sensors

MDPI

Article

Deep Q-Learning-Based Buffer-Aided Relay Selection for Reliable and Secure Communications in Two-Hop Wireless Relay Networks

Cheng Zhang [1,2], Xuening Liao [1,2,3,*], Zhenqiang Wu [1,2], Guoyong Qiu [1,2], Zitong Chen [2] and Zhiliang Yu [4]

[1] Key Laboratory of Modern Teaching Technology, Ministry of Education, Xi'an 710062, China; zhang_cheng@snnu.edu.cn (C.Z.); zqiangwu@snnu.edu.cn (Z.W.); qgyqgy@snnu.edu.cn (G.Q.)
[2] School of Computer Science, Shaanxi Normal University, Xi'an 710119, China; zitong_@snnu.edu.cn
[3] Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China
[4] School of Mathematics and Computer Science, Shaanxi University of Technology, Hanzhong 723001, China; yu@snut.edu.cn
* Correspondence: liaoxuening@snnu.edu.cn

Abstract: This paper investigates the problem of buffer-aided relay selection to achieve reliable and secure communications in a two-hop amplify-and-forward (AF) network with an eavesdropper. Due to the fading of wireless signals and the broadcast nature of wireless channels, transmitted signals over the network may be undecodable at the receiver end or have been eavesdropped by eavesdroppers. Most available buffer-aided relay selection schemes consider either reliability or security issues in wireless communications; rarely is work conducted on both reliability and security issues. This paper proposes a buffer-aided relay selection scheme based on deep Q-learning (DQL) that considers both reliability and security. By conducting Monte Carlo simulations, we then verify the reliability and security performances of the proposed scheme in terms of the connection outage probability (COP) and secrecy outage probability (SOP), respectively. The simulation results show that two-hop wireless relay network can achieve reliable and secure communications by using our proposed scheme. We also performed comparison experiments between our proposed scheme and two benchmark schemes. The comparison results indicate that our proposed scheme outperforms the max-ratio scheme in terms of the SOP.

Keywords: physical-layer security; buffer-aided relay selection; Markov decision process; deep Q-learning; secrecy outage probability; connection outage probability

Citation: Zhang, C.; Liao, X.; Wu, Z.; Qiu, G.; Chen, Z.; Yu, Z. Deep Q-Learning-Based Buffer-Aided Relay Selection for Reliable and Secure Communications in Two-Hop Wireless Relay Networks. *Sensors* **2023**, *23*, 4822. https://doi.org/10.3390/s23104822

Academic Editor: Yang Yue

Received: 10 April 2023
Revised: 8 May 2023
Accepted: 12 May 2023
Published: 17 May 2023

1. Introduction

With the development of 5G and beyond, wireless networks are widely used in various fields, such as wireless sensor networks (WSNs) [1], cognitive radio networks (CRNs) [2], and the Internet of Things (IoTs) [3]. With the wide use of wireless networks, a large amount of confidential information is transmitted over each network every day. However, signals may be undecodable at the receiver end due to the fading of wireless signals, and may be intercepted by an eavesdropper due to the broadcast nature of wireless channels, leading to critical reliability and especially security issues in wireless networks. Any unauthorized attacker within the transmission range of a transmitter can receive the transmitted information, which can easily cause information leakage [4]. Therefore, the problem of reliable and secure communications in wireless networks urgently needs to be solved.

The traditional method to achieve secure communications in wireless networks is based on a cryptographic mechanism. The principle of cryptography is to encrypt confidential information with a secret key at the legitimate sender's end and then decrypt it with a secret key at the legitimate receiver's end [5]. As the secret key is deployed only on

the legitimate transmitter and receiver, eavesdroppers cannot decrypt the encrypted information because of the lack of the secret key [6]. The disadvantage of cryptography is that its implementation requires the deployment of devices with a high level of computational performance due to the high level of computational complexity associated with encrypting and decrypting. It is impossible to require all devices connected to wireless networks to have high computational ability. In recent years, a new method called physical-layer security (PLS), which has low computational complexity, has been proposed and is used to aid cryptography in achieving secure communications in wireless networks with low computing capability. The principle of PLS is based on information theory, and uses the randomness of noise and wireless channels to achieve secure communications [7]. Compared with the cryptography method, PLS has a lower network resource overhead and computational complexity [8]. Therefore, PLS has become a promising technique that can help in enhancing security performance in wireless networks.

Common PLS techniques include beamforming [9], artificial noise [10], and relay selection [11]. The principle of beamforming is to achieve the directional transmission of signals by adjusting the transmission direction of the antennas to achieve PLS [12]. Artificial noise interferes with eavesdropping by sending noise [13]. Relay selection achieves PLS by selecting the appropriate relay nodes to which to transmit confidential information [14]. Compared with beamforming and artificial noise, the implementation complexity of relay selection is lower. Depending on whether the relay nodes are equipped with buffers or not, the relay selection technique is divided into conventional and buffer-aided relay selection [15]. In conventional relay selection, once relay nodes without buffers receive the signals, they have to immediately forward them to the next hop [16]. In contrast, buffer-aided relay selection can temporarily store the received signals in buffers instead of transmitting them immediately [17]. So, buffer-aided relay selection can achieve better security performance than that of the conventional relay selection without buffers, especially when the channel quality is poor [18]. Due to its low implementation complexity and good security performance, we used the buffer-aided relay selection technique to achieve reliable and secure communications in this paper.

Traditional buffer-aided relay selection selects the best relay by adopting a central node that collects the network information (e.g., the channel state information (CSI) of the legitimate link and the CSI of the eavesdropping links) and then selects the relays online on the basis of this information. However, it is difficult to achieve CSI of eavesdropping links, as the eavesdroppers always transmit no information, and too much energy, storage, and time are needed to conduct the relay selection, as there are several transmission patterns (i.e., source–relay, relay–destination, and source–destination transmissions) and many possible relay buffer states during the transmission. This is challenged when the central node is resource-limited. Unlike traditional methods, traditional Q-learning (TQL) and DQL define the Q-function, which can simplify the modeling of information transmission in buffer-aided relay selection by evaluating the gain of choosing a particular link to which to transmit signals in the current state in an integrated manner, especially DQL. By using neural networks to fit the Q-function, DQL can create the Q-function without storing it in a Q table, reducing both spatial and temporal complexity for buffer-aided relay selection. Therefore, we used the DQL method to propose a new buffer-aided relay scheme to achieve reliable and secure communications in two-hop wireless relay networks.

2. Related Work

For two-hop buffer-aided relay networks without eavesdroppers, the authors in [19] consider the reliability of wireless communications and proposed a novel buffer-aided relay selection scheme called the max-link scheme. In the max-link scheme, the signals at each hop are transmitted by the link with the maximum signal-to-noise ratio (SNR) to achieve reliable communications. The authors in [19] also established a theoretical analysis framework on the basis of a Markov chain (MC) for analyzing the outage performance of their proposed buffer-aided relay selection scheme. The authors in [20] combine social net-

works with two-hop wireless relay networks and investigate how to design a buffer-aided relay selection scheme to achieve reliable communications when there are untrusted relays in the network. Due to the introduction of buffers, the queuing delay of data packets at buffer-aided relay increases. To achieve reliable communication and reduce delay, the authors in [21] proposed a delay-sensitive buffer-aided relay selection based on channel-based greedy scheduling in vehicular networks. Because of the low implementation complexity of buffer-aided relay selection, the authors in [22] use buffer-aided relay selection to improve the reliability of bidirectional wireless sensor network communications. These buffer-aided relay selection schemes described above are all based on MP. The authors in [23] model buffer-aided relay selection as a Markov decision process (MDP) rather than the MP and exploit TQL to design a buffer-aided relay scheme. TQL evaluates all links by Q-function and selects the link with the maximum Q-function each time to transmit the signals and thus achieves reliable communications. Due to the excellent reliability performance of the proposed scheme based on TQL, the authors in [24,25] extend this work to vehicular networks, D2D communications and achieve reliable communications in vehicular networks and D2D communications. Although these TQL-based schemes can achieve reliable communications, too much storage space is needed to store the Q-table and a high time cost to look up the Q-function in the Q-table as all Q-functions have to be stored in the Q-table.

