Li *et al. EURASIP Journal on Wireless Communications and Networking* (2020) 2020:7
https://doi.org/10.1186/s13638-019-1595-x

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Cache-enabled physical-layer secure game against smart uAV-assisted attacks in b5G NOMA networks

Chao Li[1], Zihe Gao[2], Junjuan Xia[1*], Dan Deng[3*] and Liseng Fan[1*]

Abstract

This paper investigates cache-enabled physical-layer secure communication in a no-orthogonal multiple access (NOMA) network with two users, where an intelligent unmanned aerial vehicle (UAV) is equipped with attack module which can perform as multiple attack modes. We present a power allocation strategy to enhance the transmission security. To this end, we propose an algorithm which can adaptively control the power allocation factor for the source station in NOMA network based on reinforcement learning. The interaction between the source station and UAV is regarded as a dynamic game. In the process of the game, the source station adjusts the power allocation factor appropriately according to the current work mode of the attack module on UAV. To maximize the benefit value, the source station keeps exploring the changing radio environment until the Nash equilibrium (NE) is reached. Moreover, the proof of the NE is given to verify the strategy we proposed is optimal. Simulation results prove the effectiveness of the strategy.

Keywords: Cache, UAV, B5G, NOMA, Physical-layer security, Reinforcement learning

1 Introduction

In recent years, ultra-reliable and low-latency have been a very important requirement for supporting the wireless services for the B5G wireless communications [1–4]. To support this requirement, caching technique can pre-store the wireless data during non-peak traffic time and hence reduce the load traffic significantly [5–8]. In addition, non-orthogonal multiple access (NOMA) can provide much higher capacity and spectrum efficiency than that of orthogonal multiple access, and hence, it is one of the most promising candidate for supporting ultra-reliable and low-latency services. Moreover, NOMA protocol enables the source station to allocate the same spectrum and time resource to multiple users with power-domain multiplexing. In particular, NOMA protocol can serve different kinds of users, and it can flexibly support ultra-reliable and low-latency services for both far and near users.

Although NOMA technology can provide a reliable performance in enhancing wireless transmission, its transmission security is threatened by the eavesdroppers due to the broadcasting nature of wireless communications [9–13]. The authors in [14] have studied the protection of physical-layer security and proposed strategies for wireless communication networks which have been confirmed to perform efficiently. In [15], the authors studied the antenna selection algorithm to protect physical-layer security in NOMA network with an eavesdropper. However, the conventional strategies for protecting the physical-layer security in NOMA system work well, only when the attacker just has one work mode. Intelligent attacker with multiple work mode is proposed in [16–20] to reduce the data rate of communication systems by freely switching between eavesdropping, jamming, deception, and silent. If the networks continue to adopt the conventional strategies, the intelligent attacks will not be suppressed.

To tackle this problem, the authors in [21–24] proposed a transmission policy based on reinforcement learning. As a special branch of artificial intelligence, the reinforcement learning proposed in [25] can be regarded as

*Correspondence: xiajunjuan@gzhu.edu.cn; dengdan@mail.ustc.edu.cn; lsfan2019@126.com
[1]The School of Computer Science, Guangzhou University, Guangzhou, China
[3]Guangzhou Panyu Polytechnic, Guangzhou, China
Full list of author information is available at the end of the article

a Markov decision-making process. The agent trained by reinforcement learning can decide the action to be executed according to the environment state at the current moment, and maximize the long-term cumulative rewards to obtain the optimal action set. However, the state transition probability is generally unknowable for the agent. The Q-learning is proposed in [26] to solve the problem. Combining dynamic programming with the Monte Carlo method, Q-learning can make the agent learn optimal strategies without knowing the state transition probability. As far as we know, no previous work has used the Q-learning algorithm to protect secure transmission in the NOMA system, which is threatened by the intelligent attacker.

Due to mobility and ease of deployment, unmanned aerial vehicles (UAVs) have arisen as a new type of communication nodes in the wireless networks, for example, the UAVs can perform as a relay or base station under extreme natural conditions. However, a UAV can be a mobile intelligent attacker if it is equipped with attack module. In this paper, we investigate a NOMA network with two users in the presence of an UAV attacker which can execute multiple attack modes. The source station sends the composite signals to two users at the same time; therefore, the total transmit power is divided into two parts. We dynamically allocate the proportions of transmit power to confront the intelligent attacker. In the wireless communication process, it is hard to know the work mode transition probability of intelligent attacker. As a model-free learning method without depending on the state transition probability, the Q-learning is adopted to obtain a learning-based adaptive policy. Furthermore, we formulate the confrontation between the source station and intelligent attacker as a dynamic game, and we derive the Nash equilibrium (NE) of the dynamic game. Simulation results show that the strategy we proposed significantly improved the data rate of NOMA system.

2 Methods/experimental

Consider one cache-enabled source station S can pre-store a certain amount of information. There exists one cell-edge user U_1 and one central user U_2 in the coverage of S, where U_2 is closer to S than U_1. When the request signals from users are received, S transmits cached messages based on NOMA protocol to users. Furthermore, there exists a UAV which performs as an intelligent attacker E in this area. We suppose that the UAV is more likely to attack cell-edge user U_1, and the UAV remains in the same position when attacking. Programmable radio equipment on E can flexibly select to overheard information from S, send jamming or deception signals to U_1, or keep silent. We denote these four work modes of E as $m = 0, 1, 2$, and 3, respectively. In the experiment, the purpose of E is to attempt to decrease the system data rate and reduce the correctness of user decoding. For simplicity, all the devices in this experiment are equipped with single antenna.

