A DRL-Based Hierarchical Game for Physical Layer Security Aware Cooperative Communications

Denghui Liu, Ruoyang Chen, Tong Zhang, and Changyan Yi

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics,
Nanjing, China
{liudeghui,ruoyangchen,zhangt,changyan.yi}@nuaa.edu.cn

Abstract. This paper investigates a dynamic game framework for physical layer security (PLS)-aware wireless network with third-party collaborative users (TCUs). In the considered system, TCUs choose to assist legitimate users (LUs) in resisting eavesdropping attacks or opt to assist eavesdroppers (EVs) in eavesdropping on legitimate communication, in exchange for rewards from LUs and EVs. Due to the unpredictability of wireless systems, coalitions among the three parties may be dynamically changing. A hierarchical game integrating a matching sub-game with a coalition formation sub-game is formulated to model the interactions among LUs, EVs, and TCUs. In order to derive the optimal long-term strategies while solving the equilibrium of the hierarchical game, we develop a deep reinforcement learning (DRL)-based solution, which includes a matching process and a coalition selection process for two subgames. Simulation evaluates the performance of the proposed scheme and demonstrates its superiority over counterparts.

Keywords: Physical layer security · Cooperative communication · Hierarchical game · Deep reinforcement learning.

1 Introduction

Cooperative communication enhances wireless performance through collaboration among devices, often by optimizing network systems via signal relaying and shared antennas. However, wireless signals' broadcast nature exposes legitimate users (LUs) to threats from eavesdroppers (EVs), while physical layer security (PLS) technology can utilize the physical characteristics of wireless communication to secure LU's communication [3]. Due to the characteristics of wide coverage, a large number of access users, and substantial data transfer volume in cooperative communication, it is accompanied by potential transmission security issues. As the number of nodes in the cooperative wireless network increases, PLS serves as a powerful method to improve the secrecy transmission rates of the target devices through the appropriate cooperative schemes [3], [9].

In cooperative network, devices who share antennas resources or participate in relaying are referred to as collaborative users (CUs). Due to signal interference on the channels, CUs also can be jammers. Specifically, they can assist LUs by relaying signals and jamming EVs with artificial noise, enhancing the secrecy of LU transmissions [3]. Meanwhile, CUs also can assist EVs by eavesdropping on LUs and transmitting interference signals, improving EV eavesdropping efficiency [8]. As a viable PLS technology, current researches assume that CUs are dominated by either LUs or EVs, without considering self-centered manner of CUs. Such CUs are referred to as third-party collaborative users (TCUs) and thus they will have a tendency to choose the appropriate ally that can bring higher profits [10]. Since either LUs or EVs may be self-centric devices in real-world scenarios, LUs and EVs need to balance the incentives they offer to TCUs with performance increment brought by TCUs. Therefore, LUs and EVs will choose to ally with TCUs only when the assistance provided by TCUs is actually beneficial to their performances, and such selfish behavior consequently results in unfixed alliance relationship among LUs, EVs and TCUs in a network.

In order to provide assistance to LUs and EVs in exchange for higher profits, self-centric TCUs need to jointly determine their operation modes and ally selection. This implies that TCUs have a double-leveled role selection. At the same time, the strategy of any party in LUs, EVs and TCUs will have an impact on that of the others, leading to fluctuations in alliances relationships of the three parties, i.e., changeable coalition partitions among LUs, EVs and TCUs. This need to be carefully modeled and analyzed, which are important and very challenging due to the following reasons.

1. When determining the mode selection of TCUs, their power allocation and appropriate channel selection to avoid co-channel interference also need to be determined. LUs and EVs will pay different incentives based on the different strategies of decision tuple consisting of mode switching, power allocation and channel selection. Therefore, when TCUs ally with LUs or EVs, TCUs and their allies will have their own preferences for the possible values of this decision tuple, which incurs a matching relationships between TCUs and their allies. We employ a double-sided matching game to model the negotiation between TCUs and their allies regarding this decision tuple.
2. LUs, EVs, and TCUs are all self-centric devices, i.e., they always choose the appropriate ally that can maximize their own benefits with minimum cost. The incentives provided by LUs and EVs affect TCUs' choice of ally, and TCUs' cooperative strategies also influence the alliance intentions of LUs and EVs, which further leads to a unfixed coalition relationship among the three parties. The coalition formation among LUs, EVs and TCUs is mutually determined by each of them, which motivates us to employ a coalition formation game to model and analyze the unfixed coalition formation process.
3. Due to the uncertainty in wireless cooperative networks, i.e., time-varying channel conditions [6], the double-leveled role selection of TCUs can dynamically evolve, and thus results in dynamic matching and coalition formation process. LUs, EVs, and TCUs will react to the decisions of each other in the system and the strategies of the three parties have a sequential nature [1]. We construct a hierarchical game, which includes a matching subgame and a coalition formation subgame, to model the long-term strategic interactions among the three parties and optimize their long-term performance.

To address the above challenges, we respectively quantifies the utility functions of LUs, EVs and TCUs, and model their long-term performance optimization problems. Then, a novel hierarchical game framework with a matching subgame and a coalition formation subgame is proposed for modeling the dynamic decision-making process among LUs, EVs and TCUs. In each time slot, the matching subgame models negotiation between allied parties on the decision tuple, and the coalition formation subgame models unfixed relationship on the alliances and confrontations.We propose a deep reinforcement learning (DRL) based approach for the long-term decision optimization problems of three parties, which includes a constrained matching process and a coalition selection process for two subgames. The main contributions of this paper are summarized in the following.

1. We proposed a hierarchical game framework to model the dynamic interactions among LUs, EVs, and TCUs in a PLS-aware cooperative network, and achieve long-term strategy optimization for the three parties while considering the uncertainty of wireless systems.

2. We proposed a double-sided matching subgame to model the matching relationship between TCUs and their allies (i.e., LUs or EVs) and a coalition formation subgame for modeling the dynamically evolving alliance relationship among LUs, EVs and TCUs. We designed a DRL based approach which includes a constrained matching process and a coalition selection process for two subgames.
3. Through numerical simulations, our proposed solution, compared to existing benchmarks, e.g., non-changeable modes scheme and fixed coalition relationship scheme, can converge quickly achieve higher long-term performance in terms of the respective cumulative utilities of LUs, EVs and TCUs.

The rest of this paper is organized as follows: Section 2 introduces the network model and the long-term optimization problems of the three parties in PLS-aware cooperative communications. In Section 3, a hierarchical game with the matching subgame and the coalition formation subgame is formulated and a DRL based approach is proposed. Simulation results are given in Section 4, followed by the conclusion in Section 5.

