applied sciences

MDPI

Article

Network Attack Path Selection and Evaluation Based on Q-Learning

Runze Wu, Jinxin Gong *, Weiyue Tong and Bing Fan

State Key Laboratory of Alternate Electrical Power System with Renewable Energy Sources, North China Electric Power University, Beijing 102206, China; wurz@ncepu.edu.cn (R.W.); twymistwy@163.com (W.T.); fanbing@ncepu.edu.cn (B.F.)

* Correspondence: gjxncepu@163.com; Tel.: +86-199-7500-3890

Abstract: As the coupling relationship between information systems and physical power grids is getting closer, various types of cyber attacks have increased the operational risks of a power cyber-physical System (CPS). In order to effectively evaluate this risk, this paper proposed a method of cross-domain propagation analysis of a power CPS risk based on reinforcement learning. First, the Fuzzy Petri Net (FPN) was used to establish an attack model, and Q-Learning was improved through FPN. The attack gain was defined from the attacker's point of view to obtain the best attack path. On this basis, a quantitative indicator of information-physical cross-domain spreading risk was put forward to analyze the impact of cyber attacks on the real-time operation of the power grid. Finally, the simulation based on Institute of Electrical and Electronics Engineers (IEEE) 14 power distribution system verifies the effectiveness of the proposed risk assessment method.

Keywords: power CPS; data tampering attack; risk assessment; Q-Learning algorithm; Fuzzy Petri Net

Citation: Wu, R.; Gong, J.; Tong, W.; Fan, B. Network Attack Path Selection and Evaluation Based on Q-Learning. *Appl. Sci.* **2021**, *11*, 285. https://doi.org/10.3390/app11010285

Received: 23 October 2020
Accepted: 26 December 2020
Published: 30 December 2020

1. Introduction

Smart grid is a typical cyber-physical system (CPS), which uses intelligent terminals such as massive sensors and advanced metering equipment to realize remote monitoring, control, and protection of the grid [1,2]. Advanced Persistent Threat (APT) attacks use the openness and easy accessibility of smart terminals to invade, carry out multi-step attacks through network security vulnerabilities, and, finally, enter the main station and destroy power production on a large scale [3]. For example, Ukraine suffered a "Black Energy" attack in 2015 [4] and there was also Israel's power outage in 2009 [5]. Therefore, predicting the path of a multi-step attack and analyzing the cross-layer risk of the power CPS under the attack will help ensure the safe and stable operation of the power CPS [6].

When attacking, the attacker tends to choose the path with low attack cost and high attack profit to carry out the invasion, as this is the best attack path. The purpose of the best attack path discovery is to analyze the attacking behavior by alert correlation technology, reveal the hidden logic, construct attack scenarios, and then infer the subsequent attack steps of attackers, providing important evidence for active defense of network security [7]. It has been an important method of dealing with the multi-step attacks [8]. Current research is usually static analysis based on experience. Literature used a forward search strategy to find hidden attack paths. Reference [9] adopted heuristic search algorithms to generate attack graphs. Reference [10] proposed pruning attack graph branches to exploit Greedy search strategy finds the attack path. Reference [11] made use of an attack tree and a genetic algorithm to solve the optimal attack path problem, and found the solution through a genetic algorithm. However, the above-mentioned research has high computational complexity and is difficult to apply to large-scale networks. At the same time, it cannot reflect the influence of the different attackers on the path selection.

In order to solve the above problems, the Q-Learning algorithm is introduced to discover the best attack path. Q-Learning belongs to a category of semi-supervised learning

algorithms [12]. Because of its simplicity and convergence advantages [13,14], it has been widely used in various fields of robot path finding and planning [15]. In recent years, it has also begun to be applied in finding attack paths. Reference [16] put forward a Q-Learning method to identify key attack sequences in consideration of physical system behavior, with good results. However, the research mainly focused on network topology attacks, without considering the security vulnerabilities of the network itself. Reference [17] proposed a method of using Q-Learning to analyze the attack path based on the attack graph. However, when quantifying the threat of a network attack, only the benefit of the attack is considered and the attack cost is not considered. There is also a problem of computational efficiency.

Based on the above problems, this paper establishes an FPN-Q learning algorithm to find the best attack path. Fuzzy Petri Net (FPN) [18] combines the ability of Petri Net to describe asynchronous concurrency and graphical representation with the fuzzy reasoning ability of fuzzy systems. Therefore, this algorithm uses FPN to model the network attack process with uncertain characteristics, and use the fuzzy inference parameters of FPN to improve the Q-Learning algorithm, which improves the learning efficiency of the algorithm. In addition, considering the attack cost and the attack reward, the attack gain is proposed to quantify the threat of the attack path to the system. In order to evaluate the impact of the discovered best attack path on the real-time operation of the power grid, an information-physical system coupling model is established based on the function of load control, and the risk indicator is proposed. Finally, the efficiency of the FPN-Q Learning algorithm under multiple attack modes and the impacts on CPS operation are simulated based on the Institute of Electrical and Electronics Engineers (IEEE) 14-nodes power distribution system. The results show that the method of attack path discovery has high efficiency and accuracy. It provides a feasible analysis scheme for judging the operation of the system under multi-step attacks, and provides a reliable basis for ensuring the stable operation of power CPS.

2. FPN-Q Learning Algorithm to Determine the Best Attack Path

2.1. Attack Model

Cyber attacks have complex and random characteristics. FPN has the capability of describing concurrent events and graphical representation of Petri nets. Moreover, FPN can express this transition process concisely and clearly, avoiding the problem of state space explosion. In addition, FPN also has the fuzzy inference ability of fuzzy systems. Its place credibility and transition credibility can well represent the process of network attacks and the ambiguity of the attacks. Since the above characteristics of FNP meet the modeling needs, this study uses FPN to establish a network attack model.

