Computers, Materials & Continua CMC, vol.65, no.2, pp.1487-1505, 2020

A Robust Resource Allocation Scheme for Device-to-Device Communications Based on Q-Learning

Azka Amin[1], Xihua Liu[2], Imran Khan[3], Peerapong Uthansakul[4, *], Masoud Forsat[5] and Seyed Sajad Mirjavadi[5]

Abstract: One of the most effective technology for the 5G mobile communications is Device-to-device (D2D) communication which is also called terminal pass-through technology. It can directly communicate between devices under the control of a base station and does not require a base station to forward it. The advantages of applying D2D communication technology to cellular networks are: It can increase the communication system capacity, improve the system spectrum efficiency, increase the data transmission rate, and reduce the base station load. Aiming at the problem of co-channel interference between the D2D and cellular users, this paper proposes an efficient algorithm for resource allocation based on the idea of Q-learning, which creates multi-agent learners from multiple D2D users, and the system throughput is determined from the corresponding state-learning of the Q value list and the maximum Q action is obtained through dynamic power for control for D2D users. The mutual interference between the D2D users and base stations and exact channel state information is not required during the Q-learning process and symmetric data transmission mechanism is adopted. The proposed algorithm maximizes the system throughput by controlling the power of D2D users while guaranteeing the quality-of-service of the cellular users. Simulation results show that the proposed algorithm effectively improves system performance as compared with existing algorithms.

Keywords: 5G, D2D communications, power allocation algorithm, resource optimization.

1 Introduction

With the rapid development of the mobile Internet and the continuous updating of smart terminal technology, the number of wireless mobile users and wireless network traffic has exploded [Jameel, Hamid, Jabeen et al. (2019); Ahmad, Li, Waqas et al. (2018)]. It is predicted that by 2020, the wireless network traffic will show a thousand-fold increase

[1] School of Business, Qingdao University, Qingdao, 266061, China.

[2] School of Economics, Qingdao University, Qingdao, 266061, China.

[3] Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan.

[4] School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima, Thailand.

[5] Department of Mechanical and Industrial Engineering, College of Engineering, Qatar University, Doha, Qatar.

[*] Corresponding Author: Peerapong Uthansakul. Email: uthansakul@sut.ac.th.

Received: 27 May 2020; Accepted: 16 June 2020.