2 System Model and Problem Formulation

2.1 Network Model

Fig.1 illustrates a PLS-aware wireless cooperative communication system that consists of LUs, EVs and TCUs. In this system, the transmission pairs of LUs, denoted as $\mathcal{L} = \{1, 2, \ldots, L\}$ transmit confidential information, and each pair $l \in \mathcal{L}$ consists of a transmitter LU_l^T and a receiver LU_l^R. Proactive EVs can intercept confidential information and interfere with transmission of all LUs, denoted as $\mathcal{M} = \{1, 2, \ldots, M\}$. TCUs are denoted as $\mathcal{N} = \{1, 2, \ldots, N\}$, and each TCU can choose to operate in relaying or jamming mode, and a decode-and-forward approach is employed. They can choose to collectively secure the communication of LUs, assist in active eavesdropping for EVs, or not to participate in cooperation. In this system, the devices operate in full-duplex mode and employ orthogonal channels to avoid interference, which makes the number of sub-channels be equal to the number of pairs of LUs and different TCUs can only access different sub-channels.

When TCUs choose not to participate in any cooperation, the signal to interference plus noise ratio (SINR) of the pair l can be expressed as

$$\gamma_l^{(0)}(t) = \frac{P_L |h_l(t)|^2}{\sigma^2}, \tag{1}$$

where $t \in \{0, 1, \ldots, T-1\}$ denotes the index of time slots, P_L is the transmit power of LU transmitters, $|h_l(t)|^2$ is the channel gain of this pair, and there is zero-mean additive white Gaussian noise with variance σ^2 at each receiver. The SINR of the m's eavesdropping link to pair l in time slot t can be written as

$$\gamma_{ml}^{(0)}(t) = \frac{P_L |h_{lm}(t)|^2}{\sigma^2}, \tag{2}$$

where $|h_{lm}(t)|^2$ is the channel gain of the eavesdropping link in time slot t.

When considering the assistance of TCUs on LUs' secrecy transmission, in cooperative jamming mode, the SINR of m's eavesdropping link to the pair $l \in \mathcal{L}$ in time slot t will be reduced and

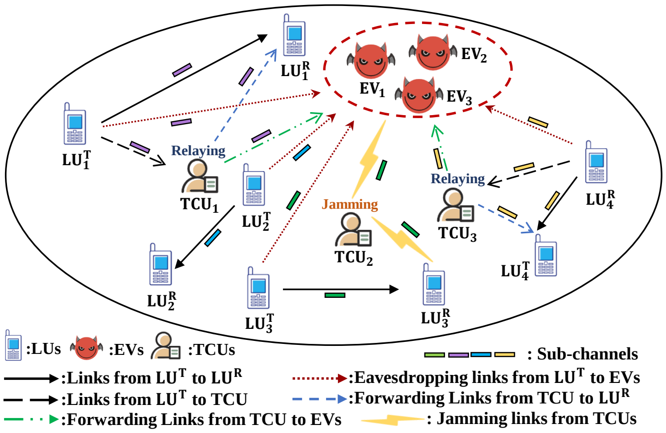
LU_1^R
EV_1
EV_2
EV_3
LU_1^T
Relaying
TCU_1
LU_2^T
Jamming
Relaying
TCU_3
LU_4^R
TCU_2
LU_2^R
LU_3^T
LU_3^R
LU_4^T
:LUs
:EVs
:TCUs
: Sub-channels
:Links from LU^T to LU^R
:Eavesdropping links from LU^T to EVs
:Links from LU^T to TCU
:Forwarding Links from TCU to LU^R
:Forwarding Links from TCU to EVs
: Jamming links from TCUs

Fig. 1. PLS-aware cooperative communication system.

calculated as

$$\gamma_{ml}^{J}(t) = \frac{P_L|h_{lm}(t)|^2}{\sigma^2 + P_n(t)|h_{nm}(t)|^2}, \tag{3}$$

where $|h_{lm}(t)|^2$ and $|h_{nm}(t)|^2$ are the channel gain of m's eavesdropping links to l and n, and $P_n(t)$ is the transmit power of TCU $n \in \mathcal{N}$. In relaying mode, the SINR of the linkof pair $l \in \mathcal{L}$ in time slot t will be elevated and obtained as

$$\gamma_{l}^{R}(t) = \frac{P_L|h_l(t)|^2}{\sigma^2} + min\left(\frac{P_L|h_{ln}(t)|^2}{\sigma^2}, \frac{P_n(t)|h_{nl}(t)|^2}{\sigma^2}\right), \tag{4}$$

where $|h_{ln}(t)|^2$ and $|h_{nl}(t)|^2$ are the channel gains of the link from l to n and the link from n to l in time slot t.

When considering TCUs assisting EVs in active eavesdropping, in jamming mode, the SINR of the link of the pair $l \in \mathcal{L}$ will be lower and written as

$$\gamma_{l}^{J}(t) = \frac{P_L|h_l(t)|^2}{\sigma^2 + P_n(t)|h_{nl}|^2}, \tag{5}$$

In the relaying mode, the SINR of m's eavesdropping link to the pair l will be calculated as

$$\gamma_{ml}^{R}(t) = \frac{P_L|h_{lm}(t)|^2}{\sigma^2} + min\left(\frac{P_L|h_{ln}(t)|^2}{\sigma^2}, \frac{P_n(t)|h_{nm}(t)|^2}{\sigma^2}\right). \tag{6}$$

We use symbol $\theta_n(t) \in \{0, 1\}$ to represent the switching of TCUs' modes in time slot t, i.e., $\theta_n(t) = 1$ if TCU n operates in jamming mode; $\theta_n(t) = 0$ if it operates in relaying mode. The sub-channel selections of TCUs in time slot t denoted as $\omega_{nl}(t) \in \{0, 1\}$, i.e., $\omega_{nl}(t) = 1$ if TCUs choose to access the sub-channel of the pair $l \in \mathcal{L}$; otherwise, $\omega_{nl}(t) = 0$. The formation of alliance between LUs and TCUs in time slot t is represented by $col_{\mathcal{NL}}(t) \in \{0, 1\}$, i.e., $col_{\mathcal{NL}}(t) = 1$ if

LUs $\mathcal{L}$ and TCUs $\mathcal{N}$ form an alliance; otherwise, $col_{\mathcal{NL}}(t) = 0$. Similarly, the formation of alliance between EVs and TCUs is represented by $col_{\mathcal{NM}}(t) \in \{0, 1\}$. According to equation (1), (4) and (5), the SINR of the link of pair $l \in \mathcal{L}$ with the help of TCUs can be expressed as

$$\begin{aligned}\gamma_l(t) = {} & col_{\mathcal{NL}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)\theta_n(t)\gamma_l^{(0)}(t) + col_{\mathcal{NL}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)(1-\theta_n(t))\gamma_l^R(t) \\ & + col_{\mathcal{NM}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)\theta_n(t)\gamma_l^J(t) + (1 - \sum_{n\in\mathcal{N}} \omega_{nl}(t))\gamma_l^{(0)}(t).\end{aligned} \tag{7}$$