As research continued, researchers began to investigate how to achieve secure communications using buffer-aided relay selection when considering possible eavesdroppers in the network [26–32]. Based on the work in [19], the authors in [26] consider the case where a passive eavesdropper is present and proposed a new buffer-aided relay selection scheme to achieve secure communications by selecting the link with the maximum instantaneous secrecy capacity at each hop to transmit the signals. In real scenarios, not only illegal eavesdroppers eavesdrop signals, but also untrusted relay nodes can intercept the transmitted signals as well. These untrusted relay nodes are both cooperators and potential eavesdroppers of information transmission. In response to the presence of untrusted relay nodes, the authors in [27] propose a secure buffer-aided relay selection scheme that uses the AF mode to avoid the decoding of confidential information by untrusted relay nodes. The authors in [28] extend this work to a more general scenario where both potential eavesdropping nodes and passive eavesdroppers are present. The authors in [29] extend secure communications to bidirectional wireless relay network and design a buffer-aided relay selection scheme based on an achievable rate. In addition, to resist the eavesdroppers, buffer-aided relay selection is often combined with full duplex (FD) [30], cooperative jamming (CJ) [31] and energy harvesting (EH) [32] to achieve secure communications. Although these schemes can realize secure communications, they also increase the implementation complexity, which conflicts with the original intent of adopting buffer-aided relay selection.

All of the above related works only consider the reliability or the security performances of wireless communications, without considering both the security and reliability issues. In fact, it is very challenging to simultaneously achieve reliable and secure communications by using buffer-aided relay selection. It requires simultaneously taking into account the legitimate channel quality, eavesdropping channel quality, buffer queues, secrecy rate, etc. Therefore, buffer-aided relay selection based on traditional methods is difficult to achieve reliable and secure communications while maintaining low implementation complexity. With the development of deep learning (DL), DL has been applied to wireless relay networks [33–35]. A large number of researchers have started to use deep learning to study buffer-aided relay selection [36–43]. The authors in [36] model the buffer-aided relay selection as a multi-classification problem and uses a deep neural network (DNN) to predict the suitable link to transmit the signals. Inspired by [23], the authors in [37] utilize DQL to solve the buffer-aided relay selection problem, where a modified version of TQL is used. Different from TQL, DQL uses DNN to fit the Q-function instead of storing Q-function in the Q-table. Therefore, DQL has lower time complexity and space complexity compared to TQL [38]. The comparison experiments in [39] demonstrated that DQL has

better learning results and lower complexity than those of TQL, and the implemented scheme via DQL is more suitable for practical scenarios, as the implemented scheme via DQL could work without prior information. On this basis, the authors in [40,41] realized reliable communications for IoTs [40] and CRNs [41] by using the DQL-based buffer-aided relay selection schemes. The authors in [40,41] extend their work further and use the proposed DQL-based buffer-aided relay selection scheme to realize reliable and secure communications in CRNs [42,43].

DQL makes it possible to achieve reliable and secure communications using buffer-aided relay selection. However, the works in [42,43] use DQL to address the issue of power allocation (PA) to achieve reliable and secure communications and they did not consider possible eavesdroppers in the network. Therefore, this paper explores how to achieve reliable and secure communications using only a DQL-based buffer-aided relay selection scheme in the more common two-hop wireless relay networks rather than CRNs. To highlight the contributions of this paper, we give a comparison of our work with related works in Table 1. The work of this paper is summarized as follows:

Table 1. The main features of our work and related works.

References and Our Work	Feature				
	System Model	Eavesdropper	Method	Reliability	Security
[19]	Two-hop DF relay network	✘	MC	✓	✘
[23]	Two-hop AF relay network	✘	TQL	✓	✘
[26]	Two-hop AF relay network	✘	MC	✘	✓
[40]	Delay-constrained DF relay IOT	✘	DQL	✓	✘
[43]	RF relay CRN	✓ (untrusted users)	DQL+PA	✓	✓
Our Work	Two-hop AF relay network	✓	DQL	✓	✓

✓ indicates that the factor is considered in the paper, and ✘ indicates that the factor is not considered in the paper.

- To propose a DQL-based buffer-aided relay selection scheme, we first analyze the communication model of a two-hop AF buffer-aided relay network with the presence of a passive eavesdropper and then model the information transmission process as an MDP.
- We then propose a DQL-based buffer-aided relay selection scheme to optimize the above MDP. In the proposed scheme, we consider both the legal channel states and eavesdropping channel states, buffer states, target rate and target secrecy rate and use DNNs to fit the Q-function and select the link with the maximum Q-function value each time.
- Finally, we verify the reliability and security performances of the proposed scheme by using Monte Carlo simulations. The reliability and security performances are measured by the COP and the SOP, respectively. Simulation results demonstrate that the proposed scheme can achieve reliable and secure communications. We also compare the COP and SOP of the proposed scheme with the max-link and max-

ratio schemes, respectively. The comparison results show that the proposed scheme outperforms max-ratio schemes in terms of security performance.

The remainder of this paper is organized as follows: Section 3 introduces the system model; Section 4 introduces the framework of information transmission based on MDP; Section 5 describes the proposed buffer-aided relay selection scheme; Section 6 shows the simulation results of proposed scheme; Section 7 concludes the contributions of this paper.

3. System Model

As depicted in Figure 1, this paper considers a two-hop AF buffer-aided relay network, which is composed of a source node S, a cluster of AF buffer-aided relay nodes R_k ($k \in \{1, 2, \cdots, K\}$), a destination node D and a passive eavesdropper node E. The source node S cannot communicate with the destination node D directly due to path loss and the long distance, so the signals from S must be forwarded by the buffer-aided relay node R_k. The number of AF buffer-aided relay nodes is K. Every relay node is in half-duplex (HD) mode and is equipped with a buffer queue Q_k of length L, so these relay nodes can store the received signals instead of forwarding them immediately to D. This paper assumes that the eavesdropping node E only eavesdrops the signals from R_k to D, and does not eavesdrop the signals from S to R_k.

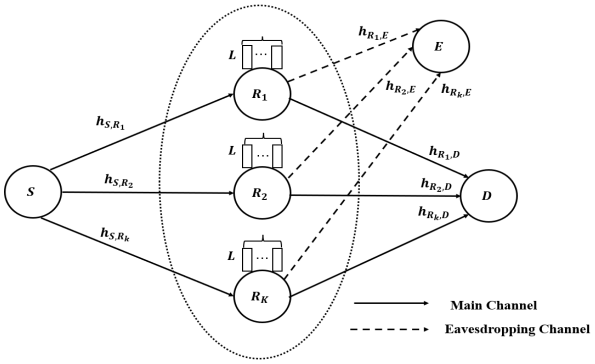


Figure 1. Illustration of the system model.

Without the decoding process at relays, AF relays can thus decrease the probability of being intercepted by potential eavesdroppers for transmitted signals [44]. Thus, we assume that all relays are AF relays in this paper to enhance the security of signals transmitted in the network.

We assume that all channels are independent and non-identically distributed quasi-static Rayleigh fading channels, including eavesdropping channels. In this paper, we use $h_{m,n}$ and $g_{m,n}$ to denote the channel coefficient and the channel gain between node m and node n, respectively, where $g_{m,n} = |h_{m,n}|^2$. Since all channels are Rayleigh channels [45], the channel gain follows the exponential distribution, which means that $E[|h_{m,n}|^2] = E[g_{m,n}] = \Omega_{m,n}$, where $E[.]$ is the expectation operator and $\Omega_{m,n}$ is the average channel gain. This paper assumes that the real-time CSI is completely known and sets the source node S as the central node, which receives the real-time CSI of all channels and buffers state information of all buffer-aided relay nodes then selects an appropriate link to transmit the signals according to relay selection schemes. Supposing at a time slot t, the central node selects an S to R_k link to transmit signals, the received signals $y_{R_k}(t)$ at R_k can be expressed as

$$y_{R_k}(t) = \sqrt{P_s} h_{S,R_k}(t) x_s(t) + n_{R_k}(t), \tag{1}$$

where P_s is the transmission power of the source node S, $x_s(t)$ is the signal sent by S at time t, and $n_{R_k}(t)$ is the additive white Gaussian noise (AWGN) noise with variance power σ^2 at R_k. According to (1), the instantaneous SNR of S to R_k link at time t is given by

$$\psi_{S,R_k}(t) = \frac{P_s|h_{S,R_k}(t)|^2}{\sigma^2}, k \in \{1, 2, \cdots, K\}, \tag{2}$$

and the channel capacity of S to R_k link is $C_{S,R_k}(t) = \frac{1}{2}\log_2(1 + \psi_{S,R_k}(t))$, $k \in \{1, 2, \cdots, K\}$. The received signal $y_{R_k}(t)$ is stored in the corresponding buffer queue Q_k waiting for the transmission to the next hop. After waiting for t_1 time slots, the received signal $y_{R_k}(t)$ is amplified to resist path fading and then forwarded to the destination node D by the buffer-aided relay node R_k. Thus, at time slot $t' = t + t_1$, the signal $x_{R_k}(t')$ sent by the buffer-aided relay node R_k is represented as

