Q-learning Enabled Intelligent Energy Attack in Sustainable Wireless Communication Networks

Long Li*,Yu Luo†, Lina Pu*
*Department of CS, The University of Alabama, Tuscaloosa, AL 35487
†Department of ECE, Mississippi State University, Mississippi State, MS 39762
Email: lli90@crimson.ua.edu, yu.luo@ece.msstate.edu, lina.pu@ua.edu

***Abstract*— In this paper, we identify a new security issue, called the malicious energy attack, in sustainable wireless communication networks (SWCNs). We show that by providing extra energy to specific nodes, a malicious energy source (MES) can intentionally manipulate the routing path of SWCNs. The efficiency of energy attack depends on which nodes to be attacked. To enhance the efficiency of energy attack, a reinforcement learning technique, Q-Learning, is used to develop an intelligent energy attack (Q-IEA) policy for MES. Through interacting with the network environment, the Q-IEA can intelligently take attack actions without having to know the details of the routing method at the network layer. This function can greatly enhance the adaptability of MES to different routing protocols and network topologies. Simulation results verify that Q-IEA can significantly manipulate the routing path of the targeted traffic on demand.**

***Index Terms*—Sustainable wireless communication networks, security, malicious energy attack, Q-Learning**

I. Introduction

The rapid development of the Internet of things (IoT), body area network (BAN), and smart infrastructures involves an ever increasing number of sensors and actuators. Powering the large number of low power devices in these applications is a great challenge, as battery replacement is time consuming and cost inefficient. This encourages us to utilize the renewable energy to meet the clean and self-sustainable requirements of the coming green revolution [1], [2]. Through scavenging energy such as sunlight, wind, electromagnetic waves, and biothermal energy from surrounding environment, an energy harvesting node (EHN) in sustainable wireless communication networks (SWCNs) can run semi-perpetually without any battery replacement [3].

In this work, we focus on radio frequency (RF) energy harvesting as RF energy radiated from cellular base stations, TV towers, and Wi-Fi access points is widely available in both indoor and outdoor environments. The energy harvesting ability greatly extends the sustainability and scalability of SWCNs; nevertheless, it causes some new security issues. The information disclosure problem in simultaneous information and power transfer [4] and the unauthorized commands attack in a body energy driven implantable medical device [5] have been investigated in the literature. Although extensive research has been conducted on protecting the wireless sensor network security [6], there is no research about using energy as a tool of attack to threat the information security in SWCNs.

In this paper, we propose a new attack method, called the *malicious energy attack*, that utilizes the energy-aware feature of the routing protocols in SWCNs [7]. Energy-aware routing protocols have been widely adopted in power-constrained networks to extend the lifespan of wireless networks with limited energy supply. In the energy attack method, the malicious energy source (MES) manipulates routing paths in the network layer by consciously charging specific EHNs [8]. The infected nodes that receive extra energy from the energy attacker will become more active than ordinary nodes to work as data forwarders or information aggregators. As shown in Fig. 1, if the MES is able to select the infected nodes properly, it can manipulate the routing path and "encourage" most of the data traffic into passing through a compromised node who was originally deviated from the source and the destination.

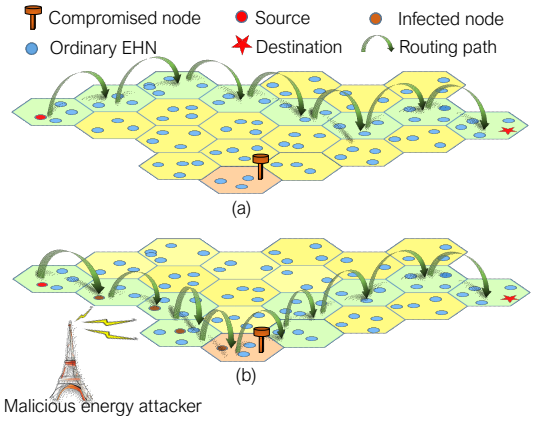

Figure 1. An example of malicious energy attack. (a) An ordinary SWCN without energy attack chose a shorter path from Source to Destination. (b) An SWCN attacked by the MES chose a path through the Compromised node.

