A zero-sum Markov defender-attacker game for modeling false pricing in smart grids and its solution by multi-agent reinforcement learning

Daogui Tang, Yiping Fang, Enrico Zio

A zero-sum Markov defender-attacker game for modeling false pricing in smart grids and its solution by multi-agent reinforcement learning

Daogui Tang

Chaire System Science and the Energy Challenge, Laboratoire Génie Industriel, CentraleSupélec, Université Paris-Saclay, France. E-mail: daogui.tang@centralesupelec.fr

Yi-Ping Fang

Chaire System Science and the Energy Challenge, Laboratoire Génie Industriel, CentraleSupélec, Université Paris-Saclay, France. E-mail: yiping.fang@centralesupelec.fr

Enrico Zio

Mines ParisTech, PSL Research University, CRC, Sophia Antipolis, France.
Energy Department, Politecnico di Milano, Milano, Italy E-mail: enrico.zio@mines-paristech.fr, enrico.zio@polimi.it

Abstract-Consumers in smart grids are expected to engage demand-response programs by two-way communication. This makes smart grids vulnerable to cyber attacks. In this paper, we study the false pricing attacks and model the interaction between attackers and defenders using a zero-sum Markov game, where neither player has full knowledge of the game model. A multi-agent reinforcement learning method is used to solve the Markov game and find the Nash Equilibrium policies for both players. An application to a simple radial power distribution system is worked out. The results show that the proposed algorithm can help the players find mixed strategies to maximize their long-term return.

Keywords: smart grids, demand-response, false pricing attack, game theory, multi-agent reinforcement learning.

1. Introduction

Smart grids (SG) are critical infrastructures (CI), which integrate information and communication technology (ICT) onto the power grid. The ICT enables two-way communications, facilitating the control and operation of the SG. Besides all the benefits that this brings to the efficient and economic operation of SG, it is important to consider potential risks associated with it. One of the most challenging risks that is emerging with the increased usage of ICT is cyber-attacks (Liang et al. 2017; Dehghanpour et al. 2019).

Load altering attack targets the consumer's demand-side management (DSM) and demand-response (DR) programs (Mohsenian-Rad and Leon-Garcia 2011). DR, as one type of DSM, is a consumer-driven activity, which improves energy system at the side of consumption and enables to reshape consumers' consumption patterns by, e.g., real-time pricing (RTP) and time-of-use (TOU) pricing (Palensky and Dietrich 2011; Deng et al. 2015). With DR programs, many types of load are vulnerable to load altering attacks, e.g., price-response load (Mohsenian-Rad and Leon-Garcia 2011) and frequency-response load (Amini, Pasqualetti, and Mohsenian-Rad 2018). In the present work, we focus on attacks aiming at the price-response load, namely false pricing attacks (FPA). The assumption behind the price-based DR programs is that consumers are rational and eager to save money by reshaping their consumption patterns with the automated energy management system (EMS). Exploiting this, attackers can launch effective attacks with false prices, resulting in an increase of loads of some consumers and eventually cause circuit overflow or other malfunctioning in the power grids (Mishra et al. 2017).

In price-based DR programs, the price signal is transferred by the operator through the Internet to a central computer in a local substation where the price information is, then, broadcasted to the smart meters located at the side of the consumers via the Internet or Wi-Fi networks (Liu, Hu, and Ho 2014). This makes various parts of the communication infrastructure vulnerable to attacks, e.g., the central computer, the access point of the Wi-Fi network, and the smart meters. Interested readers may refer to (Liu, Hu, and Ho 2014) for more information about the communication infrastructure. Smart meters are easiest to be attacked because of their weakness such as physical exposure, relatively simple authentication and encryption procedures, etc. (Tellbach and Li 2018). Besides, the attacks can be easily carried out with automated and distributed software intruding agents (Mishra et al. 2017). For these reasons, in the present study we specify the attack path in terms of injection of false prices by smart meters.