compared with now, "Internet of Everything" will become the development trend of the future wireless communication world [Shaikh and Wesissmuller (2018); Saraereh, Mohammed, Khan et al. (2019); Alemaishat, Saraereh, Khan et al. (2019)]. To meet the future development of wireless services, mobile communication systems need to find new technologies to significantly increase network capacity and spectrum efficiency [Jabeen, Ali, Khan et al. (2019); Bakht, Jameel, Ali et al. (2019)]. At present, for the future wireless communication networks, many new key technologies have appeared, such as millimeter-wave communication [Lee, Patil, Hunt et al. (2019); Jameel, Risaniemi, Khan et al. (2019); Saraereh, Alsaraira, Khan et al. (2019); Saraereh, Alsaraira, Khan et al. (2020); Shafi, Molisch, Smith et al. (2017); Alemaishat, Saraereh, Khan et al. (2019)], Massive MIMO [Saraereh, Khan, Lee et al. (2019); Khan, Rodrigues, Al-Muhtadi et al (2019)], heterogeneous networks [Stamou, Dimitriou, Kontovasilis et al. (2019)], Device-to-Device (D2D) communication [Orsino, Ometo, Fodor et al. (2017)], etc., to explore the efficient use and promotion of spectrum and energy from two perspectives: system and network. Among these new technologies, D2D communication has attracted widespread attention from scholars and the industry. The basic principle of D2D communication is that users which are nearby can multiplex cellular user spectrum resources and establish direct links for communication to achieve low-power transmission, which can not only improve the system throughput, reduce the load on cell base stations, reduce terminal delay but can also achieve higher energy efficiency and spectral efficiency. Therefore, it has become one of the hotspots in future wireless communication technology research [Orsino, Ometo, Fodor et al. (2017); Wang, Zhang, Leung et al. (2018); Wang, Tang, Wu et al. (2017); Salehi, Mohammadi, Haenggi et al. (2017)]. D2D communication is divided into two types according to different ways of using spectrum resources. One is the overlay method. In the overlay method [Wang, Zhang, Leung et al. (2018)], the frequency band resources of the cellular user equipment and the D2D communication equipment are orthogonal. Users in the two communication modes will not cause interference, but the resource utilization of this resource allocation method is low. The other is the underlying method. The spectrum of D2D users and cellular users is non-orthogonal, and the spectrum utilization rate is high. However, the randomness of D2D user locations and multiplexing cellular spectrum will bring serious interference between D2D communication and cellular systems [Wang, Tang, Wu et al. (2017)]. So, interference management has become an important research direction in D2D communication [Sun, Shin, Zhang et al. (2017); Yang, Li, Semasinghe et al. (2017); Xu, Huang, Yang et al. (2017); Li and Huang (2017)]. As we all know, the intensity of mutual interference between D2D communication and cellular systems is closely related to the transmit power of D2D user pairs. Therefore, D2D users can control the inter-link interference by dynamically adjusting the transmit power to maximize the system throughput while ensuring the communication quality of cellular users using various mechanisms such as deep learning [Li, Cao, Chen et al. (2017); Sun, Shi, Yin et al. (2019); Wang, Jiang, Luo et al. (2019); Zhang, Wang, Lu et al. (2019)], CNN [Liu, Yang, Lv et al. (2019); Zeng, Dai, Li et al. (2019); Luo, Qin, Xiang et al. (2020); Zhang, Jin, Sun et al. (2018)] and resource allocations [Li, Li, Zhang et al. (2019); Jiang, Tang, Gu et al. (2020); Zhang, Li, Wang et al. (2018)]. It can be seen that power control is the key technology to solve the problem of cross-layer interference in D2D communication

and cellular systems and has become a research hotspot of D2D communication in recent years [Wang, Wang, Jin et al. (2015); Ren, Liu, Liu et al. (2015); Huang, Nasir, Durrani et al. (2016); Lin, Ouayang, Zhu et al. (2016)]. Because the power control problem is a non-linear objective optimization problem, more and more researchers apply mathematical models such as game theory [Chen, Li, Jiang et al. (2015)], stochastic optimization [Sakr and Hossain (2015)], graph theory [Ni, Collings, Lipman et al. (2015)], mixed integer programming [Alfa, Maharaj, Lall et al. (2016)], etc. to research on power control issues. For example, the authors in Chen et al. [Chen, Yu, Shan et al. (2016)] proposed a distributed D2D communication power control algorithm and compared it with the traditional centralized open-loop power control. The authors in Zhang et al. [Zhang, Zhang, Yan et al. (2015)] summarized the power control problem in a hybrid D2D and cellular network as a Stackelberg game model, with the cellular user as a leader and the D2D user as a follower. A price-based distributed power control method is proposed in Ji et al. [Ji, Caire, Molisch et al. (2016)] based on graph theory, has two types of centralized and distributed power control and channel allocation joint optimization schemes. The authors in Maghsudi et al. [Maghsudi and Stanczak (2015)] modeled the spectrum and power allocation problem as a convex optimization problem. Based on the random geometry theory, a joint channel selection and power allocation optimization algorithm was proposed. In these studies, much prior knowledge (such as channel state information) is assumed to be known to D2D users. But, because the traditional pilot signal (in the cellular link) is implemented in D2D communication, it is difficult to know the precise interference characteristics information between the D2D terminal and the base station, especially when the number of D2D users increases, the algorithm complexity also rises sharply. Therefore, the open-loop power control and resource optimization algorithms based on the a priori assumption that D2D users and BS have no information exchange, and no channel state information for D2D users have research significance and application prospects. Some of the recent research hotspots on key issues of D2D communication based on prior knowledge such as no channel state information (CSI) are: i) how to predict the network load status, b) select the optimal channel access, c) dynamically control the power of D2D communication users, d) reduce the interference between D2D users and cellular users, e) obtain maximum system throughput. In Asheralieva et al. [Asheralieva and Miyanaga (2016)], the authors proposed an autonomous Q-learning algorithm for channel selection by D2D pairs which do not require any information of all D2D pairs. The minimum SINR constraints are utilized and the stochastic non-cooperative game mechanism is used to represent this optimization problem. However, this algorithm requires an optimal value of Boltzmann temperature and optimized locally-observed throughput and state. In Yuan et al. [Yuan, Yuang, Feng et al. (2019)], the authors proposed a cooperative algorithm in which the D2D transmitters acts as relays to assist the cellular users for utilization of the licensed spectrum and aimed to consider a realistic scenario with incomplete CSI and perform the one-to-one matching game. However, this scheme requires conversion from cooperative to non-cooperative game and requires synchronized time slots for each one-to-one pair for obtaining the payoff and feedback. The convergence of this scheme is also degraded by an increasing number of cellular and D2D users. Moreover, the learning process is