Although the MES cannot directly profit from the energy attack, it can collaborate with other network attack methods such as eavesdropping and create opportunities for an eavesdropper to sniff confidential data from any target node. Since energy attack is an indirect attack method that disrupts the network traffic through energy, a "harmless" or even "beneficial" resource in the environment. It's immune to many security mechanisms that will greatly threaten the security of SWCNs and not been fully studied.

The amount of traffic that can be "lured" to the compromised node heavily depends on the energy distribution of EHNs with the energy-aware routing. To optimize the efficiency of malicious energy attack, MES needs to make a wise decision on which nodes in an SWCN should be attacked.

This question may not be difficult to answer if we know the global network status of SWCN (i.e., network topology, energy harvesting rate and instant remaining energy of each node, traffic rate on each node, and etc.) and the parameters of the routing protocol in path selection. The optimal energy attack can be formulated as a deterministic optimization problem [9]. Unfortunately, in practice, an attacker is very likely to face an unknown network environment. For this reason, we need to develop some strategies for the energy attacker to make intelligent charing decisions without knowing network and routing configurations.

Recent development of the reinforcement learning (RL) technique provides a promising solution to tackle the above challenges [10]. Inspired by the powerful ability of RL interacting with an unknown environment, we propose an RL-enabled intelligent energy attack strategy in this paper. The Q-learning algorithm is implemented on the MES to help the attacker find the optimal attack strategy in order to maximize the amount of traffic that can be lured to the compromised node. The malicious energy attacker will train itself intelligently to improve its attack pattern by interacting with the SWCNs.

The main contributions of this paper are two folds. First, we identify a new attack method, called malicious energy attack, in energy-aware SWCNs. The malicious energy attack manipulates the routing path at the network layer by intentionally charging specific EHNs. As an emerging attack method, the malicious energy attack is immune to many security mechanisms since it is an indirect attack method that disrupts the network protocols through energy. Secondly, we study how to enhance the efficiency of malicious energy attack via the RL technique. Through applying the Q-Learning method, the MES can attack the network intelligently without knowing the global network settings or the routing protocols. The Q-learning enabled intelligent energy attack significantly outperforms the energy attack methods without learning ability.

II. Background

In this section, we briefly introduce the background about the energy-aware routing and Q-learning algorithm.

A. Energy Aware Routing

In the power-constrained wireless networks, energy awareness is an essential property of routing protocols to extend network lifespan. The shortest path may not be an optimal route, especially if nodes on that path run low on power. The early depletion of energy will cause serious consequences on network connectivity. Node with low energy usually stand-by to conserve energy. By contrast, nodes with sufficient energy tend to be more active and more likely to be selected as data forwarders [9] or as information aggregators [11]. In networks with energy harvesting capabilities, the harvesting potential is another important consideration in the route selection. To prolong the network lifetime, the harvesting aware routing protocols are designed to align the traffic load with the harvesting potency at different nodes [12].

A representative routing solution to the energy harvesting wireless networks is energy-opportunistic weighted minimum energy (E-WME) routing [7] that is both energy and harvest aware. The forwarding cost is formulated as an exponential function of the nodal residual energy (i.e., λ_n), a linear function of the transmit and receive energies (i.e., e_n), and an inversely linear function of the harvesting rate (i.e., r_n) as shown in (1). E-WME selects the route with the lowest sum weight for data delivery. In the simulation, we use E-WME as the routing protocol in the tested SWCN.

$$C_n = \frac{1}{r_n \log \mu}(\mu^{1-\lambda_n} - 1)e_n \tag{1}$$

Routing security plays a critical rule crucial to protect the data privacy and to maintain the stability of a network. In the literature, adversary users attempt to threat the network security through a variety of attack methods, like the blackhole, wormhole, selective forwarding, sybil attack, and acknowledgement spoofing. There have been extensive research on the design of secure routing for SWCNs to protect the network security at the information plane [13]–[15].