Various research works have studied the problem of FPA. In (Tan et al. 2013), the authors

Proceedings of the 29th European Safety and Reliability Conference.
Edited by Michael Beer and Enrico Zio

Published by Research Publishing, Singapore.
ISBN: 978-981-11-2724-3; doi:10.3850/978-981-11-2724-3_0743-cd

analyze the stability of the power grid subject to scaling and delay FPA, from the perspective of control theory. The authors in (Giraldo, Cardenas, and Quijano 2017) further extend the work to arbitrary price signals. In (Zhang et al. 2017), integrity attacks to the RTP process are considered. The attacks are carried out by altering the preference parameters of renewable and traditional power resources. Ref. (Mishra et al. 2017) assumes that the prices can be modified with a price change rate, and the behavior of the attacker and protector is modelled as a static game.

The common assumption underpinning the above works is that the behavior of the consumers is deterministic, and both the attackers and protectors have full knowledge of the consumers' response behavior to the electricity prices. This is unrealistic since the consumers demand-response behavior has an intrinsic stochastic element coming from the increasing penetration of distributed energy resources (DER) and renewable energies (Aghajani, Shayanfar, and Shayeghi 2017). Besides, the consumers may not respond to prices as it would be expected, because of the lack of knowledge about how to respond to time-varying prices (Mohsenian-Rad and Leon-Garcia 2010). However, consumers may receive electricity prices also from other channels like social networks (Tang et al. 2019), and determine their consumption coordinately. All these uncertainties make it hard for attackers and protectors to get full knowledge of the demand-response model. For this reason, in the present work, we consider situations in which both attackers and defenders have no knowledge of the consumers' response models.

In recent research, (partially observable) Markov decision processes (MDP) and game theory have been applied to model the attacker and defender's strategies. In (Hao, Wang, and Chow 2018), the intruder's strategy is modelled by an MDP and the optimal attack policy is solved from the intruder's perspective, based on which the attack likelihood is, then, analyzed. In the case where the attackers are only able to observe part of the environment, the behavior of the attacker can be modelled as a partially observable MDP (Chen et al. 2018), and single-agent reinforcement learning can be adopted to solve the attack policies. However, these works have not incorporated the defender's policies.

In other research works, game-theoretic approaches have been used to model the strategic interaction between attackers and defenders (Deng, Xiao, and Lu 2017; Chen, Hong, and Liu 2018; Wei et al. 2018). In (Deng, Xiao, and Lu 2017), a two-player zero-sum game is proposed to model the strategies of the attacker and defender. The attacker tries to minimize the attack cost whereas the defender aims to maximize the least budget of attacking by allocating the defending resources. However, this static game is one-shot and ignores the dynamic evolvement. In (Chen, Hong, and Liu 2018), a Markov game is adopted to model the competitive intrusion and defense policies for control of the substations. A stochastic game is introduced to determine the optimal resources allocation of attackers and defenders (Wei et al. 2018). All these game-theoretic approaches are based on the assumption that both the attacker and the defender have full knowledge of the environment. In this paper, we propose a stochastic game where the players have only partial knowledge of the environment and solve the optimal policies of the players by multi-agent reinforcement learning (MARL).

The rest of the paper is organized as follows: the power grid model and the Markov game model are introduced in Section 2; Section 3 presents the proposed multi-agent reinforcement learning algorithm to solve the Markov game and applied to an illustrative power system in Section 4. The work is concluded in Section 5.