This paper is based on the network attack model established by FPN, which is a four-tuple:

$$M = \{H, T, \alpha, \mu\}$$

(1) $H = \{h1, h2, h3, \ldots, hn\}$ is a finite set of places h, which represents the host of the information system in the model;
(2) $T = \{t1, t2, t3, \ldots, tm\}$ is a finite set of transitions t, which represents the exploitable vulnerabilities of the system host in the model;
(3) α represents the risk value caused by the system host represented by the place after being invaded, that is, the threat index; and
(4) μ: $T \times H \rightarrow (1,10)$ represents the confidence of the transition rule, that is, the probability that a certain transition is triggered. In the network attack model, it represents the complexity of the attack process. The higher the attack complexity, the lower the possibility of being attacked. Attack complexity is affected by many factors such as attack tools and attacker experience. Its value is given according to the Common Vulnerability Scoring System (CVSS) [19].

This method uses the FPN place to represent the information system host, and uses transitions to represent the exploitable vulnerabilities of the system. This method makes the complex network attack processes more concise and intuitive while reasonably considering

the actual network attack. In addition, the concept of the FPN place's credibility and transition confidence is also used, which can well represent the ambiguity and uncertainty of the network attack process and its impact. Moreover, it can make subsequent analysis more reasonable and effective.

2.2. FPN-Q Learning Algorithm to Determine the Best Attack Path

This algorithm introduces the Q-Learning algorithm to analyze the network attack model established in 1.1, and uses the parameters of FPN to improve: (1) Transition confidence μ of FPN is used to define the attack cost of a single-step attack; The place credibility α of FPN is used to define the attack revenue, that is, the threat of each attack to the system. The algorithm in this paper starts from the attacker's point of view, and comprehensively considers the attack cost and attack benefit, which are used to define the attack gain indicator. Attackers tend to choose attack paths with low attack costs but high threats to the system, that is, the path with the highest attack gain is the best attack path. (2) As described in 1.1, μ can well represent the randomness of the attacker's selection of vulnerabilities. The algorithm uses μ to optimize the exploration process of the Q-Learning algorithm, and accelerates the convergence speed of the Q function without changing the final result. The algorithm is divided into two phases: the learning phase and the attack phase.

2.3. Learning Stage

In the traditional Q-Learning algorithm, the agent selects an action to act on the unknown environment during each iteration. After the environment receives the action, it generates an enhanced signal (reward or punishment) to feed back to the node. The node chooses the next action based on the enhanced signal and the new state of the environment. The principle of action selection is to increase the probability of receiving a positive reward. After continuous learning and trial and error, the node finds the optimal action control strategy and obtains cumulative returns. It is worth noting that the traditional Q-Learning algorithm randomly selects actions with equal probability in the exploration phase. However, in a network attack, the attacker has mastered all or part of the information about the security vulnerabilities of the information system before the attack. Therefore, the attack path will be selected based on experience rather than randomly selected with equal probability. The probability of each vulnerability being selected is related to the attack complexity: the higher the attack complexity, the smaller the chance of being selected. Therefore, this paper introduces the transition confidence of the FPN model μ to optimize the exploration process of Q-Learning. In the exploration phase, the probability that the vulnerability j of host i is exploited p_{ij} is:

$$p_{ij} = \frac{\frac{1}{\mu_{ij}}}{\sum_{j=1}^{n} \frac{1}{\mu_{ij}}} \tag{1}$$

where, n represents the number of exploitable vulnerabilities of host i, μ_{ij} is the attack complexity of exploitable vulnerability j of host i, and its value range is (1,10).

In the attack model, the attack cost is related to the attack complexity of the security vulnerability. Generally, the more complex the attack, the higher the attack cost. Therefore, this paper defines the attack complexity of security vulnerabilities as the attack cost. The attack proceeds are the threats to information systems caused by network attacks, which are related to the nature of the vulnerability itself. According to the FPN attack model established in Section 2.1, the following definitions are given:

Definition 1. *Initial single-step attack gain.*

The initial single-step attack gain is the threat to the system caused by the attacker invading host j through host i before the start of the learning process, expressed by the reward function g_{ij}:

$$g_{ij} = r_{ij} / \mu_{ij} \tag{2}$$

$$r_{ij} = \alpha_i \cdot \alpha_j \tag{3}$$

where, r_{ij} is the single-step attack reward, α_i is the threat value caused by the intrusion system host i to the network, and α_j is the threat value caused by the intrusion system host j to the network.

Definition 2. *Single-step cumulative attack gain.*

The cumulative single-step attack gain is the attack gain obtained by the attacker from host i invading host j after multiple intrusion learning, denoted by $Q(h_i, t_{ij}, h_j)$.

$$Q(h_i, t_{ij}, h_j) \leftarrow (1-\alpha)Q(h_i, t_{ij}, h_j) + \alpha[g_{ij} + \gamma \max_{t_{jk} \in T_j} Q(h_j, t_{jk}, h_k)] \tag{4}$$

where, β represents the learning factor, γ represents the discount factor between delayed return and immediate return, T_j represents the optional vulnerability of the next attack after the intrusion of host j, and $\max_{t_{jk} \in T_j} Q(h_j, t_{jk}, h_k)$ represents the maximum gain that the attacker can obtain in the next attack after the host invades host j.