According to equation (2), (3) and (6), the SINR of m's eavesdropping link to the pair l can be obtained as

$$\begin{aligned}\gamma_{ml}(t) = {} & col_{\mathcal{NM}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)\theta_n(t)\gamma_{ml}^{(0)}(t) + col_{\mathcal{NM}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)(1-\theta_n(t))\gamma_{ml}^R(t) \\ & + col_{\mathcal{NL}}(t) \sum_{n\in\mathcal{N}} \omega_{nl}(t)\theta_n(t)\gamma_{ml}^J(t) + (1 - \sum_{n\in\mathcal{N}} \omega_{nl}(t))\gamma_{ml}^{(0)}(t).\end{aligned} \tag{8}$$

where $\omega_{nl}(t)$ indicates whether TCU $n \in \mathcal{N}$ eavesdrops the pair of LUs $l \in \mathcal{L}$, specifically.

2.2 Problem Formulation

For LUs, in order to enhance their long-term secure transmission performance, in each time slot t, they need to determine i) the TCU association for higher security, expressed as $\varepsilon_{\mathcal{L}}(t) \in \{0, 1\}$, i.e., $\varepsilon_{\mathcal{L}}(t) = 1$ if LUs wish to form an alliance with TCUs; otherwise, $\varepsilon_{\mathcal{L}}(t) = 0$; and ii) the unit incentive $\eta_{\mathcal{L}}(t) \in [0, \eta_{\mathcal{L}}^{max}]$ to attract TCUs' assistance. The secrecy rate [11] is the maximum achievable rate for transmitting the signals which EVs cannot perfectly intercept. And thus the secrecy rate of the pair l can be calculated as $R_l(t) = [Wlog_2(1 + \gamma_l(t)) - max_{m\in\mathcal{M}} Wlog_2(1 + \gamma_{ml}(t))]^+$, where function $[x]^+ = max(x, 0)$. The secrecy rate without the TCUs can be written as $R_l^{(0)}(t) = [Wlog_2(1 + \gamma_l^{(0)}(t)) - max_{m\in\mathcal{M}} Wlog_2(1 + \gamma_{ml}^{(0)}(t))]^+$. Then, the utility function of the pair of LUs $l \in \mathcal{L}$ in time slot t can be calculated as $r_l(t) = R_l(t) - \eta_{\mathcal{L}}(t)\varepsilon_{\mathcal{L}}(t)\varepsilon_{NL}(t)[R_l(t) - R_l^{(0)}(t)]^+$, where $\varepsilon_{NL}(t) \in \{0, 1\}$ indicates whether TCUs agree to ally with LUs in time slot t. Denoting the strategy of all pairs of LUs as $\pi_{\mathcal{L}} = \{\varepsilon_{\mathcal{L}}(t), \eta_{\mathcal{L}}(t)\}_{\forall t}$, the long-term optimization problem for LUs can be formulated as

$$[\mathcal{LP}] : \max_{\pi_{\mathcal{L}}} \frac{1}{T} \sum_{t\in[0,T]} \sum_{l\in\mathcal{L}} r_l(t) \tag{9}$$

$$s.t., C_l(t) \geqslant C_{min}, \forall l \in \mathcal{L}, \tag{9b}$$

$$0 \leqslant \eta_{\mathcal{L}}(t) \leqslant \eta_{\mathcal{L}}^{max}, \tag{9b}$$

$$\varepsilon_{\mathcal{L}}(t) \in \{0, 1\}, \tag{9c}$$

where constraint (9a) states that the transmission rate of each LU transmitter in any time slot t should not be smaller than a minimum value for guaranteeing its quality of service.

For EVs, in order to enhance their long-term eavesdropping performance, in each time slot t, they need to determine $\varepsilon_{\mathcal{M}}(t) \in \{0, 1\}$. They also need to provide their unit incentive $\eta_{\mathcal{M}}(t) \in [0, \eta_{\mathcal{M}}^{max}]$ to attract TCUs' assistance. We measure the eavesdropping performance of EVs by calculating the successful intercepting rate which is specifically defined as the difference of the maximum eavesdropping rate and the transmission rate of LUs. With the participation of TCUs, EVs' maximum successful intercepting rate of eavesdropping the LU transmitter of pair $l \in \mathcal{L}$ can be written as $I_l(t) = max_{m \in \mathcal{M}} C_{ml}(t) - C_l(t)$. Otherwise, this maximum successful intercepting rate without TCUs can be represented as $I_l^{(0)}(t) = max_{m \in \mathcal{M}} C_{ml}^{(0)}(t) - C_l^{(0)}(t)$. Then, the utility function of EVs for eavesdropping on the pair $l \in \mathcal{L}$ at time slot t can be calculated as $r_{\mathcal{M}l}(t) = I_l(t) - \eta_{\mathcal{M}}(t)\varepsilon_{\mathcal{M}}(t)\varepsilon_{\mathcal{N}\mathcal{M}}(t)[I_l(t) - I_l^{(0)}(t)]^+$, where $\varepsilon_{NM}(t) \in \{0, 1\}$ indicates whether TCUs agree to ally with EVs in time slot t. Denoting the strategy of EVs as $\pi_{\mathcal{M}} = \{\varepsilon_{\mathcal{M}}(t), \eta_{\mathcal{M}}(t)\}_{\forall t}$, the long-term optimization problem for EVs can be formulated as

$$[\mathcal{MP}]: \ \max_{\pi_{\mathcal{M}}} \frac{1}{T} \sum_{t \in [0,T]} \sum_{l \in \mathcal{L}} r_{\mathcal{M}l}(t) \tag{10}$$

$$s.t., \ 0 \leqslant \eta_{\mathcal{M}}(t) \leqslant \eta_{\mathcal{M}}^{max}, \tag{10a}$$