3 NOMA networks

Now, we depict the NOMA network system model which is shown in Fig. 1. We suppose that S transmits a composite signal consisting of x_1 and x_2, which contains messages requested by U_1 and U_2, respectively. According to NOMA protocol, S divides the total transmit power P_S into two portions, i.e., αP_S and βP_S, where α and β are the power allocation factors for x_1 and x_2, respectively. In order to satisfy the requirements of different transmission

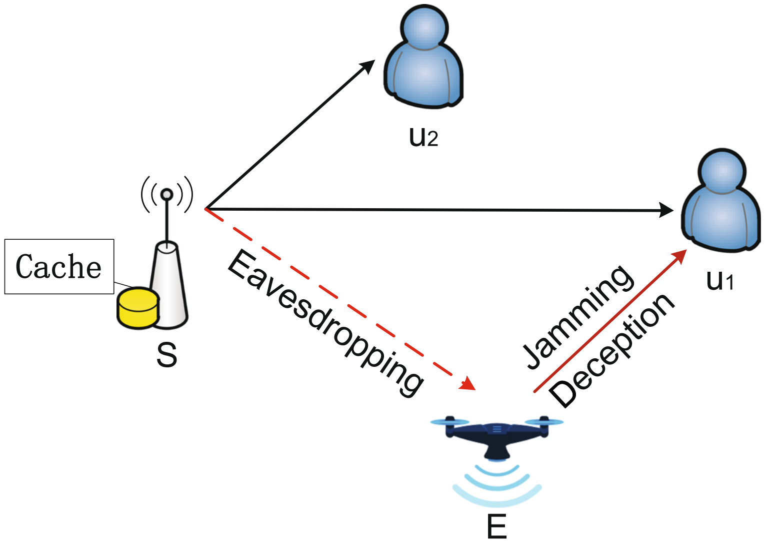

Fig. 1 Cache-assisted NOMA network of two users in different locations against intelligent attacks from UAV

distance, the two factors αP_S and βP_S have to meet the following constraint conditions:

$$\begin{cases} \alpha \gg \beta, \\ \alpha + \beta \leq 1. \end{cases} \tag{1}$$

In order to fight against the intelligent UAV attacker E, S works on improving system data rate by consciously changing its power allocation factor α. For the first step of the transmission process, S chooses a value for the power allocation factor α to transmit the mixture signal x_1, x_2, and then, the received signal at U_1 denoted by y_{U_1} can be given as:

$$y_{U_1} = h_{SU_1}(\sqrt{\alpha P_S}x_1 + \sqrt{\beta P_S}x_2) + n_{U_1} \tag{2}$$

where $h_{SU_1} \sim \mathcal{CN}(0, \nu^2)$ is the instantaneous channel coefficient of $S - U_1$ link. $n_{U_1} \sim \mathcal{CN}(0, \sigma^2)$ represents the additive white Gaussian noise (AWGN) received at U_1 [27–30]. The resultant SINR for x_1 at U_1 can be written as:

$$\mathrm{SINR}_{U_1}^{x_1} = \frac{\alpha P_S |h_{SU_1}|^2}{\beta P_S |h_{SU_1}|^2 + \sigma^2}. \tag{3}$$

when $m = 0$ holds, i.e., E shuts down radio equipment and stays silent. In this case, the achievable rates of x_1 at U_1 denoted by C_{U_1} is exactly the system data rate $C_{sys,0}$. Thus, the system data rate is acquired by [31]:

$$\begin{aligned} C_{sys,0} &= \log_2(1 + \frac{\alpha P_S |h_{SU_1}|^2}{\beta P_S |h_{SU_1}|^2 + \sigma^2}) \\ &= \log_2(1 + \frac{\alpha \widetilde{P}_S |h_{SU_1}|^2}{\beta \widetilde{P}_S |h_{SU_1}|^2 + 1}), \end{aligned} \tag{4}$$

where $\widetilde{P}_S = P_S/\sigma^2$. When $m = 1$ holds, E executes to overhear information from S; the received signal at E can be given as:

$$y_E = h_{SE}(\sqrt{\alpha P_S}x_1 + \sqrt{\beta P_S}x_2) + n_E, \tag{5}$$

we assume that perfect SIC receiver is applied at E; thus, according to [32], the achievable rate of x_1 at E denoted by C_E can be written as:

$$C_E = \log_2(1 + \frac{\alpha \widetilde{P}_S |h_{SE}|^2}{\beta \widetilde{P}_S |h_{SE}|^2 + 1}), \tag{6}$$

where $h_{SE} \sim \mathcal{CN}(0, \mu^2)$ is the instantaneous channel coefficient of $S - E$ link. $n_E \sim \mathcal{CN}(0, \sigma^2)$ represents AWGN received at E. Consequently, according to [17], the system data rate $C_{sys,1}$ can be computed by:

$$C_{sys,1} = [\, C_{sys,0} - C_E]^+, \tag{7}$$

where $[X]^+$ returns X if X is positive, while returns 0 otherwise. When $m = 2$ holds, E selects to transmit a jamming signal to U_1; the received signal y_{U_1} at U_1 can be acquired by:

$$y_{U_1,J} = h_{SU_1}(\sqrt{\alpha P_S}x_1 + \sqrt{\beta P_S}x_2) + h_{EU_1}\sqrt{P_J}x_J + n_{U_1} \tag{8}$$

where $h_{EU_1} \sim \mathcal{CN}(0, \lambda^2)$ is the instantaneous channel coefficient of $E - U_1$ link. P_J is the jamming power of E, and x_J represents the jamming signal transmitted by E. Therefore, in this case, the system data rate $C_{sys,2}$ can be computed by:

$$C_{sys,2} = \log 2(1 + \frac{\alpha \widetilde{P}_S |h_{SU_1}|^2}{\beta \widetilde{P}_S |h_{SU_1}|^2 + \widetilde{P}_J |h_{EU_1}|^2 + 1}) \tag{9}$$

where $\widetilde{P}_J = P_J/\sigma^2$. When $m = 3$ holds, S does not send signal to U_1 while E transmits the deception signal x_D. The received signal at U_1 becomes:

$$y_{U_1,D} = h_{EU_1}\sqrt{P_D}x_D + n_{U_1}, \tag{10}$$

where P_D is the deception power. The increase of the deception signal received by U_1 is bound to cause more loss in the achievable rate at U_1. Thus, the system data rate $C_{sys,3}$ can be formulated as a linear function and given by:

$$C_{sys,3} = C_{sys,0} - \gamma \log_2(1 + \widetilde{P}_D |h_{EU_1}|^2), \tag{11}$$

where $\widetilde{P}_D = P_J/\sigma^2$. $\gamma \in (0, 1)$ is the deception factor which quantifies the probability of the influence of each deception signal.