Based on the above definition, the attacker is regarded as the agent of the Q-Learning algorithm, and the information system is regarded as the environment where the attacker's attack behavior is given feedback. The basic idea of the learning process is: the agent first starts from the initial intrusion node according to the scanning of the network environment, and then selects one from the current intrusive system vulnerabilities to invade according to Formula (1), and finally updates the single-step cumulative attack gain of this attack according to Formula (4) The attacker takes the attack path to the target host as a scenario-based learning until the Q value of each optional intrusion step reaches the maximum and converges. The specific learning process is shown in the Algorithm 1 environment:

Algorithm 1 environment: FPN-Q Learning algorithm: learning stage

Initialization: Initialize the Q-table
Determine the attack access host h_0 and the target host h_t.
for current number of episodes $\leq$ maximal episodes
do
Reset: Current host: $h_c = h_0$;
Target host = h_t;
while h_c ~= h_t **do**
Obtain all vulnerabilities t from the attackable hosts h;
Obtain the probability p_{ei} of each vulnerability being selected according to (1)
Choose a vulnerability t_{ei} according to p_{ei};
Obtain evaluative feedback: Generate the reward r_{t+1} according to (2);
Update the current host: $h_c = h_e$;
Update the value of Q(h_e, v_i) according to (4).
end while
end for

2.4. Attack Stage

The basic idea of the attack stage is: after learning, the attacker will select the host vulnerability with the highest cumulative gain in a single step for each step of the attack, until it invades the target host. Therefore, the optimal attack path and its attack gain G for multi-step attack are obtained, and the algorithm is shown in the Algorithm 2 environment.

Algorithm 2 environment: FPN-Q Learning algorithm: learning stage

Initialization: Initialize the G
Determine the attack access host h_0 and the target host h_t.
Set: Current host: $h_c = h_0$;
Target host = h_t;
while h_c ~= h_t **do**
Obtain all attackable hosts h;
Obtain $Q(h_i,t_{ij},h_j)$ of each attackable host;
Find the max $Q(h_i,t_{ij},h_e)$ from attackable hosts;
Update G: $G \leftarrow G + g_{ie}$
Update the current host: $h_c = h_e$;
end while

3. Information-Physical Cross-Layer Risk Spread Model

In order to evaluate the impact of the best attack path found on the power CPS, the following information-physical cross-layer risk propagation model is established. Power CPS is a multi-dimensional heterogeneous system which fully integrates the physical network and information network of the power system [20]. Through the coordination of computing equipment, sensor equipment, and communication equipment. The overall operating performance of the power system is optimized by physical equipment, etc. The CPS structure under the power Internet of Things is shown in Figure 1 which can be divided into three levels from bottom to top: user load, terminal sensing, and control and decision. The physical system and information system realize the interaction between information flow and energy flow through intelligent terminals. On the one hand, the terminal equipment collects the electricity consumption data and equipment status data of different power users, which are used by the control center to analyze the operation status of the power system and formulate appropriate control plans [21]. On the other hand, they receive commands from the control center to regulate electrical primary equipment, such as increasing or decreasing generator output, adjusting transformer taps, etc. [22].

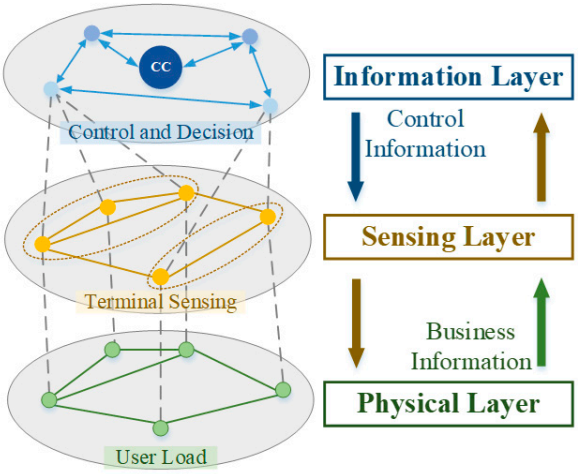


Figure 1. Cyber-physical system (CPS) structure under the power Internet of Things.

Different from traditional physical grid cascading failures, information-physical dual-network cascading failures caused by cyber-attacks will cause more serious consequences [23,24]. In terms of security risks in the information space, smart terminal equipment is the only way to spread to the power space. Therefore, attackers usually choose widely distributed smart terminals as access points to attack at present. After obtaining permission, scanning software is used to obtain the security vulnerabilities that can be used by each host of the system. Eventually attackers initiate a multi-level invasion, modify the configuration file or business data in the server after entering the control center.

It will cause the control center to incorrectly perceive the current state of the physical power grid [25], and affect the operation of the physical power grid by issuing wrong control commands. This paper mainly researches the attacks of tampering with system data in this situation.

4. Security Risk Assessment of Electric Power CPS under Cyber Attack

This paper selects the function of load control for risk assessment, and its propagation process is as follows:

(1) The attacker uses a certain strategy to launch an attack through the smart terminal to enter the control center, and then randomly or deliberately tamper with the business data according to the knowledge of the physical power grid, so that the load of some physical nodes exceeds the predetermined quota.

(2) The control center considers that the load on the node exceeds the capacity, and judges that the node is faulty. Therefore, the control center performs load reduction according to Formulas (5)–(10) with the goal of minimizing load loss, and issues a control command to cut off part of the load of the node and its neighboring nodes to ensure the safe and stable operation of the system.

$$\min I = \sum_{i=1}^{N} Ls_i \tag{5}$$

where, I is the total load loss of the physical system, N is the number of load shedding nodes in the physical system, and Ls_i is the load loss of node i.

