$, and is able to learn the channels' packet arrival rate with or without attacks (i.e., $\underline{\epsilon_i}$ and ϵ_i) from realization of $\{\gamma_k^{(i)}\}_{k \in \mathbb{N}}$. At each time, the attacker determines the subset of the communication channels to be attacked based on all the information it collects. Let $\gamma_k = (\gamma_k^{(1)}, \ldots, \gamma_k^{(\mathtt{M})})$ and $\gamma_{1:k} = (\gamma_1, \ldots, \gamma_k)$; η_k and $\eta_{1:k}$ are defined in the same way. Define a feasible attack attention allocation decision rule at time k as a stochastic kernel π_k from $\gamma_{1:k-1}$ and $\eta_{1:k-1}$ to Ω [3], where Ω is the set of all feasible η_k:

$$\Omega \triangleq \left\{ \eta \in \{0, 1\}^{\mathtt{M}} : \sum_{i=1}^{\mathtt{M}} \eta_{[i]} \leq \mathtt{N} \right\}.$$

[2] The steady state estimation error covariance $\hat{P}^{(i)}$ thus can be obtained by solving a discrete-time algebraic Riccati equation.

[3] We say π_k is a stochastic kernel from $\gamma_{1:k-1}$ and $\eta_{1:k-1}$ to Ω if the map $\pi_k : \wp(\Omega) \times \{0, 1\}^{\mathtt{M}(k-1)} \times \Omega^{k-1} \mapsto [0, 1]$ with $\wp(\Omega)$ being the power set of Ω has the following properties:

(1) For any realization of $\gamma_{1:k-1} \in \{0, 1\}^{\mathtt{M}(k-1)}$ and $\eta_{1:k-1} \in \Omega^{k-1}$, $\pi_k(\cdot | \gamma_{1:k-1}, \eta_{1:k-1})$ is a probability measure on $\wp(\Omega)$.

(2) For any set $\mathbb{B} \in \wp(\Omega)$, $\pi_k(\mathbb{B} | \cdot)$ is a measurable function on $\{0, 1\}^{\mathtt{M}(k-1)} \times \Omega^{k-1}$.

This kernel-form definition includes the possibility that the attack policy is randomized. Nevertheless, in Section 3 we prove that there exists a deterministic optimal attack policy.

Let $\pi = (\pi_1, \ldots, \pi_k, \ldots)$ be the infinite-horizon attack policy. A policy π is feasible only if $\pi_k, k \geq 1$ are feasible. Let Π be the set of all feasible policies. The reward (from the perspective of the attacker) associated with an attack policy π is the averaged infinite-horizon estimation error at the centers defined as

$$\boldsymbol{R}(\pi) = \liminf_{\mathtt{T}\to\infty} \frac{1}{\mathtt{T}} \boldsymbol{E}\left[\sum_{k=1}^{\mathtt{T}} \sum_{i=1}^{\mathtt{M}} \mathrm{Tr}(\tilde{P}_k^{(i)})\right]. \tag{2}$$

The goal of the attacker is to seek a feasible policy maximizing the above reward:

Problem 1

$$\sup_{\pi\in\Pi} \boldsymbol{R}(\pi). \tag{3}$$

To avoid trivial problems, we assume $\underline{\epsilon_i} > 1 - \frac{1}{|A_i|^2}, \forall i$. Otherwise, the attacker may consistently attack the communication channel of the i-th system to gain an infinite reward since $\tilde{P}_k^{(i)} \to \infty$ as $k \to \infty$ in the presence of consistent attacks.

3 Main Results

In this section, we solve Problem 1 by formulating it as a MDP problem. We show that, without any performance loss, the attack decision rule can be restricted to a smaller class: the optimal policy is deterministic (i.e., the stochastic kernel π_k is reduced to a measurable function), stationary (independent of time index k) and Markovian (the argument is not the whole history $\gamma_{1:k-1}$). We further prove that the optimal policy has a threshold structure. For the asymptotic regime (i.e., $\mathtt{M} \to \infty$ and $\mathtt{N} \to \infty$), an explicit form of the optimal policy is provided, which is quite easy to compute and implement.

3.1 MDP Formulation

Before proceeding, we define a random variable $\tau_k^{(i)}$ as

$$\tau_k^{(i)} = k - \max\{k^* : \gamma_{k^*}^{(i)} = 1, 0 \leq k^* \leq k\},$$

which indicates the time duration from the last successful transmission time to time k. Let $\tau_k = (\tau_k^{(1)}, \ldots, \tau_k^{(\mathtt{M})})$.

For ease of exposition, except for the myopic policy and asymptotic analysis, in the remainder of this section we assume that $\mathtt{M} = 2$ and $\mathtt{N} = 1$. We remark that the following MDP formulation and the existence of a deterministic and stationary optimal policy (Theorem 1) can be extended trivially to the cases with general $\mathtt{M}$ and $\mathtt{N}$. While for the threshold structure, see Remark 1.

Now we describe the formulated infinite-horizon discrete-time MDP by a quadruplet $(\mathbb{S}, \mathbb{A}, \boldsymbol{P}(\cdot|\cdot,\cdot), r(\cdot,\cdot))$. Each item in the tuple is elaborated as follows.

(1) The state at time step $k \geq 1$ is defined as $s_k \triangleq (\tau_{k-1}^{(1)}, \tau_{k-1}^{(2)})$. Therefore, the state space $\mathbb{S} = \mathbb{N}^2$.

(2) The action space $\mathbb{A} \triangleq \{\mathbf{0}, e_1, e_2\}$, where $\mathbf{0} = (0,0)$ means that none of the systems is attacked, $e_1 = (1,0)$ and $e_2 = (0,1)$ means that *only* the first and *only* the second is attacked, respectively.

(3) The transition probability is stationary. Let $s = (j_1, j_2), s' = (j_1', j_2')$ with $j_i, j_i' \in \mathbb{N}, i = 1,2$ and $a \in \mathbb{A}$, then $\forall k \geq 1$,

$$\begin{aligned}\boldsymbol{P}(s'|s,a) &\triangleq \boldsymbol{Pr}(s_{k+1} = s'|s_k = s, a_k = a)\\ &\triangleq p_1(j_1'|j_1, a_{[1]}) p_2(j_2'|j_2, a_{[2]}),\end{aligned}$$

where for $i = 1, 2$,

$$p_i(j_i'|j_i, a_{[i]}) = \begin{cases} \epsilon_i, & \text{if } j_i' = 0, a_{[i]} = 0,\\ \underline{\epsilon_i}, & \text{if } j_i' = 0, a_{[i]} = 1,\\ 1-\epsilon_i, & \text{if } j_i' = j_i + 1, a_{[i]} = 0,\\ 1-\underline{\epsilon_i}, & \text{if } j_i' = j_i + 1, a_{[i]} = 1,\\ 0, & \text{otherwise.}\end{cases}$$

(4) The one-stage reward is independent of the action and defined as

$$r(s = (j_1, j_2), a) = \mathrm{Tr}(h_1^{j_1}(\hat{P}^{(1)})) + \mathrm{Tr}(h_2^{j_2}(\hat{P}^{(2)})). \tag{4}$$

Let $\mathbb{H}_k \triangleq (s_1, a_1, \ldots, s_k)$ be the history of states and actions up to time k, and $\theta = (\theta_1, \ldots, \theta_k, \ldots)$ be an admissible policy with θ_k as a stochastic kernel from $\mathbb{H}_k$ to $\mathbb{A}$. Let Θ be the class of all such admissible policies. Define the reward associated with initial state $s_1 = s$ and policy θ by

$$\boldsymbol{J}(s,\theta) = \liminf_{\mathtt{T}\to\infty} \frac{1}{\mathtt{T}} \boldsymbol{E}_s^{\theta}\left[\sum_{k=1}^{\mathtt{T}} r(s_k, a_k)\right].$$

Let $s_{1:k} \triangleq (s_1, \ldots, s_k)$. It is evident that $s_{1:k-1}$ is equivalent to $\gamma_{1:k-1}$, and thus θ is also equivalent to π (specialized to the case $\mathtt{M} = 2, \mathtt{N} = 1$). One thus verifies that Problem 1 (specialized to the case $\mathtt{M} = 2, \mathtt{N} = 1$) can be equivalently transformed to the following problem.