4 Secure game in NOMA network

The interaction between S and E in the NOMA network performs in a rivalry way, which is formulated as a secure game. To discuss the process of the secure game, we need to first quantify the variety range of α. While ensuring that U_1 can decode the received information correctly, we must also ensure that U_2 can correctly decode x_2. We denote the minimum data rate requirement for U_1 and U_2 as $C_{\min}^{U_1}$ and $C_{\min}^{U_2}$. Thus, α and β satisfy the following constraint:

$$\log_2(1 + \frac{\alpha \widetilde{P}_S |h_{SU_1}|^2}{\beta \widetilde{P}_S |h_{SU_1}|^2 + 1}) \geq C_{\min}^{U_1}, \tag{12}$$

$$\log_2(1 + \beta \widetilde{P}_S |h_{SU_2}|^2) \geq C_{\min}^{U_2}, \tag{13}$$

according to (1), the threshold value of α is given by:

$$\begin{cases} \alpha_{\max} = 1 - \frac{2^{C_{\min}^{U_2}-1}}{\widetilde{P}_S|h_{SU_2}|^2}, \\ \alpha_{\min} = \frac{(2^{C_{\min}^{U_1}}-1)(\beta\widetilde{P}_S|h_{SU_1}|^2+1)}{\widetilde{P}_S|h_{SU_1}|^2}. \end{cases} \tag{14}$$

where $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum power allocation factor for x_1. We now turn to discuss the process of the secure game. S is adaptively adjusting its power allocation factor in the range of $[\alpha_{\min}, \alpha_{\max}]$, while E selects to execute an attack modes $m \in \{0, 1, 2, 3\}$, which represents keeping silent, eavesdropping, jamming, or deception, respectively. In each time slot, E attempts to reduce the system data rate, i.e., $C_{sys,1}$, $C_{sys,2}$, or $C_{sys,3}$. S devotes to increase the system data rate by controlling α and meanwhile suppressing the probability of attacking. In view of this, we regard the confrontation between S and E as a zero-sum game. Depending on the system data rate and power consumption, the reward function of S denoted by R_S in the zero-sum game is formulated as:

$$R_S(\alpha, m) = \ln 2 C_{sys,m} - \alpha\theta, \tag{15}$$

where θ is the total power consumption. We introduce coefficient $\ln 2$ to simplify the subsequent derivation process. According to the distinguishing feature of zero-sum game, the reward function of E denoted by R_E is defined as:

$$R_E(\alpha, m) = -\ln 2 C_{sys,m} - \varphi_m, \tag{16}$$

where $\varphi_{m=0,1,2,3}$ denotes the consumption of E in mode m. In the secure game, S tries to find an optimal power allocation factor in $[\alpha_{\min}, \alpha_{\max}]$ to maximize R_S, and E is dynamically adjusting its work modes to maximize R_E. The purpose of the game between S and E is to achieve their own optimal strategies α^* and m^*, respectively. Then, we define the set of strategies $\{\alpha^*, m^*\}$ as the Nash equilibrium (NE) of the secure game, where S and E gain the maximize reward value. Thus, the NE strategy is given by:

$$R_S(\alpha^*, m^*) \geq R_S(\alpha, m^*), \tag{17}$$

$$R_E(\alpha^*, m^*) \geq R_E(\alpha^*, m). \tag{18}$$

Through analytical derivation, we obtain one NE solution $\{\alpha^*, 0\}$. That is to say, if S keeps choosing a power allocation factor α^*, E will obtain the maximized reward value by keeping silent, and it has no motivation to execute any attack modes. Specifically, the NE solution is given and proved in the following Lemma 1 and Proof.

Lemma 1 *: The secure game in the NOMA network has one NE solution* $\{\alpha^*, 0\}$, *which is acquired by*

$$\alpha^* = \frac{\widetilde{P}_S|h_{SU_1}|^2 - \theta}{\widetilde{P}_S|h_{SU_1}|^2\theta} - \beta \qquad \alpha_{\min} < \alpha^* \leq \alpha_{\max}. \tag{19}$$

if the following constraints are met:

$$\frac{\widetilde{P}_S|h_{SU_1}|^2}{(\alpha_{\max}+\beta)\widetilde{P}_S|h_{SU_1}|^2+1} < \theta < \frac{\widetilde{P}_S|h_{SU_1}|^2}{(\alpha_{\min}+\beta)\widetilde{P}_S|h_{SU_1}|^2+1}, \tag{20a}$$

$$\varphi_1 \geq \ln(1 + \frac{\alpha^*\widetilde{P}_S|h_{SE}|^2}{\beta\widetilde{P}_S|h_{SE}|^2 + 1}), \tag{20b}$$

$$\varphi_2 \geq \ln \tag{20c}$$

$$-\ln(1 + \frac{\alpha^*\widetilde{P}_S|h_{SU_1}|^2}{\beta\widetilde{P}_S|h_{SU_1}|^2 + \widetilde{P}_J|h_{EU_1}|^2 + 1}), \tag{20d}$$

$$\varphi_3 \geq \gamma\ln(1 + \widetilde{P}_D|h_{EU_1}|^2). \tag{20e}$$

Proof The proof of this Lemma is given in the Appendix

5 NOMA power allocation algorithm

In order to suppress the attack probability efficiently in the secure game, S must adopt appropriate power allocation strategy. However, because of the complexity and variability of radio signals in the NOMA network, S can barely predict the channel state information and the work modes of E. For this reason, we propose a power allocation algorithm based on Q-learning. By incorporating the Monte Carlo and dynamic programming methods, Q-learning is regarded as one of the most effective algorithms in model-free reinforcement learning. Without knowing the state of the environment and its transition probability, the agent is constantly exploring the environment and making trial-and-error experiments. After many independent repetitive experiments and the average is obtained, the Q-learning-based agent will acquire the optimal strategy.