2. model description

Nomenclature

$G\ (V, E)$	power grid with vertex set V and edge set E
P_{ij}	power flow from vertex i to j
m, n, m_d	numbers of nodes, edges and demand nodes
d	power demand
λ	electricity price
pcr	price change rate
p	probability of response to electricity price
s	load shedding
$EENS$	expected energy not supplied
$\mathcal{S}$	set of states of the system
$\mathcal{A}_1, \mathcal{A}_2$	action sets of the attacker and defender
a, a'	actions of attacker and defender
$\mathcal{R}_1, \mathcal{R}_2$	reward sets of the attackers and defenders
$r(s, a, a')$, $r'(s, a, a')$	reward of attacker and defender at state s with joint action (a, a')
C_1, C_2	resources allocation sets of attacker and defender

Acronym

SG	smart grids
CI	critical infrastructures
ICT	information and communication technology
DSM	demand-side management
DR	demand-response
RTP	real-time pricing
TOU	time-of-use
FPA	false pricing attacks
EMS	energy management system
DER	distributed energy resources
MDP	Markov decision processes
MARL	multi-agent reinforcement learning

2.1 *Power grid model*

We consider a power grid represented by a directed graph $G(V, E)$, where the vertex set $V = \{v_1, v_2, \dots, v_m\}$ represents the generators, transformers and demand nodes in the power grid and the edges set $E = \{e_1, e_2, \dots, e_n\}$ represents the distribution lines. The numbers of the nodes and edges are m and n, respectively, and the number of demand nodes are represented by m_d. The power flow can be modeled by the classical *LinDistFlow* model (Baran and Wu 1989):

$$P_{ij} = \sum\nolimits_{k:(j,k)\in E} P_{jk} + d_j \quad \forall (i,j) \in E \quad (1)$$

where P_{ij} is the power flow through distribution line (i, j); d_j represents the power demand at node j. Normally, the power demand of a consumer is dependent to the electricity price λ and it is automatically controlled by the EMS located in the smart meters.

In this paper, we follow the demand response rule in (Mishra et al. 2017) **:**

$$(1 + k_i) \cdot B_i \geq d_j \cdot \lambda \quad (2)$$

where B_i represents the consumer's targeting bill amount and k_i is the sensitivity of the consumers to the billing amount.

As a consequence of the renewable energy penetration and the uncertain behavior of consumers, the consumer's energy consumption can be modeled as a binary state at each time, i.e., elastic (responsive to price signals) or inelastic (unresponsive to price signals) (Ghosh, Sun, and Zhang 2012). In the present study, for simplicity, we assume that there is a probability p_i that at each time the consumers respond to the price signals in one way or the other. In the case of $p_i = 0$, the consumers demand is independent to the prices and determined only according to the consumers' needs.

The attacker can control consumers' consumption indirectly by FPA, where the electricity price is modified with a price change rate (pcr). For instance, an attacker can inject a false price lower than the real one, i.e., $\lambda \cdot (1 - pcr)$ to a consumer's smart meter and the consumer may change the load according to the false price and the probability p_i.

In case of overload, the operator will respond immediately to shed some load and minimize the impact of the attack:

$$\min \sum_{i=1}^{m_d} s_i \quad (3.1)$$

$$\text{s.t.} \quad 0 \leq s_i \leq d_i \quad (3.2)$$

$$-P_{ij}^{\max} \leq P_{ij} \leq P_{ij}^{\max} \quad (3.3)$$

$$Equation\ (1)$$

The impact of the attack can be quantified by the Expected Energy Not Supplied (*EENS*).

$$EENS = \sum_{i=1}^{m_d} \frac{s_i}{d_i} \quad (4)$$

2.2 *Markov game model*

In reality, both the attacker and the defender have limited available resources to attack and protect the power grid. The resources that the attacker can utilize include the hackers, technical and economic resources. Similarly, the resources of a defender include the personnel, the technical resources, e.g., software and hardware security elements (Y. Yan et al. 2012) and economic resources. The problem for the attacker and defender is how to utilize their resources to achieve their goals at most. The decision process of the attacker and defender can be modeled as a two-player zero-sum Markov Game, which is represented as a tuple $MG = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} \rangle$.