Problem 2 *Find the optimal policy $\theta^* \in \Theta$ such that*

$$\boldsymbol{J}((0,0), \theta^*) = \sup_{\theta\in\Theta} \boldsymbol{J}((0,0), \theta).$$

3.2 Structural Results

We first show that the optimal policy is stationary and deterministic, and satisfies an equality. We say that $\theta = (\theta_1, \ldots, \theta_k, \ldots)$ is stationary and deterministic, if there exists a measurable function $f : \mathbb{S} \mapsto \mathbb{A}$ satisfying $\forall k \geq 1$, $\theta_k(f(s)|\mathbb{H}'_k) = 1$ for any $\mathbb{H}'_k \triangleq (s_1, a_1, \ldots, s_k = s)$. Therefore, in the following, with abuse of notations, we use f to represent a stationary and deterministic policy and let $\mathbb{F}$ be the set of all admissible stationary and deterministic policies. For a measurable function $q : \mathbb{S} \mapsto \mathbb{R}$, denote

$$\boldsymbol{G}(q, s, a) \triangleq \sum_{s' \in \mathbb{S}} q(s')\boldsymbol{P}(s'|s, a). \tag{5}$$

We then have the following theorem.

Theorem 1 *There exists an optimal stationary and deterministic policy $f^* \in \mathbb{F}$ such that*

$$\boldsymbol{J}(s, f^*) \geq \boldsymbol{J}(s, \theta), \quad \forall s \in \mathbb{S}, \theta \in \Theta.$$

Moreover,

$$\begin{aligned} f^*(s) =& \arg\max_{a \in \mathbb{A}} \{r(s, a)) - \varrho^* + \boldsymbol{G}(q, s, a)\}, \qquad (6)\\ \boldsymbol{J}(s, f^*) =& \varrho^*, \end{aligned}$$

where $q : \mathbb{S} \mapsto \mathbb{R}$ and $\varrho^* \in \mathbb{R}$ satisfy

$$q(s) = \max_{a \in \mathbb{A}} \{r(s, f(s)) - \varrho^* + \boldsymbol{G}(q, s, a)\}. \tag{7}$$

Theorem 1 says that deterministic and stationary optimal policy exists and can be computed as (6) with a differential value function (i.e., $q(s)$) satisfying the Bellman equation (7). This provides a theoretic basis for further analysis (structural properties of optimal policies) and computation methods. In particular, with some additional technical requirements [4], the value iteration algorithm converges. Furthermore, following the ideas in [Sennott, 2009, Chapter 8], one can use a value iteration algorithm for finite states to approximate the countable state space in our case, and compute the optimal policy f^*, the differential value function q and the optimal averaged reward ϱ^*.

We now present a nice structure of the optimal policy f^*, which helps reduce the computational complexity of the MDP algorithm significantly.

[4] One may verify that all requirements in [Zhu and Guo, 2005, Assumption 3.8] are satisfied in our case. Due to the limited space, we omit the verification here.

Theorem 2 *There exists a critical curve $l_c(j_1, j_2) = 0$, of which the function $l_c(j_1, j_2)$ is non-decreasing (and non-increasing) with respect to j_1 (j_2), dividing $\mathbb{N}^2$ into disjoint regions such that*

(1) $f^*(s = (j_1, j_2)) = e_1$, if $l_c(j_1, j_2) > 0$;
(2) $f^*(s = (j_1, j_2)) = e_2$, if $l_c(j_1, j_2) \leq 0$.

Due to their ease in implementation and enabling efficient computation, structural results of the optimal deterministic and stationary policy are very much appealing to decision makers Puterman [2005]. Thanks to the threshold structure, one only needs to store the transition points *a priori*, and the online implementation is simply by comparisons. Specialized algorithms can be developed to search among a special class (much smaller) of policies instead of general backward induction algorithms (less efficient) Puterman [2005].

Remark 1 *The threshold structure can be extended to cases with general $\mathtt{M}$ and $\mathtt{N}$. For $1 \leq i \leq \mathtt{M}$, define $j_i^- \triangleq (j_1, \ldots, j_{i-1}, j_{i+1}, \ldots, j_{\mathtt{M}})$ as the state of the whole system except for the i-th system. Then the optimal policy has the following threshold structure. Let state $s = (j_1, \ldots, j_{\mathtt{M}})$, there exist measurable functions $l_i : \mathbb{N}^{\mathtt{M}-1} \mapsto N$ such that for any $1 \leq i \leq \mathtt{M}$, the optimal policy f^* has the form:*

(1) if $j_i \geq l_i(j_i^-)$, $f^*(s) \in \mathbb{E}_i$;
(2) if $j_i < l_i(j_i^-)$, $f^*(s) \in \Omega \backslash \mathbb{E}_i$,

where $\mathbb{E}_i$ represents the feasible attack attention allocation subset such that the i-th system is under attack:

$$\mathbb{E}_i \triangleq \left\{ \eta \in \{0, 1\}^{\mathtt{M}} : \sum_{i=1}^{\mathtt{M}} \eta_{[i]} \leq \mathtt{N}, \eta_{[i]} = 1 \right\}.$$

What is more, the functions $l_i, 1 \leq i \leq \mathtt{M}$ are such that at each time there are exactly $\mathtt{N}$ systems to be attacked.

We now consider homogeneous models where the system dynamics are the same and $\epsilon_i, \underline{\epsilon_i}, 1 \leq i \leq \mathtt{M}$ are identical. For the homogeneous models with general $\mathtt{M}$ and $\mathtt{N}$, we propose a myopic policy as follows. *At each time k, the attacker attacks the $\mathtt{N}$ systems with largest $\tau_{k-1}^{(i)}$*. Denote this myopic policy by π_{m}. Then based on the above threshold structure and the symmetry of homogeneous models, one easily obtains the following corollary, the proof of which is omitted.

Corollary 1 *The myopic policy π_{m} is optimal to Problem 1 for homogeneous models, i.e., $\boldsymbol{R}(\pi_{\mathrm{m}}) = \sup_{\pi \in \Pi} \boldsymbol{R}(\pi)$.*

Note that to implement the myopic policy π_{m}, no specific model knowledge is required. Instead, one only needs to know the realization of the packet arrival process.

3.3 Explicit Asymptotic Optimal Policy

When $\mathtt{M}$ is large, the "curse of dimensionality" will render MDP numerical algorithms impractical. Then for heterogeneous models, one may ask whether or not there exists an algorithm that resembles the above myopic policy. The answer is positive. In the following, we provide an algorithm that is quite easy to compute and implement. Furthermore, it is proved to be asymptotically optimal as $\mathtt{M}$ and $\mathtt{N}$ go to infinity.

3.3.1 Virtual Attack Model

To present the algorithm, we introduce an virtual attacker. Consider the i-th system *in isolation*. Assume that an (virtual) attacker is able to attack the i-th system *all the time*, while if the attacker refuses to launch an attack at some time, it receives an extra *constant* "subsidy" z_i (which is independent of the system state $\tau_{k-1}^{(i)}$). In other words, the one-stage reward is given by

$$r_i(\tau_{k-1}^{(i)}, \eta_k^{(i)}) = \mathrm{Tr}(h_i^{\tau_{k-1}^{(i)}}(\hat{P}^{(i)})) + (1-\eta_k^{(i)})z_i.$$

The goal of the attacker is to maximize the averaged infinite-horizon accumulated reward as in Problem 1 for the sole i-th system: $\liminf_{\mathtt{T}\to\infty} \frac{1}{\mathtt{T}} \boldsymbol{E}\left[\sum_{k=1}^{\mathtt{T}} r_i(\tau_{k-1}^{(i)}, \eta_k^{(i)})\right]$. Denote the optimal rule for the state $\tau_{k-1}^{(i)} = j$ with $j \in \mathbb{N}$ when the subsidy is z_i as $d_i^*(j, z_i)$ [5]: $d_i^*(j, z_i) = 0$ if no attacks and $d_i^*(j, z_i) = 1$ otherwise.