$$x_{R_k}(t') = A_{R_k}(t')y_{R_k}(t), \tag{3}$$

where

$$A_{R_k}(t') = \frac{1}{\sqrt{P_s|h_{S,R_k}(t)|^2 + \sigma^2}} \tag{4}$$

is the amplification factor of the buffer-aided relay node R_k at time t', it is determined by the quality of the channel between source node S and the buffer-aided relay node R_k at time t. Due to the broadcast nature of wireless channel, eavesdropping nodes within the transmission range can also receive the transmitted signals. In this paper, we assume that the eavesdropping node only eavesdrops the signals sent by the buffer-aided relay nodes R_k to the destination node D. So the signals received by S and E can be expressed as

$$\begin{aligned} y_D(t') &= \sqrt{P_{R_k}}h_{R_k,D}(t')x_{R_k}(t') + n_D(t'), \\ y_E(t') &= \sqrt{P_{R_k}}h_{R_k,E}(t')x_{R_k}(t') + n_E(t'), \end{aligned} \tag{5}$$

respectively, where P_{R_k} is the transmission power of R_k, $n_D(t')$ and $n_E(t')$ are AWGN noises at D and E, respectively. According to (5), the instantaneous end-to-end SNR from S to D and from S to E can be derived as

$$\begin{aligned} \psi_{S,D}(t') &= \frac{P_s P_{R_k}|h_{S,R_k}(t)|^2|h_{R_k,D}(t')|^2}{(P_s|h_{S,R_k}(t)|^2 + P_{R_k}|h_{R_k,D}(t')|^2 + \sigma^2)\sigma^2}, \\ \psi_{S,E}(t') &= \frac{P_s P_{R_k}|h_{S,R_k}(t)|^2|h_{R_k,E}(t')|^2}{(P_s|h_{S,R_k}(t)|^2 + P_{R_k}|h_{R_k,E}(t')|^2 + \sigma^2)\sigma^2}, \end{aligned} \tag{6}$$

respectively. Thus, the end-to-end channel capacity from S to D and S to E can be given by

$$\begin{aligned} C_{S,D}(t') &= \frac{1}{2}\log_2(1 + \psi_{S,D}(t')), \\ C_{S,E}(t') &= \frac{1}{2}\log_2(1 + \psi_{S,E}(t')), \end{aligned} \tag{7}$$

respectively. The end-to-end secrecy rate from S to D is given by

$$C_{S,D}^{(s)}(t') = [\theta - C_{S,E}(t')]^+, \tag{8}$$

where $[z]^+ = \max(o, z)$, and θ is the target rate of the two-hop AF buffer-aided relay network.

4. The Framework of Information Transmission Based on MDP

To design a buffer-aided relay selection scheme that enables reliable and secure communications in two-hop wireless relay networks, we need to first analyze the information transmission process in two-hop wireless relay networks. Due to the Markovian property of the process of receiving and forwarding information in the buffers, the information transmission process in two-hop wireless relay networks can be modeled as an MDP to analyze. As shown in Figure 2, a complete MDP consists of a five-tuple (state s_t, action a_t, policy $\pi(a_t|s_t)$, reward $r(s_t, a_t)$), return U_t, environment and an agent. This section describes in detail how to model the process of information transmission in two-hop wireless relay networks as an MDP.

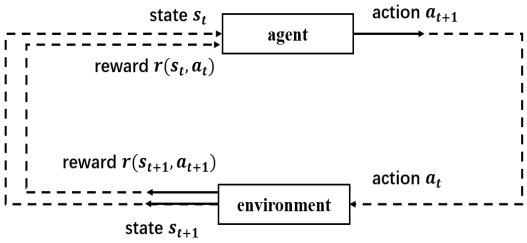


Figure 2. An MDP, which consists of state s_t, action a_t, policy $\pi(a_t|s_t)$, reward $r(s_t, a_t)$, return U_t, environment and an agent.

4.1. Agent and Environment

In the MDP, the agent can perceive the state of the environment, take actions according to the state and adjust the decisions based on the feedback of the environment. In the two-hop AF buffer-aided relay network, the central node is regarded as the agent in the MDP and the whole two-hop AF buffer-aided relay network is modeled as the environment in the MDP. The state of the environment will be changed by action of the agent, which can be perceived by the agent. In addition, the environment will give the agent feedback after each decision made by the agent.

4.2. State

For the two-hop AF buffer-aided relay network, this paper defines the state space $s(t)$ at time slot t as $s(t) = \{l(t), b(t)\}$, where $l(t)$ and $b(t)$ are the link states of all links and the buffer states of all buffer queues at time slot t, respectively. The link states $l(t)$ at time t are defined as

$$l(t) = \{l_{0,1}(t), l_{0,2}(t), \cdots, l_{0,K}(t), l_{1,1}(t), \cdots, l_{1,K}(t)\}, \tag{9}$$

where $j = 0, l_{0,k}(t)$ is the link state of S to the corresponding R_k link; $j = 1, l_{1,k}(t)$ is the link state of the corresponding R_k to D link. As we assume that the eavesdropping node E only intercept signals from the R_k to D link, only the reliability issue of the transmission link needs to be considered in the first hop. The value of $l_{0,k}(t)$ is taken as follows.

- $l_{0,k}(t) = 0$ denotes $C_{S,R_k}(t) \leq \theta$ and the corresponding link is unreliable. When $l_{0,k}(t) = 0$, the corresponding link can not transmit the signals at the target rate θ.
- $l_{0,k}(t) = 2$ denotes $C_{S,R_k}(t) \geq \theta$ and the corresponding link is reliable. When $l_{0,k}(t) = 2$, the corresponding link can transmit the signals at the target rate θ.

For an R_k to D link, the reliability and security of the link are both considered due to eavesdropping by E. The value of $l_{1,k}(t)$ is taken as follows.

- $l_{1,k}(t) = 0$ denotes $C_{S,D}(t) < \theta$ and the corresponding link is unreliable. When $l_{1,k}(t) = 0$, the corresponding link can not transmit the signals at the target rate θ.
- $l_{1,k}(t) = 1$ denotes $C_{S,D}(t) \geq \theta$, $C_{S,D}^{(s)}(t) < \zeta$ and the corresponding link is reliable but not secure, where ζ is the target secrecy rate. When $l_{1,k}(t) = 1$, the corresponding link

can transmit the signals with the target rate θ but cannot transmit the signals at the target secrecy rate ζ.

- $l_{1,K}(t) = 2$ denotes $C_{S,D}(t) \geq \theta$, $C_{S,D}^{(s)}(t) \geq \zeta$ and corresponding link is reliable and secure. When $l_{1,k}(t) = 2$, the corresponding link can transmit the signals at the target secrecy rate ζ.

Regarding one buffer-aided relay node R_k, there are two links, i.e., an S to R_k link and an R_k to D link, so the buffer state $b(t)$ at time t are defined as

$$b(t) = \{b_{0,1}(t), \cdots, b_{0,K}(t), b_{1,1}(t), \cdots, b_{1,K}(t)\}, \tag{10}$$

where $b_{j,k}(t) \in \{0, 1, \cdots, L\}$, $j \in \{0, 1\}$, $k \in \{1, 2, \cdots, K\}$, because the length of buffer queue is L. If the selected link is an S to R_k link and $b_{0,k}(t) = L$, the corresponding buffer-aided relay node R_k is unavailable at this time because its buffer queue Q_k is full, it can not receive the signals from S. If selected link is an R_k to D link and $b_{1,k}(t) = 0$, the corresponding buffer-aided relay node R_k is also unavailable at this time because its buffer queue Q_k is empty, it can not forward the signals to D.

According to the above analysis, we can conclude that the size of link state space $l(t)$ and buffer state space $b(t)$ are 6^K and $(L+1)^{2K}$, respectively. As $s(t) = \{l(t), b(t)\}$, the size of state space $s(t)$ is $(6(L+1)^2)^K$.

4.3. Action and Policy

In two-hop AF buffer-aided relay networks, the selection of a link for transmitting the signals is modeled as action in the MDP. The set of links that the agent can choose at time t is modeled by the action space $a(t)$.

At state s_t, if the agent selects an S to R_k link to transmit the signals, we denote $a_t = l_{(0,k)}$. If the agent selects an R_k to D link to transmit the signals, we denote $a_t = l_{(1,k)}$. It is worth noting that when the states of all S to R_k links are 0 and the states of all R_k to D links are not equal to 2, the agent will select no link to transmit the signals (i.e., a connection outage event occurs directly) and this case is denoted as $a_t = \emptyset$. Based on the analysis above, we also can deduce that the size of the action space $a(t)$ is $2K + 1$.