$$\varepsilon_{\mathcal{M}}(t) \in \{0, 1\}. \tag{10b}$$

For TCUs, in order to enhance their long-term utility, they need to determine the transmit power allocation $P_n(t)$ of each TCU $n \in \mathcal{N}$, and determine the sub-channel selection of TCU $n \in \mathcal{N}$ $\omega_{nl}(t) \in \{0, 1\}$ for secure transmission or eavesdropping on the pair of LUs $l \in \mathcal{L}$ in time slot t. They also need to determine each TCU's mode $\theta_n(t) \in \{0, 1\}$ between cooperative jamming and relaying. Considering the provided unit incentive of LUs $\eta_{\mathcal{L}}(t)$ and EVs $\eta_{\mathcal{M}}(t)$, they should determine $\varepsilon_{\mathcal{N}\mathcal{L}}(t)$ and $\varepsilon_{\mathcal{N}\mathcal{M}}(t)$ for the potential coalition formation with minimum power consumption. Then, we can calculate the rewards of TCU $n \in \mathcal{N}$ obtained from LUs or EVs as $r_n(t) = \varepsilon_{\mathcal{L}}(t)\varepsilon_{\mathcal{N}\mathcal{L}}(t)\eta_{\mathcal{L}}(t)\sum_{l \in \mathcal{L}} \omega_{nl}(t)(R_l(t) - R_l^{(0)}(t)) + \varepsilon_{\mathcal{M}}(t)\varepsilon_{\mathcal{N}\mathcal{M}}(t)\eta_{\mathcal{M}}(t)\sum_{l \in \mathcal{L}} \omega_{nl}(t)(I_{\mathcal{M}l}(t) - I_{\mathcal{M}l}^{(0)}(t))$. Then, the utility function of the set of all TCUs $\mathcal{N}$ in time slot t can be expressed as $U_{\mathcal{N}}(t) = \sum_{n \in \mathcal{N}}(r_n(t) - \mu_{\mathcal{N}} P_n(t)) - \eta_{conf}(col_{\mathcal{N}\mathcal{L}}(t), col_{\mathcal{N}\mathcal{M}}(t)) \oplus (col_{\mathcal{N}\mathcal{L}}(t-1), col_{\mathcal{N}\mathcal{M}}(t-1))$, where $\mu_{\mathcal{N}}$ is the unit power consumption of TCUs, η_{conf} denotes the potential configuration cost caused by the additional connection established by TCUs to notify the coalition changes, and $col_{\mathcal{L}}(t) = \varepsilon_{\mathcal{L}}(t)\varepsilon_{\mathcal{N}\mathcal{L}}(t)$, $col_{\mathcal{M}}(t) = \varepsilon_{\mathcal{M}}(t)\varepsilon_{\mathcal{N}\mathcal{M}}(t)$. Denoting TCUs' strategy as $\pi_{\mathcal{N}} = \{\omega_{nl}(t), P_n(t), \theta_n(t), \varepsilon_{\mathcal{N}\mathcal{L}}(t), \varepsilon_{\mathcal{N}\mathcal{M}}(t)\}_{\forall t}$, the long-term optimization problem for TCUs can be formulated as

$$[\mathcal{NP}]: \ \max_{\pi_{\mathcal{N}}} \frac{1}{T} \sum_{t \in [0,T]} U_{\mathcal{N}}(t) \tag{11}$$

$$s.t., \ \sum_{l \in \mathcal{L}} \omega_{nl}(t) \leqslant 1, \forall n \in \mathcal{N}, \tag{11a}$$

$$\theta_n(t) \in \{0, 1\}, \forall n \in \mathcal{N}, \tag{11b}$$

$$\varepsilon_{\mathcal{N}\mathcal{L}}(t) + \varepsilon_{\mathcal{N}\mathcal{M}}(t) \leqslant 1, \varepsilon_{\mathcal{N}\mathcal{L}}(t), \varepsilon_{\mathcal{N}\mathcal{M}}(t) \in \{0, 1\}, \tag{11c}$$

$$P_n(t) \in P_{\mathcal{K}}, P_{\mathcal{K}} = \{P_0, P_1, \ldots, P_{k-1}\}, \forall n \in \mathcal{N}, \tag{11e}$$

where constraint (11a) signifies that each TCU $n \in \mathcal{N}$ can only assist or eavesdropping on one pair of LUs $l \in \mathcal{L}$ in each time slot t, and constraint (11e) represents the transmit power of TCU, i.e., $P_n(t)$ is leveled into k power levels.

It can be observed that the long-term optimization problems $[\mathcal{LP}]$, $[\mathcal{MP}]$, $[\mathcal{NP}]$ are tightly coupled. When LUs or EVs provide unit incentives $\eta_{\mathcal{L}}(t), \eta_{\mathcal{M}}(t)$ to TCUs for improving their utilities, TCUs' decisions on whether assist LUs or EVs apparently influence the actual payment of LUs or EVs. Meanwhile, when TCUs ally with LUs or EVs in previous time slot $t-1$, if TCU $n \in \mathcal{N}$ changes its power $P_n(t)$ and mode $\theta_n(t)$, though its consumption may decrease, LUs or EVs are unwilling to ally with TCUs since their utilities may decrease. Such behavior stems from the inherent selfishness of LUs and TCUs, and can further decrease their utilities. Therefore, to analyze the interactions among LUs, EVs and TCUs considering the mutual influences among their strategies, we applied game theory to analyze and solve these optimization problems.

3 Game Analysis and Approach

In this section, to optimize the long-term strategies of LUs, EVs and TCUs, i.e., address optimization problems [$\mathcal{LP}$], [$\mathcal{NP}$] and [$\mathcal{MP}$], we proposed a hierarchical game $\mathcal{H}$ consisting of a matching subgame ($\mathcal{G}^M$) and a coalition formation subgame ($\mathcal{G}^C$) to analyze the strategic interaction.

3.1 Hierarchical Game for Three Parties

To capture and analyze the interdependent decision-making process of LUs, EVs anf TCUs, the three parties' hierarchical game (TPHG) is defined as $\mathcal{H} = \{\mathcal{D}, \{\pi_d\}_{d\in\mathcal{D}}, \{U_d\}_{d\in\mathcal{D}}\}$, where $\mathcal{D} = \{\mathcal{L}, \mathcal{M}, \mathcal{N}\}$ refers to the set of the three parties, $\{\pi_d\}_{d\in\mathcal{D}}$ is the strategy of each party, and $\{U_d\}_{d\in\mathcal{D}}$ is the game utility of the three parties. According to equation (15) and (19), $U_{\mathcal{L}}(t) = \sum_{l\in\mathcal{L}} r_l(t)$ and $U_{\mathcal{M}}(t) = \sum_{l\in\mathcal{L}} r_{\mathcal{M}l}(t)$ in time slot t. TCUs first determine the decisions $\theta_n(t)$, $P_n(t)$ and $\omega_{nl}(t)$, and then reactions, LUs determine $\eta_{\mathcal{L}}(t)$, $\varepsilon_{\mathcal{L}}(t)$ and EVs determine $\eta_{\mathcal{M}}(t)$, $\varepsilon_{\mathcal{M}}(t)$, and finally TCUs can determine $\varepsilon_{\mathcal{NL}}(t)$ and $\varepsilon_{\mathcal{NM}}(t)$.