In conventional Network layer attack methods, the attacker usually needs to gain a full control of at least one legitimate node to insert illegal routing information to the network. The legality of information can be verified by inserting artificial imprints (e.g., cryptography, packet identification, and preamble). In adversary energy attack, however, the routing path is intentionally manipulated without injecting bogus routing information or creating artificial high-quality links, but by changing the energy level of EHNs. Cryptographic techniques can not prevent this type of attack since energy attacker does not modify or fabricate routing information; by contrast, the energy attack takes place at the energy plane. In addition, unlike the wormhole attack that tunnels the packets to a distant node in the network and thus can be detected by measuring the distance of a single hop, the geographic information won't help detect the energy attack. These features make the malicious energy attack particularly difficult to defend against.

B. Q-learning Algorithm

Q-learning is a representative model-free reinforcement learning algorithm. In the process of Q-learning algorithm, the agent seeks an optimal action that produces the maximal cumulative rewards via a trial-and-error manner [16]. The agent interacts with the environment through action and the reward from the environment. Each action is evaluated by a reward which helps the learning agent to learn from its past experiences. Q-learning algorithm maintains a Q table to record the learned experience. A typical equation used to update the Q value is depicted in (2), where s_t and a_t are the state and action at time t, respectively. $r(s_t, a_t)$ represents the measured reward.

$$\begin{aligned} Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha[r(s_t, a_t) \\ + \gamma * \max_a Q(s_{t+1}, a)] \end{aligned} \tag{2}$$

The learning rate α and discount factor γ in (2) are two important parameters that affect learning. The learning rate controls the aggressiveness of learning. As the learning rate increase, the agent will rely more on the current reward and less on knowledge learned from previous experiments. It can cause unnecessary oscillations when a is too large. Discount factor control agent to predict future reward, let the agent take a longer view. In practice, γ is commonly set slightly smaller than 1.0 to facilitate the convergence of Q value during the learning process [17]. In Section V-B, we will investigate the impact of learning rate and discount factor on the Q-learning enabled intelligent energy attack.

III. System and Attack Model

A. System Model

Consider a typical RF energy harvesting powered SWCN, as shown in Fig. 1. EHNs harvest ambient RF energy from air and store the harvested energy for future computation and communications. Since the ambient RF energy density is very thin, the EHNs usually maintain low energy level. We assume E-WME routing is implemented in the Network layer to prolong the lifetime of this energy-constrained network. The EHNs with higher harvest rate and more energy will have a lower forwarding cost; the path with least accumulative forwarding cost will be selected for the traffic transmission from source to destination. When there is no external energy sources (i.e., energy attackers), all EHNs in the network have comparable energy harvest rate and the path with fewer hops and shorter distance tends to be selected as the preferred forwarding route.

B. Attack Model

Different from the conventional Network layer attack methods, the MES has no information exchange with the targeted SWCN. Generally, the MES is not interested in collecting the information either. The role of MES is, through intentionally charging specific EHNs, to encourage the network traffic to the compromised nodes in assisting with other attack methods such as eavesdropping or blackhole attacks. We assume that there are some observer nodes and one compromised node scattered in the SWCN deployed area. These nodes can communicate with the MES. For malicious charging without location information, we evenly divide the SWCN area into many small cells which contain approximate 0-2 EHNs. At present, we assume that the cell division is optimal that each cell contains one EHN. The MES equipped with beamforming antennas can directional charge each cell. In the current stage, our attack model is on-off attack; in other words, the battery will be charged to nearly full when the nodes are under energy attack[1]. The reinforcement learning algorithm is implemented on MES to select the best nodes to attack in order to maximize the amount of traffic lured to the compromised node, which is discussed in the next section.

[1]Due to the nonlinear charging feature of the battery, it will take infinite time to charge the EHN to full battery. In this paper, we assume the EHN can be charged to 99% under energy attack.