To maximize the average infinite-horizon reward for the sole i-th system, one can also formulate it as an MDP problem and prove the existence of optimal deterministic and stationary policy. Furthermore, as for Theorem 2, one can prove the monotonicity of the differential value function as well, based on which the threshold structure of $d_i^*(j, z_i)$ can be proved. Specifically, for any $1 \le i \le \mathtt{M}$, given z_i, $d_i^*(j, z_i)$ has a form as

$$d_i^*(j, z_i) = \begin{cases} 1, \text{ if } j \ge \ell_i(z_i), \\ 0, \text{ if } j < \ell_i(z_i), \end{cases} \tag{8}$$

where $\ell_i(z_i)$ is a function of z_i.

[5] We use this notation to emphasize the dependence on z_i. It quite easy to show that the optimal rule is stationary, we thus omit the time index k.

3.3.2 Index-based Policy

We introduce an index $o_i(\cdot) : \mathbb{N} \mapsto \mathbb{R}$ associated with $\tau_{k-1}^{(i)} = j$, which satisfies that, for $1 \le i \le \mathtt{M}$,

$$\begin{aligned} & v_i(j)\left[\frac{1-(1-\epsilon_i)^j}{\epsilon_i} o_i(j) + \sum_{n=0}^{j} \mathrm{Tr}(h_i^n(\hat{P}^{(i)}))(1-\epsilon_i)^n \right. \\ & \left. + (1-\epsilon_i)^j \sum_{n=1}^{\infty} \mathrm{Tr}(h_i^{n+j}(\hat{P}^{(i)}))(1-\underline{\epsilon_i})^n \right] \\ =& v_i(j+1)\left[\frac{1-(1-\epsilon_i)^{j+1}}{\epsilon_i} o_i(j) + \sum_{n=0}^{j} \mathrm{Tr}(h_i^n(\hat{P}^{(i)}))(1-\epsilon_i)^n \right. \\ & \left. + (1-\epsilon_i)^{j+1} \sum_{n=0}^{\infty} \mathrm{Tr}(h_i^{n+j+1}(\hat{P}^{(i)}))(1-\underline{\epsilon_i})^n \right], \end{aligned} \tag{9}$$

where $v_i(j)$ is computed by

$$v_i(j) = \frac{1}{{\epsilon_i}^{-1} - (1-\epsilon_i)^j {\epsilon_i}^{-1} + (1-\epsilon_i)^j {\underline{\epsilon_i}}^{-1}}.$$

Notice that $o_i(\cdot)$ only depends on the i-th system and is irrelative with the others. Notice also that $o_i(j)$ in (9) can be interpreted as the subsidy such that when the i-th system state $\tau_{k-1}^{(i)} = j$, the action "attack" and "not attack" are equally attractive if the single i-th system is considered. We propose an index-based policy, denoted by π_{d}, as follows. *At each time k, the attacker attacks the $\mathtt{N}$ systems of greatest index $o_i(\tau_{k-1}^{(i)})$.* We then have the following theorem.

Theorem 3 *The index-based policy π_{d} is asymptotically optimal to Problem 1. That is, as $\mathtt{M} \to \infty$ and $\mathtt{N} \to \infty$ with $\mathtt{N} < \mathtt{M}$, $\boldsymbol{R}(\pi_{\mathrm{d}}) \to \boldsymbol{R}^*$, where $\boldsymbol{R}^* = \sup_{\pi\in\Pi} \boldsymbol{R}(\pi)$.*

Remark 2 *Numerical simulations in Section 4 show that the index-based policy π_{d} works quite well even when $\mathtt{M}$ and $\mathtt{N}$ are small.*

Remark 3 *In some scenarios, the attacker might get a larger reward for attacking one system than the other. Then one may add different weight to attacks on different channels, i.e., the reward in* (2) *is replaced with*

$$\boldsymbol{R}(\pi) = \liminf_{\mathtt{T}\to\infty} \frac{1}{\mathtt{T}} \boldsymbol{E}\left[\sum_{k=1}^{\mathtt{T}} \sum_{i=1}^{\mathtt{M}} w_i \mathrm{Tr}(\tilde{P}_k^{(i)})\right],$$

with $w_i \in \mathbb{R}_+$ being weight coefficients. The main results in this paper, Theorems 1–3, still hold. Amending the reward function by adding into the coefficients, the analysis in the appendices remains valid.

4 Numerical Examples

In this section, we use numerical examples to illustrate the threshold structure of the optimal policy (Theorem 2), the optimality of the myopic policy for homogeneous models (Corollary 1) and the asymptotic optimality of the index-based policy (Theorem 3).

Example 1 *We let* $\mathtt{M} = 2$ *and* $\mathtt{N} = 1$. *The parameters involved are as follows:*

$$A_1 = \begin{bmatrix} 1.2 & 0.2 \\ 0.3 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1.2 & 0.15 \\ 0 & 1.1 \end{bmatrix},$$
$$Q_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

$C_1 = [1, 0], C_2 = [1, 0.2], R_1 = 1, R_2 = 3, \epsilon_1 = 0.95, \underline{\epsilon_1} = 0.5, \epsilon_2 = 0.9$ *and* $\underline{\epsilon_2} = 0.4$. *Notice that the steady-state local estimation error covariances are*

$$\hat{P}^{(1)} = \begin{bmatrix} 0.79 & 0.54 \\ 0.54 & 8 \end{bmatrix}, \quad \hat{P}^{(2)} = \begin{bmatrix} 1.54 & -0.49 \\ -0.49 & 11.87 \end{bmatrix}.$$

We compute the optimal policy and optimal averaged reward using the value iteration algorithm. To cope with the countable infinity of the state space, the ideas in [Sennott, 2009, Chapter 8] are borrowed. The details of the algorithm are as follows. We truncate the state space with $N \in \mathbb{N}$, *i.e., the truncated state space* $\mathbb{S}_N \triangleq \{0, \ldots, N\}^2$. *Compute the value function (defined on* $\mathbb{S}_N$*) iteratively by*

$$\boldsymbol{J}_n^N(s) = \max_{a \in \mathbb{A}} \{r(s, a) + \boldsymbol{G}(\boldsymbol{J}_{n-1}^N, s, a)\}, \ \forall s \in \mathbb{S}_N$$

with $\boldsymbol{J}_0^N(s) = 0$. *Since the value iteration algorithm converges in our case (see Zhu and Guo [2005]), then for any* N, *let*

$$\varrho_N^* \triangleq \lim_{n \to \infty} \boldsymbol{J}_n^N((0,0)) - \boldsymbol{J}_{n-1}^N((0,0)),$$
$$q_N(s) \triangleq \lim_{n \to \infty} \boldsymbol{J}_n^N(s) - \boldsymbol{J}_n^N((0,0)).$$

One thus obtain the differential value function $q(s) = \lim_{N \to \infty} q_N(s), \forall s \in \mathbb{S}$. *The* N *is chosen such that* $|\varrho_N^* - \varrho_{N-1}^*|/\varrho_{N-1}^*$ *is smaller than a prescribed tolerance error. In our simulation, we let* $N = 19$ *and the error is 0.01. We obtain that the optimal averaged reward is* 50.21 *and the optimal policy is depicted as in Fig. 2. One may see that the optimal policy has the threshold structure stated in Theorem 2.*

