Secure Routing in Multihop Ad-Hoc Networks With SRR-Based Reinforcement Learning

Jianzhong Lu, Dongxuan He, and Zhaocheng Wang, *Fellow, IEEE*

***Abstract*—In this letter, a reinforcement learning-assisted secure routing methodology is proposed for multihop ad-hoc networks in the presence of multiple eavesdroppers. Specifically, secure relay region (SRR) is firstly proposed, which depicts the distribution of the relays forwarding the information securely. Moreover, a SRR-based on-policy Monte Carlo methodology is derived, aiming at accelerating the convergence of routing. The secrecy connection probability is also calculated, which indicates the secure performance of different routes. Simulation results show that our proposed SRR-based reinforcement learning methodology can select the secure route efficiently and fast, which is also robust to the time-varying available relays.**

***Index Terms*—Reinforcement learning, secure routing, secure relay region, on-policy Monte Carlo.**

I. Introduction

DUE TO the absence of centralized administration, the communication security of ad-hoc networks could not be guaranteed [1]. Besides, conventional encryption techniques, which rely on key generation, distribution and management, are costly and hard to realize. To tackle this issue, physical layer security is emerging as an attractive security paradigm to prevent illegal interception by utilizing the characteristics of wireless channel [2].

Recently, physical layer security has been introduced into the multihop ad-hoc networks to enhance the transmission security. For instance, Wang *et al.* optimized the routing and transmit power in a multihop ad-hoc network to minimize the end-to-end connection outage probability under a secrecy outage probability constraint [1]. A revised Bellman-Ford algorithm was proposed to find the optimal path based on an upper bound approximation of secrecy connection probability (SCP) Chen *et al.* proposed an improved secure routing algorithm for a full-duplex receiver in the presence of random eavesdropper clusters [4]. These works show a satisfying routing performance in time-invariant wireless networks. However, they cannot realize the reliable routing in time-varying networks since a new routing search is required when the network changes, which is extremely troublesome and time-consuming.

Reinforcement learning, which is capable of perceiving and adapting to the change of network, has been regarded as a potential technology to realize real-time physical layer security. For example, an algorithm containing Q-learning and expert advice methods has been utilized to select the optimal transmission parameters against active eavesdroppers in dynamic environment [5]. Reference [6] proposed a hot-booting Q-learning to realize the power allocation scheme in dynamic nonorthogonal multiple access secure transmission. In [7], deep reinforcement learning was utilized to improve the system secrecy rate in the presence of multiple eavesdroppers under time-varying channels.

Moreover, reinforcement learning has also shown to be an appropriate solution for routing problems, especially in dynamic networks. A Q-learning based routing protocol was proposed to improve the energy efficiency in underwater sensor networks [8]. Based on the ON-policy Monte Carlo (ONMC) method, an energy-efficient path selection scheme was proposed to balance the network lifetime and energy consumption [9]. However, the performance of reinforcement learning based routing is limited when transmission security is considered, since slow convergence of reinforcement learning will induce high risk of information disclosure.

In this letter, we consider secure routing for a multihop ad-hoc network in the presence of homogeneous Poisson point process (PPP) distributed eavesdroppers, where SCP is utilized to evaluate the route secrecy. Secure relay region (SRR) is firstly proposed and a SRR-based ONMC-assisted learning algorithm is derived to facilitate accurate and adaptive secure routing. Our main contributions are summarized as follows: 1) SRR is newly proposed to depict the distribution of relays satisfying the secure transmission requirements, which provides prior knowledge to accelerate the convergence. 2) Based on our explicitly derived SCP, the SRR-based ONMC learning methodology is derived to select the optimal route maximizing the SCP, which only requires the statistical channel state information (CSI). Simulation results demonstrate that our proposed SRR-based ONMC-assisted routing methodology is capable of realizing secure routing adaptively in the time-varying networks. Benefit from SRR-based initialization, our proposed method converges faster when compared to its conventional counterparts [9].

Manuscript received October 1, 2021; revised November 9, 2021; accepted November 14, 2021. Date of publication November 17, 2021; date of current version February 17, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62101306, and in part by the Postdoctoral Science Foundation of China under Grant 2020M670332. The associate editor coordinating the review of this article and approving it for publication was X. Zhou. *(Corresponding author: Dongxuan He.)*

Jianzhong Lu and Zhaocheng Wang are with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: ljz19@mails.tsinghua.edu.cn; zcwang@tsinghua.edu.cn).