- $\mathcal{S} = \{s_1, s_2, \dots, s_t\}$ is the finite set of environment states;
- $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2\}$ represents the joint action of the attacker and defender, where $\mathcal{A}_1$ and $\mathcal{A}_2$ represent the attacker's and defender's action spaces, respectively, and n_a represents the number of actions that the player can choose;
- $\mathcal{T}$ represents the state transition probability function.
- $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2\}$ represents the player's reward function, where $\mathcal{R}_1$ and $\mathcal{R}_2$ represent the attacker's and defender's rewards, respectively.

The game is played for a sequence of discrete time steps. At each time step, some representation of the state of the environment, $s \in \mathcal{S}$, can be observed by the players. Based on the state of the environment, the players choose their actions $(a, a') \in \mathcal{A}$ independently; then, as a consequence, in the next time step, the environment's state transits to s' with probability $\mathcal{T}(s'|s, a, a')$. Meanwhile, the attacker and the defender get a reward $r(s, a, a')$ and $r'(s, a, a')$, respectively. Since the game is a zero-sum game, $r(s, a, a') + r'(s, a, a') = 0$. Each part of the game is defined as follows:

2.2.1 *The state of the environment*

In the perspective of the attacker, the goal is to inject false price information in the demand response process of consumers and cause failure to the power grid. For this reason, we use the *EENS* to define the state of the power system:

$$s = \begin{cases} s_1, & EENS = 0 \\ s_2, & EENS > 0 \end{cases} \quad (5)$$

The state of the power grid can be solved by Eq. (3). The information that the attacker requires to know is the consumers' usages at each time step, the topology information of the power grid and the capacity of the distribution lines. We assume that the attacker has access to all the consumers and knowledge of the related information.

2.2.2 *The action space of the agents*

Suppose the maximum price change rate that cannot be detected is pcr_{max} and an action of the attacker can be modeled as $a = \{pcr_1, pcr_2, \dots, pcr_{m_d}\}$, where $pcr \in [0, pcr_{max}]$. As in (Mishra et al. 2017), we assume that there is a cost associated to the price change, i.e. $c_i = f(pcr_i)$, and consider a linear cost function $f(\cdot)$, Then, the attack resource targeting each consumer is $C_1 = \{c_1, c_2, \dots, c_{m_d}\}$ and the total resource of the attacker at any time is limited as c_{max}, i.e., $\sum_{i=1}^{m_d} c_i \le c_{max}$. Similarly, the defender can allocate his defense resources to protect the smart meters from being attacked: $C_2 = \{c'_1, c'_2, \dots, c'_{m_d}\}$ and $\sum_{i=1}^{m_d} c'_i \le c'_{max}$.

2.2.3 *The immediate reward of players*

In the present work, the attacker aims to find the optimal resource allocation policies to cause the power system overload, whereas the defender tries to allocate the limited defending resources to protect it. Thus, the immediate reward of the attacker at a state $s \in S$ with the joint action of the attacker can be calculated as:

$$r(s, a, a') = EENS \quad (6)$$

and the immediate reward of the defender is the negative value of that of the attacker.

2.2.4 *The state transition probability*

In previous works, the state transition probability is derived from empirical information. However, practically, the players have no historical information to get the explicit expression of the transition probability. Theoretically, in order to obtain this probability, both the demand response models of all the consumers and the probability of demand response need to be known. However, the response model is normally private (Samadi et al. 2010) and the probability of demand-response are difficult to know. Therefore, in the present work, we assume that the state transition probability is unknown to the players.

3. Minimax-Q learning algorithm

Generally, the players in the game try to find the optimal policy to maximize the expected long-term return from every state of the environment:

$$Val^{\pi}(s) = E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k \cdot r_{k+1} | s_0 = s\right\} \quad (7)$$

where $\pi = \{\pi_1, \pi_2\}$ is the joint policy of the players and $\gamma \in [0,1]$ is the discount factor, which represents the degree of interest of the future reward. Choosing a value of γ close to 1 means that the agent regards every immediate reward nearly equal important. On the contrary, a γ close to 0 favors the immediate reward.