IV. Intelligent Energy Attack Policy

In light of the unknown network and routing information, there are two main challenges in the design of an optimal energy attack strategy.

First of all, without knowing the network state information (e.g., traffic distribution, routing protocols and energy level of EHNs), it is difficult to mathematically model the relationship between the network states and the amount of traffic traveling through a compromised node. Secondly, the network environment is highly dynamic. If the MES attacks the SWCN based on an instantaneous status of the network, the solution may be only optimal for a snapshot of the network. Therefore, how to adjust the attack pattern to accommodate the dynamics of the SWCN is challenging.

In order to tackle the above challenges, we propose a Q-learning enabled intelligent energy attack method (Q-IEA). The Q-IEA is a model-free method that doesn't need a mathematical model to describe the interaction between the attacking actions and the network environments. Instead it directly observes the states of the network, and improves an attack policy based on the learning experience.

The key for the Q-IEA to handle the highly dynamic challenge is how to construct the $state$ space. Reasonable design of $state$ space can greatly determine the agent's performance. The best $state$ information should contain the energy level of all nodes in the network. However, these information are private to each EHN and wouldn't be obtained by the attacker. Instead, we deployed some observer nodes that cooperate with the MES. These nodes will act as spy nodes to monitor the local harvesting rate and traffic distribution, then provide it to the attacker. Since harvesting rate and traffic of spied nodes is known, we can estimate the rough battery level[2] of nodes in the local areas of the spy nodes. The estimated energy level of spied nodes will construct the $state$ space in Q-IEA.

In each period, the MES takes action a_t to select k number of victim EHNs to charge aiming to manipulate the routing paths and encourage the most traffic through the compromised node. We measure the amount of targeted traffic lured through the compromised node as the reward r_t. In each period, the reward information is reported to the MES. The agent aims at finding the best policy mapping the state to the most appropriate action so that the total rewards can be maximized.

In addition, a pre-train is employed to abstract essential nodes for reducing the action space of Q-IEA. Since charging most of the nodes that are far away from the compromised node, N_c, will not receive any reward and leads to inefficient and lengthy training for Q-IEA; we call these EHNs invalid nodes. Pre-train efficiently eliminates those invalid nodes and greatly improve the performance of Q-IEA. We will show more details in the evaluation section.

Moreover, to pick up the key nodes that truly contributes to the positive rewards in every state, we revised the $action$

[2]Although at the beginning, there will be errors in the estimations of battery level due to the heterogenous initial energy among different nodes. But as experiment goes, the error will gradually decrease, which converges to $\leq 5\%$ in our simulation.

to select a single node in the training stage. The value in the Q-table will represent the effectiveness of attacking a node in manipulating the traffic route at the corresponding state. In the attack stage, the most k essential EHNs are selected[3]. The single-attack training can greatly improve training efficiency and reduce training time.

V. Simulation And Analysis

In this section, we evaluate the performance of the proposed Q-IEA and compare the performance of Q-IEA with benchmark solutions.

A. Simulation Settings

The SWCN network is built based on the Python wsnsimpy, a dedicated simulator for wireless sensor networks. We have modified the package accordingly to fully support energy harvesting and malicious energy attacks. We constructed a large SWCN with 90 EHNs deployed in 700 meters by 700 meters area, as shown in Fig. 2. The average distance among neighboring nodes is 70 meters, while the maximum transmission range of each node is 100 meters. In the test, each EHN generates data packets following Poisson process with mean value $\lambda = 0.2$ packets per slot. We suppose the compromised node located at the bottom left corner is interested in the target traffic from the source at the top left corner to the destination at the bottom right corner. The blue arrows in the graph constitute the preferred main path in the ordinary network when the energy attacker is not present.