Dongxuan He is with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: dongxuan_he@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/LWC.2021.3128582

II. System Model

We consider a large-scale multihop wireless ad-hoc network consisting of multiple eavesdroppers and *M* legitimate nodes including a source, relays and a destination. The passive eavesdroppers Φ_e are randomly distributed following a homogeneous PPP with intensity λ_e, which is usually adopted to model the uncertainty of eavesdroppers' location [10], [11]. Particularly, both legitimate nodes and eavesdroppers are equipped with a single omnidirectional antenna. The source aims to send its confidential information to the destination through a secure multihop path. Assuming that there is a route containing *N* decode-and-forward (DF) relays, each hop can be denoted by $l_n, n = 1, \ldots, N+1$ with the transmitter and the receiver at the *n*-th hop denoted by T_n and R_n, respectively. As a result, the entire route can be denoted by $\Pi = \{l_1, l_2, \ldots, l_{N+1}\}$.

Specifically, all the wireless channels are assumed to be subjected to large-scale path loss along with small-scale Rayleigh fading [1], [3], [4]. The path loss exponent is denoted as α and each Rayleigh fading channel coefficient $h_{i,j}$ is modeled as independent complex Gaussian variable with zero mean and unit variance, i.e., $h_{i,j} \sim \mathcal{CN}(0,1)$, where $i \in \{T_1, T_2, \ldots, T_{N+1}\}$ and $j \in \{R_1, R_2, \ldots, R_{N+1}, \Phi_e\}$.

According to the DF protocol, the signal-to-noise ratio (SNR) of the path Π can be derived as[3]

$$SNR^{\Pi} = \min_{n=1,\ldots,N+1} \frac{P_{T_n}|h_{T_nR_n}|^2}{d^{\alpha}_{T_nR_n}\sigma^2_{R_n}}, \tag{1}$$

where P_{T_n} denotes the transmit power at T_n, $\sigma^2_{R_n}$ denotes the noise power at R_n following $\mathcal{CN}(0,\sigma^2)$ and $d_{T_nR_n}$ denotes the distance between T_n and R_n. Eavesdroppers are assumed to operate at colluding mode so that the SNR of the wiretap channels can be calculated by the maximal ratio combining protocol as [3]

$$SNR^{\Pi}_{E} = \sum_{E_i \in \Phi_e} \sum_{n=1}^{N+1} \frac{P_{T_n}|h_{T_nE_i}|^2}{d^{\alpha}_{T_nE_i}\sigma^2_{E_i}}, \tag{2}$$

where $\sigma^2_{E_i}$ denotes the noise power at eavesdropper E_i following $\mathcal{CN}(0,\sigma^2)$. To evaluate the secrecy performance, SCP is adopted as the performance metric [11], which can be derived as

$$\mathcal{P}^{\Pi}_{SCP} = \Pr\left\{\frac{\log_2(1+SNR_{\Pi}) - \log_2(1+SNR_E)}{N+1} > 0\right\}. \tag{3}$$

Substituting (1) and (2) into (3), we arrive at (4), shown at the bottom of the page. Without loss of generality, the transmit powers of all nodes are assumed to be the same[1] and (5), shown at the bottom of the page, can be obtained, where (r_j, θ_j) denotes the polar coordinate of T_j, the source locates at the origin point and *R* is the radius for homogeneous PPP distribution.

To maximize the secrecy performance, the route selection can be formulated as

$$\Pi^* = \arg\max_{\Pi \in \Psi} \mathcal{P}^{\Pi}_{SCP}, \tag{6}$$

where Ψ denotes the set of all feasible routes from the source to the destination.

III. Improved ONMC-Assisted Secure Routing

Since the conventional routing algorithms with statistical CSI, such as Bellman-Ford [14] and Dijkstra's algorithm [15], use the approximated SCP, their performance is limited. To meet this challenge, we propose an improved reinforcement leaning-assisted approach that is capable of properly selecting the optimal route in time-varying networks. Furthermore, SRR is introduced firstly to accelerate the convergence of learning.

A. ONMC-Assisted Routing

In this subsection, ONMC-assisted routing is presented, where the route selection is regarded as an episodic task [13] based on the explicit SCP. When compared to the conventional off-policy scheme, our proposed ONMC method could improve the learning speed significantly, which focuses on exploiting the previous experience.