At the same time, considering the power flow constraints of the distribution network and the observable and controllable nodes, the following constraints are obtained:

$$\begin{cases} P_i = G_{ii}U_i^2 + \sum\limits_{j \in s(i)} U_i U_j (G_{ij} cos\theta_{ij} + B_{ij} sin\theta_{ij}) \\ Q_i = -B_{ii}U_i^2 - \sum\limits_{j \in s(i)} U_i U_j (B_{ij} cos\theta_{ij} - G_{ij} sin\theta_{ij}) \end{cases} \tag{6}$$

$$\begin{cases} U_{\min} \le U_i \le U_{\max} \\ I_{\min} \le I_{ij} \le I_{\max} \end{cases} \tag{7}$$

where, P_i and Q_i are the active power and reactive power of node i, respectively, $s(i)$ is the set of nodes connected to node i, G_{ii} and B_{ii} are the self-conductance and self-susceptance of node I, respectively; G_{ij} and B_{ii} are the conductance and susceptance between nodes i and j, respectively; U_i and U_j are nodes, respectively The voltages of i and j; θ_{ij} is the phase angle difference between nodes i and j; $U_{\min}$ and $U_{\max}$ are the lower and upper limits of the voltage of node i, respectively; $I_{\min}$ and $I_{\max}$ are the lower and upper limits of the line current.

$$\sum_{i \in NG} PG_i - \sum_i PD_i + \sum_i Ls_i = 0 \tag{8}$$

$$PG_i^{\min} \le PG_i \le PG_i^{\max} \tag{9}$$

$$0 \le Ls_i \le PD_i \tag{10}$$

where, PG_i is the power generation of the controllable power generation equipment connected to node i, $PG_i^{\min}$, and $PG_i^{\max}$ are the lower limit and upper limit of the generator's power generation capacity, and PD_i is the load of node i.

(1) The intelligent terminal adjusts the load of the physical system according to the wrong instruction issued by the control system;

(2) Each node of the physical power grid adjusts the load according to the control command, and some nodes will lose the load of normal operation. Therefore, the physical power grid trend will change, and new business data will be transmitted to the control center.

CPS security risks under data attacks are related to the threat value of the attack path to the information system and the consequences of cascading failures caused. Section 2.4 discusses the attack benefit G of the best attack path, that is, the threat to the information system caused by the attack through this path. Based on the above, the risk of CPS under data attack is defined as:

$$R = G \cdot \sum_{i=1}^{n} Ls_i \tag{11}$$

$$ls_i = \sum_{j=1}^{n} ls_{ij} \tag{12}$$

$$ls_{ij} = \frac{Ls_{ij}}{Load_j} \tag{13}$$

where, R_i represents the risk of power CPS when the load data of node i is tampered with. G represents the threat to the information system that an attacker launches an attack on the network through a predetermined path. p_i represents the probability that the data of node i is modified, which is related to the attacker's familiarity with the operation of the system. Ls_{ij} indicates that the load change of node i causes the load loss of node j. $Load_j$ represents the original load of node j. ls_{ij} represents the load loss rate of node j caused by the load change of node i. ls_i Indicates the load loss rate of the entire system caused by tampering of the load data of node i, which is the sum of load loss rates of all nodes.

5. Simulation Evaluation

5.1. Establishment of Simulation Environment

In order to verify the feasibility and effectiveness of the proposed algorithm, the Supervisory Control And Data Acquisition (SCADA) power distribution system was selected to establish a network attack model based on FPN, as shown in Figure 2. The model mainly includes Demilitarized Zone (DMZ) domains, work stations, and control centers. The network area is divided by installing firewalls, and communication rules between domains are formulated to ensure that external access cannot reach the intranet area. The specific access rules are introduced as follows: (1) The data collected by the terminal can only be accessed to the master station through the DMZ domain; (2) The workstation can realize two-way communication with the DMZ domain and the control center; and (3) The DMZ domain and control center can only communicate with workstations. Before quantitative modeling, one makes the following assumptions about the attacker's capabilities: (1) The attacker understands Direct-Attached Storage (DAS) and has the latest DAS vulnerability information and (2) Attackers can deliberately and effectively use social engineering to attack.

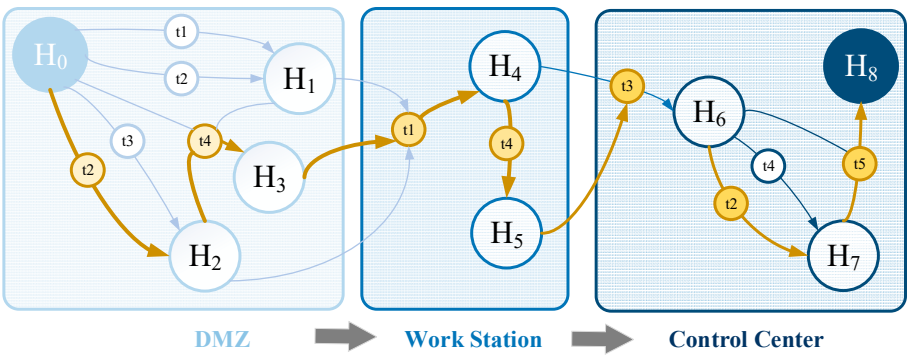


Figure 2. Attack model. t1: Input validation error; t2: Boundary condition error; t3: Buffer error; t4: Information leakage; t5: Permission acquisition and access control.