very slow because there is no sub-grouping of the cellular and D2D pairs and also the convergence rate is slower.

Machine learning is an important core of artificial intelligence. The main idea is to simulate human learning behavior through computers. After participants are stimulated by the environment, they continuously change their behavior based on experience accumulation to better adapt to the environment and achieve their interests. In recent years, more and more researchers have applied machine learning methods to solve key problems of wireless communication systems [Ge, Song, Wu et al. (2019); Peng, Li, Abboud et al. (2017)]. Q-learning is a branch of machine learning (Reinforcement Learning, RL), and its main elements include the environment, reward, action, and state [Maghsudi and Stanczak (2015)]. The learner interacts with the environment through the historical state, and through the learning algorithm, calculates and optimizes a certain target decision value of the system. The specific process of Q-learning is to establish a Q value list. By learning what action each state takes to maximize the Q value, the action with the highest Q value is selected as the final action. The learner repeatedly interacts with the control environment and uses the reward value to evaluate its performance, thereby achieving an optimal decision.

This paper introduces Q-learning ideas into the study of dynamic resource allocation strategies for D2D communication. In a hybrid D2D and cellular network, a mathematical model for power control and resource optimization is constructed based on Q-learning. Multiple D2D user pairs in a cellular network are considered as a symmetric multi-agent system. The D2D user action interacts with the environment, and the final target Q value function converges to the maximum value, and the user action in this state is the optimal resource allocation strategy. During the Q learning process, the D2D terminal and the base station do not need to obtain accurate channel state information and mutual interference. The user learns the distributed optimal power allocation strategy through historical throughput and power values to optimize the overall system throughput.

The rest of the paper is organized as follows. In Section 2, the system model is described. In Section 3, the proposed Algorithms and their principle are analyzed. Section 4 provides the simulation results, while Section 5 concludes the paper.

2 System model and problem description

2.1 System model

The network structure used in this paper is shown in Fig. 1. D2D users and cellular users use the underlying method to multiplex the spectrum. Within a single cell, there is a macro base station BS. The cell comprises of N number of cellular users and M D2D user pairs. The sets $m \in \{1,2,\dots,M\}$ and $k \in \{1,2,\dots,K\}$ denote the index set of cellular users and D2D pairs, respectively.

Assume that users are randomly distributed in the cell. There is no signal exchange between the D2D user and the BS, that is, the D2D user does not know the channel state information, and both the transmitting user and the receiving user are single antennas. Suppose there are N orthogonal wireless channels in a BS, where the channels for cellular users are represented as the set $L_j \in L, j \in \{1,2,\dots,N\}$ and the channels used for D2D

users are represented as the set $K_j \in \kappa$. The average transmit power of each cellular user is a fixed value p_c, and the power value of the D2D user pair is $p_d^k \in \{p_d^1, p_d^2, \dots, p_d^K\}$, $p_d^k \ll p_c$, $(k \in \{1,2,\dots,K\})$. The channel of any user is represented by c_j^m for $(U_j, U_{j'})$, and $j \in \{1,2,\dots,N\}$. The binary channel selection vector space is N-dimensional and $c^m = [c_1^m, c_2^m, \dots, c_N^m]^T$, $(m \in \{1,2,\dots,M\})$.