Based on above ideas, we propose the power allocation algorithm of NOMA for the secure game. In consideration of the inherent relation between S and E, the work mode of E determines the state of S; similarly, S can influence the environment of E by adjusting α. In the first step of the algorithm, we initialize the Q-table denoted by $Q(m, \alpha)$ which is used for updating the reward values of state-action pairs. For each experiment, E first selects a work mode randomly, which determines S to adopt an instantaneous α_t accordingly, where α_t denotes the power allocation factor at time t. It should be emphasized that we do not expect that S always selects the appropriate power allocation factor by searching in the Q-table. To avoid getting the local optimal solution, we use $\epsilon -$greedy policy when S chooses a value of α. Specifically, S searches for the current optimal α in Q-table with probability ϵ, otherwise chooses a value in the range of $[\alpha_{\min}, \alpha_{\max}]$ randomly. At this time slot, S transmits a signal with power $\alpha_t P_S$ and computes the system data rate as reward value R_S from the environment. Then, E changes the work mode from m to m_{t+1} according to the system data rate. By incorporating the instantaneous reward value R_S and the accumulated experience in Q-table, the update process of

Q-table presented by the authors in [33] can be formulated as:

$$Q(m_t,\alpha_t)\leftarrow Q(m_t,\alpha_t)+\zeta\left[R_S + \rho \max Q(m_{t+1},\alpha) - Q(m_t,\alpha_t)\right], \tag{21}$$

where $\zeta \in (0,1]$ is the parameter to control the rate of learning. $\rho \in [0,1]$ represents the proportion of accumulated experience. To solve the problem of not knowing the state transition probability, we repeat the experiment multiple times and compute the average reward value. After enough updates and repeated experiments, the Q-table converges to be optimal. From the optimal Q-table, S can obtain a learning-based optimal power allocation strategy. Algorithm 1 describes the learning process:

Algorithm 1:NOMA Power Allocation Algorithm.

```
Initialize $Q(m,\alpha)$,
for all $m \in 0,1,2,3$, $\alpha \in [\alpha_{\min},\alpha_{\max}]$ at random
Loop for each episode:
    Initialize $m$
    loop for each time slot of episode:
        Choose $\alpha_t$ from $m_t$ using $Q(\epsilon - greedy)$ policy
        Take $\alpha_t$, observe $R_S, m_{t+1}$
        $Q(m_t,\alpha_t)\leftarrow Q(m_t,\alpha_t)+\zeta[R_S+\rho \max_\alpha Q(m_{t+1},\alpha)$
        $-Q(m_t,\alpha_t)]$
        $m_t \leftarrow m_{t+1}$
    until time slot is terminal
```

6 Results and discussion

In this section, we simulate the communication process to verify the effectiveness of the proposed algorithms. The links in the network experience the Rayleigh flat fading [34–37], and the nodes are equipped with a single antenna. We set the parameter as follows: $\{\nu^2,\mu^2,\lambda^2\} = \{1.2, 0.5, 2\}$, $\varphi_{m=\{0,1,2,3\}} = \{0, 1.8, 2.0, 2.1\}$, $\gamma = 0.6$, $\widetilde{P}_J = 2$, $\widetilde{P}_D = 2.1$. We set the power allocation factor α to vary from 0.6 to 0.9 with a change interval of 0.02, and β is set to a constant value 0.1. Specifically, we set 10,000 time slots for each experiment, and then, we repeat 5000 experiments to find the average.

Figure 2 reflects the variation of the average reward value of S and E from 0 to 10,000 time slots. From this figure, we can see that the average reward value of S and E both increases rapidly between 0 and 1000 time slots. In the subsequent process, the two curves rise slowly and reach their peak value at 3000 time slot point, respectively. Then, the two curves remain steady until the terminal of the experiment. In the learning-based algorithms, we expect agents to select specific actions to improve their long-term cumulative rewards, which is consistent with the experimental results.

The purpose of our proposed power allocation strategy is to improve the average data rate of the system, which is well reflected in Fig. 3. From 0 to 1000 time slot, the average system data rate dramatically grows from the initial value 0.76 to a temporary value 1.23. After that, the average system data rate continues to rise slowly until it converges to 1.31 at 3000 time slot point, and then keeps a steady level from 3000 to the terminal. The change trend