The policy of an agent is a mapping from the state of the environment to a probability distribution over the associated actions in the state. For instance, the policy of the attacker $\pi_1(s, a)$ is a probability distribution of each action in every state and satisfies:

$$\sum_{a \in \mathcal{A}_1} \pi_1(s, a) = 1 \quad \forall s \in \mathcal{S} \quad (8)$$

As the definition of the concepts in the game is similar for both players, in this paper we analyze the game from the perspective of the attacker. In reinforcement learning, the attacker acts to the environment and receives one-step feedback. Then, the quality of the action taken by the attacker, i.e., the immediate reward, the action of the defender $a' \in \mathcal{A}_2$ can be observed. Thus, the agent can maximize its long-term return by calculating the optimal value of the Q function at each time step:

$$Q(s, a, a') = E_{\pi}\left\{\sum_{k=0}^{N} \gamma^k \cdot r(s, a, a')\right\} \quad (9)$$

In the present study, the defender tries to minimize the return of the attacker whereas the attacker tries to maximize it. The best policy of them is to achieve a Nash Equilibrium. At the Nash Equilibrium, each player is maximizing its rewards and any changes in strategy would make it worse or stay the same.

To achieve Nash Equilibrium, the Minimax - Q algorithm (Littman 1994) and the temporal-difference (TD) learning method (Sutton and

Barto 2011) can be adopted. Minimax-Q is proven to converge to the equilibrium value function (Littman and Szepesvári 1996). In the maximum-Q algorithm, the $Q(s,a,a')$ is recursively updated:

$$Q(s,a,a') = Q(s,a,a') + \alpha \cdot \left(r\left(s,a,a'\right) + \gamma \cdot Val\left[s'\right] - Q\left(s,a,a'\right)\right) \quad (10)$$

where α is the learning rate.

Thus, the policy of the attacker can be derived:

$$Val(s) = \max_{\pi_1} \min_{a' \in \mathcal{A}_2} \sum_{a \in \mathcal{A}_1} \pi_1(s,a) \cdot Q(s,a,a') \quad (11)$$

The aim of the defender is to minimize the Q value of the attacker, whereas the attacker wants to maximize it. Thus, the policy of the defender can be derived by:

$$Val(s) = \min_{\pi_2} \max_{a \in \mathcal{A}_1} \sum_{a' \in \mathcal{A}_2} \pi_2(s,a') \cdot Q(s,a,a') \quad (12)$$

At each time step, the agent can choose its action according to the policy already learned (exploitation) or randomly (exploration). A challenge in reinforcement learning is to balance exploration and exploitation. A common method is the ϵ- greedy policy (Sutton and Barto 2011), where ϵ represents the probability of exploration and the probability of exploitation is $1-\epsilon$. The proposed algorithm is presented in Table 1.

Table 1. Proposed algorithm.

Algorithm 1: Proposed MARL for FPA process
Initialization:
$\forall s \in \mathcal{S}, a \in \mathcal{A}_1, a' \in \mathcal{A}_2$
$Q(s,a,a') = 1$, $Val(s) = 1$, $\pi_1(s,a) = 1/\lvert\mathcal{A}_1\rvert$
Let $\alpha = 1$
Loop
In state s
Choose a random action a from $\mathcal{A}_1$ with probability ϵ
Otherwise, choose action a with probability $\pi_1(s,a)$
Derive consumers' demand-response according to Eq. (2)
In state s'
Observe the defender's action a', derive $r(s,a,a')$ according to Eq. (3)-(4)
Update $Q(s,a,a')$ according to Eq. (10)
Derive the optimal attack policy according to Eq. (11).
$\alpha = \alpha \cdot decay$
End loop

4. Case study

We use a 5-bus power distribution system modified from a 6-bus test feeder (Kocar and Lacroix 2012), as an example to illustrate the application of our proposed model. In Fig. 1, node 1 is the infinite bus and nodes 2, 3, 4 and 5 are demand nodes. The capacities of lines L1, L2, L3 and L4 are 1600 kW, 3500 kW, 8000 kW, and 4000 kW, respectively. The electricity price at the initial time of learning is 35 cents/kW. The probability of consumers responding to the price is uniformly assumed to be 0.7.