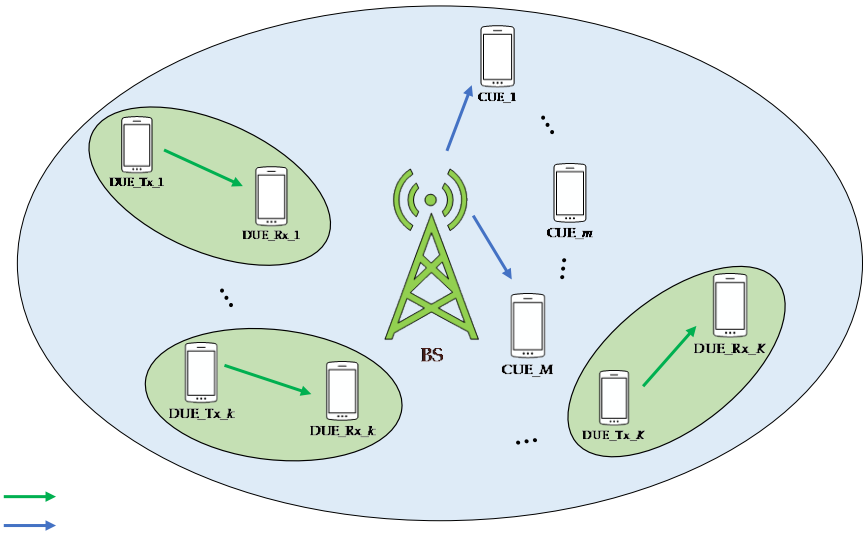

Figure 1: Proposed system model

When the user selects the jth channel for $(U_m, U_{m'})$ at time t, $c_j^m = 1$, otherwise $c_j^m = 0$. That is, each user selects at most one channel.

$$\sum_{j \in N} c_j^m \leq 1, \forall\, m \in M \tag{1}$$

For a cellular user U_m $(m \in \{1,2,\dots,M\})$, given the transmit power and occupying the channel c_j^m, the rate $R_{j,m}$ of the cellular link can be expressed as

$$R_{j,m}^c = \log_2\left(1 + \frac{p_c^m c_j^m G_m^j}{\sum_{k \in \kappa} p_d^n c_j^n G_k^j + \delta^2}\right), \forall\, m \in M, j \in \{1,2,\dots,N\} \tag{2}$$

where p_c^m is the power of the cellular user, p_d^n is the power of the D2D user, δ^2 is the variance of the white Gaussian noise power, G_m^j is the channel gain of the cellular user U_m on the channel c_j^m at any time t and G_k^j is the channel gain from the interfering D2D users to the cellular users. similarly, for each D2D user pair $(U_j, U_{j'})$, $(j \in \{1,2,,\dots,N\})$, given the transmission power, the channel c_j^m is occupied, and the D2D link rate $R_{j,k}^D$ can be expressed as

$$R_{j,k}^D = \log_2\left(1 + \frac{p_d^k c_j^m G_{k,k'}^j}{\sum_{k \in \kappa} p_d^n c_j^m G_{m,m'}^j + \delta^2 + p_c^m c_j^m}\right), \forall\, m \in M, j \in \{1,2,,\dots,N\} \tag{3}$$

where $G_{k,k'}^{j}$ is the channel gain of the D2D user pair on the channel c_j^m at any time t. Since D2D users do not know the precise channel state information, the values of $G_{k,k'}^{j}$ are unknown.