For each party $d \in \mathcal{D}$, let $\pi_d^{H_*}$ indicate its optimal strategies, the equilibrium of game $\mathcal{H}$ is equivalent to the long-term optimal solution of [$\mathcal{LP}$], [$\mathcal{NP}$] and [$\mathcal{MP}$], and such equilibrium can be defined as follows.

Definition 1 (Equilibrium of $\mathcal{H}$). *In the proposed game $\mathcal{H}$, a strategy profile $\pi_d^{H_*}, \forall d \in \mathcal{D}$ that each party can not achieve higher utility by unilaterally deviating from this strategy, constitute the equilibrium of $d \in \mathcal{D}$ if and only if the following inequality holds:*

$$\frac{1}{T}\sum_{t=0}^{T-1} U_d^{\pi_d^{H_*},\pi_{-d}^{H_*}}(t) \geqslant \frac{1}{T}\sum_{t=0}^{T-1} U_d^{\pi_d^{H},\pi_{-d}^{H_*}}(t), \tag{12}$$

where $\pi_{-d}^{H_*}$ denotes the equilibrium strategies of all other parties except d.

Proof. This proof is omitted due to the page limit.

For the decision tuple that includes selections of TCUs' modes, powers and sub-channels, we have formulated a double-sided matching subgame $\mathcal{G}^M = \{\mathcal{S}, \mathcal{L}, \varPhi, \mathcal{I}\}$. $\mathcal{S} = \mathcal{N} \times P_{\mathcal{K}} \times \varTheta$, $\varTheta = \{0, 1\}$ is the players set of TCUs participating in the subgame, each palyer is a tuple $s = (n, p_k, \theta)$. When the alliances among LUs, EVs and TCUs are fixed, $\mathcal{L}$ is the players set of LUs while TCUs and LUs form an alliance, or $\mathcal{L}$ is the players set of EVs while TCUs and EVs form an alliance. $\varPhi$ is the matching between $\mathcal{S}$ and $\mathcal{L}$ which should satisfy the constraint that each TCU can only select one sub-channel, which is expressed as $\forall s^i, s^j \in \varPhi(\mathcal{L}), s^i \neq s^j$ if and only if $n^i \neq n^j$. $\mathcal{I}$ is the preference lists of all players. In this subgame $\mathcal{G}^M$, we focus on finding a stable matching between $\mathcal{S}$ and $\mathcal{L}$ in time slot t, which can be defined as

Definition 2 (Stable Matching of $\mathcal{G}^M$). *In each time slot t, a matching $\varPhi^*$ is stable, if there is no such blocking pair (s, l) that the matched players $s \in \mathcal{S}$, $l \in \mathcal{L}$ prefer each other over their partners in the current matching, i.e., $l \succ_s \varPhi(l)$ with $s \neq \varPhi(l)$ and $s \succ_l \varPhi(s)$ with $l \neq \varPhi(s)$.*

Proof. This proof is omitted due to the page limit.

And we need to consider two different fixed alliance scenarios: Scenario 1, where TCUs ally with LUs; Scenario 2, where TCUs ally with EVs, resulting in different stable matchings.

For analyzing the changes in alliance relationship among LUs, EVs and TCUs, we formulated the coalition formation subgame $\mathcal{G}^C = \{\mathcal{D}, \Delta, \mathcal{U}\}$, where $\mathcal{D} = \{\mathcal{L}, \mathcal{M}, \mathcal{N}\}$ represents the set of all parties, $\Delta = \{\{\mathcal{L}\}, \{\mathcal{M}\}, \{\mathcal{N}\}, \{\mathcal{L}, \mathcal{N}\}, \{\mathcal{M}, \mathcal{N}\}\}$ is the set of all available coalitions which can be temporally formed or break down, $\mathcal{U} = \{U_{\mathcal{L}}, U_{\mathcal{M}}, U_{\mathcal{N}}\}$ is the utility function of LUs, EVs and TCUs. In each time slot t, such utility function can be defined as $U_{\mathcal{L}} = \sum_{l \in \mathcal{L}} r_l(t)$, $U_{\mathcal{M}} = \sum_{l \in \mathcal{L}} r_{\mathcal{M}l}(t)$ and $U_{\mathcal{N}} = U_{\mathcal{N}}(t)$.

For any self-centric party $d \in \mathcal{D}$, it has the preference to choose a coalition that can enhance its own utility and prefers to leave a coalition that would decrease their utility. The preference order for player $d \in \mathcal{D}$ is given as $f \succ_d^t f' \Leftrightarrow U_d^{f \cup d}(t) \geqslant U_d^{f' \cup d}(t), \forall d \notin f, f'$ and $\forall f, f' \in \Delta \cup \{\varnothing\}, f \neq f'$, where $U_d^{f \cup d}(t)$ and $U_d^{f' \cup d}(t)$ represent the utilities of player d after joining coalition f and f' in time slot t. The three parties may choose to join a coalition that enhance their utilities without harming other members of the original and new coalitions. We define the switch operation $\rightarrow_d^t$ to represent the conditions under which coalition changes occur, given a partition $\mathcal{F}(t) = \{f_1, f_2, \ldots\}$, a party $d \in \mathcal{D}$ wants to leave its current coalition f and join a new coalition f', $f \rightarrow_d^t f'$ if and only if $f' \succ_d^t f \backslash d$ and $f' \cup d \backslash d' \succ_{d'} f' \backslash d', \forall d' \in f'$. Specifically, if a party can achieve the highest utility by forming a separate coalition, it can be expressed as $f \rightarrow_d^t \varnothing \Leftrightarrow \varnothing \succ_d^t f \backslash d, \forall d \in f, f \in \Delta$. Based on the introduction of switch operation, the stable coalition partition can be defined as

Definition 3 (Stable coalition partition of $\mathcal{G}^C$). *In time slot t, if no party can increase its utility by switch operation to change its coalition, the coalition partition $\mathcal{F}^*(t)$ is said to be stable.*

Proof. This proof is omitted due to the page limit.

3.2 The Approach to Equilibrium

In this subsection, we propose a DRL-based approach for the equilibrium of TPHG to obtain the optimal strategies for LUs, EVs and TCUs. In each time slot t, preference lists are established based on DRL to each party, we propose a constrained matching process for a stable matching in $\mathcal{G}^M$, and a coalition selection process for stable coalition partition in $\mathcal{G}^C$.