As a kind of reinforcement learning, ONMC method can learn the optimal policy gradually, where the agent can make proper decisions based on the experience obtained in the past interactions. Specifically, in the *n*-th episode, the agent firstly generates a feasible route denoted by $\{s^n_0, a^n_0, s^n_1, a^n_1, \ldots, s^n_{T_n}\}$ according to the existing policy π, where $s^n_t \in \mathcal{S}$, $a^n_t \in \mathcal{A}$ denotes the state and the action at the *t*-th hop that reflect the transfer node and receiving node, respectively. $\mathcal{S}$ and $\mathcal{A}$ denote the state set and the action set consisting of serial numbers of transfer nodes and receiving nodes in the route. The SCP of the selected route can be calculated based on (5) and the state-action value $Q(s,a)$ of passing nodes pair (s,a) can be updated by

$$Q(s,a) \leftarrow Q(s,a) + \frac{1}{C(s,a)}(R^n - Q(s,a)), \tag{7}$$

[1]The identical relays are assumed to be deployed in the networks, where their transmit powers are the same [3], [4].

$$\mathcal{P}^{\Pi}_{SCP} = \exp\left[-\lambda_e \int_{\mathbb{R}^2} \left(1 - \prod_{j=1}^{N+1} \frac{1}{1 + \frac{P_{T_j}}{d^{\alpha}_{T_jE_i}} \sum_{n=1}^{N+1} \frac{d^{\alpha}_{T_nR_n}}{P_{T_n}}}\right) dx_{E_i}\right] \tag{4}$$

$$\mathcal{P}^{\Pi}_{SCP} = \exp\left\{-\lambda_e \int_0^{2\pi} \int_0^{R} \left[1 - \prod_{j=1}^{N+1} \frac{1}{1 + \left(r^2 + r_j^2 - 2rr_j\cos(\theta - \theta_j)\right)^{-\frac{\alpha}{2}} \sum_{n=1}^{N+1} d^{\alpha}_{T_nR_n}}\right] rdrd\theta\right\} \tag{5}$$

where $C(s,a)$ denotes the number of executed transmissions from node s to node a, and R^n denotes the SCP of the selected route. Hereby, the discount factor $\frac{1}{C(s,a)}$ makes $Q(s,a)$ converge to the average of the rewards, which is helpful to improve the computational efficiency [13]. To make a good tradeoff between environment exploration and experience exploitation, ϵ-soft greedy policy is adopted based on $Q(s,a)$ [13], and we have

$$\Pr(a_t^n = a) = \begin{cases} 1-\epsilon+\epsilon/|\mathcal{A}|, & a = A_t^* \\ \epsilon/|\mathcal{A}|, & a \neq A_t^* \end{cases}, \quad (8)$$

where $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$, ϵ is a constant that scales the probability distribution of actions and A_t^* denotes the action maximizing the $Q-values$ when the state is s_t^n.

With the help of the ONMC-assisted routing, secure route with maximal SCP could be obtained even in time-varying ad-hoc networks. However, due to huge dimension of the state-action pairs, the convergence of such on-policy methodology is time-consuming. To guarantee a satisfied performance with fast convergence, the SRR is proposed firstly to facilitate accurate and adaptive RL based secure routing.

B. Secure Relay Region

In this subsection, to accelerate the convergence of our adopted ONMC-assisted routing, SRR is introduced to depict the distribution of the relays forwarding the information securely, which is defined as the geometry region where the expectation of probability that the nodes to be selected exceeds a threshold $\varepsilon \in [0,1)$, given by

$$\mathcal{R}_{srr}^{(\varepsilon)} \triangleq \left\{(r_r,\theta_r) : \mathbf{E}_{(r_r,\theta_r)} \geq \varepsilon, \forall (r_r,\theta_r) \in \mathbb{R}^2\right\}, \quad (9)$$

where (r_r,θ_r) denotes the polar coordinate of relay, $\mathbb{R}^2$ denotes the rectangle region of relay deployment with source and destination locating at the diagonal points, and $\mathbf{E}_{(r_r,\theta_r)}$ is defined as

$$\mathbf{E}_{(r_r,\theta_r)} = \sum_{n=3}^{\infty} \Pr\{M=n\}\Pr\{(r_r,\theta_r) \in S | M=n, \mathcal{P}_{SCP}^S \geq \eta\}, \quad (10)$$

where S is a route consisting of the location set of relays and η is a threshold, namely the accepted SCP of the selected route. Particularly, M is larger than 3 since at least one relay is required to forward the information, owing to the lack of direct link from source to destination. Without loss of generality, the location of relays is assumed to follow a homogeneous PPP distribution with intensity λ_r [11]. Therefore, the number of relays falling within $\mathbb{R}^2$ is subject to Poisson distribution with mean $\lambda_r A$, where A is the area of $\mathbb{R}^2$ and each relay is randomly and uniformly distributed over the $\mathbb{R}^2$ [16].