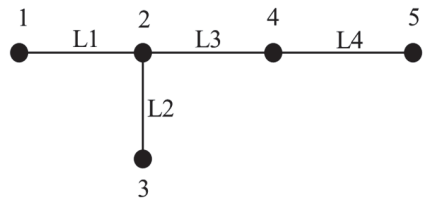

Fig. 1 5-bus test feeder considered.

We set the exploration probability ϵ as 0.3, the learning rate α to be 0.1 and the discount factor γ to be 0.9, following the common setting in (J. Yan et al. 2017). We assume the resources of the attacker and defender are 4 and 3, respectively. For feasibility, we assume that the attacker chooses pcr from {0, 0.1, 0.3} and the defender chooses c' from {0, 0.3}. If no defense is allocated to a consumer, the attack is successful; otherwise, the attack fails. In each trial, the attacker tries the action randomly or following the learned policies until the state of the system becomes s_2. This process is denoted as an episode. We try 100 episodes and get the following results.

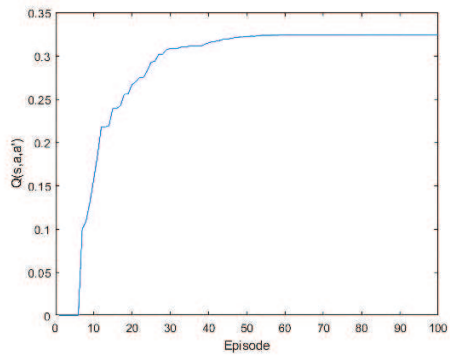

Fig. 2 Result of learned $Q(s,a,a')$ of the attacker.

The result of the learned $Q(s,a,a')$ of the attacker is presented in Fig. 2. It can be seen that at the beginning, the attacker searches the policy

randomly and the learning efficiency is quite low since the value of $Q(s,a,a')$ is nearly 0. After 6 episodes of learning, the attacker agent starts to learn the policy quickly and when the episode reaches 50, the agent learns the best policy and the game converges to the approximate Nash Equilibrium.

The mixed policies of attacker and defender are shown in Fig. 3 and Fig. 4, respectively. There are totally 15 actions available to the attacker and 5 actions available to the defender. The contents of the mixed policy of attacker and defender are listed in Table 2 and Table 3, respectively. It can be observed that the attacker distributes his resources to nodes 2 and 3 or nodes 1 and 4, or just attack node 2. This is the best policy for the attacker since the defender can only protect one node at each time and the districting the attack resources to different consumers with small pcr rather than only on one consumer with a large pcr can avoid getting nothing if the attacked node is protected by the defender. To minimize the value of $Q(s,a,a')$ of the attacker, the defender can protect node 3 or node 4 or do nothing. Since the defender can only allocate the defense resources to one consumer at most, the defender can choose one of the attackers' policies to protect, i.e., choose to protect node 3 considering the action 8 of the attacker, and choose to protect node 4 considering the action 12 of the attacker.

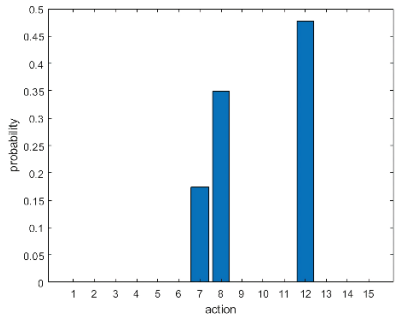

Fig. 3 Policy of attacker.