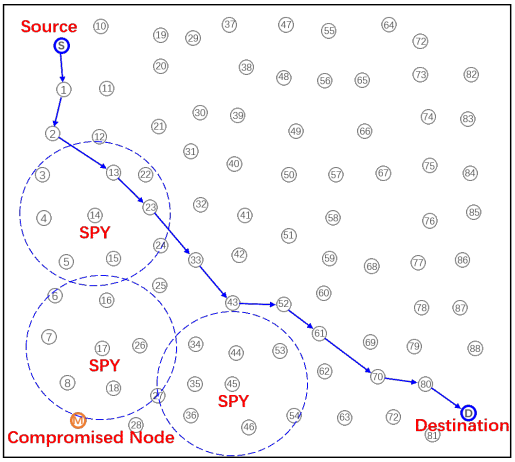

Figure 2. An example of a large SWCN with 90 EHNs

The EHNs among the network have heterogenous average energy harvest rate, r_n, which follows a uniform distribution. In average, it takes 75 slots to charge the battery (i.e., supercapacitor) to nearly full[4]. Once fully charged, it can send roughly 17 packets before battery reaches the low-energy threshold. The EHNs in low-energy mode will stay at inactive mode and avoid forwarding packets for neighboring nodes. Due to the randomness of the network traffic, the battery level of EHNs are highly dynamic, which further results in different selected paths from the source to the destination.

We suppose the MES selects three victim EHNs to attack and charges them to full battery in each attack. We set 20 slots as one observation period. There are three spy nodes evenly placed around the compromised node. The spied areas are marked by blue circles in Fig. 2. At the end of each period, the spy node will take the average of the estimated energy on all monitored nodes as the average energy of the spied area. In the Q-IEA implementation, we evenly discretized the energy level, 0 to 100%, to level 0 to 9, because Q-learning only deal with discrete states. After collected the discrete state information from spy nodes, the attacker, which is not shown in Fig. 2 will update Q-table based on (2) and select the top k EHNs that achieves highest rewards for energy attack.

Since the compromised node is largely deviated from both the source and destination, the amount of targeted information captured by the compromised node is little in the ordinary network without energy attack. In our simulation, we choose random attack as the baseline. The random attack is an inefficient but simplest attack method. In each period, it selects random nodes to attack without any learning capability. As discussed earlier, the optimal energy attack can be easily solved using deterministic optimization algorithms if the routing information and the instant energy level of each EHN are known. Although the optimal energy attack is infeasible to implement, we consider it as the upper bound of the malicious energy attack. In the performance evaluation, we compare Q-IEA with baseline and upper bound performance in different settings.

B. Performance Evaluation

In this section, we evaluate the performance of Q-IEA and analyze the impact of training time, number of nodes attacked, traffic rate on the effectiveness of Q-IEA.

1) Performance Comparisons and Improvements: In order to evaluate the effectiveness of the Q-IEA, we ran the random attack and Q-IEA with the same simulation settings and presented their rewards that are normalized by the upper bound in Fig. 3. Each bar plot is the average performance achieved with ten independent tests. By comparing the performance of Baseline and Q-IEA in Fig. 3, we observe that with the assistance of Q-learning, the intelligent attack achieves 5 times higher reward than the random attack, which verifies the effectiveness of Q-IEA.

However, the performance gap between Q-IEA and the Upper is significant; the Q-IEA only achieves 7.4% of the UpperBound performance in average. In the UpperBound attack, we assume the MES knows the global and instantaneous network states and can choose the most appropriate three nodes to attack. On average, it is able to lure 44% of the total target traffic to the compromised node. In reality, only

[3]It might be possible that attacking the most k essential EHNs is not be the optimal solution. However, through simulation evaluations, we found that compared with sequentially training multiple Q-Tables for each EHN attacked, training a single Q-table is more efficient and effective with less training time.

[4]Note that, due to the nonlinearity of the battery, the actual amount of energy that can be captured by the EHN depends on the instant residual energy of the battery [18]. Therefore, the amount of harvested energy in each slot will not be a constant but calculated based on Equation (6) of [18].

the average energy of nodes in the local spied area is known to the energy attacker in Q-IEA. Due to the imperfect state information, Q-IEA only attracts about 3.3% of total traffic by attacking three nodes, which is 7.4% of the UpperBound performance.