Markov decision process (MDP) for each party: Considering the local information, there involves the channel states of all current links, the coalition partition situation in the system, and the channel selections of TCUs. In each time slot t, the strategies of the three parties are only related to the current state of the system and their decisions. Thus the state transitions of our system satisfy Markov properties, so the strategy making problem of each party is depicted as a MDP. For each party $d \in \mathcal{D}$, its MDP can be written as $\mathbb{M}_d = \{\mathbb{O}, \mathbb{A}_d, \mathbb{P}_d, \mathbb{R}_d\}$. $\mathbb{O}$ is the environment state in this system. Note that the state for each party is the same. The states of party d at time slot t include its current available coalition partition $\mathcal{F}(t) \in \Delta$, channel gains of all links $\mathcal{CH}(t) = \{|h_{xy}(t)|^2 | \forall x \in \mathcal{L} \cup \mathcal{N}, \forall y \in \mathcal{L} \cup \mathcal{N} \cup \mathcal{M}\}$ and the sub-channel selections $\Omega(t) = \{l | \omega_{nl}(t) = 1, \forall n \in \mathcal{N}\}$, then the environment state of each party is $\mathbb{O} = \{o^t\}_{\forall t}$ where $o^t = \{\mathcal{F}(t), \mathcal{CH}(t), \Omega(t)\}$. $\mathbb{A}_d = \{a_d\}_{\forall d \in \mathcal{D}}$ is the action space of party d. For LUs and EVs, their actions are a set of preference lists $a_d = \{I_1, \ldots, I_L\}_{d \in \{\mathcal{L}, \mathcal{M}\}}$ representing the preferences of $\mathcal{L}$ and $\mathcal{M}$ for $\mathcal{S}$ described on $\mathcal{G}^M$. Conversely, we have $a_{\mathcal{N}} = \{I_1, \ldots, I_{|\mathcal{S}|}\}$. $\mathbb{P}_d = Pr(o'|o, a_d)$ is the state transition probabilities of party $d \in \mathcal{D}$. Representing the probability of d executing an action a_d

Algorithm 1: DRL-Based Solution for TPHG $\mathcal{H}$

```
Input: Initial channel gains of all links.
Output: The equilibrium, i.e., $\pi^*_{\mathcal{L}}$, $\pi^*_{\mathcal{M}}$, $\pi^*_{\mathcal{N}}$.
Initialize: $\mathcal{CH}(0)$, $\Omega(0)$, $\mathcal{F}(0) = \{\{\mathcal{L}\}, \{\mathcal{M}\}, \{\mathcal{N}\}\}$ , $\rho^d$ and $\vartheta^d$;
for training step $= 0, 1, 2, ..., max$ do
    for $t = 1, 2, ..., T$ do
        $\mathcal{F}(t) = \mathcal{F}^*(t-1)$; $\mathbb{O} = \{\mathcal{CH}(t), \Omega(t), \mathcal{F}(t)\}$;
        $agt_{\mathcal{L}}$, $agt_{\mathcal{M}}$, $agt_{\mathcal{N}}$ generate action $\mathcal{I}_{\mathcal{L}}, \mathcal{I}_{\mathcal{M}}, \mathcal{I}_{\mathcal{N}}$ based on state $\mathbb{O}$;
        for two coalition scenarios : $\{\mathcal{L}, \mathcal{N}\}$ and $\{\mathcal{M}, \mathcal{N}\}$ do
            Players of $\mathcal{S}$ initiate invitations to $\mathcal{L}$ in sequence based on their preference lists; Players
              of $\mathcal{L}$ choose the inviter with a higher ranking based on their preference lists; The
              constraint which is $\forall s^i, s^j \in \Phi(\mathcal{L}), s^i \neq s^j$ if and only if $n^i \neq n^j$ need to be satisfied
              when accepting invitations;
        end
        Calculate $\eta_{\mathcal{L}}$ and $\eta_{\mathcal{M}}$ of LUs and EVs according to (13);
        while $\mathcal{F}(t)$ is different from it in the previous loop do
            Calculate the utilities $U_d^{f_i}(t)$ and $U_d^{f_j \cup \{d\}}(t)$ according to utility functions defined in
              subgame $\mathcal{G}^C$; Each $d$ decides whether to leave $f_i$ and join $f_j$; Adjust the coalition
              partition $\mathcal{F}(t)$ ;
        end
        Update $\pi_{\mathcal{L}}(t)$, $\pi_{\mathcal{M}}(t)$ and $\pi_{\mathcal{N}}(t)$ $\mathbb{O} = \{\mathcal{CH}(t), \Omega(t), \mathcal{F}(t)\}$;
        Calculate each party's rewards $r_d^t, d \in \mathcal{D}$;
        Store $(\mathbb{O}^{t-1}, a_d^t, r_d^t, \mathbb{O}^t)$ in replay buffer of $agt_d$.
    end
    Update critic $\rho^d$ and actor $\vartheta^d$ networks of $agt_d$ by Adam w.r.t $L_\rho^d$ and $L_\vartheta^d$.
end
```

causing state $o \in \mathbb{O}$ to transition to state $o' \in \mathbb{O}$. $\mathbb{R}_d = \{r_d^t\}_{\forall t}$ is the reward for party $d \in \mathcal{D}$. In time slot t, the immediate rewards of $\mathcal{L}$, $\mathcal{N}$ and $\mathcal{M}$ are respectively defined as $r_d^t = U_d(t), \forall d \in \{\mathcal{N}, \mathcal{M}\}$, $r_{\mathcal{L}}^t = U_{\mathcal{L}}(t) - \varphi \sum_{l \in \mathcal{L}} (C_l(t) - C_{min})$, where φ is the penalty of QoS constraint for $\mathcal{L}$.

Preference List Based on DRL: We have a large state space and a multidimensional discrete action space. Therefore, we propose a proximal policy optimization (PPO) based algorithm and adopt the actor-critic (AC) [2] framework to improve the training efficiency. The AC framework includes an actor network with parameter ϑ, responsible for interacting with the environment and learning a better policy, and a critic network with parameter ρ, responsible for learning a value function $V_\rho^d(o) = \mathbb{E}\{\sum_{t=0}^{T} \gamma^t r_d^t | o, \pi_d\}$ to evaluate the action.