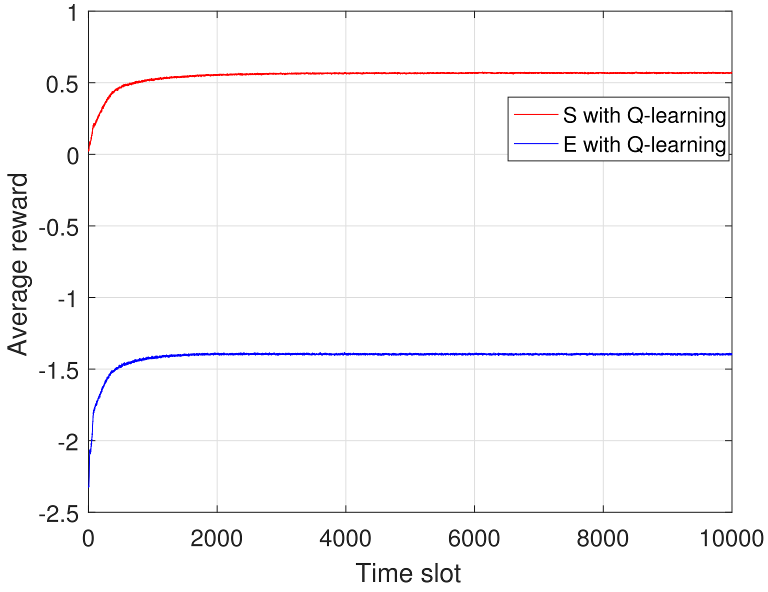

Fig. 2 The average reward of the power allocation algorithm

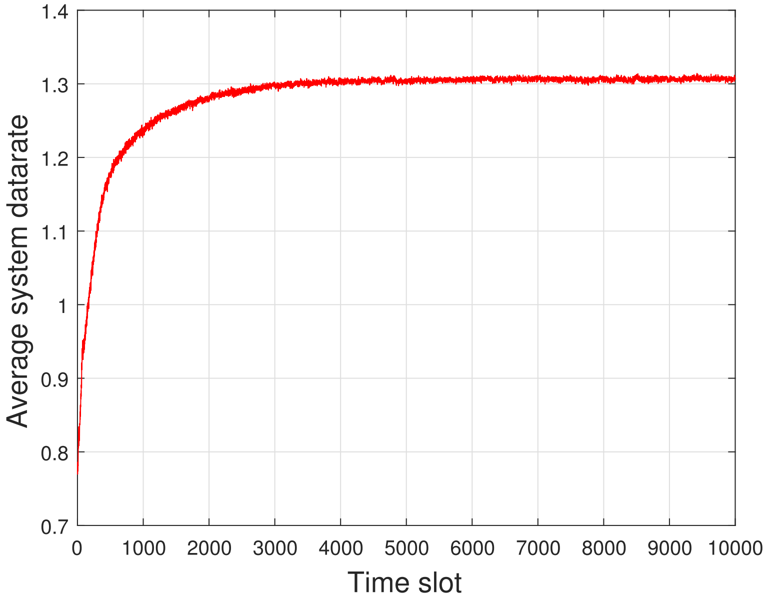

Fig. 3 The system data rate of the power allocation algorithm

of system data rate is basically consistent with the average reward value, which also proves that the increase of system data rate will bring more rewards to agents.

Figure 4 shows a dynamic programming process of average power allocation factor in the reinforcement learning process. As can be seen from the figure, the power allocation factor has a random initial value of 0.75. After the start of the experiment, the work mode of E begins to change, and S dynamically adjusts the power allocation factor according to the environment transformation. In the first 500 time slots, the average power allocation factor gradually decreases to a temporary value of 0.708. Between 500 and 4000, the average power allocation gradually increases and then remains stable around 0.737.

Figure 5 indicates the average attack probabilities of E versus the time slot varying from 0 to 10,000. We find that the average attack probabilities fall quickly from 0 to 1000. After that, the three curves decrease slowly and

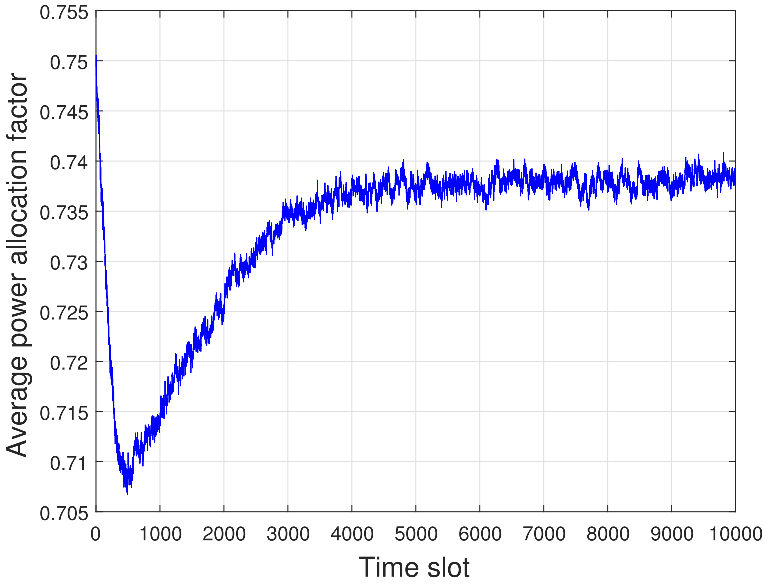

Fig. 4 The average power allocation factor of the power allocation algorithm

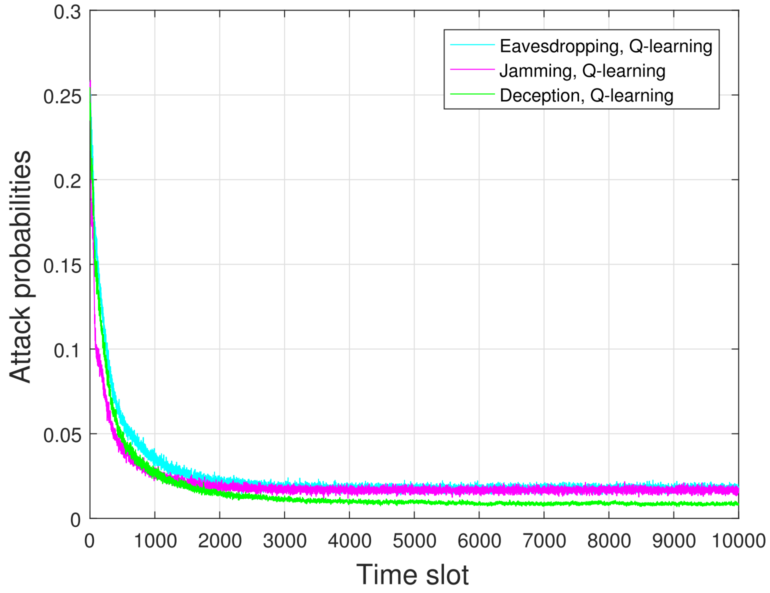

Fig. 5 The average attack probabilities of the power allocation algorithm with $\{\nu^2, \mu^2, \lambda^2\} = \{1.2, 0.5, 2\}$