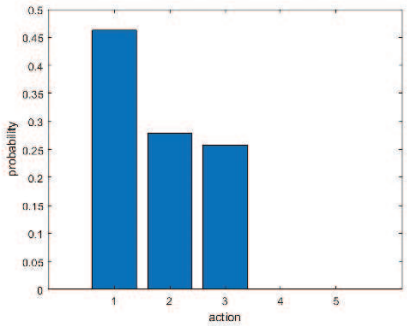

Fig. 4 Policy of defender.

Table 2 Policy of attacker.

action	pcr_1	pcr_2	pcr_3	pcr_4
7	0	0.2	0	0
8	0	0.2	0.2	0
12	0.2	0	0	0.2

Table 3 Policy of defender.

action	c'_1	c'_2	c'_3	c'_4
1	0	0	0	0
2	0	0	0	0.3
3	0	0	0.3	0

5. Conclusion

In this paper, the interaction between the attacker of a power grid by FPA and the defender who allocates protections with limited resources is modeled as a Markov game, where neither of the players has full knowledge of the game model as a result of the uncertainty in the consumers' consumption of energy. The Minimax-Q learning algorithm is adopted to find the approximate Nash Equilibrium solution for each player. The proposed framework is demonstrated on an illustrative 5-bus radial power system. The result shows that by learning from the interaction with the environment, the players can get mixed strategies to maximize their long-term return given the other player's policy.

Reference

Aghajani, GR, HA Shayanfar, and H Shayeghi. 2017. "Demand side management in a smart micro-grid in the presence of renewable generation and demand response." *Energy* 126: 622-637.

Amini, S., F. Pasqualetti, and H. Mohsenian-Rad. 2018. "Dynamic Load Altering Attacks Against Power System Stability: Attack Models and Protection Schemes." *IEEE Transactions on Smart Grid* 9 (4): 2862-2872.

Baran, Mesut E, and Felix F Wu. 1989. "Optimal capacitor placement on radial distribution systems." *IEEE Transactions on power Delivery* 4 (1): 725-734.

Chen, Ying, Junho Hong, and Chen-Ching Liu. 2018. "Modeling of intrusion and defense for assessment of cyber security at power substations." *IEEE Transactions on Smart Grid* 9 (4): 2541-2552.

Chen, Ying, Shaowei Huang, Feng Liu, Zhisheng Wang, and Xinwei Sun. 2019. "Evaluation of reinforcement learning based false data injection attack to automatic voltage control." *IEEE Transactions on Smart Grid* 10 (2): 2158-2169..

Dehghanpour, Kaveh, Zhaoyu Wang, Jianhui Wang, Yuxuan Yuan, and Fankun Bu. 2019. "A survey on state estimation techniques and challenges in smart distribution systems." *IEEE Transactions on Smart Grid* 10 (2): 2312-2322.

Deng, Ruilong, Gaoxi Xiao, and Rongxing Lu. 2017.

"Defending against false data injection attacks on power system state estimation." *IEEE Transactions on Industrial Informatics* 13 (1): 198-207.

Deng, Ruilong, Zaiyue Yang, Mo-Yuen Chow, and Jiming Chen. 2015. "A survey on demand response in smart grids: Mathematical models and approaches." *IEEE Transactions on Industrial Informatics* 11 (3): 570-582.

Ghosh, Soumyadip, Xu Andy Sun, and Xiaoxuan Zhang. 2012. "Consumer profiling for demand response programs in smart grids." IEEE PES Innovative Smart Grid Technologies.

Giraldo, J., A. Cardenas, and N. Quijano. 2017. "Integrity Attacks on Real-Time Pricing in Smart Grids: Impact and Countermeasures." *IEEE Transactions on Smart Grid* 8 (5): 2249-2257.

Hao, Yingshuai, Meng Wang, and Joe H Chow. 2018. "Likelihood analysis of cyber data attacks to power systems with Markov decision processes." *IEEE Transactions on Smart Grid* 9 (4): 3191-3202.