Depending on the ability of the attacker, the attacker can launch attacks on different system hosts. There are two most common modes:

(1) The attacker uses the power distribution terminal equipment as the access point to further invade the DMZ area. By invading the DMZ area, The system's security vulnerabilities are continuously used to increase the authority until entering the control center application server because of the invasion. Deliberate tampering of business data will cause greater losses to the system. At this time, the attacker's target is H_8.

(2) The attacker uses the power distribution terminal equipment as the access point to invade the DMZ area and continuously use the system security vulnerabilities to increase the authority until entering the operator's Human Machine Interface (HMI). The business data is randomly tampered on the HMI side, because the attacker does not have detailed physical power grid parameters and data. At this time the attacker's target is H4.

At the same time, in order to analyze the risks caused by network attacks to the power CPS, the IEEE14-node system shown in Figure 3 is selected as the experimental model. Nodes 1, 2, 5, and 7 of the system are distributed power sources, and nodes 4, 8, and 13 are important load nodes. The total power generation capacity of the distributed power generation is 3.7 MW, and the sum of the power requirements of each load is 3.19 MW.

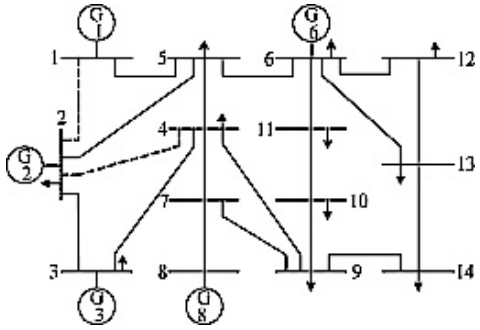


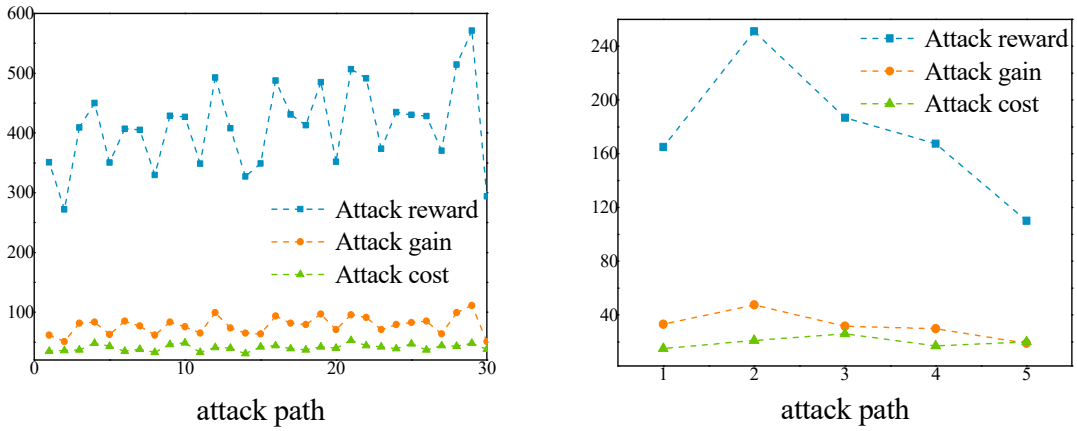
Figure 3. The Institute of Electrical and Electronics Engineers (IEEE) 14 node system.

5.2. Analysis of Experimental Results

5.2.1. Experimental Results—Security Analysis of the Information Layer

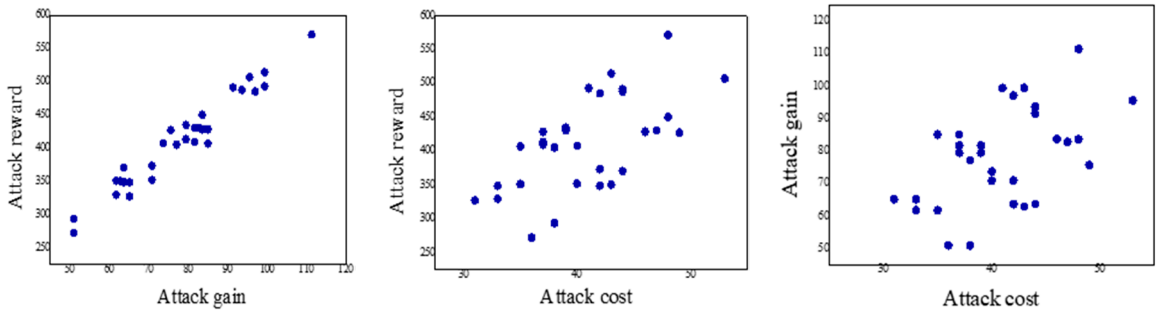
The attack gain index in this paper is based on the attack reward and attack cost. As a result, the relationship of the three should be studied first. There are 30 attack paths in attack mode 1, and five attack paths in attack mode 2. Figure 4 (its abscissa variables are the attack path number) shows the attack gain, attack reward and attack cost of each attack path in the two attack modes. It can be seen that the value trend of the attack gain is roughly the same as the attack reward, but the attack cost reduces the attack gain to a greater extent. When the attack reward is small, the attack gain may even appear to be smaller than the attack cost (path 5 of attack mode 2), which is not good for the attacker. In order to further illustrate the relationship of the three, a scatter plot between the three under two attack modes is drawn in Figures 5 and 6. Figures 5a and 6a show that the attack gain is closely related to the attack reward, and a high attack reward will bring a high attack gain. Figure 5b,c and Figure 6b,c show that high attack rewards and attack gains often require high attack costs. However, the path with the highest attack cost will not get the highest attack reward and highest attack gain. This is due to the different nature of each security vulnerability. The attack complexity of the security vulnerability that poses the greatest threat to the system may not be high (such as Common Vulnerabilities and Exposures (CVE)-2004-0893, it is a user privilege escalation vulnerability. According to the CVSS, its threat index reaches 7.2 (the highest is 10), but the attack complexity is only moderate.). Therefore, there is an optimal attack path, which reduces the cost of the

attack and obtains a higher attack return. At this time, the attacker obtains the maximum attack gain.