To guarantee the reliable and secure communications between legitimate users, if the link selected to transmit the signals is unreliable, then a connection outage event occurs. If the selected link is not secure, then a secrecy outage event occurs. Therefore, after the agent acts the action a_t, the environment may enter a new state s_{t+1}, or remain in the current state s_t due to the connection outage or secrecy outage. In addition, if the selected link is reliable and secure but the corresponding buffer is unavailable, a connection outage event also happens. Transmission is considered successful only if the selected link is reliable and secure (for an S to R_k link, the selected link is only required to be reliable) and the corresponding buffer is available. Table 2 shows the results of performing actions in different link states and buffer states.

In the MDP, the policy function $\pi(a_t|s_t)$ is the probability that the agent acts action a_t at state s_t and is denoted by

$$\pi(a_t|s_t) = P(a_t|s_t). \tag{11}$$

From (11), we can observe that $\pi(a_t|s_t)$ will affect the choice of an action and also the reward for the action.

Table 2. The results of performing actions in different link states and buffer states [1].

Action	Link State	Buffer State	Result
$l_{0,k}$	0	full	connection outage
$l_{0,k}$	0	not full	connection outage
$l_{0,k}$	2	not full	successful transmission
$l_{0,k}$	2	full	connection outage
$l_{1,k}$	2	empty	connection outage
$l_{1,k}$	2	not empty	successful transmission
$l_{1,k}$	1	empty	secrecy outage
$l_{1,k}$	1	not empty	secrecy outage
$l_{1,k}$	0	empty	connection outage
$l_{1,k}$	0	not empty	connection outage
$\emptyset$	$\forall k \in \{1, 2, \cdots, K\}, l_{0,k} = 0, l_{1,k} \neq 2$	any	connection outage

[1] Since TQL and DQL discussed in this paper are based on value iterations rather than policy iterations, the policy is not described in detail in this paper.

4.4. Reward and Return

The reward is the feedback given to the agent by the environment after the agent acts an action a_t in a state s_t, and is noted as $r(s_t, a_t)$. The reward can be divided into three categories: positive reward, negative reward and neutral reward.

- Positive reward: the selected link satisfies the transmission requirements, in which the target transmission rate θ and target secrecy transmission rate ζ are both considered, and the corresponding buffer-aided relay node is available.
- Negative reward: the selected link can not satisfy the transmission requirements or the corresponding buffer-aided relay node is unavailable.
- Neutral reward: no link is selected.

In the MDP, the accumulated reward from the beginning time t to the end time $t + n$ is called as the return, denoted by U_t. The expression of return U_t is given by

$$U_t = r(s_t, a_t) + \gamma * r(s_{t+1}, a_{t+1}) + \cdots + \gamma^n * r(s_{t+n}, a_{t+n}) = r(s_t, a_t) + \gamma * U_{t+1}, \quad (12)$$

where γ is the discount factor in the MDP. Moreover, the conditional expectation of the return U_t of acting action a_t in state s_t is defined as the action-value function $Q_\pi(s_t, a_t)$ and

$$Q_\pi(s_t, a_t) = E|U_t|s_t, a_t|, s_t \in s(t), a_t \in a(t), \quad (13)$$

which is used to evaluate the value of state s_t and action a_t. However, the action-value function is also influenced by the policy function π, and to eliminate the influence of the policy function π, we use the optimal action-value function $Q_*(s_t, a_t)$ (also known as the Q-function) to evaluate the value of state s_t and action a_t. The optimal action-value function is obtained by

$$Q_*(s_t, a_t) = \max_\pi Q_\pi(s_t, a_t), s_t \in s(t), a_t \in a(t). \quad (14)$$

In the MDP, the goal of the agent is to make the return U_t on each episode as high as possible, so the agent should select the link corresponding to the action with the maximum Q-function to transmit the signals each time.

With the above methods, we can model the process of information transmission in two-hop wireless relay networks as an MDP. Subsequently, we can use Q-learning algorithms to optimize the MDP for reliable and secure communications.

5. The Proposed Buffer-Aided Relay Selection Scheme

After modeling the process of information transmission as an MDP, we use Q-learning algorithms to optimize the transmission process and propose a new buffer-aided relay selection scheme based on it. Most of the existing schemes use TQL based on Q-table to

optimize the transmission process and only consider reliability or security. Our proposed scheme utilizes DQL based on DNN to optimize the transmission process and considers both reliability and security. This section describes the principle of Q-learning algorithms and the steps of the proposed scheme based on DQL, respectively.

The principle of Q-learning algorithms including TQL and DQL is shown in Figure 3. The goal of the MDP is to make the return U_t of each episode as high as possible, so the agent should perform the action with the largest Q-function value each time. In TQL, the values of Q-function are stored in Q-table and updated by the Q-learning algorithms updates Q-function by

$$Q_*(s_t, a_t) = r(s_t, a_t) + \gamma * \max_{a \in a(t+1)} Q_*(s_{t+1}, a). \tag{15}$$

In the MDP, the size of state space $s(t)$ and action space $a(t)$ is $(6(L+1)^2)^K$ and $2K+1$, respectively. If we use the TQL based on Q-table to optimize the transmission process, it needs to occupy a lot of space to store a Q-table of $(6(L+1)^2)^K$ by $2K+1$ as in Table 3 and consume a lot of time to update Q-function and search the action with the maximum Q-function. In order to reduce the space occupation and lookup time, this paper uses DQL based on DNN rather than TQL based on Q-table to optimize the transmission process. DQL uses neural network to fit the Q-function without storing the Q-function in the Q-table, so DQL can save storage space.

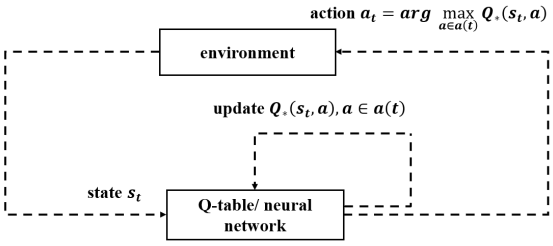


Figure 3. Framework of the Q-learning.

Table 3. The structure of the Q-table. The rows represent state space $s(t)$ and the columns represent action space $a(t)$.

Q-Table					
	s_1	s_2	s_3	$\cdots$	$s_{(6(L+1)^2)^K}$
a_1	$Q_*(s_1, a_1)$	$Q_*(s_2, a_1)$	$Q_*(s_3, a_1)$	$\cdots$	$Q_*(s_{(6(L+1)^2)^K}, a_1)$
a_2	$Q_*(s_1, a_2)$	$Q_*(s_2, a_2)$	$Q_*(s_3, a_2)$	$\cdots$	$Q_*(s_{(6(L+1)^2)^K}, a_2)$
a_3	$Q_*(s_1, a_3)$	$Q_*(s_2, a_3)$	$Q_*(s_3, a_3)$	$\cdots$	$Q_*(s_{(6(L+1)^2)^K}, a_3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_{2K+1}	$Q_*(s_1, a_{2K+1})$	$Q_*(s_2, a_{2K+1})$	$Q_*(s_3, a_{2K+1})$	$\cdots$	$Q_*(s_{(6(L+1)^2)^K}, a_{2K+1})$

The proposed scheme based on DQL is divided into three phases, which are experience collection, training the network model and deploying it online. The steps of the proposed scheme are as follows.

5.1. Experience Collection

This phase focuses on collecting the experience needed to train the network model. Firstly, a DNN, which is called a prediction network, is initialized and used to fit the Q-function. The structure of the prediction network is shown in Figure 4.

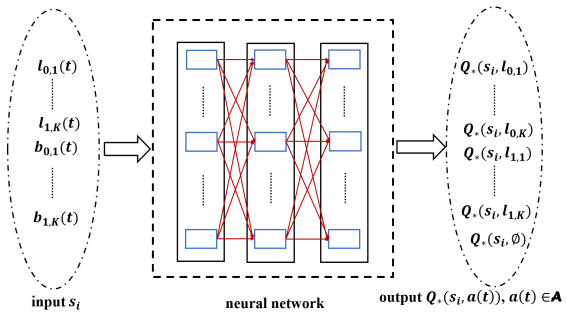


Figure 4. The structure of the prediction network. The input is state s_t, the output is Q-function for each action in action space $a(t)$, $Q_*(s_t, a), a \in a(t)$.