Example 2 *We shall show that the myopic policy is optimal for homogeneous models. To this end, each system is the same as the* 2*nd system in Example 1, and the state space is also truncated with* $N = 19$. *In the first case, we let* $\mathtt{M} = 2, \mathtt{N} = 1$*; the second case* $\mathtt{M} = 3, \mathtt{N} = 2$ *and the third case* $\mathtt{M} = 5, \mathtt{N} = 2$. *The averaged reward obtained by the MDP algorithm and the myopic policy are shown in Table 1. As a baseline, we also simulate a random policy: at each time,* $\mathtt{N}$ *out of the* $\mathtt{M}$ *systems are randomly and uniformly chosen to be attacked. One sees that the averaged rewards obtained by the optimal MDP algorithm and the myopic policy are quite close, which verifies the optimality of the myopic policy. Also, compared with the random policy, the myopic policy has a significant performance improvement.*

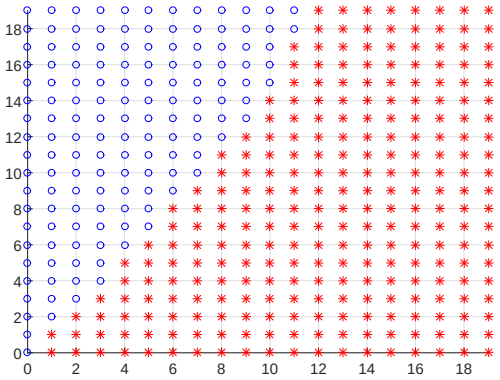

Fig. 2. Optimal action of state $s = (j_1, j_2)$ with x-axis presenting j_1 and y-axis j_2. The red stars and blue circles indicate the action e_1 and e_2, respectively.

Table 1
Averaged reward obtained by the MDP algorithm (denote by the symbol ♠), the myopic policy (♣) and random policy (♡) in different cases for homogeneous models.

Case No.	♠	♣	♡
1	40.98	40.82	29.94
2	71.49	71.37	55.91
3	93.75	93.46	71.45

Example 3 *We do simulations for four cases with heterogeneous models: in the first case, we let* $\mathtt{M} = 2, \mathtt{N} = 1$*; the second case* $\mathtt{M} = 3, \mathtt{N} = 2$*, the third case* $\mathtt{M} = 5, \mathtt{N} = 2$ *and the fourth case* $\mathtt{M} = 6, \mathtt{N} = 3$[6]. *In each case the first* $\lfloor \mathtt{M}/2 \rfloor$ *systems are the same as the* 1*-st system in Example 1, while the remaining are the same as the* 2*nd system. We truncate the state space with* $N = 12$, *which is mainly due to computation accuracy of index* $o_i(\cdot)$ *defined in* (9). *Specifically, since as* $j \to \infty$, $v_i(j)[1 - (1 - \epsilon_i)^j]/\epsilon_i \to v_i'(j)[1 - (1 - \epsilon_i)^{j+1}]/\epsilon_i$, *then when* j *is large enough (*$N = 13$ *for* 1*-st system and*

[6] We do not simulate asymptotic cases (i.e., $\mathtt{M}$ and $\mathtt{N}$ are sufficiently large) since state space size increases exponentially with respect to $\mathtt{M}$, the memory required would be beyond our capabilities.

$N = 17$ for 2-th system), numerical computing software (Matlab in our simulation) cannot provide accurate value of $o_i(\cdot)$. The averaged reward obtained by the MDP algorithm, the index-based policy and the random policy (the same as in the second example) are shown in Table 2, from which one sees that the index-based policy approximates the MDP algorithm surprisingly well even in these non-asymptotic cases. As in Example 2, the index-based policy has a significant performance gain over the random policy. To better illustrate this performance gain, we further simulate the index-based policy and the random policy for some large `M`*'s and* `N`*'s (we do not simulate the MDP algorithm due to capacity limitation). The results are shown in Table 3.*

Table 2
Averaged reward obtained by the MDP algorithm (denote by the symbol ♠), the index-based policy (♢) and random policy (♡) in different cases for heterogeneous models.

Case No.	♠	♢	♡
1	44.88	42.72	28.15
2	80.50	78.97	51.97
3	106.37	103.4	69.03
4	136.22	131.94	84.5

Table 3
Averaged reward obtained by the index-based policy (denote by the symbol ♢) and random policy (♡) in different cases (with large `M`'s and `N`'s) for heterogeneous models.

Case No.	♢	♡
1	4 446	2 816
2	18 971	12 658
3	22 269	14 082
4	25 733	16 885

5 Conclusion

In this paper, attack allocation on remote state estimation in multi-systems was considered. The problem was solved by formulating it as an MDP problem, of which an optimal deterministic and stationary policy exists. Threshold structure of the optimal policy was proved, by which both online implementation and off-line computation overhead can be reduced. To overcome the curse of dimensionality, an asymptotically optimal index-based policy, which is quite easy to compute and implement, was provided. The results were verified by numerical simulations. In particular, our numerical examples illustrated that the index-based policy works well even when the number of systems is small. An interesting direction of future works is to investigate the problem in a game-theoretic way, where the sensors (which have limited communication energy) are aware of the presence of the attacker.

Appendix A
Proof of Theorem 1

We first show that our MDP model has some "nice" properties, by which Theorem 1 can be proved. To this end, we define a function $\boldsymbol{W}:\mathbb{S}\mapsto[1,\infty)$ as

$$\begin{aligned}&\boldsymbol{W}(s=(j_1,j_2))\\&=\begin{cases}2, & \text{if } j_1=0, j_2=0,\\ \boldsymbol{W}_1(j_1)+\boldsymbol{W}_2(j_2), & \text{otherwise,}\end{cases}\end{aligned}\tag{10}$$

with $\boldsymbol{W}_1,\boldsymbol{W}_2:\mathbb{N}\mapsto[1,\infty)$ as

$$\boldsymbol{W}_1(j)=\begin{cases}\phi\lambda_1^j, & \text{if } j\le N_1,\\ \phi\lambda_1^{N_1}|A_1|^{2(j-N_1)}, & \text{if } j>N_1,\end{cases}$$
$$\boldsymbol{W}_2(j)=\begin{cases}\phi\lambda_2^j, & \text{if } j\le N_2,\\ \phi\lambda_2^{N_2}|A_2|^{2(j-N_2)}, & \text{if } j>N_2,\end{cases}$$

where ϕ,λ_i,N_i are parameters satisfying the following: for each $i=1,2$,

$$\lambda_i>1,$$
$$(1-\underline{\epsilon_i})(\lambda_i-1)\le\frac{1}{2}\underline{\epsilon_1\epsilon_2},\tag{11}$$
$$\phi[\beta-(1-\frac{1}{2}\underline{\epsilon_1\epsilon_2})]\ge1,\tag{12}$$
$$\phi\lambda_i^{N_i}[\beta-(1-\underline{\epsilon_i})|A_i|^2]\ge\phi+1,\tag{13}$$

with a constant $\beta<1$, which is bounded below by

$$\beta>\max\left(1-\frac{1}{2}\underline{\epsilon_1\epsilon_2},(1-\underline{\epsilon_i})|A_i|^2\right),\ i=1,2.\tag{14}$$

One may see that since $\phi>1,\lambda_i>1$, $\boldsymbol{W}_i$ together with $\boldsymbol{W}$ are well defined (i.e., they are all greater than 1).

About $\boldsymbol{W}$, we have the following two lemmas. Before proceeding, we need the following definition.

Definition 1 *Given a function $\boldsymbol{W}:\mathbb{S}\mapsto[1,\infty)$, for a function $u:\mathbb{S}\mapsto\mathbb{R}$, define its $\boldsymbol{W}-$norm as*

$$\|u\|_W=\sup_{s\in\mathbb{S}}|u(s)|/\boldsymbol{W}(s).$$

Let $\mathbb{B}_{\boldsymbol{W}}(\mathbb{S})$ be the normed linear space of measurable functions u on $\mathbb{S}$ with $\|u\|_{\boldsymbol{W}}<\infty$.