2.2 Problem description

In heterogeneous networks, users tend to choose the network that can always get the best service. There is no information exchange between cellular users and D2D users, and D2D users are unknown about channel availability and channel quality. Therefore, it is more difficult to achieve fair and efficient cross-network resource allocation. For users, whether to communicate through D2D or cellular base stations, the principle of choice is to obtain the best system performance. For users of D2D and cellular heterogeneous networks, the maximum throughput is obtained with the smallest power consumption, so the resource allocation and power control problems are modeled as utility problems, and select the overall rate (i.e., throughput), channel allocation, and power control as the utility functions. The goal is to maximize the system user rate (throughput). Therefore, channel allocation and power control problems can be modeled as

$$\max_{L_j \in L, K_j \in \kappa, p_d^m, p_c} \sum_{j=1} \left(R_{j,m}^{c} + R_{j,k}^{D}\right)$$

Subject to

$$p_k \le p_d^M, \forall\, k \in \kappa$$

$$p_d^m \ll p_c, m \in \{1,2,\dots,M\} \tag{4}$$

among them, the channel of the cellular user is represented as the set $L_j \in L$ and the channel of the D2D user is represented as the set $K_j \in \kappa, j \in \{1,2,\dots,N\}$, the power value of the D2D user pair is $p_d^m \in \{p_d^1, p_d^2, \dots, p_d^M\}$, and the power of the cellular network users is p_c. In the D2D system, Eq. (4) is difficult or even impossible to solve. The reason is: first, D2D users do not know the precise channel state information, which means that the values of $G_{k,k'}^{j}$ are unknown, and the objective function (4) cannot be solved; second, Eq. (4) does not take into account the higher priority (QoS) of cellular users; third, the solution of the maximum value of Eq. (4) depends on the set $L_j \in L$ and the set $K_j \in \kappa$, that is, the channel selection and power allocation of the cellular users and D2D users are related. Due to $p_k \le p_d^M, \forall\, k \in \kappa$:

$$\sum_{j=1}^{N} \sum_{m \in M} \log_2 \left(\frac{p_j^m c_j^m G_{k,k'}^{j}}{1 + \sum_{n \in \kappa} p_d^n c_d^n G_{m,m'}^{j} + \delta^2 + p_c^m c_i^m G_{m,m'}^{j}} \right)$$

$$\ge \sum_{j=1}^{N} \sum_{l \in M} \log_2 \left(\frac{p_c^m c_j^m G_{k,k'}^{j}}{1 + \sum_{n \in \kappa} p_k c_j^n G_{m,m'}^{j} + \delta^2} \right) \tag{5}$$

based on the basic logarithmic properties, Eq. (5) can be written as

$$\sum_{j=1}^{N} \sum_{m \in M} \log_2 \left(p_j^m c_j^m G_{k,k'}^{j} \right) - \sum_{j=1}^{N} \sum_{m \in M} \log_2 \left(1 + \sum_{n \in \kappa} p_d^n c_d^n G_{m,m'}^{j} + \delta^2 + p_c^m c_i^m G_{m,m'}^{j} \right)$$

$$\ge \sum_{j=1}^{N} \sum_{m \in M} \log_2 \left(p_c^m c_j^m G_{m,m'}^{j} \right) - \sum_{j=1}^{N} \sum_{m \in M} \log_2 \left(1 + \sum_{n \in \kappa} p_k c_j^n G_{m,m'}^{j} + \delta^2 \right) \tag{6}$$

when above Eq. (6) takes the lower limit, it is the worst scenario of the network environment, that is, all D2D users send signals with the maximum available power, causing the largest mutual interference. Therefore, for power control, the lower limit of the Eq. (6) is maximized. That is

$$\max\left[\sum_{j=1}^{N}\sum_{m\in M}\log_2\left(p_d^k c_j^m G_{k,k'}^j\right)-\sum_{j=1}^{N}\sum_{m\in M}\log_2\left(1+\sum_{n\in M}p_k^n c_j^n G_{m,m'}^j+\delta^2\right)\right] \quad (7)$$