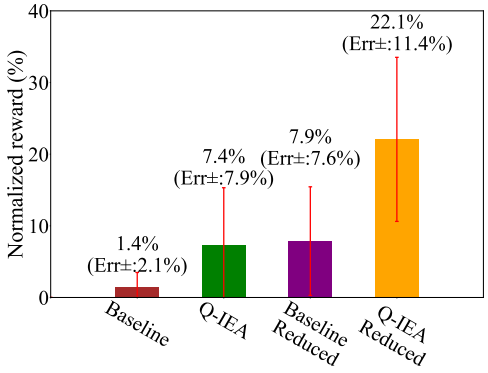

Figure 3. Q-IEA performance comparisons and improvements

Another reason that accounts for the relatively low performance of Q-IEA is the large amount of invalid attacks in the action space, as most of EHNs are invalid nodes. In order to improve the attack efficiency, a pre-train is used to solve this issue as we discussed in section III. In the pre-train phase, the reward when each EHN is attacked is recorded. By eliminating the nodes with no positive rewards, the action space is significantly reduced resulting in an improved efficiency of the energy attack. In Fig. 3, we apply the reduced action space to both baseline random attack and the Q-IEA. The reduced action space significantly improves the baseline attack and makes it comparable to the original Q-IEA. However, the Q-learning can further improve the attack efficiency as Q-IEA can eventually converge to "most" effective actions. The Q-IEA Reduced achieves 22.1% normalized reward, which is three times of the Baseline Reduced attack.

2) Train Time: We plotted the training curve and demonstrated the effectiveness of Q-IEA Reduced with training steps in Fig. 4. The x-axis is the training time in terms of thousands of steps and y-axis is the percentage of target traffic traveled through the compromised node representing the reward. We can observe that the performance grows as the training time increases, since with longer training time, the model is trained more sufficiently to focus on the truly effective actions. However, the performance will not continue improve significantly as time steps exceed $8,000$ after the model well trained.

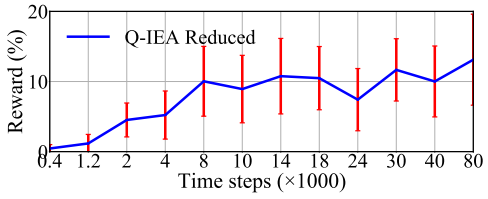

Figure 4. Impact of training time on the performance of Q-IEA.

3) Learning Rate & Discount Factor: Fig. 5 demonstrates the impact of learning Rate (LR) and the Discount Factor (DF) on the performance of Q-IEA with reduced action space. We conducted independent tests with three combined settings. Comparing the results between first and second settings, we observe slight performance improvements with a larger DF. The existence of DF can help the model to enhance those actions that gained higher reward in the long term rather than just the best action at the current time. By comparing the results between the second and third settings, we noted that as the LR increases, the performance becomes worse. With a larger LR, the model will depend more on the Q-value of current step but less on the updated Q-value. As a result, the Q-value is more likely to oscillate rather than converge to the optimal solution. Therefore, we choose LR$=0.02$ and DF$=0.9$ for the remaining tests in the paper.

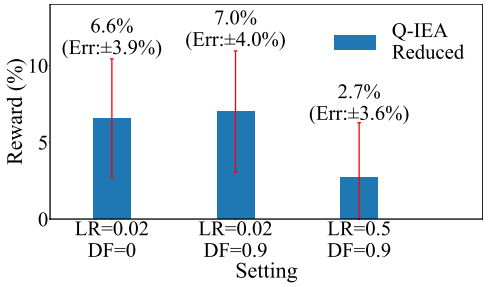

Figure 5. Impact of LR and DF on the performance of Q-IEA.