The input of the prediction network is state s_t, and the output is Q-function $Q_*(s_t, a)$, where $a \in a(t)$, which corresponds to a state action at time t. In this phase, the ϵ-greedy policy is used to select actions to balance the exploration–exploitation dilemma. The agent chooses the action with the Q-function with $1 - \epsilon$ probability, and randomly chooses an action with ϵ probability, as shown in (16)

$$a_t = \begin{cases} \arg\max\limits_{a \in a(t)} Q_*(s_t, a), & prob.(1-\epsilon) \\ \text{random an action}, & prob.\epsilon \end{cases}, \tag{16}$$

where $0 < \epsilon \leq 1$. In this phase, the agent needs to explore the action space as much as possible, so ϵ is set as 1. In the training network model phase, the agent needs to train the network model by exploiting the collected experience, so as the number of training episodes increases, ϵ decreases to $\epsilon_{min} = 0.1$, and the attenuation factor $\varphi = 0.998$.

After the agent selects an action and enacts the selected action a_t, the state s_t moves to s_{t+1}, and the environment returns the reward $r(s_t, a_t)$, a sample $\{s_t, a_t, r(s_t, a_t), s_{t+1}\}$ is generated. The prediction network does not learn the sample immediately, but stores the sample in a buffer called as the replay buffer, which is depicted in Figure 5 and is used to store the generated experiences. The above steps of generating and collecting experience are repeated until the replay buffer is full and the prediction network starts to learn the experience.

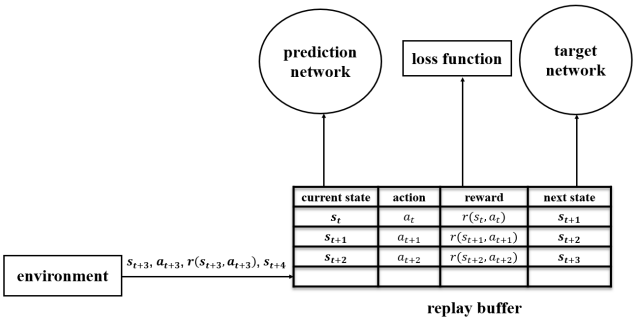


Figure 5. The structure of the replay buffer.

5.2. Training the Network Model

This phase uses the experience collected in the previous phase to train and update the network model. When the replay buffer is full, the agent starts to randomly select a batch of samples for training the network model. The trick is called experience replay, which can effectively reduce the correlation between samples and improve the convergence speed of the prediction network. To avoid bootstrapping of the prediction network, this paper introduces another neural network called the target network, which has the same structure as the prediction network. The input and the output of the prediction network are s_t and $Q_*(s_t, a), a \in a(t)$, respectively. Similarly, the input and output of the target network are s_{t+1} and $Q_*(s_{t+1}, a), a \in a(t+1)$, respectively. As the action a_t acts at state s_t and the reward $r(s_t, a_t)$ are available according to the sample $\{s_t, a_t, r(s_t, a_t), s_{t+1}\}$, we can obtain $Q_*(s_t, a_t)$ and $r(s_t, a_t) + \max_{a \in a(t+1)} Q_*(s_{t+1}, a)$. Next, we calculate the error between $Q_*(s_t, a_t)$ and $r(s_t, a_t) + \max_{a \in a(t+1)} Q_*(s_{t+1}, a)$ by using the loss function. This paper uses the mean square error (MSE) as the loss function of the prediction network and the target network. According to (16), the expression of the MSE loss function is obtained as

$$\varpi = \sum_{1}^{N} (r(s_t, a_t) + \gamma * \max_{a \in a(t+1)} Q_*(s_{t+1}, a) - Q_*(s_t, a_t))^2, \tag{17}$$

where N is the batch size. Then, we update the weights of the prediction network by using the MSE loss function and copy the weights of the prediction network to the target network periodically. Finally, we repeat the above steps of learning and updating for many episodes until the prediction network and target network converge. The framework of the experience collection phase and the training network model phase of the proposed scheme is shown in Figure 6.

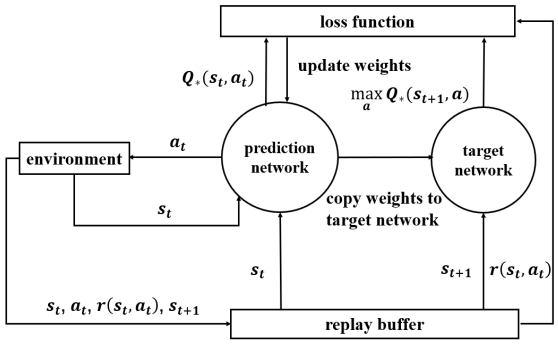


Figure 6. The framework of DQL with target network and experience replay.

5.3. Deployment Online

After the prediction network and the target network converge, we deploy the network model online. It is worth noting that both the experience collecting and training the network model phases are offline. In this phase, the prediction network directly estimates the $Q_*(s_t, a)$ corresponding to each action a, $a \in a(t)$ based on the current state s_t, and selects the action with the maximum $Q_*(s_t, a), a \in a(t)$ without training and updating the weights.

Finally, all the steps of the proposed buffer-aided relay selection scheme based on DQL are shown in Algorithm 1, where N_e is the number of training episodes and N_c is the capacity of the replay buffer.

Algorithm 1 The proposed buffer-aided relay scheme based on DQL

Initialize the environment for the two-hop AF buffer-aided relay network
Repeat:
for $i = 1, 2, \cdots, N_e$ **do**
  for $j = 1, 2, \cdots, N_c$ **do**
    (First phase: experience collection)
    At current state s_t, select action a_t according to ϵ-greedy policy.
    Act the selected action a_t, and return reward $r(s_t, a_t)$ and next state s_{t+1}.
    Generate a sample $s_t, a_t, r(s_t, a_t), s_{t+1}$, and store it in replay buffer.
  end for
  (Second phase: training the network model)
  Randomly select a batch of samples from replay buffer.
  According to s_t and a_t, get $Q_*(s_t, a_t)$ from the prediction network.
  According to $r(s_t, a_t)$ and s_{t+1}, get $r(s_t, a_t) + \max_{a \in a(t+1)} Q_*(s_{t+1}, a)$ from the target network.
  Calculate the loss between $Q_*(s_t, a_t)$ and $r(s_t, a_t) + \max_{a \in a(t+1)} Q_*(s_{t+1}, a)$ by the MSE loss function.
  Update the weights of prediction network.
  if i%100=0 **then**
    Copy the weights of the prediction network to the target network.
  end if
end for
(Third phase: deployment online)
Deploy the prediction network online.

6. Simulation Results and Discussion

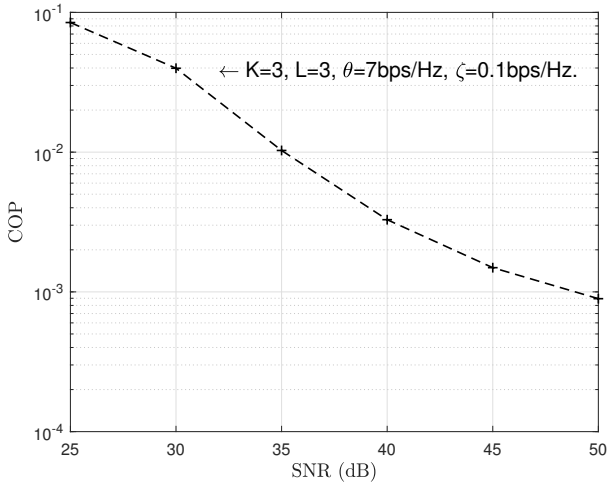
This section verifies the reliability and security performances of the proposed scheme by using Monte Carlo simulations, and uses the COP and SOP to measure the reliability and security performances of the proposed scheme. In the two-hop AF buffer-aided relay network, the number of buffer-aided relay nodes is $K = 3$, the length of the buffer queue is $L = 3$, the average channel gain is set as $\Omega_{S,R_k} = \Omega_{R_k,D} = 30$ dB, and $\Omega_{R_k,E} = 5$ dB. Since the power of AWGN is normalized to unity, the ratio of transmitting power to noise is set as $P_s/\sigma^2 = R_{R_k}/\sigma^2 = 30$ dB. Furthermore, the target rate θ is set as 7 bps/Hz and the target secrecy rate ζ is set as 0.1 bps/Hz. In the proposed buffer-aided relay selection scheme based on DQL, the discount factor γ and learning rate υ of the DQL are set as 0.9 and 0.1, respectively. The capacity N_c of replay buffer which stores samples and the batch size is set as 2000. In the phase of training the network model, the training episodes N_e is set as 20,000, the batch size in each episode is set as 128 and the target network updates its parameters every 100 episodes. After the training of the network model is completed, the reliability and security performances of the proposed scheme are verified by 1 million Monte Carlo simulations. The lower the COP and SOP, the higher the reliability and security performances. The expressions of COP and SOP are obtained by

$$\begin{aligned} COP &= \frac{n_c^{'}}{1{,}000{,}000}, \\ SOP &= \frac{n_s^{'}}{1{,}000{,}000}, \end{aligned} \tag{18}$$