Kocar, Ilhan, and Jean-Sébastien Lacroix. 2012. "Implementation of a modified augmented nodal analysis based transformer model into the backward forward sweep solver." *IEEE Transactions on Power Systems* 27 (2): 663-670.

Liang, Gaoqi, Junhua Zhao, Fengji Luo, Steven R Weller, and Zhao Yang Dong. 2017. "A review of false data injection attacks against modern power systems." *IEEE Transactions on Smart Grid* 8 (4): 1630-1638.

Littman, Michael L. 1994. "Markov games as a framework for multi-agent reinforcement learning." In *Machine learning proceedings 1994*, 157-163. Elsevier.

Littman, Michael L, and Csaba Szepesvári. 1996. "A generalized reinforcement-learning model: Convergence and applications." ICML.

Liu, Yang, Shiyan Hu, and Tsung-Yi Ho. 2014. "Vulnerability assessment and defense technology for smart home cybersecurity considering pricing cyberattacks." 2014 IEEE/ACM International Conference on Computer-Aided Design (ICCAD).

Mishra, Subhankar, Xiang Li, Tianyi Pan, Alan Kuhnle, My T Thai, and Jungtaek Seo. 2017. "Price modification attack and protection scheme in smart grid." *IEEE Transactions on Smart Grid* 8 (4): 1864-1875.

Mohsenian-Rad, Amir-Hamed, and Alberto Leon-Garcia. 2010. "Optimal residential load control with price prediction in real-time electricity pricing environments." *IEEE Transactions on Smart Grid* 1 (2): 120-133.

A.-H. Mohsenian-Rad and A. Leon-Garcia. 2011. "Distributed internet-based load altering attacks against smart power grids." *IEEE Transactions on Smart Grid* 2 (4): 667-674.

Palensky, Peter, and Dietmar Dietrich. 2011. "Demand side management: Demand response, intelligent energy systems, and smart loads." *IEEE transactions on industrial informatics* 7 (3): 381-388.

Samadi, Pedram, Amir-Hamed Mohsenian-Rad, Robert Schober, Vincent WS Wong, and Juri Jatskevich. 2010. "Optimal real-time pricing algorithm based on utility maximization for smart grid." Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on.

Sutton, Richard S, and Andrew G Barto. 2011. "Reinforcement learning: An introduction."

Tan, Rui, Varun Badrinath Krishna, David KY Yau, and Zbigniew Kalbarczyk. 2013. "Impact of integrity attacks on real-time pricing in smart grids." Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security.

Tang, Daogui, Yiping Fang, Enrico Zio, and Jose Emmanuel Ramirez-Marquez. 2019. "Analysis of the Vulnerability of Smart Grids to Social Network-Based Attacks." 2018 3rd International Conference on System Reliability and Safety (ICSRS).

Tellbach, Denise, and Yan-Fu Li. 2018. "Cyber-Attacks on Smart Meters in Household Nanogrid: Modeling, Simulation and Analysis." *Energies* 11 (2): 316.

Wei, Longfei, Arif I Sarwat, Walid Saad, and Saroj Biswas. 2018. "Stochastic games for power grid protection against coordinated cyber-physical attacks." *IEEE Transactions on Smart Grid* 9 (2): 684-694.

Yan, Jun, Haibo He, Xiangnan Zhong, and Yufei Tang. 2017. "Q-learning-based vulnerability analysis of smart grid against sequential topology attacks." *IEEE Transactions on Information Forensics and Security* 12 (1): 200-210.

Yan, Ye, Yi Qian, Hamid Sharif, and David Tipper. 2012. "A survey on cyber security for smart grid communications." *IEEE Communications Surveys & Tutorials* 14 (4): 998-1010.

Zhang, Xialei, Xinyu Yang, Jie Lin, Guobin Xu, and Wei Yu. 2017. "On Data Integrity Attacks Against Real-Time Pricing in Energy-Based Cyber-Physical Systems." *IEEE Transactions on Parallel and Distributed Systems* 28 (1): 170-187.