To properly set an acceptable threshold of SCP in (10) when the location and numbers of relays are uncertain, the upper bound of SCP is provided as a guideline.

Proposition 1: The SCP for N relays multihop network is upper bounded by

$$\mathcal{P}_{SCP} \leq \exp\left[-\frac{\pi\lambda_e\Gamma\left(1-\frac{2}{\alpha}\right)\Gamma\left(\frac{2}{\alpha}+N+1\right)D^2}{\Gamma(N+1)(N+1)^{\frac{2(\alpha-1)}{\alpha}}}\right], \quad (11)$$

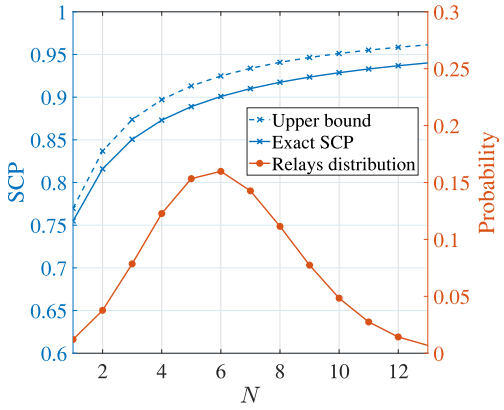

Fig. 1. Exact SCP and its upper bound versus numbers of relays (left axis), and the probability of relays distribution (right axis), where $\alpha = 4$, $R = 400m$, $\lambda_r = 2.5 \times 10^{-3}$, $\lambda_e = 2.0 \times 10^{-5}$ and $D = 50\sqrt{2}m$.

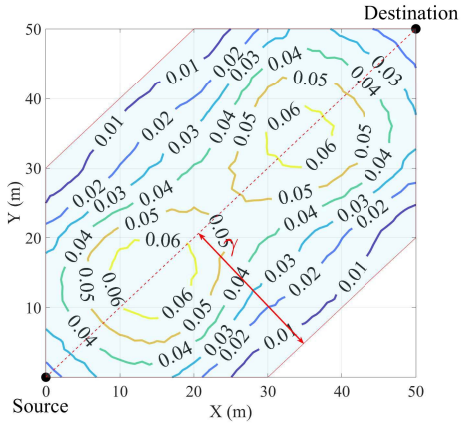

Fig. 2. Secure relay region in a 50×50 m^2 area, where $\lambda_r = 2.5 \times 10^{-3}$, $\lambda_e = 2.0 \times 10^{-5}$ and $\eta = 0.77$.

where D is the distance between source and destination, $\Gamma(\cdot)$ is the gamma function, and the equality holds when these N relays are uniformly distributed along the straight line.

Proof: Please refer to the Appendix. ■

Based on Proposition 1, the upper bound of SCP could be obtained, as shown in Fig. 1, where $\alpha = 4$, $R = 400m$, $\lambda_r = 2.5 \times 10^{-3}$, $\lambda_e = 2.0 \times 10^{-5}$ and $D = 50\sqrt{2}m$. It can be observed that the gap between the upper bound and the exact SCP is small, indicating that the approximation could be taken as a guide for η setting. In addition, the probability of relays distribution is also illustrated in Fig. 1, which shows that the probability that 12 relays locate in the network is as small as 0.01, and can be ignored actually.

For the numerical analysis of SRR, η is set to be 0.77, which is close to the upper bound when $N = 1$, and the maximum N is set to be 12. The numerical result has been presented in Fig. 2, which shows that the SRRs symmetrically distribute around the straight line from source to destination and the area of SRR decreases as the increase of ε. It can also be seen that the expectation value relies on the location of relays, and those nodes with higher $\mathbf{E}_{(r_r,\theta_r)}$ locate closely to the diagonal. Such observations could be used to guide the initialization of our adopted ONMC-assisted secure routing methodology to explore paths with relatively higher SCP.