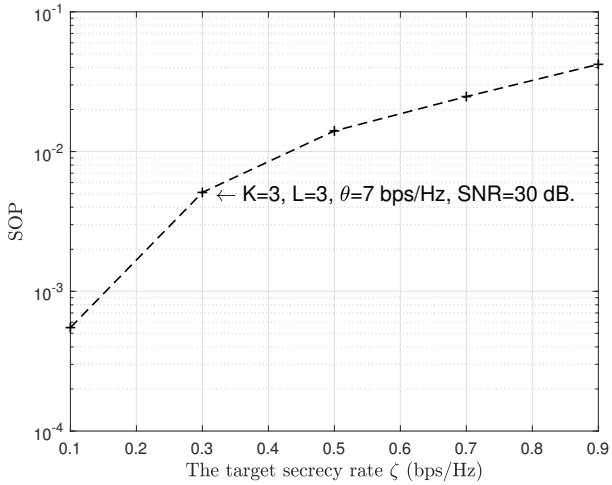
where $n_c^{'}$ and $n_s^{'}$ are the number of connection outage events and secrecy outage events that occurred in 1 million Monte Carlo simulations, respectively.

First, we verify the reliability and security performances of the proposed scheme. The simulation results are shown in Figure 7, where Figure 7a illustrates how the COP varies with the SNR P/σ^2 and Figure 7b shows how the SOP varies with the target secrecy

rate ζ. From Figure 7a, we can observe that the COP decreases as the SNR increases. This is because an increase of SNR means a better transmission link and thus a lower COP. We can further see from Figure 7b that when the target secrecy rate ζ is set as 0.1 bps/Hz, the SOP can reach 10^{-4}, and the SOP increases as the target secrecy rate ζ increases. This is because as ζ increases, fewer legitimate channels can meet the requirements to secure transmissions, which will lead to more secrecy outage events. In conclusion, the simulation results in Figure 7 confirm that our proposed scheme can achieve reliable and secure communications in two-hop wireless relay networks.



(**a**) COP vs. SNR.

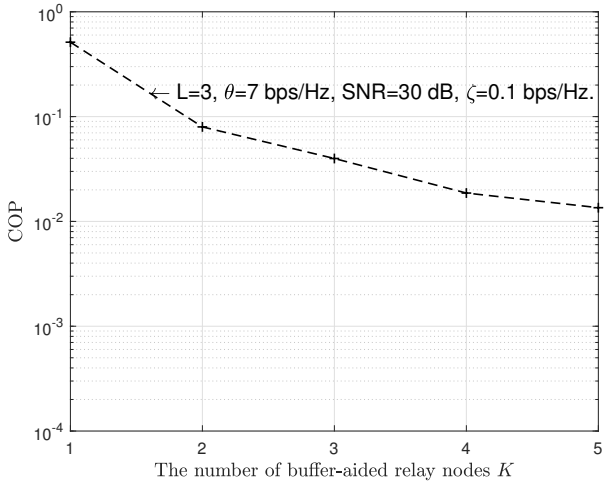


(**b**) SOP vs. the target secrecy rate ζ.

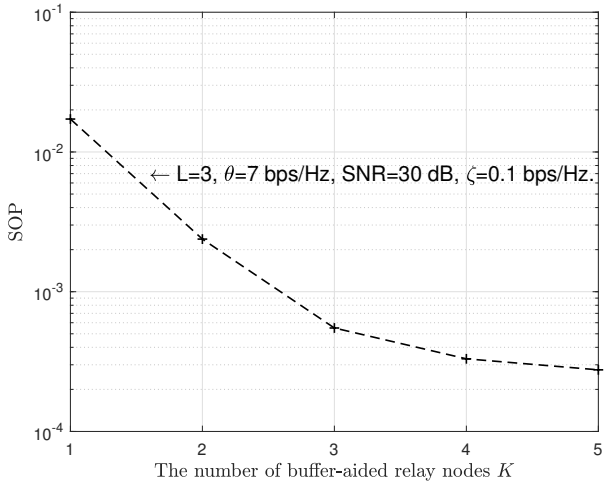
Figure 7. Analysis of the COP and SOP of the proposed scheme.

Then, we investigate the effect of the number of buffer-aided relay nodes K and the buffer length L on the reliability and security performances of the proposed scheme. Figures 8a and 9a investigate the effect of the number of buffer-aided relay nodes K and the buffer length L on the reliability performance of the proposed scheme, respectively. In Figures 8b and 9b, we, respectively, investigate the effect of the number of buffer-aided relay nodes k and the buffer length L on the security performance of the proposed scheme. We can observe from Figures 8a and 9a that the COP decreases gradually as the number of buffer-aided relay nodes K and the buffer length L increase. In Figures 8b and 9b, SOP also decreases as the number of buffer-aided relay nodes K and the buffer length L increase, respectively. The lower the COP and SOP, the higher the reliability and security. Resulting in Figures 8 and 9 indicate that the increase in the number of buffer-aided relay nodes K and

the buffer length L can improve the reliability and security performances of the proposed scheme. This is because the increase in the number of buffer-aided relay nodes K implies an increasing number of legitimate channels, and an increase in the buffer length L implies a lower probability of buffer-aided relay unavailability (the probability that a buffer is full).



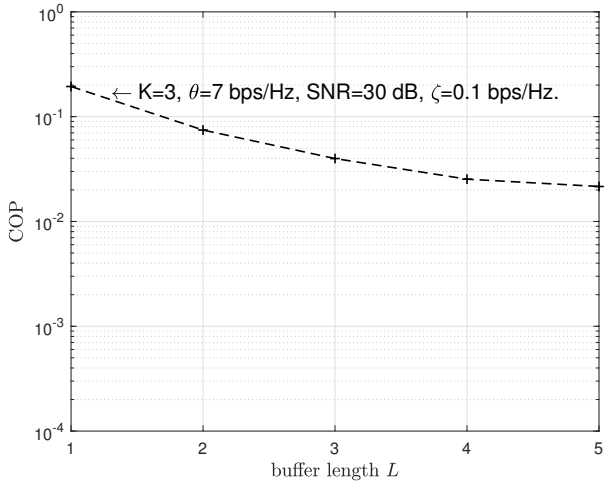
(**a**) The COP vs. K.



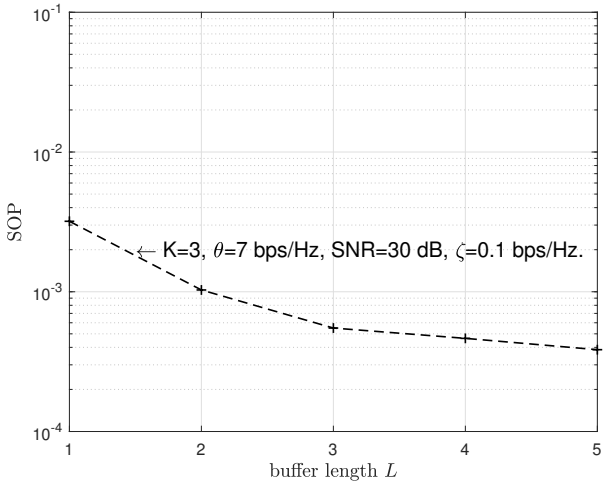
(**b**) The SOP vs. K.

Figure 8. The impact of the number of buffer-aided relay nodes K on the COP and SOP of the proposed scheme.

Finally, we made a comparison between our proposed scheme and two benchmark schemes (i.e., the max-link scheme and the max-ratio scheme) regarding the COP and SOP, respectively. By setting $K = 3$, $L = 3$, $\theta = 7$ bps/Hz, and $\zeta = 0.1$ bps/Hz, we show in Figure 10a how the COP changes by varying the SNR from 25 dB to 50 dB. The results in Figure 10a show that the COP of our scheme is always lower than that of the max-link scheme. By setting $K = 3$, $L = 3$, $\theta = 7$ bps/Hz, and SNR = 30 dB, we then illustrate in Figure 10b how the SOP varies by varying the target secrecy rate ζ from 0.1 to 0.9. The results in Figure 10b show that the SOP of our scheme is always lower than that of the max-ratio scheme, indicating that the security performance of the two-hop buffer-aided wireless network can be improved by adopting the DQL. In addition, we investigate the differences between the proposed scheme implemented by DQL and TQL. The comparison results are also shown in Figure 10a and Figure 10b, respectively. The comparison results clearly show that, under the same conditions, the COP and SOP of the proposed scheme implemented by DQL are both lower than those of the proposed scheme implemented by TQL.