4) Number of nodes attacked: Intuitively speaking, the more nodes that are attacked (i.e., a larger k), the greater impact of the MES on SWCN, which can potentially bring higher rewards. This is confirmed by the simulation results presented in Fig. 6. With more nodes attacked, the MES gains stronger forces to lure the targeted traffic to the compromised node. Especially if all nodes along the source$\rightarrow N_c \rightarrow$destination path are charged to nearly full, the most majority of the traffic can be lured to the compromised node. However, it's neither practical nor efficient in reality. From the UpperBound curve we can see that, 38% of total traffic is encouraged to the N_c with only three nodes attacked. But it only attracts 20% more traffic with another three nodes attacked. The Q-IEA Reduced curve also verifies that the performance improvement brought by charging more nodes significantly reduces when k is large. For this reason, we suggest attacking 3 to 5 nodes with the given SWCN setting for a balanced performance between reward and energy efficiency. In this paper, the MES attacks 3 victim EHNs in each step. We also notice that no matter how many nodes are attacked, the performance of Q-IEA Reduced is always better than Baseline Reduced because of learning capability.

5) Main Path Traffic: In this test, we investigate the impact of traffic rate on the performance of Q-IEA. We set the default packet generation rate as 0.5 packets per slot and changes the traffic rate based on the default value. The upper bound of 3-node attack and the performance of Q-IEA are presented Fig. 7. We can observe from Fig. 7 that as the traffic rate increases, the

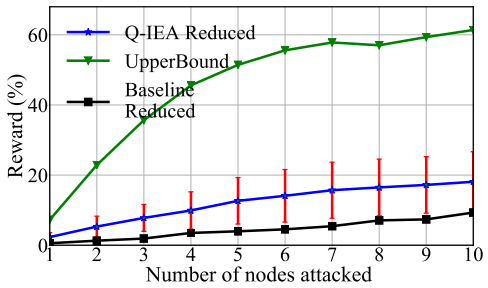

Figure 6. Impact of the number of attacked nodes on malicious energy attacks.

performance of Q-IEA remains almost the same. The traffic lured to the compromised node by the optimal 3-node attack in upper bound, however, decreases as traffic rate grows.

As traffic rate grows, the increased packet transmissions will drain the battery of EHNs in a faster manner. The nodes that are not being attacked will become the bottleneck in the energy attack. When the battery on those EHNs are drained, the source$\rightarrow N_c \rightarrow$destination path is likely to become disconnected and the source node tends to switch to other routes where N_c is not involved. While other nodes along the source$\rightarrow N_c \rightarrow$destination path become severe limitation, the performance gap between the UpperBound and Q-IEA caused by more "wisely" selecting three nodes to attack reduces as traffic increases. This is a benefit from the DF parameters in the training process as described before.

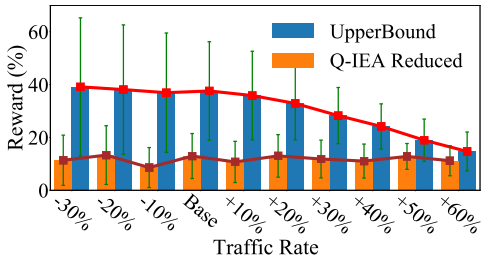

Figure 7. Impact of the traffic rate on the performance of Q-IEA.

VI. Conclusion

In this work, we introduced a new security issue in SWCNs where an adversarial energy source can intentionally provide specific nodes extra energy to manipulate the data path in the network layer. Malicious energy attack is a brand new attack method in SWCNs and worth more investigations in the future. In addition, we proposed a Q-learning enabled intelligent energy attack algorithm and some training tricks to find an efficient attack pattern. Through simulation results, we verify that the Q-IEA can adapt to the dynamic network environment and select appropriate victim EHNs for good attack performance. However, because of the little information obtained by the attacker, the constructed states and the real network environment is not one-to-one corresponding. Learning algorithms like Q-learning that use deterministic method are not the best solutions in this scenario. In the future work, we will use non-deterministic method for action selection to revise our algorithm. In addition, the energy efficiency will be considered to reduce the energy waste and achieve balanced performance between attack and energy efficiency.