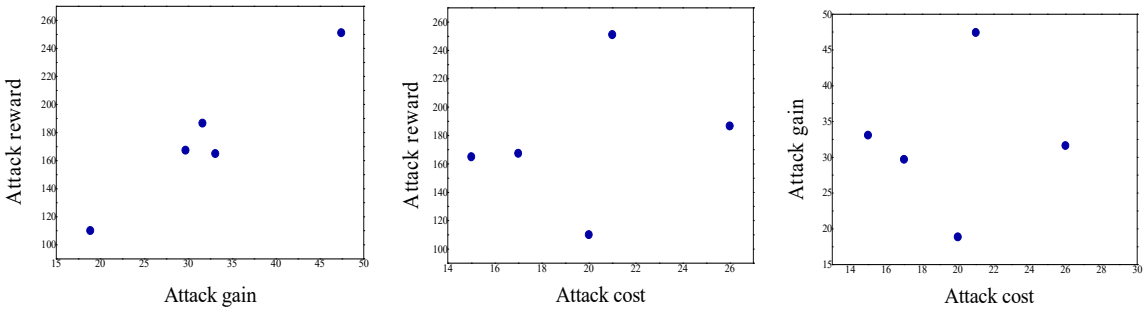
(**a**) Attack mode 1 (**b**) Attack mode 2

Figure 4. The relationship between attack gain, attack reward, and attack cost: (**a**) attack mode 1, (**b**) attack mode 2.



(**a**) (**b**) (**c**)

Figure 5. The relationship between attack gain, attack reward, and attack cost in mode 1: (**a**) The relationship of attack gain and attack reward; (**b**) The relationship of attack cost and attack reward; and (**c**) The relationship of attack cost and attack gain.

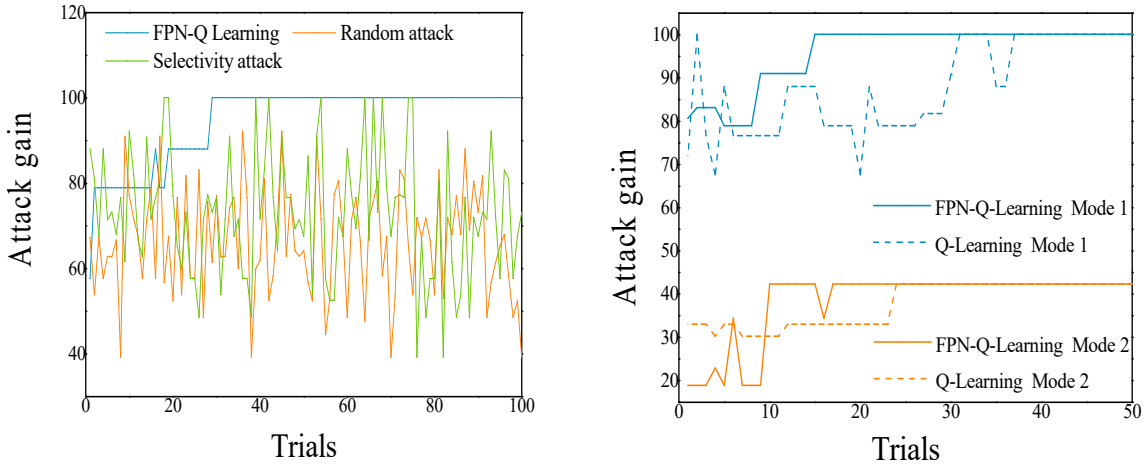


(**a**) (**b**) (**c**)

Figure 6. The relationship between attack gain, attack reward, and attack cost in mode 2: (**a**) The relationship of attack gain and attack reward; (**b**) The relationship of attack cost and attack reward; and (**c**) The relationship of attack cost and attack gain.

In the two attack modes, the attacker uses the FPN-Q Learning algorithm to perform attack learning as shown by the solid line in Figure 7. It can be seen that the attack gain

obtained by the attacker invading the control center is much greater than that obtained by invading the control center. This is because once the information host in the control center is destroyed, the security threat to the system is greater. Therefore, it is necessary to strengthen the monitoring and protection of the control center host.



(**a**) Attack mode 1 (**b**) Attack mode 2

Figure 7. (**a**) Comparison with random attack and selective attack and (**b**) Comparison with unimproved Q-Learning algorithm

5.2.2. Experimental Results—Algorithm Comparison

(1) Compare the algorithm proposed in this paper with random attacks and selective attacks. Taking attack mode 1 as an example, the attack gains obtained after 100 attacks are shown in Figure 7a (its abscissa variables are the experiment times). It can be seen from the figure that compared with random attacks, selective attacks have a greater redirection through the best attack path to get the greatest attack gain, but the algorithm proposed in this paper shows obvious advantages. After learning 15 times, you can find the best attack path and get the maximum attack gain.