where $p_d^k \in \{p_d^1, p_d^2, \ldots, p_d^K\}$, $p_d^k \ll p_c$, $(k \in \{1,2,\ldots,K\})$. Eq. (7) is an NP-hard problem, and it is difficult to solve it directly. There are many methods for solving NP difficult problems, such as branch and bound algorithm, genetic algorithm, etc., but these methods need to consider all D2D users and base stations at the same time, all are centralized optimization algorithms. When the number of users increases, the complexity of the algorithm also rises sharply, and for the non-linear objective function Eq. (7), it is necessary to know the characteristics of interference between D2D terminals and base stations, but this information is difficult to know. To solve this problem, joint resource allocation and power control algorithm based on Q-learning is proposed. D2D terminals and base stations do not need to obtain accurate channel state information and mutual interference. Users learn the best power allocation strategy to optimize the overall system throughput through historical characteristics and power values. That is, multiple D2D user pairs in a cellular network are considered as multi-agent systems, and a joint optimization algorithm for distributed channel selection and power control based on Q-learning is designed. The advantage of this distributed algorithm is that it reduces the complexity of the algorithm, and only needs its information to perform power control, avoiding the complexity of the above calculation.

3 Proposed algorithm

3.1 Q-learning

This article uses one of the most widely used algorithms in reinforcement learning, Q Learning [Wang, Wang, Jin et al. (2015); Ren, Liu, Liu et al. (2015)]. The basic principle of Q-learning is that the agent, that is, the initiator of the action, causes a change in the state of the environment after taking an action. The impact of this change can be quantified as a reward. The value or magnitude of the reward value can reflect the reward or punishment for evaluating the learner's actions. The learner then chooses the next action based on the reward value and the current state of the environment. The selection principle is to increase the probability of receiving a positive reward value (award) until convergence. Therefore, the action chosen depends not only on the immediate return value, but also on the historical environmental status, and has an impact on the environmental status and final return value at the next moment.

Therefore, suppose action set A and state set S. The learner chooses an action $\alpha \in A$. After the environment accepts the action, the state changes and generates an instantaneous return value $Re(s)$. Assuming the current state $s \in S$, the next action $\alpha' \in A$, α' is related to the state s' at the next moment and the cumulative return value $Re_c(s)$. The learning goal of Q-Learning is to dynamically adjust the next action $\alpha' \in A$ so that $Re_c(s)$ takes the maximum value. $Re_c(s)$ can be expressed as follows [Wang, Wang, Jin et al. (2015)]:

$$Re_c(s) = Re(s,\alpha) + \varepsilon\sum_{s\in S}P(s'|s,\alpha)\,Re_c(s') \quad (8)$$

where $0 < \varepsilon < 1$ is the return factor and $P(s'|s,\alpha)$ is the state transition probability from state s to state s' when the learner performs the action α'. According to Bellman theory [Kiumarsi; Vamvoudakis; Modares et al. (2018)], the maximum value of the cumulative reward value $Re_c(s)$ is expressed as

$$Re_c^{\text{opt}}(s) = \max\{Re(s,\alpha) + \varepsilon \sum_{s\in S} P(s'|s,\alpha)\, Re_c(s')\} \tag{9}$$

that is, Q-Learning is used to learn $Re(s,\alpha)$ and $P(s'|s,\alpha)$ values. The Q function can be expressed as

$$Q(s,\alpha) = Re(s,\alpha) + \varepsilon \sum_{s\in S} P(s'|s,\alpha)\, Re_c(s') \tag{10}$$

3.2 Problem mapping

This article introduces the idea of Q-Learning into the power control algorithm. The three major elements of Q-learning are: learner agent, state s, action $\alpha \in A$, and reward signals are mapped into the actual power control model. The specific mapping process is described below.

D2D users are mapped as learners. Suppose there are K D2D users in a cell. For D2D users U_i, the state set is as follows

$$s^t = \left\{d_i, i \in \{i, 2, \dots, K\}, G_{k,k'}^j\right\} \tag{11}$$

where d_i is the straight-line distance between the user and the base station, and $G_{k,k'}^j$ is the direct link gain of the user pair $\left(U_j, U_{j'}\right)$. In D2D communication, at time t, it is assumed that the return function Re_{ct_p} represents the instantaneous return value if the user accesses the cellular network base station for communication; $Re_d t$ represents the instantaneous return value of users communicating through D2D. For each unit time slot t, the mobile user has an action α_p in the state s^t, resulting in an instantaneous return value Re_{ct_p} and $Re_d t$. The user learns to derive the best strategy for each s^t by interacting based on the user's location and environment. The return value $Q(s,\alpha)$ is the largest. For the network scenario in this paper, the rate is used as a return function, and the instantaneous change of the rate can intuitively and accurately reflect the network congestion, thereby deriving the throughput. Based on the rate of state s^t, by calculating the cumulative return value of the rate, find the best power control to maximize the system throughput and QoS. Therefore, the instantaneous return function Re_{ct_p} can be expressed as