Algorithm 1 Proposed SRR-Based ONMC-Assisted Secure Routing Algorithm

Set suboptimal SRR with $\gamma = \frac{D}{4}$.
Initial $Q(\mathcal{S}, \mathcal{A})$ according to suboptimal SRR.
Set small $\epsilon > 0$, generate a ϵ-soft policy π, $C(\mathcal{S}, \mathcal{A}) = 0$.
for episode $n = 1, 2, \ldots, I$ **do**
  if n = 1 **then**
    Generate a route randomly: $\left\{s_0^1, a_0^1, ..., a_{T-1}^1, s_T^1\right\}$.
  else
    Generate a route according to policy π: $\left\{s_0^n, a_0^n, s_1^n, a_1^n, ..., s_{T-1}^n, a_{T-1}^n, s_T^n\right\}$.
  end if
  for each state-action pair (s^n, a^n) in episode n **do**
    $C(s^n, a^n) = C(s^n, a^n) + 1$.
    Update $Q(s^n, a^n)$ via (7).
  end for
  $A_t^* = \text{argmax}_{a \in \mathcal{A}}\, Q(s_t^n, a)$.
  $\pi(a|s^n) = \begin{cases} 1 - \epsilon + \epsilon/|\mathcal{A}|, & a = A_t^* \\ \epsilon/|\mathcal{A}|, & a \neq A_t^* \end{cases}$.
end for

C. Proposed SRR-Based ONMC-Assisted Secure Routing

Typically, the conventional initializations like all-zero and all optimistic (all-opt) are widely used. However, all-zero schemes may lead to large bias but converge fast, and all-opt schemes may lead to slow convergence but high accuracy. To address this issue, SRR-based initialization is introduced whereby the values of those nodes in the SRR are initialized to be optimistic while others are initialized to be zero.

Nevertheless, the exact SRR is not possible due to the limitation of calculation ability in the source. As shown in Fig. 2, a relay deployed closely to the diagonal tends to be selected in the optimal route, which could be an intuitive guideline to estimate the SRR. Afterwards, a modified suboptimal SRR is proposed, which is a strap region that distributes symmetrically around the diagonal with width 2γ, as the blue shadow area depicted in Fig. 2.

According to [17], the computational complexity of ONMC-assisted routing grows with the total number of the convergence steps to the optimal policy. Let K be the number of steps per episode, and Z be the number of the episodes. The complexity of our proposed algorithm is $\mathcal{O}(KZ)$. Specifically, empirical works have suggested that the prior experience-based initialization may reduce the useless random explorations and the learning samples size, which could accelerate the learning speed [18], [19]. Therefore, the SRR is introduced into the initialization to improve the convergence rate. To take full advantage of SRR, the value of γ needs to be properly selected, where large γ may lead to all-opt scheme and small γ may lead to all-zero scheme. When some but not all relays are located in the selected SRR, the SRR-based initialization could explore these relays frequently and reduce the useless explorations, which guarantees the fast convergence of our routing algorithm.

The detail of our proposed SRR-based ONMC-assisted secure routing algorithm is summarized in Algorithm 1, where $\gamma = \frac{D}{4}$ is set.

IV. Simulation Results

In this section, numerical results are presented to illustrate the effectiveness of the proposed methodology. Similar parameters as [3] are adopted, where $M = 12$ and legitimate nodes are deployed in a 50×50 m^2 square area, the source and destination are located at (0, 0), $(50\sqrt{2}, 0.25\pi)$ while the eavesdroppers are randomly distributed in a disk region around the source with $R = 400m$. The density of eavesdroppers is $\lambda_e = 2.0 \times 10^{-5}$, the path loss exponent $\alpha = 4.0$, $\epsilon = 0.1$ and $D = 50\sqrt{2}m$.