Lemma 1 *For any $f\in\mathbb{F}$, the transition kernel $\boldsymbol{P}(\cdot|\cdot,f(\cdot))$ is* uniformly *$\boldsymbol{W}-$geometrically ergodic*[7]*,*

[7] Interested readers are referred to Meyn and Tweedie [1993] to see a more elegant definition, which, however, requires more background knowledge, and is thus omitted here.

i.e., for any $f \in \mathbb{F}$ and any measurable function $u \in \mathbb{B}_{\boldsymbol{W}}(\mathbb{S})$, there exists a probability measure μ_f (depending on f) and constants L and $\delta < 1$, which are independent of f, such that for any $s \in \mathbb{S}, k \in \mathbb{N}$,

$$\left| \boldsymbol{G}(u, s, f(s)) - \int u \mathrm{d}\mu_f \right| \leq \|u\|_{\boldsymbol{W}} \boldsymbol{W}(s) L \delta^k. \quad (15)$$

Proof 1 *We prove that for each $f \in \mathbb{F}$, there exist constant $0 < \varpi < 1$ and b, which are independent of f, such that*

$$\boldsymbol{P}((0,0)|(0,0), f((0,0))) \geq \varpi \quad (16)$$

and for any $s \in \mathbb{S}$

$$\boldsymbol{G}(\boldsymbol{W}, s, f(s)) \leq \beta \boldsymbol{W}(s) + b \mathbf{1}_{\{(0,0)\}}(s) \quad (17)$$

where $\boldsymbol{W}(\cdot)$ and β are defined in (10) *and* (14)*, respectively. Then by [Meyn and Tweedie, 1994, Theorem 2.1 and 2.2], for each f, L and δ in* (15) *can be chosen in terms of ϖ, β, b (which are independent of f). The uniform ergodicity in Lemma 1 thus can be established.*

Equation (16) *is trivial. To show* (17)*, notice that when $s = (0,0)$, one may choose a sufficiently large b such that* (17) *is satisfied. Let $s \triangleq (j_1, j_2) \neq (0,0)$, suppose the action is e_1, then*

$$\begin{aligned}
&\boldsymbol{G}(\boldsymbol{W}, s, f(s)) \\
=&(1-\underline{\epsilon_1})\boldsymbol{W}_1(j_1+1) + (1-\epsilon_2)\boldsymbol{W}_2(j_2+1) \\
&+ \underline{\epsilon_1}(1-\epsilon_2)\phi + (1-\underline{\epsilon_1})\epsilon_2\phi + 2\underline{\epsilon_1}\epsilon_2 \\
\leq&(1-\underline{\epsilon_1})\boldsymbol{W}_1(j_1+1) + \underline{\epsilon_1}(1-\epsilon_2)\phi + 1 \quad (18) \\
&+ (1-\epsilon_2)\boldsymbol{W}_2(j_2+1) + (1-\underline{\epsilon_1})\epsilon_2\phi + 1 \quad (19)
\end{aligned}$$

Denote the term in (18) *and* (19) *by Λ_1 and Λ_2, respectively. We show $\Lambda_1 \leq \beta \boldsymbol{W}_1(j_1)$ by examining cases.*

Case $j_1 < N_1$:

$$\begin{aligned}
\Lambda_1 =&(1-\underline{\epsilon_1})\lambda_1 \boldsymbol{W}_1(j_1) + \underline{\epsilon_1}(1-\epsilon_2)\phi + 1 \\
\leq&(1-\underline{\epsilon_1})\lambda_1 \boldsymbol{W}_1(j_1) + \underline{\epsilon_1}(1-\epsilon_2)\boldsymbol{W}_1(j_1) + 1 \\
\leq&(1-\frac{1}{2}\underline{\epsilon_1\epsilon_2})\boldsymbol{W}_1(j_1) + 1 \\
\leq&\beta \boldsymbol{W}_1(j_1),
\end{aligned}$$

where the second inequality follows from (11) *and the last one* (12)*.*

Case $j_1 \geq N_1$:

$$\begin{aligned}
\Lambda_1 =&(1-\underline{\epsilon_1})|A_1|^2 \boldsymbol{W}_1(j_1) + \underline{\epsilon_1}(1-\epsilon_2)\phi + 1 \\
\leq&\beta \boldsymbol{W}_1(j_1),
\end{aligned}$$

where the inequality follows from (13)*. Using similar arguments, one may prove $\Lambda_2 \leq \beta \boldsymbol{W}_2(j_2)$, which completes the case when action e_1 is used. When e_2 or $\mathbf{0}$, similar results can be proved in the same way. The proof thus is complete.* □

Lemma 2 *There exists a constant α such that*

$$\left\|\bar{r}(s)\right\|_{\boldsymbol{W}} \leq \alpha,$$

with $\bar{r}(s) \triangleq \sup_{a \in \mathbb{A}} r(s,a)$.

Proof 2 *Let $\boldsymbol{W}_i'(j) = |A_i|^{2j}, j \in \mathbb{N}, i = 1, 2$. Since $\boldsymbol{W}(s) \geq 1, \forall s$, we only need to check asymptotic case of $\bar{r}(s)/\boldsymbol{W}(s)$. Since for $i = 1, 2$,*

$$\lim_{j \to \infty} \frac{\boldsymbol{W}_i(j)}{\boldsymbol{W}_i'(j)} = \phi \lambda_i^{N_i} |A_i|^{-2N_i}$$

is a constant, it suffices to prove for $i = 1, 2$,

$$\limsup_{j \to \infty} \frac{\mathrm{Tr}(h_i^j(\hat{P}^{(i)}))}{\boldsymbol{W}_i'(j)} < \infty. \quad (20)$$

Since the arguments are exactly the same, we do not distinguish $i = 1$ and $i = 2$ and suppress subscript i in the remainder of this proof. Let φ be a constant such that $\hat{P} \preceq \varphi I$ and $Q \preceq \varphi I$. Define a function

$$g(X) = AXA^\top + \varphi I.$$

One then obtains that

$$h^j(\hat{P}) \preceq g^j(\varphi I) \preceq \varphi \sum_{k=0}^{j} A^k (A^\top)^k,$$

which yields that $\mathrm{Tr}(h^j(\hat{P}))/|A|^{2j}$ is bounded. Equation (20) *thus follows and the proof is complete.* □

We are ready to prove Theorem 1 using the results in Guo and Zhu [2006] [8]. Since our state space is denumerable, by Remark 4.1(b) thereof, to prove Theorem 1, it suffices to verify Assumptions 3.1,3.2 [9] and 3.3 thereof. Since our action space is finite, Assumption 3.2 holds trivially. Assumption 3.1 and 3.3 follows directly from Lemma 1 and 2 (see Remark 3.3(b) thereof). The proof thus is complete.

[8] Notice that the distinguished feature of our MDP model is that the one-stage reward function is *unbounded above*, while the conventional MDP models (including the model in Leong et al. [2015]) have the reward (cost) function being bounded above (below).

[9] Notice that in Guo and Zhu [2006], the goal is to minimize an average cost, while we aims to maximize a reward function. Assumption 3.2 thereof should be adjusted accordingly, i.e., the requirement that the one-stage cost function is lower semicontinuous should be replaced with that the one-stage reward function is upper semicontinuous.

Appendix B
Proof of Theorem 2

To present structure of the optimal action, we give the following supporting lemma about the structure of so-called differential value function $q(s)$ in (7). To this end, we define a partial order on $\mathbb{S}$. Let $s = (j_1, j_2), s' = (j_1', j_2') \in \mathbb{S}$, we say that $s \preccurlyeq s'$ if $j_1 \leq j_1'$ and $j_2 \leq j_2'$. This partially ordered set is a lattice. Let $s \uparrow (\downarrow) s'$ denote the join (meet) on $(\preccurlyeq, \mathbb{S})$.