(2) Compare the algorithm proposed in this article with the traditional Q-Learning algorithm. Use the traditional Q-Learning algorithm to perform path learning and attack gain calculation for attack mode 1 and attack mode 2. The comparison of the result with two methods is shown in Figure 7b. It can be seen from the figure that the attack gain obtained by the improved algorithm is consistent with the traditional algorithm, indicating that the algorithm in this paper has high accuracy. However, in terms of learning speed, the algorithm proposed in this paper can find the best attack path faster. Especially for the more complex attack mode (mode 1), the traditional algorithm needs 30 times of learning to get the best gain and reach convergence, while the algorithm in this paper only needs 15 times. The learning efficiency has nearly doubled, showing obvious advantages.

5.2.3. Experimental Results—Cross-Layer Risk Communication Analysis

Assuming that the attacker only modifies the load of a single node each time, and the offset is 1 MW, the ls_i of each node is obtained as shown in the Figure 8. It can be seen from the figure that when the load shedding amount is the same, the load loss rate caused by changing the data of No. 6 and No. 9 nodes is higher. Therefore, these two nodes are defined as high-risk nodes.

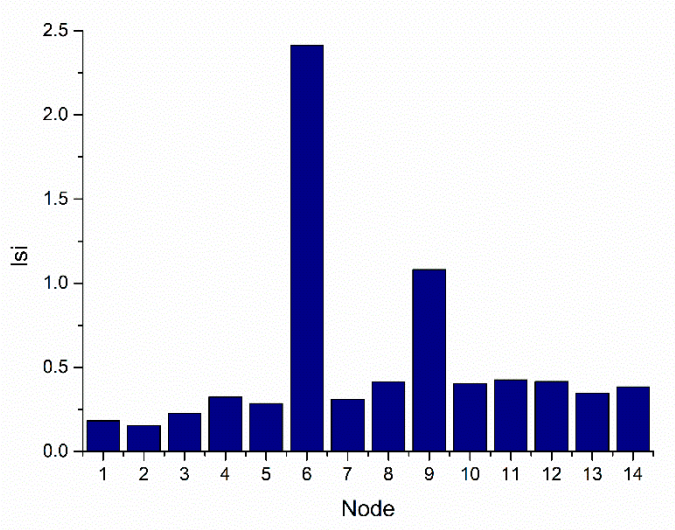


Figure 8. The load loss of every node.

In the mode 1, the attacker enters the control master station and masters the operation of the power grid. At this time, the possibility p_i of tampering with data of nodes with high load loss rate is higher. In the mode 2, the attacker does not enter the control center and does not understand the operation of the power grid. At this time, only business data can be modified randomly, so the probability of any physical node data being modified is the same. The risks faced by each node in the two attack modes are shown in Table 1:

Table 1. Node risk under different attack modes.

Node	Risk		Node	Risk	
	Attack Mode 1	Attack Mode 2		Attack Mode 1	Attack Mode 2
1	0.05571	0.16091	8	0.28763	0.36563
2	0.0402	0.13669	9	1.96169	0.95486
3	0.08665	0.20068	10	0.27507	0.35756
4	0.17548	0.28559	11	0.30509	0.37656
5	0.13683	0.25218	12	0.29032	0.36734
6	9.77915	2.13194	13	0.20045	0.30523
7	0.15986	0.27258	14	0.24754	0.33920

It can be seen from the above table that high-risk nodes (No. 6 and No. 9) face significantly higher risks in attack mode 1 than in attack mode 2. This is because in mode 1, the attacker enters the control master station and masters the operation of the power grid. At this time, the possibility p_i of high-risk node data to be tampered is greater, and the possibility of ordinary node data being tampered is less. However, the opposite is true in mode 2. The attacker does not enter the control center in mode 2, so he does not know the operation of the system and will randomly tamper node data. In this case, the probability p_i of each node's data being tampered is the same. Therefore, the possibility of data tampering of ordinary nodes in mode 2 is greater than that in mode 1, and the risks faced are also increased. However, on the whole, mode 1 poses a greater risk to the system than mode 2, because the attacker in mode 1 has more power grid operating data and can change the operating data in a targeted manner. From the above analysis, we can see that the risks faced by the power CPS system are closely related to the attacker's attack mode. Therefore, for information systems, the protection of confidential data and control centers needs to be strengthened. For physical nodes, high-risk nodes are protected, and the power

load needs to be allocated reasonably to minimize the risk of the power CPS system when the information system is invaded.

As can be seen from the above table, the risk to the system caused by the attack mode 1 is greater than the attack mode 2. This is because the attacker under the attack mode can change the operation data in a targeted manner with more power grid operation data, causing greater risks to the system. At the same time, it can be seen from the table that ordinary nodes (such as No. 6 and No. 9) have greater operational risks under the same load shedding amount. The reason for this is these nodes have less output and are not located in important positions of the system, and the attacker is less difficult to attack. Therefore, it is necessary to strengthen the protection of ordinary nodes to avoid major losses to the power CPS.

6. Conclusions

This research uses the fuzzy reasoning ability of FPN to improve the Q-Learning algorithm, and uses Q-Learning to solve the shortcomings of FPN's inability to self-learn at the same time, thereby finding the most vulnerable path in the network system (the attacker can obtain the highest gain route for). Compared with traditional methods, this algorithm saves computing resources and can better reflect the impact of the difference between the attackers and the attack targets on the network. The experimental results show that this method has high accuracy, which can better help the study of defense measures against network attacks. In addition, this paper establishes an information-physical risk propagation model to evaluate the risks brought by different attack modes to the operation of the power grid. This allows the research in this paper applicable to the identification and protection of key nodes of the CPS system, cascading failure analysis, and the transmission of confidential data in management and other fields.