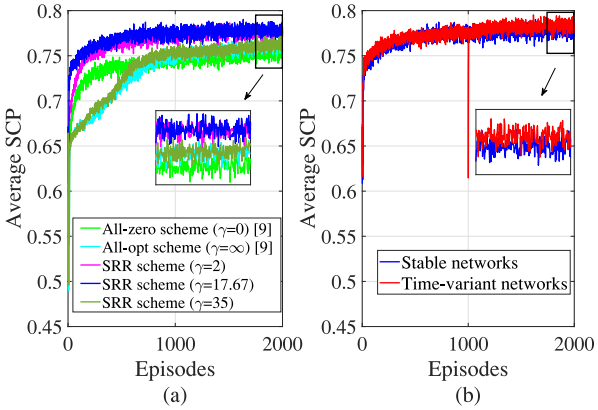

Fig. 3. (a) Average SCP over five different initial schemes (all-zero, all-opt and suboptimal SRR with $\gamma = 2m, 17.67m$, and $35m$). (b) Average SCP over stable networks with $M = 12$ and time-varying networks that deploying a new node in suboptimal SRR at the 1000-th episode.

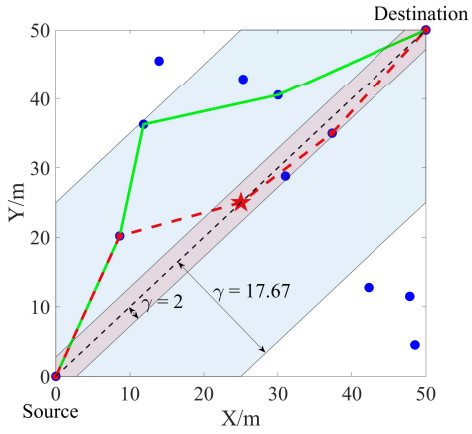

Fig. 4. Proposed SRR-based ONMC-assisted secure routing algorithm in time-varying networks (The green solid line depicts the selected route when $M = 12$, the star represents the newly deployed relay and the red dashed line represents the new selected route).

To investigate the impact of state-action values of initialization on routing, Fig. 3(a) is presented to illustrate the average SCP of five kinds of initial schemes. As the source and destination locate at the diagonal points of a 50×50 m^2 square area, we set $\gamma = 2m, 17.67m$ and $35m$, respectively. It can be seen that our proposed SRR initialization with proper γ outperforms two other traditional initial schemes, which achieves a faster convergence and higher average SCP. For the suboptimal SRR with a tiny $\gamma = 2m$, our proposed methodology performs well since some relays are still deployed inside, as shown in Fig. 4. Besides, a huge $\gamma = 35m$ makes SRR cover all relays, leading to the all-opt scheme. Furthermore, the suboptimal SRR with $\gamma = 17.67m$ achieves a 0.02 higher average SCP compared to the all-opt scheme and all-zero scheme [9].

To verify the adaptation of our proposed SRR-based secure routing, the average SCPs over stable and time-varying

networks are shown in Fig. 3(b). The newly deployed relay in the SRR increases the amounts of paths, which results in the drastic fluctuations and it is evident that our proposed methodology could re-converge fast and obtain a new route with increased SCP performance. Fig. 4 depicts the selected routes in a time-varying network. It is illustrated that when a new relay is deployed, the selected route changes correctly for achieving better secure performance with the aid of SRR.

V. Conclusion

In this letter, an improved reinforcement learning for multihop ad-hoc networks in the presence of homogeneous PPP distributed eavesdroppers was proposed. Specifically, SRR initialization was introduced into ONMC-assisted secure routing methodology in order to accelerate the learning convergence and provide the potential region of relay deployment. Simulation results validate that our proposed SRR-based secure routing methodology could select the optimal path maximizing secure connection probability in a time-varying dynamic network, and converge faster in comparison to the state-of-the-art methodologies.

Appendix
Proof of Proposition 1

According to the approximation of SCP for colluding eavesdroppers case in [3], we have

$$\mathcal{P}_{SCP}^{appro} = \exp\left[-K(N+1)\left(\sum_{i=1}^{N+1} d_{T_nR_n}^{\alpha}\right)^{\frac{2}{\alpha}}\right], \quad (12)$$

where $K(N+1) = \frac{\lambda_e \pi \Gamma(1-\frac{2}{\alpha})\Gamma(\frac{2}{\alpha}+N+1)}{\Gamma(N+1)}$. To obtain the upper bound of SCP and the optimal location when N relays are deployed, the SCP maximization problem can be expressed by

$$\max_{d_n} \ \exp\left[-K(N+1)\left(\sum_{i=1}^{N+1} d_n^{\alpha}\right)^{\frac{2}{\alpha}}\right]$$
$$\text{s.t.} \ -\sum_{n=1}^{N+1} d_n \le -D, \quad -d_n \le 0, \forall n \in [1, N+1], \quad (13)$$