Lemma 3 *Let $s, s' \in \mathbb{S}$, for function $q(\cdot)$, the followings hold:*

Monotonicity: *If $s \preccurlyeq s'$, $q(s) \leq q(s')$.*
Submodularity: *$q(s) + q(s') \geq q(s \downarrow s') + q(s \uparrow s')$.*

Proof 3 *Let $0 < \alpha < 1$. Define the* discounted *reward associated with the initial state $s_1 = s$ and policy θ by*

$$\boldsymbol{J}_\alpha(s, \theta) = \liminf_{\mathtt{T} \to \infty} \frac{1}{\mathtt{T}} \boldsymbol{E}_s^\theta \left[\sum_{k=1}^{\mathtt{T}} \alpha^k r(s_k, a_k) \right],$$

and $\boldsymbol{J}_\alpha^*(s) \triangleq \sup_{\theta \in \Theta} \boldsymbol{J}_\alpha(s, \theta)$. With the existence of stationary and deterministic optimal policy proved in Theorem 1, one may let

$$q(s) = \lim_{\alpha \to 1} \boldsymbol{V}_\alpha(s).$$

with $\boldsymbol{V}_\alpha(s) = \boldsymbol{J}_\alpha^*(s) - \boldsymbol{J}_\alpha^*((0,0))$.

Then we show the monotonicity and submodularity of $q(s)$ by examining $\boldsymbol{V}_a(s)$. We do this by value iteration. To this end, we define a dynamic programming operator $\boldsymbol{T}_\alpha$: given a measurable function $u : \mathbb{S} \mapsto \mathbb{R}$, let

$$\boldsymbol{T}_\alpha u(s) \triangleq \max_{a \in \mathbb{A}} \left[r(s, a) + \alpha \boldsymbol{G}(u, s, a) \right], \ s \in \mathbb{S}.$$

Given $0 < \alpha < 1$, we define a function $\boldsymbol{W}_\alpha' : \mathbb{S} \mapsto [1, \infty)$ (depending on α) that has exactly the same form as $\boldsymbol{W}(s)$ in (10) *but the parameters involved have a different constraint. Specifically, the equations* (11)-(13) *are replaced with*

$$\begin{aligned} (1 - \underline{\epsilon_i})(\lambda_i - 1) &< \frac{1}{\alpha} - 1, \\ 1 \leq \phi &< \frac{1}{\alpha}, \\ \phi \lambda_i^{N_i} [1 - (1 - \underline{\epsilon_i}) |A_i|^2] &\geq \phi + 1. \end{aligned}$$

Using the same arguments as for Lemma 2, it is easy to see that $\left\| \sup_{a \in \mathbb{A}} r(s, a) \right\|_{\boldsymbol{W}_\alpha'} < \infty$. Thus, for any $0 < \alpha < 1$, $\|\boldsymbol{J}_\alpha^*(s)\|_{\boldsymbol{W}_\alpha'} < \infty$. Furthermore, by some basic calculations, one obtains that $\boldsymbol{W}_\alpha'$ satisfies [Hernández-Lerma and Lasserre, 1999, Assumption 8.3.2]. It then follows from Proposition 8.3.9 thereof, $\boldsymbol{T}_\alpha$ is contraction operator on $\mathbb{B}_{\boldsymbol{W}_\alpha'}(\mathbb{S})$. By Banach's Fixed Point Theorem, for any $u \in \mathbb{B}_{\boldsymbol{W}_\alpha'}(\mathbb{S})$, $0 < a < 1$,

$$\lim_{n \to \infty} \boldsymbol{T}_\alpha^n u = \boldsymbol{J}_\alpha^*(s). \tag{21}$$

Since given α, $\boldsymbol{J}_\alpha^*((0,0))$ is a constant, the structure (montonicity or submodularity) of $\boldsymbol{V}_\alpha(s)$ can be proved by showing that $\boldsymbol{J}_\alpha^*(s)$ has the same structure. By (21)*, it suffices to prove that the structure is preserved by the dynamic operator $\boldsymbol{T}_\alpha$.*

Monotonicity: *Suppose $s \preccurlyeq s'$ and $u(s) \leq u(s')$, since for any f, $r(s, f(s)) \leq r(s', f(s'))$, it holds that*

$$r(s, f(s)) + \alpha \boldsymbol{G}(u, s, f(s)) \leq r(s', f(s')) + \alpha \boldsymbol{G}(u, s', f(s'))$$

for any f, which yields $\boldsymbol{T}_\alpha u(s) \leq \boldsymbol{T}_\alpha u(s')$.

Submodularity: *By the monotonicity of $q(s)$, without any performance loss one may eliminate action $\mathbf{0}$. In the remainder, we let the action space $\mathbb{A} = \{e_1, e_2\}$. Suppose $u \in \mathbb{B}_{\boldsymbol{W}_\alpha'}(\mathbb{S})$ is monotonic, and for any $s, s' \in \mathbb{S}$*

$$u(s) + u(s') \geq u(s \downarrow s') + u(s \uparrow s'), \tag{22}$$

we need to prove $\boldsymbol{T}_\alpha u(s) + \boldsymbol{T}_\alpha u(s') \geq \boldsymbol{T}_\alpha u(s \downarrow s') + \boldsymbol{T}_\alpha u(s \uparrow s')$. By the definition of one stage reward function $r(s, a)$, it suffices to prove

$$\begin{aligned} &\max_{a \in \mathbb{A}} \boldsymbol{G}(u, s, a) + \max_{a \in \mathbb{A}} \boldsymbol{G}(u, s', a) \\ \geq &\max_{a \in \mathbb{A}} \boldsymbol{G}(u, s \downarrow s', a) + \max_{a \in \mathbb{A}} \boldsymbol{G}(u, s \uparrow s', a). \end{aligned} \tag{23}$$

Let $s = (j_1, j_2), s' = (j_1', j_2')$ with $j_1 \leq j_1', j_2 \geq j_2'$. Without loss of any generality, we assume $(1 - \underline{\epsilon_1})(1 - \epsilon_2) \geq (1 - \epsilon_1)(1 - \underline{\epsilon_2})$. For the function u, define the optimal action associated with state s by

$$a^*(s) \triangleq \arg\max_{a \in \mathbb{A}} \boldsymbol{G}(u, s, a).$$

In the following, we prove (23) *by cases.*

Case $a^*(s \downarrow s') = a^*(s \uparrow s')$: *Without loss of generality, we let $a^*(s \downarrow s') = a^*(s \uparrow s') = e_1$. Let $\varepsilon_1 = (1 - \underline{\epsilon_1})(1 -$*

ϵ_2), *one then obtains that*

$$\begin{aligned}
&\max_{a\in\mathbb{A}} \boldsymbol{G}(u,s,a) + \max_{a\in\mathbb{A}} \boldsymbol{G}(u,s,a)\\
\geq&\boldsymbol{G}(u,s,e_1) + \boldsymbol{G}(u,s',e_1)\\
=&\varepsilon_1\Big(u((j_1+1,j_2+1)) + u((j_1'+1,j_2'+1))\Big)\\
&+(1-\underline{\epsilon_1})\epsilon_2\Big(u((j_1+1,0)) + u((j_1'+1,0))\Big)\\
&+\underline{\epsilon_1}(1-\epsilon_2)\Big(u((0,j_2+1)) + u((0,j_2'+1))\Big)\\
&+2\underline{\epsilon_1}\epsilon_2 u((0,0))\\
\triangleq&\varepsilon_1\Big(u((j_1+1,j_2+1)) + u((j_1'+1,j_2'+1))\Big) + \Lambda\\
\geq&\varepsilon_1\Big(u((j_1+1,j_2'+1)) + u((j_1'+1,j_2+1))\Big) + \Lambda\\
=&\boldsymbol{G}(u,s\downarrow s',e_1) + \boldsymbol{G}(u,s\uparrow s',e_1)\\
=&\max_{a\in\mathbb{A}} \boldsymbol{G}(u,s\downarrow s',a) + \max_{a\in\mathbb{A}} \boldsymbol{G}(u,s\uparrow s',a),
\end{aligned}$$

where the second inequality follows from (22).