Author Contributions: Conceptualization, R.W. and J.G.; methodology, R.W. and J.G.; software, J.G.; validation, J.G.; formal analysis, J.G. and W.T.; investigation, J.G. and B.F.; resources, R.W.; data curation, J.G.; writing original draft preparation, J.G.; writing—review and editing, R.W. and J.G.; visualization, J.G. and W.T.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 51677065).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xin, S.; Guo, Q.; Sun, H.; Zhang, B.; Wang, J.; Chen, C. Cyber-Physical Modeling and Cyber-Contingency Assessment of Hierarchical Control Systems. *IEEE Trans. Smart Grid* **2017**, *6*, 2375–2385. [CrossRef]
2. Sridhar, S.; Hahn, A.; Govindarasu, M. Cyber–Physical System Security for the Electric Power Grid. *Proc. IEEE Inst. Electr. Electron. Eng.* **2011**, *100*, 210–224. [CrossRef]
3. Alshamrani, A.; Myneni, S.; Chowdhary, A.; Huang, D. A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 1851–1877. [CrossRef]
4. Liang, G.; Weller, S.R.; Zhao, J.; Luo, F.; Dong, Z.Y. The 2015 ukraine blackout: Implications for false data injection attacks. *IEEE Trans. Power Syst.* **2016**, *32*, 3317–3318. [CrossRef]
5. Staff, T. Steinitz: Israel's Electric Authority Hit by 'Severe' Cyber-Attack. *The Times of Israel*. 2016. Available online: https://www.timesofisrael.com/steinitz-israels-electric-authority-hit-by-severe-cyber-attack/ (accessed on 26 January 2016).
6. Liu, S.; Chen, B.; Zourntos, T.; Kundur, D.; Butler-Purry, K. A Coordinated Multi-Switch Attack for Cascading Failures in Smart Grid. *IEEE Trans. Smart Grid* **2014**, *5*, 1183–1195. [CrossRef]
7. Zhou, Y.; Chen, N. The LAP under facility disruptions during early post-earthquake rescue using PSO-GA hybrid algorithm. *Fresen. Environ. Bull.* **2019**, *28*, 9906–9914.
8. Liu, X. A network attack path prediction method using attack graph. *J. Ambient Intell. Humaniz. Comput.* **2020**, 1–8. [CrossRef]

9. Swiler, L.P.; Phillips, C.; Ellis, D.; Chakerian, S. Computer-attack graph generation tool. In Proceedings of the DARPA Information Survivability Conference and Exposition II, DISCEX'01, Anaheim, CA, USA, 12–14 June 2001; Volume 2, pp. 307–321.
10. Zhang, B.; Lu, K.; Pan, X.; Wu, Z. Reverse search based network attack graph generation. In Proceedings of the International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December 2009; pp. 1–4.
11. Singhal, A.; Ou, X. Security risk analysis of enterprise networks using probabilistic attack graphs. In *Network Security Metrics*; Springer: Cham, Switzerland, 2017; pp. 53–73.
12. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: London, UK, 1998.
13. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
14. Jaradat, M.A.K.; Al-Rousan, M.; Quadan, L. Reinforcement based mobile robot navigation in dynamic environment. *Robot. Comput. Integr. Manuf.* **2011**, *27*, 135–149. [CrossRef]
15. Maoudj, A.; Hentout, A. Optimal path planning approach based on Q-learning algorithm for mobile robots. *Appl. Soft Comput.* **2020**, *97*, 106796. [CrossRef]
16. Yan, J.; He, H.; Zhong, X.; Tang, Y. Q-Learning-Based Vulnerability Analysis of Smart Grid against Sequential Topology Attacks. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 200–210. [CrossRef]
17. Yousefi, M.; Mtetwa, N.; Zhang, Y.; Tianfield, H. A Reinforcement Learning Approach for Attack Graph Analysis. In Proceedings of the 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), New York, NY, USA, 1–3 August 2018; pp. 212–217.
18. Chen, S.M. Fuzzy backward reasoning using fuzzy Petri nets. *IEEE Trans. Syst. Man Cybern. Syst. Part B (Cybern.)* **2000**, *30*, 846–856. [CrossRef] [PubMed]
19. Mell, P.; Scarfone, K.; Romanosky, S. Common vulnerability scoring system. *IEEE Secur. Priv.* **2006**, *4*, 85–89. [CrossRef]
20. Xu, L.; Guo, Q.; Yang, T.; Sun, H. Robust Routing Optimization for Smart Grids Considering Cyber-Physical Interdependence. *IEEE Trans. Smart Grid* **2018**, *10*, 5620–5629. [CrossRef]
21. Li, X.; Zhou, C.; Tian, Y.C.; Xiong, N.; Qin, Y. Asset-based dynamic impact assessment of cyberattacks for risk analysis in industrial control systems. *IEEE Trans. Ind. Inform.* **2017**, *14*, 608–618. [CrossRef]
22. Ye, X.; Zhao, J.; Zhang, Y.; Wen, F. Quantitative vulnerability assessment of cyber security for distribution automation systems. *Energies* **2015**, *8*, 5266–5286. [CrossRef]
23. Zhou, X.; Yang, Z.; Ni, M.; Lin, H.; Li, M.; Tang, Y. Analysis of the Impact of Combined Information-Physical-Failure on Distribution Network CPS. *IEEE Access* **2020**, *8*, 44140–44152. [CrossRef]
24. Stellios, I.; Kotzanikolaou, P.; Psarakis, M.; Alcaraz, C.; Lopez, J. A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 3453–3495. [CrossRef]
25. Teixeira, A.; Shames, I.; Sandberg, H.; Johansson, K.H. A secure control framework for resource-limited adversaries. *Automatica* **2015**, *51*, 135–148. [CrossRef]