tend to converge gradually. The probability of eavesdropping drops from the initial value of 0.25 to the convergence value of 0.025, and the decline rate reaches 90%. The probability of jamming drops from the initial value of 0.26 to the convergence value of 0.02, and the decline rate is 92.3%. Similarly, the probability of deception drops from the initial value of 0.27 to the convergence value of 0.01; therefore, the decline rate is 96.2%. What is more, we simulate the average attack probabilities of the power allocation algorithm again with different parameters. We set the channel parameters as $\{\nu^2, \mu^2, \lambda^2\} = \{0.9, 0.3, 2\}$. That is to say, we assume that the cell-edge user u_1 is placed further away from S. Correspondingly, E is also further away from S. Compared with Fig. 5, Fig. 6 shows that the converged eavesdropping probability becomes lower; at the same time, the converged deception and jamming probabilities grow up 2% with the condition that the jamming and deception power are fixed. Alignment of Fig. 5 with Fig. 6 can find that the proposed policy performs well regardless of the location of cell-edge user and UAV.

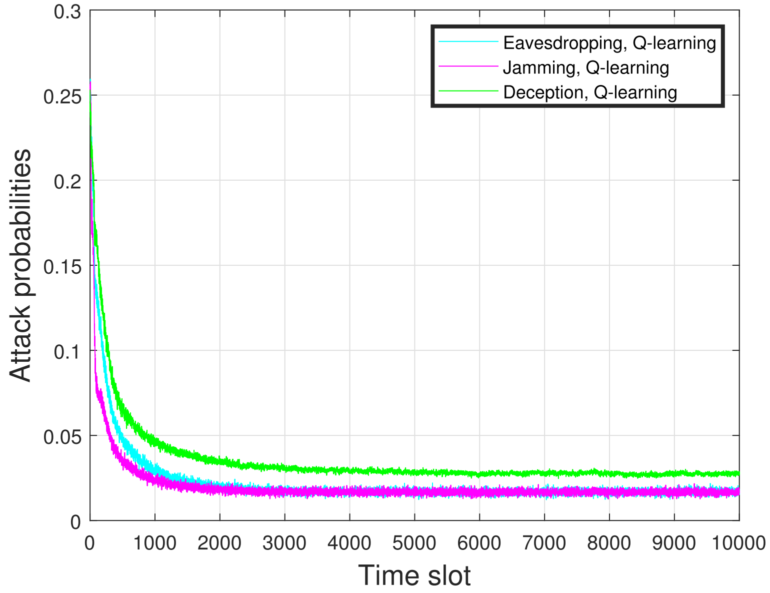

Fig. 6 The average attack probabilities of the power allocation algorithm with $\{\nu^2, \mu^2, \lambda^2\} = \{0.9, 0.3, 2\}$

7 Conclusions

In this paper, we investigated the cache-assisted physical-layer security of a NOMA communication network where there exists an intelligent attacker UAV nearby the cell-edge user. The UAV within the coverage of the network tries to reduce the system data rate of the NOMA network by flexibly switching a work mode among eavesdropping, jamming, deception, and keep silence. According to the NOMA protocol, the transmitter in the system has to allocate the total power to two users in a certain proportion. In that way, we need an immediate strategy to adjust the power allocation factor to suppress the attack motivation of the UAV. To tackle this problem, we proposed the power allocation strategy based on Q-learning to control the power allocation factor. From the simulation results, we can see that the proposed strategy can well adjust the power allocation factor in real time. Furthermore, we confirmed that this strategy has excellent performance in enhancing the system data rate and suppressing the attack probabilities. In the future works, we will apply the wireless caching technique[38–40] to the NOMA systems to further enhance the transmission reliability and security. In addition, we will consider some new materials [41–43] for enhancing the communication performance in the practical applications. Furthermore, some intelligent algorithms such as deep learning-based algorithms [44–47] will be applied into the considered system, in order to further enhance the network performance.

Appendix

Proof : By substituting $m = 0$ into (15), we have

$$R_S(\alpha, 0) = \ln(1 + \frac{\alpha \widetilde{P}_S |h_{SU_1}|^2}{\beta \widetilde{P}_S |h_{SU_1}|^2 + 1}) - \alpha\theta. \tag{22}$$

We take the partial derivative of $R_S(\alpha, 0)$ with respect to α and have

$$\frac{\partial R_S(\alpha, 0)}{\partial \alpha} = \frac{\widetilde{P}_S |h_{SU_1}|^2}{(\alpha + \beta)\widetilde{P}_S |h_{SU_1}|^2 + 1} - \theta, \tag{23}$$

by making further derivative, easy to find

$$\frac{\partial R_S^2(\alpha, 0)}{\partial \alpha^2} = -\frac{\widetilde{P}_S^2 |h_{SU_1}|^4}{[(\alpha + \beta)\widetilde{P}_S |h_{SU_1}|^2 + 1]^2} \leq 0, \tag{24}$$

showing that (22) is a convex function, i.e., $\partial R_S(\alpha, 0)/\partial \alpha = 0$. So we substitute $\alpha = \alpha^*$ into (23); thus, (19) holds on. To ensure that (23) acquires the maximum in the range of $[\alpha_{\min}, \alpha_{\max}]$, let the following inequalities hold:

$$\frac{\partial R_S(\alpha, 0)}{\partial \alpha}|_{\alpha=\alpha_{\min}} = \frac{\widetilde{P}_S |h_{SU_1}|^2}{(\alpha_{\min} + \beta)\widetilde{P}_S |h_{SU_1}|^2 + 1} - \theta > 0, \tag{25}$$