Case $a^*(s\downarrow s') = e_1, a^*(s\uparrow s') = e_2$: *Let* $\varepsilon_2 = (1-\epsilon_1)(1-\underline{\epsilon_2})$ *and* $\varepsilon_3 = (1-\underline{\epsilon_1})\epsilon_2 - (1-\epsilon_1)\underline{\epsilon_2}$, *one then obtains that*

$$\begin{aligned}
&\boldsymbol{G}(u,s,e_2) + \boldsymbol{G}(u,s',e_1) - \boldsymbol{G}(u,s{\downarrow}s',e_1) - \boldsymbol{G}(u,s{\uparrow}s',e_2)\\
=&\varepsilon_2 u((j_1+1,j_2+1)) + \varepsilon_1 u((j_1'+1,j_2'+1))\\
&-\varepsilon_1 u((j_1+1,j_2'+1)) - \varepsilon_2 u((j_1'+1,j_2+1))\\
&+\varepsilon_3\Big(u((j_1'+1,0)) - u((j_1+1,0))\Big)\\
\geq&\varepsilon_1 u((j_1+1,j_2+1)) + \varepsilon_1 u((j_1'+1,j_2'+1))\\
&-\varepsilon_1 u((j_1+1,j_2'+1)) - \varepsilon_1 u((j_1'+1,j_2+1))\\
\geq&0,
\end{aligned}$$

where the first inequality follows from the monotonicity of u *and the fact* $\varepsilon_1 \geq \varepsilon_2$, *and the second inequality is due to* (22). *Equation* (23) *thus follows.*

Case $a^*(s\downarrow s') = e_2, a^*(s\uparrow s') = e_1$: *One has the following:*

$$\begin{aligned}
&\boldsymbol{G}(u,s,e_2) + \boldsymbol{G}(u,s',e_1) - \boldsymbol{G}(u,s{\downarrow}s',e_2) - \boldsymbol{G}(u,s{\uparrow}s',e_1)\\
=&\varepsilon_2 u((j_1+1,j_2+1)) + \varepsilon_1 u((j_1'+1,j_2'+1))\\
&-\varepsilon_2 u((j_1+1,j_2'+1)) - \varepsilon_1 u((j_1'+1,j_2+1))\\
&+(\varepsilon_1-\varepsilon_2+\epsilon_2-\underline{\epsilon_2})\Big(u((0,j_2+1)) - u((0,j_2'+1))\Big)\\
\geq&\varepsilon_1 u((j_1+1,j_2+1)) + \varepsilon_1 u((j_1'+1,j_2'+1))\\
&-\varepsilon_1 u((j_1+1,j_2'+1)) - \varepsilon_1 u((j_1'+1,j_2+1))\\
&+(\varepsilon_1-\varepsilon_2)\Big(u((0,j_2+1)) + u((j_1+1,j_2'+1))\\
&-u((0,j_2'+1)) - u((j_1+1,j_2+1))\Big)\\
\geq&0,
\end{aligned}$$

which yields (23). *The proof thus is complete.* □

We are ready to prove Theorem 2. First, let fix j_2 and show that if $f^*(s=(j_1,j_2)) = e_1$, then $f^*(s=(j_1+j,j_2)) = e_1$ with $j\in\mathbb{N}$. Since $f^*(s=(j_1,j_2)) = e_1$ implies that

$$\begin{aligned}
&(\varepsilon_1-\varepsilon_2)q((j_1+1,j_2+1)) + \varepsilon_3 q((j_1+1,0))\\
\geq&\varepsilon_4 q((0,j_2+1)) + (\epsilon_1\underline{\epsilon_2} - \underline{\epsilon_1}\epsilon_2)q((0,0))\\
\triangleq&\Lambda_3.
\end{aligned}$$

where $\varepsilon_4 = \epsilon_1(1-\underline{\epsilon_2}) - \underline{\epsilon_1}(1-\epsilon_2)$ Since $\varepsilon_1-\varepsilon_2\geq 0$, ε_3 and Λ_3 is constant for a given j_2, by the monotonicity of q in Lemma 3, one obtains that

$$(\varepsilon_1-\varepsilon_2)q((j_1+j+1,j_2)) + \varepsilon_3 q((j_1+j+1,0)) \geq \Lambda_3,$$

which yields $f^*(s=(j_1+j,j_2)) = e_1$. Then it concludes that given a j_2, there is a critical curve $l_1(j_2)$ such that

$$f^*(s=(j_1,j_2)) = \begin{cases} e_1, \text{ if } j_1 \geq l_1(j_2),\\ e_2, \text{ if } j_1 < l_1(j_2).\end{cases} \tag{24}$$

Similarly, let fix j_1 and show that if $f^*(s=(j_1,j_2)) = e_2$, then $f^*(s=(j_1,j_2+j)) = e_2$ with $j\in\mathbb{N}$. Note that $f^*(s=(j_1,j_2)) = e_2$ implies that

$$\begin{aligned}
&\varepsilon_4 q((0,j_2+1)) - (\varepsilon_1-\varepsilon_2)q((j_1+1,j_2+1))\\
\geq&\varepsilon_3 q((j_1+1,0)) + (\epsilon_1\underline{\epsilon_2} - \underline{\epsilon_1}\epsilon_2)q((0,0))\\
\triangleq&\Lambda_4.
\end{aligned}$$

Then one has

$$\begin{aligned}
&\varepsilon_4 q((0,j_2+j+1)) - (\varepsilon_1-\varepsilon_2)q((j_1+1,j_2+j+1))\\
=&(\varepsilon_1-\varepsilon_2)\Big(q((0,j_2+j+1)) - q((j_1+1,j_2+j+1))\Big)\\
&+(\epsilon_2-\underline{\epsilon_2})q((0,j_2+j+1))\\
\geq&(\varepsilon_1-\varepsilon_2)\Big(q((0,j_2+1)) - q((j_1+1,j_2+1))\Big)\\
&+(\epsilon_2-\underline{\epsilon_2})q((0,j_2+1))\\
=&\varepsilon_4 q((0,j_2+1)) - (\varepsilon_1-\varepsilon_2)q((j_1+1,j_2+1))\\
\geq&\Lambda_4,
\end{aligned}$$

where the first inequality follows from the monotonicity and submodularity of $q(s)$ established in Lemma 3. Hence $f^*(s=(j_1,j_2+j)) = e_2$. Similarly, it concludes that given a j_1, there is a critical curve $l_2(j_1)$ such that

$$f^*(s=(j_1,j_2)) = \begin{cases} e_2, \text{ if } j_2 \geq l_2(j_1),\\ e_1, \text{ if } j_2 < l_2(j_1).\end{cases} \tag{25}$$

To simultaneously satisfy both (24) and (25), both functions $l_1(\cdot)$ and $l_2(\cdot)$ must be monotonically non-decreasing. Then the statements in Theorem 2 follow immediately by letting $l_c(j_1,j_2) = l_2(j_1) - j_2$.

Appendix C
Proof of Theorem 3

The byproduct of Theorem 2 is that for an optimal policy at no time the action **0** is chosen. This can be extended to a general case, i.e., the constraint that at each time the attacker can attack at most N of M systems is equivalent to the constraint that the attacker attacks *exactly* N of M systems. With this in mind, we prove the theorem using the results in Whittle [1988] on the restless multi-armed bandit problem.

Recall that $d_i^*(j, z_i)$ is the optimal rule for the state $\tau_{k-1}^{(i)} = j$ with $j \in \mathbb{N}$ when the subsidy is z_i. We then have the following definition and lemma.