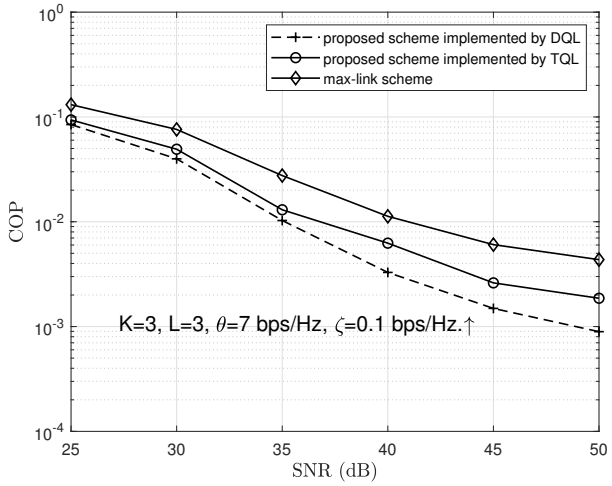


(**a**) The COP vs. L.



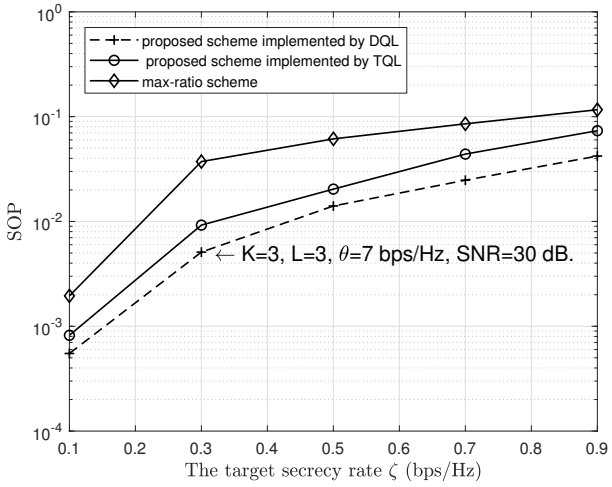
(**b**) The SOP vs. L.

Figure 9. The impact of the buffer length L on the COP and SOP of the proposed scheme.



(**a**) Comparison results of the COP.

Figure 10. *Cont*.



(**b**) Comparison results of the SOP.

Figure 10. Comparison results between the proposed scheme and other schemes.

7. Conclusions

This paper utilizes DQL to solve the problem of buffer-aided relay selection to achieve reliable and secure communications in a two-hop AF buffer-aided relay network with a passive eavesdropper. To propose the buffer-aided relay selection scheme, we first model the information transmission process in the network by applying an MDP. With the help of the MDP model, we then propose a novel buffer-aided relay selection scheme based on DQL to optimize the MDP. We finally verify the reliability and security performances of the proposed scheme by conducting Monte Carlo simulations and analyze how the network parameters affect the reliability and security performances of the concerned network in terms of the COP and the SOP. We also made a comparison between our proposed scheme and two benchmark buffer-aided relay selection schemes (i.e., the max-link scheme and the max-ratio scheme) regarding the COP and SOP, respectively. The results show that our proposed scheme can outperform the max-ratio scheme in terms of the SOP by 2.76 times.

Author Contributions: Conceptualization, C.Z. and X.L.; methodology, C.Z. and X.L.; software, C.Z. and Z.C.; writing—original draft preparation, C.Z.; writing—review and editing, X.L., Z.W., G.Q. and Z.Y.; supervision, Z.W.; funding acquisition, Z.W. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (62001273, 62002210), the Open Project Program of the Shaanxi Key Laboratory for Network Computing and Security Technology (NCST2021YB-02), the Fundamental Research Funds for the Central Universities (GK202103087) and the Scientific Research Plan of Shaanxi Provincial Department of Education (19JK0176).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study is included within the article.

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

Abbreviations

The following abbreviations are used in this manuscript:

AF	Amplify-and-forward
DQL	Deep Q-learning
COP	Connection outage probability
SOP	Secrecy outage probability
WSNs	Wireless sensor networks
CRNs	Cognitive radio networks
IoTs	Internet of Things
PLS	Physical-layer security
CSI	Channel-state information
TQL	Traditional Q-learning
MC	Markov chain
MDP	Markov decision process
FD	Full-duplex
CJ	Cooperative jamming
EH	Energy harvesting
DL	Deep learning
DNN	Deep neural network
PA	Power allocation
DF	Decode-and-forward
RF	Randomize-and-forward
HD	Half-duplex
AWGN	Additive white Gaussian noise
SNR	Signal-to-noise ratio
MSE	Mean square error

Symbols

The following symbols are used in this manuscript:

$h_{m,n}$	Channel coefficient between m and n
$g_{m,n}$	Channel gain between m and n
$\Omega m, n$	Average channel gain between m and n
$E[.]$	Expectation operator
$y_{R_k}(t)$	The received signal of R_k at time t
$x_S(t)$	The signal sent by S at time t
$n_{R_k}(t)$	AWGN noise of R_k at time t
$\psi_{S,R_k}(t)$	SNR of S to R_k link at t
$C_{S,R_K}(t)$	Channel capacity of S to R_k link
$A_{R_k}(t')$	Amplification factor of R_k at $t^{'}$
$C_{S,D}^{(s)}(t')$	The end-to-end secrecy rate
θ	The target rate
ζ	The target secrecy rate
s_t	The state at time t
a_t	The action at time t
$s(t)$	State space
$a(t)$	Action space
γ	Discount factor
$Q_\pi(s_t, a_t)$	Action-value function
$Q_*(s_t, a_t)$	The optimal action-value function
ϵ	Exploration probability
ϖ	MSE loss function
N_e	Training episodes
N_c	Capacity of replay buffer