$$\frac{\partial R_S(\alpha, 0)}{\partial \alpha}|_{\alpha=\alpha_{\max}} = \frac{\widetilde{P}_S |h_{SU_1}|^2}{(\alpha_{\max} + \beta)\widetilde{P}_S |h_{SU_1}|^2 + 1} - \theta < 0, \tag{26}$$

i.e., (20a) holds. Therefore, $(\alpha^*, 0)$ satisfies (17). To ensure that $(\alpha^*, 0)$ satisfies (18), by substituting $((\alpha^*, 0))$ into (16), we let the following inequalities hold:

$$R_E(\alpha^*, 0) - R_E(\alpha^*, 1) \geq 0, \tag{27a}$$

$$R_E(\alpha^*, 0) - R_E(\alpha^*, 2) \geq 0, \tag{27b}$$

$$R_E(\alpha^*, 0) - R_E(\alpha^*, 3) \geq 0, \tag{27c}$$

i.e., (20b)–(20d) hold. Therefore, $(\alpha^*, 0)$ also satisfies (18).

Above all, we prove the set of strategy $(\alpha^*, 0)$ meanwhile satisfies Eqs. (17) and (18), which is the strict definition of NE. With this, Lemma 1 is completely proved. □

Abbreviations

NE: Nash equilibrium; NOMA: non-orthogonal multiple access; UAV: unmanned aerial vehicle

Acknowledgements

Not applicable.

Author's contributions

LC deduced the formulas and made the simulation experiments. ZG analyzed the communication scenarios and modeled the network of this paper. JX presented the reinforcement learning algorithm in this work. DD embellished the language of this manuscript. FL improved the presentation of figure style in this work and enhanced the novelty. All the authors have read and approved the final manuscript.

Funding

This work was supported by National Natural Science Foundation of China 397 under Grant 61871139, by the Science and Technology Program of Guangzhou under Grant 201807010103, by the Natural Science Foundation of Guangdong Province under Grant 2018A030313736, by the Scientific Research Project of Education Department of Guangdong, China under Grant 2017GKTSCX045, by the Science and Technology Program of Guangzhou, China under Grant 201707010389, and by the Project of Technology Development Foundation of Guangdong under Grant 706049150203.

Availability of data and materials

The authors state the data availability in this manuscript through the email to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]The School of Computer Science, Guangzhou University, Guangzhou, China. [2]The Research Center of Institute of Telecommunication Satellite, China Academy of Space Technology, Beijing, China. [3]Guangzhou Panyu Polytechnic, Guangzhou, China.

Received: 20 September 2019 Accepted: 6 November 2019
Published online: 06 January 2020

References

1. J. Zhao, Q. Li, Y. Gong, Computation offloading and resource allocation for mobile edge computing with multiple access points. IET Commun. **PP**(99), 1–10 (2019)