Definition 2 *Whittle [1988] The i-th system is said to be indexable if for any $j \in \mathbb{N}$, $d_i^*(j, z_i) = 0$ implies $d_i^*(j, z_i') = 0$ with $z_i' \geq z_i$. The whole system is indexable if each system is indexable.*

Lemma 4 *The system introduced in Section 2 is indexable.*

Proof 4 *We show that each system is indexable. For ease of notations, throughout this proof, we omit the subscript i. Denote $p(\cdot)$ as the resulted equilibrium probability distribution of the state when $\ell(z) = j^*$ (function $\ell(\cdot)$ is, recall, introduced in (8) [10]). Then due to the threshold structure in (8), one obtains that*

$$\sum_{j=0}^{\infty} p(j) = 1, \tag{26}$$

$$p(j) = \begin{cases} (1-\epsilon)p(j-1), \text{ if } 1 \leq j \leq j^*, \\ (1-\underline{\epsilon})p(j-1), \text{ if } j > j^*. \end{cases} \tag{27}$$

Note that the averaged reward obtained by the attacker has two parts: the averaged subsidy $\boldsymbol{R}_s$ and the averaged estimation error $\boldsymbol{R}_e$:

$$\boldsymbol{R}_s = \sum_{j=0}^{j^*-1} p(j) z,$$

$$\boldsymbol{R}_e = \sum_{j=0}^{\infty} p(j) \mathrm{Tr}(h^j(\hat{P})).$$

Now fix the subsidy z and consider a suboptimal policy. The policy has a similar threshold structure as in (8) but with the switching threshold $0 \leq j^\diamond < j^*$. Denote the corresponding equilibrium probability distribution as p', which is computed in a similar way as (26)(27). Then one has

$$\sum_{j=0}^{j^*-1} p(j) > \sum_{j=0}^{j^\diamond-1} p'(j). \tag{28}$$

Denote the averaged subsidy and averaged estimation error as $\boldsymbol{R}_s'$ and $\boldsymbol{R}_e'$, respectively. Due to the optimality of $l(z) = j^*$, one obtains that $\boldsymbol{R}_s - \boldsymbol{R}_s' \geq \boldsymbol{R}_e' - \boldsymbol{R}_e$, i.e.,

$$\left[\sum_{j=0}^{j^*-1} p(j) - \sum_{j=0}^{j^\diamond-1} p'(j) \right] z \geq \boldsymbol{R}_e' - \boldsymbol{R}_e.$$

Then by (28), for any $z' \geq z$, it holds that

$$\left[\sum_{j=0}^{j^*-1} p(j) - \sum_{j=0}^{j^\diamond-1} p'(j) \right] z' \geq \boldsymbol{R}_e' - \boldsymbol{R}_e,$$

which means that for any subsidy $z' \geq z$, the optimal rule for the states $0 \leq j < j^*$ is still "not attack". The proof thus is complete. □

Asymptotic optimality of the index-based policy π_{d} stated in Theorem 3 follows immediately from the indexability established in Lemma 4 and [Whittle, 1988, Conjecture] [11].

[10] Notice that the subscript i has been omitted.

[11] Weber and Weiss [1990] presented a counterexample to the conjecture, and provided an additional technical requirement to assure the asymptotic optimality. Notice that, however, this technical requirement is inconsequential, as argued by the authors, since the cases where the indexability is not sufficient are "extremely rare" and the deviation from optimality (if exists) is "minuscule". On the other hand, it is quite difficult to verify this technical requirement (which says that a differential equation describing the fluid approximation of the index-based policy has no limit cycles or chaotic behavior), we thus omit the verification here.

References

Saurabh Amin, Alvaro A Cárdenas, and S Shankar Sastry. Safe and secure networked control systems under denial-of-service attacks. In *Hybrid Systems: Computation and Control*, pages 31–45. Springer, 2009.

Brian D O Anderson and John B Moore. *Optimal filtering*. North Chelmsford: Courier Corporation, 2012.

BBC. Hackers caused power cut in western Ukraine - us, 2016. URL `http://www.bbc.com/news/technology-35297464`. [Online; accessed 16-May-2016].

Eric Byres and Justin Lowe. The myths and facts behind cyber security risks for industrial control systems. In *Proceedings of the VDE Kongress*, volume 116, pages 213–218, 2004.

James P Farwell and Rafal Rohozinski. Stuxnet and the future of cyber war. *Survival*, 53(1):23–40, 2011.

Xianping Guo and Quanxin Zhu. Average optimality for markov decision processes in borel spaces: a new condition and approach. *Journal of Applied Probability*, pages 318–334, 2006.

Ziyang Guo, Dawei Shi, Karl H Johansson, and Ling Shi. Optimal linear cyber-attack on remote state estimation. *IEEE TCNS Special Issue on Secure Control of Cyber Physical Systems, to appear*, 2016.

Abhishek Gupta, Cédric Langbort, and Tamer Basar. Optimal control in the presence of an intelligent jammer with limited actions. In *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, pages 1096–1101. IEEE, 2010.

Onésimo Hernández-Lerma and Jean B Lasserre. *Further topics on discrete-time Markov control processes*, volume 42. New York: Springer Science & Business Media, 1999.

Alex S Leong, Subhrakanti Dey, and Daniel E Quevedo. On the optimality of threshold policies in event triggered estimation with packet drops. In *Proceedings of European Control Conference (ECC)*, pages 927–933, 2015.

Yuzhe Li, Ling Shi, Peng Cheng, Jiming Chen, and Daniel E Quevedo. Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. *IEEE Transactions on Automatic Control*, 60(10):2831–2836, 2015.

Sean P Meyn and Richard L Tweedie. Markov chains and stochastic stability. *London: Springer-Verlag*, 1993.

Sean P Meyn and Richard L Tweedie. Computable bounds for geometric convergence rates of markov chains. *The Annals of Applied Probability*, pages 981–1011, 1994.

Yilin Mo and Bruno Sinopoli. Secure control against replay attacks. In *Proceedings of 47th Annual Allerton Conference on Communication, Control, and Computing, 2009.*, pages 911–918. IEEE, 2009.

Yilin Mo, Bruno Sinopoli, Ling Shi, and Emanuele Garone. Infinite-horizon sensor scheduling for estimation over lossy networks. In *Proceedings of 51st IEEE Conference on Decision and Control (CDC), 2012*, pages 3317–3322, 2012.

Fabio Pasqualetti, Florian Dorfler, and Francesco Bullo. Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, 35 (1):110–127, 2015.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. New Jersey: John Wiley & Sons, 2005.

Xiaoqiang Ren, Yilin Mo, and Ling Shi. Optimal dos attacks on bayesian quickest change detection. In *Proceedings of 53rd IEEE Conference on Decision and Control*, pages 3765–3770, 2014a.

Zhu Ren, Peng Cheng, Jiming Chen, Ling Shi, and Huanshui Zhang. Dynamic sensor transmission power scheduling for remote state estimation. *Automatica*, 50(4):1235–1242, 2014b.

Linn I Sennott. *Stochastic dynamic programming and the control of queueing systems*, volume 504. New York: John Wiley & Sons, 2009.

André Teixeira, Daniel Pérez, Henrik Sandberg, and Karl Henrik Johansson. Attack models and scenarios for networked control systems. In *Proceedings of the 1st International Conference on High Confidence Networked Systems*, pages 55–64. ACM, 2012.

Antonio Teixeira, Kin Cheong Sou, Henrik Sandberg, and Karl H Johansson. Secure control systems: A quantitative risk management approach. *IEEE Control Systems Magazine*, 35(1):24–45, 2015.

Richard R Weber and Gideon Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, pages 637–648, 1990.

Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of Applied Probability*, pages 287–298, 1988.

Heng Zhang, Peng Cheng, Ling Shi, and Jiming Chen. Optimal denial-of-service attack scheduling with energy constraint. *IEEE Transactions on Automatic Control*, 60(11):3023–3028, 2015.

Quanxin Zhu and Xianping Guo. Value iteration for average cost markov decision processes inborel spaces. *Applied Mathematics Research eXpress*, 2005(2):61–76, 2005.