Based on the generated preference lists of LUs, EVs, and TCUs, stable matching results can be determined in two scenarios :if TCUs assist to LUs or if TCUs assist to EVs by following process. Each $s \in \mathcal{S}$ sequentially initiates requests to the highest-ranked $l \in \mathcal{L}$, and each $l \in \mathcal{L}$ temporarily reserves higher-ranked $s \in \mathcal{S}$ and rejects other $s' \neq s, s' \in \mathcal{S}$. The rejected $s' \in \mathcal{S}$ then initiates requests to the next-highest $l \in \mathcal{L}$ in its preference list. Continuing the above operations until all $s' \in \mathcal{S}$ are matched or no more $l \in \mathcal{L}$ can initiate requests. Especially, before reserving an $s(n, p, \theta)$, each $l \in \mathcal{L}$ will check whether there exists an $s'(n, p', \theta')$ reserved by another $l' \in \mathcal{L}$. If it exists, then a comparison is made between $r_n^s | \omega_{nl} = 1$ and $r_n^{s'} | \omega_{nl'} = 1$. In the case where $r_n^s \leq r_n^{s'}$, s will agree to match with l; otherwise, s will reject the match with l and be marked as matched.

According to the two stable matching result, in each time slot t, LUs and EVs calculated their unit incentives by Shapley value, which can be obtained as follows $\phi_d(t) = \sum_{f' \subseteq f_i \setminus \{d\}} (|f'|!(|f_i| - |f'| - 1)!)/|f_i|!(U^{f' \cup d}(t) - U^{f'}(t))$, which expressed the weighted contribution made by party d in its coalition. Then we can calculate the unit incentive provided by $\mathcal{L}$ and $\mathcal{M}$ according to the contribution ratio of $\mathcal{N}$ to coalitions $\{\mathcal{L}, \mathcal{N}\}$ and $\{\mathcal{M}, \mathcal{N}\}$.

$$\eta_d(t) = \frac{\eta_d^{max} \phi_{\mathcal{N}}(t)}{\phi_{\mathcal{N}}(t) + \phi_d(t)}, \forall d \in \{\mathcal{L}, \mathcal{M}\}. \tag{13}$$

Then, $d \in \mathcal{D}$ will decide whether to perform a switching operation according to its utility. Repeat the above process until the coalition partition no longer changes and we can obtain a stable coalition partition $\mathcal{F}^*(t)$.

Specifically, for each party, the dimension of their actions are no less than $div = min\{|\mathcal{L}|, |\mathcal{S}|\}$. Considering that directly exploring such a large dimensional action space is challenging, we assign an agent to each party to enhance exploration efficiency. We set agents $agt_{\mathcal{L}}$, $agt_{\mathcal{M}}$ and $agt_{\mathcal{N}}$ to generate preference lists for LUs, EVs and TCUs. During the training process, these three agents have their own actor networks, critic networks, and replay buffers. Each agent also independently updates its own network parameters ρ^d for critic network and ϑ^d for actor network. The network parameters of each agent have to be updated at a certain frequency. The updating process is mainly inspired by the PPO method, namely, when the replay buffers of all agents reach the total capacity, we update the actor networks and critic networks of all agents. For each agent agt_d, this update process includes: 1)calculation of its rewards-to-go, i.e., the discounted rewards $J_t^d = \sum_{t=0}^{T-1} \gamma^t r_t^d$, where γ^t stands for the discount factor in time slot t; 2) calculation of advantage function $A_t^d = J_t^d - V_\rho^d(s_t^d)$, where $V_\rho^d(o_t) \triangleq \mathbb{E}\{\sum_{t=0}^{T} \gamma^t r_d^t | o, \pi_d(\rho)\}$; 3) calculation of the loss function of actor network $L_\vartheta^d = \mathbb{E}[min(\frac{p_\vartheta(a_t^d|o_t)}{p_{\vartheta'}(a_t^d|o_t)} A_t^d, clip(\frac{p_\vartheta(a_t^d|o_t)}{p_{\vartheta'}(a_t^d|o_t)}, 1 - \epsilon, 1 + \epsilon) A_t^d)]$, where the function $clip()$ clips the ratio to be no more than $1 + \epsilon$ and no less than $1 - \epsilon$; 4)calculation of the loss function of critic networks $L_\rho^d = \mathbb{E}[(V_\rho^d(o_t) - J_t^d)^2]$, and parameters ρ and ϑ can be updated by minimizing the loss function via the Adam optimizer.

In summary, the above approach is summarized in Algorithm 1, which will be used to optimize ρ and ϑ of the critic networks and actor networks, and thus the strategy equilibrium solution of TPHG.

4 Simulation Results

We consider a cooperative communication system, where $\mathcal{L}$ contains 6 LUs, $\mathcal{M}$ contains 2 EVs and $\mathcal{N}$ contains 3 TCUs. The power of LUs' transmitters is 20 dBm and the range of k-leveled working power values that TCUs can choose is from 20 dBm to 30 dBm with $k = 10$ [1]. The frequency bandwidth W is set to $10MHz$ [3]. We assume that $\zeta_{\mathcal{L}} = 5$, $\mu_{\mathcal{N}} = 5$, $\varphi = 0.2$, $\eta_{conf} = 0.1$, $C_{min} = 4bps/Hz$, $\eta_{\mathcal{L}}^{max} = 10$ and $\eta_{\mathcal{M}}^{max} = 10$. Note that similar assumptions were made in [8], [4]. In addition, the uncertain channel gain is assumed to have a rayleigh distribution for flat multi-path fading environment, which is employed in [4]. The AWGN noise is assumed to have zero-mean and variance $\sigma^2 = 0.55$ which are similar to the settings in [7].

Fig.3 shows the performance of the three parties' utilities during the training stage. It can be observed that after 5000 episodes of training, the utilities of LUs, EVs and TCUs respectively converged where the fluctuation is caused by the exploration mechanism of DRL, which indicates

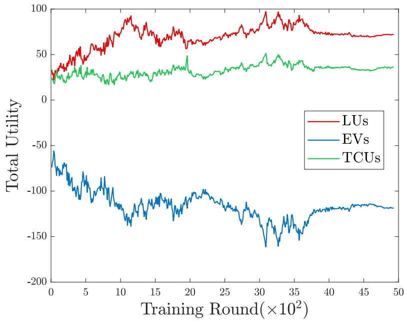

Fig. 3. Convergence of the proposed DRL-based solution.

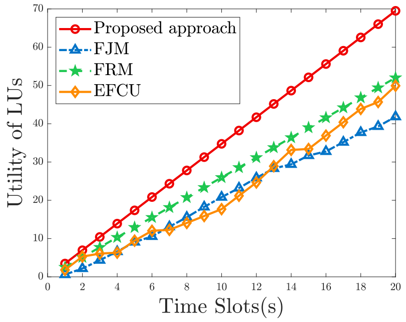

Fig. 4. Performance of LUs' cumulative utility.