2. J. Yang, D. Ruan, J. Huang, X. Kang, Y.-Q. Shi, An embedding cost learning framework using gan. IEEE Trans. Inf. Forensic. Secur. **PP**(99), 1–10 (2019)
3. B. Wang, F. Gao, S. Jin, H. Lin, G. Y. Li, Spatial- and frequency-wideband effects in millimeter-wave massive MIMO systems. IEEE Trans. Sig. Processing. **66**(13), 3393–3406 (2018)
4. X. Hu, C. Zhong, X. Chen, W. Xu, Z. Zhang, Cluster grouping and power control for angle-domain mmwave mimo noma systems. IEEE J Sel. Top. Sig. Process. **13**(5), 1167–1180 (2019)
5. L. Fan, N. Zhao, X. Lei, Q. Chen, N. Yang, G. K. Karagiannidis, Outage probability and optimal cache placement for multiple amplify-and-forward relay networks. IEEE Trans. Veh. Technol. **67**(12), 12373–12378 (2018)
6. X. Lin, Probabilistic caching placement in uav-assisted heterogeneous wireless networks. Phys. Commun. **33**, 54–61 (2019)
7. F. Shi, Secure probabilistic caching in random multi-user multi-uav relay networks. Phys. Commun. **32**, 31–40 (2019)
8. C. Li, L. Peng, Z. Chao, S. Fan, J. Cioffi, L. Yang, Spectral-efficient cellular communications with coexistent one- and two-hop transmissions. IEEE Trans. Veh. Technol. **65**(8), 6765–6772 (2016)
9. G. Gomez, F. J. Martin-Vega, F. J. Lopez-Martinez, Y. Liu, M. Elkashlan, G. Gomez, F. J. Martin-Vega, F. J. Lopez-Martinez, Y. Liu, M. Elkashlan, Uplink noma in large-scale systems: Coverage and physical layer security. CoRR. **abs/1709.04693** (2017)
10. C. Zheng, H. Xin, X. Guo, T. Ristaniemi, H. Zhu, Secure and energy efficient resource allocation for wireless power enabled full-/half-duplex multiple-antenna relay systems. IEEE Trans. Veh. Technol. **65**(12), 11208–11219 (2017)
11. X. Liang, C. Xie, M. Min, W. Zhuang, User-centric view of unmanned aerial vehicle transmission against smart attacks. IEEE Trans. Veh. Technol. **67**(4), 3420–3430 (2017)
12. C. Li, S. Zhang, P. Liu, F. Sun, J. Cioffi, L. Yang, Overhearing protocol design exploiting inter-cell interference in cooperative greennetworks. IEEE Trans. Veh. Technol. **65**(1), 441–446 (2016)
13. C. Li, H. J. Yang, S. Fan, J. Cioffi, L. Yang, Multi-user overhearing for cooperative two-way multi-antenna relays. IEEE Trans. Veh. Technol. **65**(5), 3796–3802 (2016)
14. J. Xia, Secure cache-aided multi-relay networks in the presence of multiple eavesdroppers. IEEE Trans. Commun. **PP**(99), 1–10 (2019)
15. C. Zheng, L. Lei, H. Zhang, T. Ristaniemi, H. Zhu, Energy-efficient and secure resource allocation for multiple-antenna noma with wireless power transfer. IEEE Trans. Green Commun. Netw. **2**(4), 1059–1071 (2018)
16. Y. Li, L. Xiao, H. Dai, P. H. Vincent, in *IEEE Int. Conf. Commun.* Game theoretic study of protecting mimo transmissions against smart attacks, (2017), pp. 1–6
17. C. Li, Y. Xu, Protecting secure communication under UAV smart attack with imperfect channel estimation. IEEE Access. **6**(1), 76395–76401 (2018)
18. Y. Xu, Q-learning based physical-layer secure game against multi-agent attacks. IEEE Access. **7**, 49212–49222 (2019)
19. X. Liang, Y. Li, G. Han, H. Dai, H. V. Poor, A secure mobile crowdsensing game with deep reinforcement learning. IEEE Trans. Inf. Forensic. Secur. **13**(1), 35–47 (2018)
20. C. Li, S. Fan, J. M. Cioffi, L. Yang, Energy efficient mimo relay transmissions via joint power allocations. IEEE Trans. Circ. Syst. II Express Briefs. **61**(7), 531–535 (2014)
21. X. Liang, C. Xie, et al., A mobile offloading game against smart attacks. IEEE Access. **4**, 2281–2291 (2017)
22. C. Li, W. Zhou, Enhanced secure transmission against intelligent attacks. IEEE Access. **7**, 53596–53602 (2019)
23. X. Liang, T. Chen, C. Xie, H. Dai, V. Poor, Mobile crowdsensing games in vehicular networks. IEEE Trans. Veh. Technol. **67**(2), 1535–1545 (2018)
24. X. Liang, Y. Li, C. Dai, H. Dai, H. V. Poor, Reinforcement learning-based noma power allocation in the presence of smart jamming. IEEE Trans. Veh. Technol. **67**(4), 3377–3389 (2018)
25. A. G. Barto, Reinforcement learning. A Bradford Book. **15**(7), 665–685 (1998)
26. C. J. C. H. Watkins, P. Dayan, Technical note: Q-learning. Mach. Learn. **8**(3-4), 279–292 (1992)
27. X. Lin, MARL-based distributed cache placement for wireless networks. IEEE Access. **7**, 62606–62615 (2019)
28. J. Zhao, A dual-link soft handover scheme for C/U plane split network in high-speed railway. IEEE Access. **6**, 12473–12482 (2018)
29. H. Xie, F. Gao, S. Zhang, S. Jin, A unified transmission strategy for TDD/FDD massive MIMO systems with spatial basis expansion model. IEEE Trans. Veh. Technol. **66**(4), 3170–3184 (2017)
30. X. Lai, Distributed secure switch-and-stay combining over correlated fading channels. IEEE Trans. Inf. Forensic. Secur. **14**(8), 2088–2101 (2019)
31. Z. Na, Y. Wang, Subcarrier allocation based simultaneous wireless information and power transfer algorithm in 5g cooperative OFDM communication systems. Phys. Commun. **29**, 164–170 (2018)
32. C. E. Shannon, Communication theory of secrecy systems. Bell Syst. Tech. J. **28**, 656–715 (1948)
33. E. N. Barron, H. Ishii, The bellman equation for minimizing the maximum cost. Nonlinear Anal. Theory Methods Appl. **13**(9), 1067–1090 (1989)
34. Z. Na, J. Lv, M. Zhang, M. Xiong, GFDM based wireless powered communication for cooperative relay system. IEEE Access. **7**, 50971–50979 (2019)
35. X. Lai, W. Zou, DF relaying networks with randomly distributed interferers. IEEE Access. **5**, 18909–18917 (2017)
36. J. Zhao, J. Liu, Y. Nie, S. Ni, Location-assisted beam alignment for train-to-train communication in urban rail transit system. IEEE Access. **7**, 80133–80145 (2019)
37. J. Xia, Cache-aided mobile edge computing for b5g wireless communication networks. EURASIP J. Wirel. Commun. Netw. **PP**(99), 1–5 (2019)
38. J. Xia, When distributed switch-and-stay combining meets buffer in IoT relaying networks. Phys. Commun. **PP**, 1–9 (2019)
39. S. Lai, Intelligent secure communication for cognitive networks with multiple primary transmit power. IEEE Access. **PP**(99), 1–7 (2019)
40. J. Zhao, Q. Li, Y. Gong, K. Zhang, Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks. IEEE Trans. Veh. Technol. **68**(8), 7944–7956 (2019)
41. J. Yang, Inverse optimization of building thermal resistance and capacitance for minimizing air conditioning loads. Renew. Energy. **PP**, 1–10 (2020)
42. H. Huang, Optimum insulation thicknesses and energy conservation of building thermal insulation materials in chinese zone of humid subtropical climate. Renew. Energy. **52**, 101840 (2020)
43. J. Yang, Numerical and experimental study on the thermal performance of aerogel insulating panels for building energy efficiency. Renew. Energy. **138**, 445–457 (2019)
44. G. Liu, Deep learning based channel prediction for edge computing networks towards intelligent connected vehicles. IEEE Access. **7**, 114487–114495 (2019)
45. Z. Zhao, A novel framework of three-hierarchical offloading optimization for mec in industrial IoT networks. IEEE Trans. Ind. Inform. **PP**(99), 1–12 (2019)
46. J. Xia, Intelligent secure communication for internet of things with statistical channel state information of attacker. IEEE Access. **7**(1), 144481–144488 (2019)
47. K. He, A MIMO detector with deep learning in the presence of correlated interference. IEEE Trans. Veh. Technol. **PP**(99), 1–5 (2019)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.