References

1. Ding, Y.; Yang, Y.; Jiang, W.; Liu, Y.; He, T.; Zhang, D. Nationwide Deployment and Operation of a Virtual Arrival Detection System in the Wild. *IEEE/ACM Trans. Netw.* **2023**, *31*, 574–589. [CrossRef]
2. Li, Z.; Wang, S.; Han, S.; Meng, W.; Li, C. Joint Design of Beam Hopping and Multiple Access Based on Cognitive Radio for Integrated Satellite-Terrestrial Network. *IEEE Netw.* **2023**, *37*, 36–43. [CrossRef]
3. Xie, H.; Xia, M.; Wu, P.; Wang, S.; Poor, H.V. Edge Learning for Large-Scale Internet of Things With Task-Oriented Efficient Communication. *IEEE Trans. Wirel. Commun.* **2023**, 1–16.
4. Bapatla, D.; Prakriya, S. Performance of Two-Hop Links With an Energy Buffer-Aided IoT Source and a Data Buffer-Aided Relay. *IEEE Internet Things J.* **2021**, *8*, 5045–5061. [CrossRef]
5. Yerrapragada, A.K.; Eisman, T.; Kelley, B. Physical Layer Security for Beyond 5G: Ultra Secure Low Latency Communications. *IEEE Open J. Commun. Soc.* **2021**, *2*, 2232–2242. [CrossRef]
6. Angueira, P.; Val, I.; Montalbán, J.; Seijo, O.; Iradier, E.; Sanz, P.; Fanari, L.; Arriola, A. A Survey of Physical Layer Techniques for Secure Wireless Communications in Industry. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 810–838. [CrossRef]
7. Mitev, M.; Chorti, A.; Poor, H.V.; Fettweis, G.P. What Physical Layer Security Can Do for 6G Security. *IEEE Open J. Veh. Technol.* **2023**, *4*, 375–388. [CrossRef]
8. Huynh, P.; Phan, K.T.; Liu, B.; Ross, R. Throughput Analysis of Buffer-Aided Decode-and-Forward Wireless Relaying with RF Energy Harvesting. *Sensors* **2020**, *20*, 1222. [CrossRef]
9. Lu, X.; Xiao, L.; Li, P.; Ji, X.; Xu, C.; Yu, S.; Zhuang, W. Reinforcement Learning-Based Physical Cross-Layer Security and Privacy in 6G. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 425–466. [CrossRef]
10. Arzykulov, S.; Celik, A.; Nauryzbayev, G.; Eltawil, A.M. Artificial Noise and RIS-Aided Physical Layer Security: Optimal RIS Partitioning and Power Control. *IEEE Wirel. Commun. Lett.* **2023**, 1–5.
11. Huang, H.; Hu, S. Generalized Relays Subsets Selection Algorithm in Cloud-Based 6G Large-Scale Relays Network. *IEEE Internet Things J.* **2022**, *9*, 24754–24766. [CrossRef]
12. Zhou, F.; Chu, Z.; Sun, H.; Hu, R.Q.; Hanzo, L. Artificial Noise Aided Secure Cognitive Beamforming for Cooperative MISO-NOMA Using SWIPT. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 918–931. [CrossRef]
13. Vaishnavi, K.N.; Khorvi, S.D.; Kishore, R.; Gurugopinath, S. A Survey on Jamming Techniques in Physical Layer Security and Anti-Jamming Strategies for 6G. In Proceedings of the 2021 28th International Conference on Telecommunications (ICT), London, UK, 1–3 June 2021; pp. 174–179. [CrossRef]
14. Shukla, A.K.; Bhatnagar, M.R. Differential Modulation-Based Buffer-Aided Cooperative Relaying Network. *IEEE Syst. J.* **2022**, 1–12.
15. Xu, P.; Wang, Y.; Chen, G.; Krikidis, I.; Wong, K.K. Novel Mode Selection Schemes for Buffer-Aided Cooperative NOMA System. *IEEE Trans. Veh. Technol.* **2022**, *72*, 866–880.
16. Ikhlef, A.; Michalopoulos, D.S.; Schober, R. Max-Max Relay Selection for Relays with Buffers. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 1124–1135. [CrossRef]
17. Hamamreh, J.M.; Furqan, H.M.; Arslan, H. Classifications and Applications of Physical Layer Security Techniques for Confidentiality: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 1773–1828. [CrossRef]
18. Guo, W.; Qureshi, N.M.F.; Siddiqui, I.F.; Shin, D.R. Cooperative Communication Resource Allocation Strategies for 5G and Beyond Networks: A Review of Architecture, Challenges and Opportunities. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 8054–8078. [CrossRef]
19. Krikidis, I.; Charalambous, T.; Thompson, J.S. Buffer-Aided Relay Selection for Cooperative Diversity Systems without Delay Constraints. *IEEE Trans. Wirel. Commun.* **2012**, *11*, 1957–1967. [CrossRef]
20. Gong, Y.; Chen, G.; Xie, T. Using Buffers in Trust-Aware Relay Selection Networks With Spatially Random Relays. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 5818–5826. [CrossRef]
21. Alam, M.Z.; Abkenar, F.S.; Adhicandra, I.; Murali, S.; Jamalipour, A. Low-Delay Path Selection for Cluster-Based Buffer-Aided Vehicular Communications. *IEEE Trans. Veh. Technol.* **2020**, *69*, 9356–9363. [CrossRef]
22. Adanvo, V.F.; Mafra, S.; Montejo-Sánchez, S.; Fernández, E.M.G.; Souza, R.D. Buffer-Aided Relaying Strategies for Two-Way Wireless Networks. *Sustainability* **2022**, *14*, 13829. [CrossRef]
23. Jadoon, M.A.; Kim, S. Relay selection algorithm for wireless cooperative networks: A learning-based approach. *IET Commun.* **2017**, *11*, 1061–1066. [CrossRef]
24. Wang, X.; Jin, T.; Hu, L.; Qian, Z. Energy-Efficient Power Allocation and Q-Learning-Based Relay Selection for Relay-Aided D2D Communication. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6452–6462. [CrossRef]
25. Dong, Y.; Zhang, F.; Joe, I.; Lin, H.; Jiao, W.; Zhang, Y. Learning for Multiple-Relay Selection in a Vehicular Delay Tolerant Network. *IEEE Access* **2020**, *8*, 175602–175611. [CrossRef]
26. Chen, G.; Tian, Z.; Gong, Y.; Chen, Z.; Chambers, J.A. Max-Ratio Relay Selection in Secure Buffer-Aided Cooperative Wireless Networks. *IEEE Trans. Inf. Forensics Secur.* **2014**, *9*, 719–729. [CrossRef]
27. Mekkawy, T.; Yao, R.; Qi, N.; Lu, Y. Secure Relay Selection for Two Way Amplify-and-Forward Untrusted Relaying Networks. *IEEE Trans. Veh. Technol.* **2018**, *67*, 11979–11987. [CrossRef]

28. Zhang, C.; Liao, X.; Wu, Z.; Qiu, G. Buffer-Aided Relay Selection for Wireless Cooperative Relay Networks with Untrusted Relays. In Proceedings of the 2021 International Conference on Networking and Network Applications (NaNA), Lijiang, China, 29 October–1 November 2021; pp. 69–74. [CrossRef]
29. Nie, Y.; Lan, X.; Liu, Y.; Chen, Q.; Chen, G.; Fan, L.; Tang, D. Achievable Rate Region of Energy-Harvesting Based Secure Two-Way Buffer-Aided Relay Networks. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1610–1625. [CrossRef]
30. Srirutchataboon, G.; Sugiura, S. Physical Layer Security of Buffer-Aided Hybrid Virtual Full-Duplex and Half-Duplex Relay Selection. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 19–22 June 2022; pp. 1–5. [CrossRef]
31. He, J.; Liu, J.; Su, W.; Shen, Y.; Jiang, X.; Shiratori, N. Jamming and Link Selection for Joint Secrecy/Delay Guarantees in Buffer-Aided Relay System. *IEEE Trans. Commun.* **2022**, *70*, 5451–5468. [CrossRef]
32. Wang, Y.; Yin, H.; Zhang, T.; Yang, W.; Shang, X.; Shen, Y. Secure Transmission for Energy Harvesting Sensor Networks with a Buffer-Aided Sink Node. *IEEE Internet Things J.* **2021**, *9*, 6703–6718.
33. Mao, Q.; Hu, F.; Hao, Q. Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2595–2621. [CrossRef]
34. Zhang, C.; Patras, P.; Haddadi, H. Deep Learning in Mobile and Wireless Networking: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2224–2287. [CrossRef]
35. Buenrostro-Mariscal, R.; Santana-Mancilla, P.C.; Montesinos-López, O.A.; Nieto Hipólito, J.I.; Anido-Rifón, L.E. A Review of Deep Learning Applications for the Next Generation of Cognitive Networks. *Appl. Sci.* **2022**, *12*, 6262. [CrossRef]
36. Zhang, Z.; Lu, Y.; Huang, Y.; Zhang, P. Neural Network-Based Relay Selection in Two-Way SWIPT-Enabled Cognitive Radio Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6264–6274. [CrossRef]
37. Zhou, D.; Yan, B.; Li, C.; Wang, A.; Wei, H. Relay selection scheme based on deep reinforcement learning in wireless sensor networks. *Phys. Commun.* **2022**, *54*, 101799. [CrossRef]
38. Rezwan, S.; Choi, W. A Survey on Applications of Reinforcement Learning in Flying Ad-Hoc Networks. *Electronics* **2021**, *10*, 449. [CrossRef]
39. Zhu, J.; Song, Y.; Jiang, D.; Song, H. A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things. *IEEE Internet Things J.* **2018**, *5*, 2375–2385. [CrossRef]
40. Huang, C.; Chen, G.; Gong, Y. Delay-Constrained Buffer-Aided Relay Selection in the Internet of Things With Decision-Assisted Reinforcement Learning. *IEEE Internet Things J.* **2021**, *8*, 10198–10208. [CrossRef]
41. Huang, C.; Zhong, J.; Gong, Y.; Abdullah, Z.; Chen, G. Novel deep reinforcement learning-based delay-constrained buffer-aided relay selection in cognitive cooperative networks. *Electron. Lett.* **2020**, *56*, 1148–1150. [CrossRef]
42. Huang, C.; Chen, G.; Gong, Y.; Xu, P. Deep Reinforcement Learning Based Relay Selection in Delay-Constrained Secure Buffer-Aided CRNs. In Proceedings of the GLOBECOM 2020—2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; pp. 1–6.
43. Huang, C.; Chen, G.; Gong, Y.; Han, Z. Joint Buffer-Aided Hybrid-Duplex Relay Selection and Power Allocation for Secure Cognitive Networks With Double Deep Q-Network. *IEEE Trans. Cogn. Commun. Netw.* **2021**, *7*, 834–844. [CrossRef]
44. El-Zahr, S.; Abou-Rjeily, C. Threshold Based Relay Selection for Buffer-Aided Cooperative Relaying Systems. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 6210–6223. [CrossRef]
45. Xu, P.; Quan, J.; Chen, G.; Yang, Z.; Li, Y.; Krikidis, I. A Novel Link Selection in Coordinated Direct and Buffer-Aided Relay Transmission. *IEEE Trans. Wirel. Commun.* **2022**, *22*, 3296–3309.