VII. Acknowlegement

This work is supported in part by the US National Science Foundation under Grant No. 2051356.

References

[1] T. Wu, F. Wu, J.-M. Redouté, and M. R. Yuce, "An autonomous wireless body area network implementation towards IoT connected healthcare applications," *IEEE Access*, vol. 5, pp. 11 413–11 422, 2017.

[2] F. Akhtar and M. H. Rehmani, "Energy harvesting for self-sustainable wireless body area networks," *IT Professional*, vol. 19, no. 2, pp. 32–40, 2017.

[3] D. Niyato, D. I. Kim, M. Maso, and Z. Han, "Wireless powered communication networks: research directions and technological approaches," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 88–97, 2017.

[4] X. Chen, D. W. K. Ng, and H.-H. Chen, "Secrecy wireless information and power transfer: challenges and opportunities," *IEEE Wireless Communications*, vol. 23, no. 2, pp. 54–61, 2016.

[5] S. Gollakota, H. Hassanieh, B. Ransford, D. Katabi, and K. Fu, "They can hear your heartbeats: non-invasive security for implantable medical devices," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. ACM, 2011, pp. 2–13.

[6] H. Modares, R. Salleh, and A. Moravejosharieh, "Overview of security issues in wireless sensor networks," in *2011 Third International Conference on Computational Intelligence, Modelling Simulation*, 2011, pp. 308–311.

[7] L. Lin, N. B. Shroff, and R. Srikant, "Asymptotically optimal energy-aware routing for multihop wireless networks with renewable energy sources," *IEEE/ACM Transactions on Networking (TON)*, vol. 15, no. 5, pp. 1021–1034, 2007.

[8] J. Guo, X. Zhou, and S. Durrani, "Wireless power transfer via mmwave power beacons with directional beamforming," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 17–20, 2019.

[9] G. Han, Y. Dong, H. Guo, L. Shu, and D. Wu, "Cross-layer optimized routing in wireless sensor networks with duty cycle and energy harvesting," *Wireless communications and mobile computing*, vol. 15, no. 16, pp. 1957–1981, 2015.

[10] Y. Li, "Deep reinforcement learning: an overview," *arXiv preprint arXiv:1701.07274*, 2017.

[11] Y. Dong, J. Wang, B. Shim, and D. I. Kim, "DEARER: a distance-and-energy-aware routing with energy reservation for energy harvesting wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3798–3813, 2016.

[12] A. Kansal, J. Hsu, M. Srivastava, and V. Raghunathan, "Harvesting aware power management for sensor networks," in *Proceedings of the 43rd annual Design Automation Conference*. ACM, 2006, pp. 651–656.

[13] J. Tang, A. Liu, J. Zhang, N. N. Xiong, Z. Zeng, and T. Wang, "A trust-based secure routing scheme using the traceback approach for energy-harvesting wireless sensor networks," *Sensors*, vol. 18, no. 3, p. 751, 2018.

[14] N. A. Alrajeh, S. Khan, J. Lloret, and J. Loo, "Secure routing protocol using cross-layer design and energy harvesting in wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 9, no. 1, p. 374796, 2013.

[15] T. Zhu, S. Xiao, Y. Ping, D. Towsley, and W. Gong, "A secure energy routing mechanism for sharing renewable energy in smart microgrid," in *Proceedings of International Conference on Smart Grid Communications (SmartGridComm)*. IEEE, 2011, pp. 143–148.

[16] J. Yan, H. He, X. Zhong, and Y. Tang, "Q-learning-based vulnerability analysis of smart grid against sequential topology attacks," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 200–210, 2016.

[17] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[18] Y. Luo, L. Pu, Y. Zhao, W. Wang, and Q. Yang, "A nonlinear recursive model based optimal transmission scheduling in rf energy harvesting wireless communications," *IEEE Transactions on Wireless Communications*, 2020.