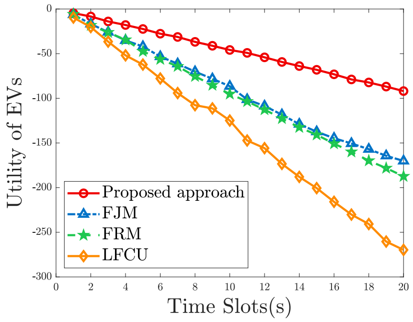

Fig. 5. Performance of EVs' cumulative utility.

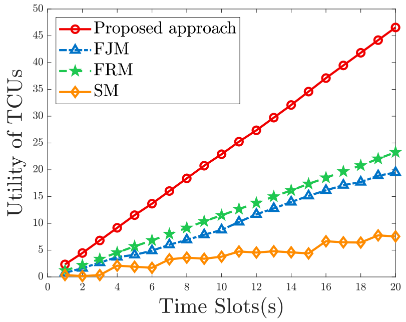

Fig. 6. Performance of TCUs' cumulative utility.

that the optimal strategy of the formulated TPHG $\mathcal{H}$ is achieved by the DRL-based solution. This also verifies the theoretical analysis of the equilibrium of TPHG, indicating that our proposed approach is practical and feasible.

The figures show the progressiveness of the cumulative total utilities of the proposed solution in LUs, EVs and TCUs compared with other approaches. Fig.4 demonstrates that our proposed approach outperforms fixed jamming mode (FJM) [7] , fixed relaying mode (FRM) [7] and EVs' frindly TCUs scheme (EFCU) [8] in terms of utility of LUs. Fig.5 shows that our proposed approach outperforms FJM, FRM and LUs' frindly TCUs scheme (LFCU) [9] in terms of utility of EVs. Fig.6 illustrates that our proposed approach outperforms FJM, FRM and static matching (SM) [5] approach in terms of utility of TCUs. This is because the proposed approach allows LUs, EVs, and TCUs to dynamically determine the formation of coalitions based on the information available in the time-varying wireless system. It helps improve the secure transmission performance of LUs or the eavesdropping efficiency of EVs and TCUs profits, which is different from LFCU and EFCU approaches. Similarly, static matching does not allow TCUs to strategically select powers, modes, and sub-channels, making it difficult for TCUs to better assist LUs and EVs, resulting in a decrease in the utilities of LUs, EVs, and TCUs. Moreover, for FJM and FRM schemes, the TCUs' fixed mode cannot fully meet the needs of different links. Specifically, for LUs with poor channel conditions, relay is needed to improve transmission rate, while for LUs closer to EVs, interference is needed to confuse eavesdroppers and improve their security.

5 Conclusion

In this paper, we propose a hierarchical game framework for PLS-aware wireless cooperative networks characterized by their dynamism and collaborative transmissions. Prompted by the framework, we build a hierarchical game model integrating a matching subgame and a coalition formation subgame to simulate interactions among LUs, EVs, and TCUs, as well as their potential coalition relationships. In order to derive the optimal long-term strategies while solving the equilibrium of the hierarchical game, we develop a DRL based solution, which includes a matching process and a coalition selection process for two subgames. Simulation results validate the feasibility of the proposed approach, demonstrating its superiority in increasing the cumulative utilities of LUs, EVs, and TCUs compared to other methods.

References

1. Chen, R., Yi, C., Zhu, K., Chen, B., Cai, J., Guizani, M.: A three-party hierarchical game for physical layer security aware wireless communications with dynamic trilateral coalitions. IEEE Trans. Wireless Commun. pp. 1–1 (2023). https://doi.org/10.1109/TWC.2023.3322776
2. Dai, C., Zhu, K., Hossain, E.: Multi-agent deep reinforcement learning for joint decoupled user association and trajectory design in full-duplex multi-uav networks. IEEE Trans. Mobile Comput. **22**(10), 6056–6070 (2023). https://doi.org/10.1109/TMC.2022.3188473
3. Fang, H., Xu, L., Zou, Y., Wang, X., Choo, K.K.R.: Three-stage stackelberg game for defending against full-duplex active eavesdropping attacks in cooperative communication. IEEE Trans. Veh. Technol. **67**(11), 10788–10799 (2018). https://doi.org/10.1109/TVT.2018.2868900
4. Han, C., Liu, A., Wang, H., Huo, L., Liang, X.: Dynamic anti-jamming coalition for satellite-enabled army iot: A distributed game approach. IEEE Internet Things J. **7**(11), 10932–10944 (2020). https://doi.org/10.1109/JIOT.2020.2991585
5. Kazmi, S.M.A., Tran, N.H., Saad, W., Han, Z., Ho, T.M., Oo, T.Z., Hong, C.S.: Mode selection and resource allocation in device-to-device communications: A matching game approach. IEEE Trans. Mobile Comput. **16**(11), 3126–3141 (2017). https://doi.org/10.1109/TMC.2017.2689768
6. Kuhestani, A., Mohammadi, A., Yeoh, P.L.: Optimal power allocation and secrecy sum rate in two-way untrusted relaying networks with an external jammer. IEEE Trans. Commun. **66**(6), 2671–2684 (2018). https://doi.org/10.1109/TCOMM.2018.2802951
7. Li, B., Yang, Z., Zou, Y., Zhu, J., Cao, W., Shi, C.: Securing multiuser communications via an energy harvesting node: Jammer or relay? IEEE Trans. Veh. Technol. **72**(7), 8755–8769 (2023). https://doi.org/10.1109/TVT.2023.3244548
8. Moon, J., Lee, S.H., Lee, H., Lee, I.: Proactive eavesdropping with jamming and eavesdropping mode selection. IEEE Trans. Wireless Commun. **18**(7), 3726–3738 (2019). https://doi.org/10.1109/TWC.2019.2918452
9. Ragheb, M., Hemami, S.M.S., Kuhestani, A., Ng, D.W.K., Hanzo, L.: On the physical layer security of untrusted millimeter wave relaying networks: A stochastic geometry approach. IEEE Trans. Inf. Forensics Security **17**, 53–68 (2022). https://doi.org/10.1109/TIFS.2021.3131028
10. Wang, K., Yuan, L., Miyazaki, T., Zeng, D., Guo, S., Sun, Y.: Strategic antieavesdropping game for physical layer security in wireless cooperative networks. IEEE Trans. Veh. Technol. **66**(10), 9448–9457 (2017). https://doi.org/10.1109/TVT.2017.2703305
11. Wyner, A.D.: The wire-tap channel. Bell Syst. Tech. J. **54**(8), 1355–1387 (1975). https://doi.org/10.1002/j.1538-7305.1975.tb02040.x