where d_n denotes $d_{T_nR_n}$ for simplicity, and (13) is equivalent to

$$\min_{d_n} \ \sum_{n=1}^{N+1} d_n^{\alpha}. \quad (14)$$

The objective function (14) is strictly convex, which is the sum of convex power functions with exponent $\alpha \ge 2$. The Lagrangian function is defined as

$$L(\mathbf{d}, \lambda) = \sum_{n=1}^{N+1} d_n^{\alpha} + \lambda_0\left(D - \sum_{n=1}^{N+1} d_n\right) - \sum_{n=1}^{N+1} \lambda_n d_n, \quad (15)$$

where $\lambda_n \ge 0, 0 \le n \le N+1$ are the Lagrange multipliers and $\mathbf{d} = (d_1, d_2, \ldots, d_{N+1})$. The Karush-Kuhn-Tucker conditions for the optimal solution of (14) are given by

$$\frac{\partial L(\mathbf{d}, \lambda)}{\partial d_n} = 0, -d_n \le 0, -\lambda_n d_n = 0 \ \forall n \in [1, N+1], \quad (16)$$

$$D - \sum_{n=1}^{N+1} d_n \le 0, \lambda_0\left(D - \sum_{n=1}^{N+1} d_n\right) = 0, \quad (17)$$

$$\lambda_n \ge 0, \ \forall n \in [0, N+1]. \quad (18)$$

Combining (16), (17) and (18), we have $d_n = \frac{D}{N+1}, n \in [1, N+1]$, and the minimum of the objective function is $\frac{D^{\alpha}}{(N+1)^{\alpha-1}}$. The proof is completed.

References

[1] H. Wang, Y. Zhang, D. W. K. Ng, and M. H. Lee, "Secure routing with power optimization for ad-hoc networks," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4666–4679, Oct. 2018.

[2] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.* vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[3] J. Yao, S. Feng, X. Zhou, and Y. Liu, "Secure routing in multihop wireless ad-hoc networks with decode-and-forward relaying," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 753–764, Feb. 2016.

[4] G. Chen, J. P. Coon, and S. E. Tajbakhsh, "Secure routing for multihop ad hoc networks with inhomogeneous eavesdropper clusters," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10660–10670, Nov. 2018.

[5] D. He, H. Wang, and H. Zhou, "Learning-based secure communication against active eavesdropper in dynamic environment," *IET Commun.*, vol. 13, no. 15, pp. 2235–2242, 2019.

[6] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[7] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[8] T. Hu and Y. Fei, "QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks," *IEEE Trans. Mobile Comput.*, vol. 9, no. 6, pp. 796–809, Jun. 2010.

[9] W. Naruephiphat and W. Usaha, "Balancing tradeoffs for energy-efficient routing in MANETs based on reinforcement learning," in *Proc. IEEE Veh. Technol. Conf.*, May 2008, pp. 2361–2365.

[10] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic geometry and random graphs for the analysis and design of wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 7, pp. 1029–1046, Sep. 2009.

[11] X. Zhou, R. K. Ganti, J. G. Andrews, and A. Hjorungnes, "On the throughput cost of physical layer security in decentralized wireless networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2764–2775, Aug. 2011.

[12] X. Zhou, R. K. Ganti, and J. G. Andrews, "Secure wireless network connectivity with multi-antenna transmission," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 425–430, Feb. 2011.

[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[14] J. Y. Yen, "An algorithm for finding shortest routes from all source nodesto a given destination in general networks," *Quart. Appl. Math.*, vol. 27, no. 4, pp. 526–530, 1970.

[15] D. E. Knuth, "A generalization of Dijkstra's algorithm," *Inf. Process. Lett.*, vol. 6, no. 1, pp. 1–5, 1997.

[16] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications*. Hoboken, NJ, USA: Wiley, 2013.

[17] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, Dec. 2018, pp. 4863–4873.

[18] X. Xiao, C. Dai, Y. Li, C. Zhou, and L. Xiao, "Energy trading game for microgrids using reinforcement learning," in *Proc. EAI Int. Conf. Game Theory Netw.*, Knoxville, TN, USA, May 2017, pp. 131–140.

[19] L. Xiao *et al.*, "Reinforcement learning-based downlink interference control for ultra-dense small cells," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 423–434, Jan. 2020.