$$Re_{ct_p} = R_{j,m}^c = \log_2\left(\frac{p_c^m c_j^m G_m^j}{1+\sum_{n\in K} p_d^n c_j^n G_n^j + \delta^2}\right), \forall\, m \in M, j \in L \tag{12}$$

The instantaneous return function $Re_d t$ can be expressed as

$$Re_d t = R_{j,m}^D = \log_2\left(\frac{p_j^m c_j^m G_{m,m'}^j}{1+\sum_{n\in K} p_d^n c_j^n G_{n,m'}^j + \delta^2 + p_c^m c_i^m G_{m,m'}^j}\right), \forall\, m \in M, j \in L \tag{13}$$

then the total return function is

$$\begin{aligned} Re(s,a) &= \max\left(Re_{ct_p} + Re_d t\right) \\ &= \max\left[\sum_{c_j=1}^M \sum_{j\in L} \log_2\left(p_d^k c_j^m G_{k,k'}^j\right) - \sum_{c_j=1}^M \sum_{j\in L} \log_2\left(1 + \sum_{n\in M} p_k c_j^n\, G_{m,m'}^j + \delta^2\right)\right] \end{aligned} \tag{14}$$

where $p_d^k \in \{p_d^1, p_d^2, \dots, p_d^K\}$, $p_d^k \ll p_c$, $(k \in \{1,2,\dots,K\})$.

The action set is denoted as $\alpha_i \in \{0,1,2,\ldots.N\}$. When $\alpha_i = 0$, it means that the user pair $(U_m, U_{m'})$ will be connected to the macro base station, and $\alpha_i = j$ ($j \in \{1,2,\ldots N\}$), which means that the user pair $(U_m, U_{m'})$ will communicate through D2D. The probability of defining a cellular link is Pr_c, and the probability of selecting D2D communication is Pr_d [Kiumarsi, Vamvoudakis, Modares et al. (2018)], can be expressed as

$$Pr_d = \frac{\frac{eRe_{dt_p}}{\tau}}{\sum_{l_i=1}^{L}\frac{eRe_{ct_p}}{\tau}+\sum_{k_i=1}^{K}\frac{eRe_dt}{\tau}} \tag{15}$$

$$Pr_c = \frac{\frac{eRe_{ct_p}}{\tau}}{\sum_{l_i=1}^{L}\frac{eRe_{ct_p}}{\tau}+\sum_{k_i=1}^{K}\frac{eRe_dt}{\tau}} \tag{16}$$

where τ is Boltzmann temperature parameter [Maghsudi and Stanczak (2015)] which is expressed as

$$\tau = \frac{\tau_0}{\log_2(1+t)} \tag{17}$$

among them, τ_0 is the initial temperature and t is the channel selection duration. When the value of τ is high, the probability distribution of channel selection is the same. When the value of τ is low, the user's probability distribution of the cellular network and D2D channel selection is different. Therefore, Eqs. (15) and (16) can be used as channel selection measures. The larger the probability value, the easier it is for the user to choose.

3.3 Algorithm description

According to the above mapping rules, the specific implementation steps are as follows: In the first step, the D2D user first initializes the instantaneous return value Re_{dt_p} and the cumulative return value Re_{ct_p} according to the current state s. In the second step, the action α is arbitrarily selected from the action set $\tau\{0,1,\ldots,N\}$. By calculating Eqs. (15) and (16), the larger the probability value, the easier it is for the user to choose. In the third step, the mobile user's next state s' is used to calculate the total return value $Re(s,a)$ according to Eq. (14). The fourth step is to calculate the value of the Q function and continue to perform the above steps until it converges to the optimal strategy to obtain the maximum cumulative return value. which is

$$Q_t(s,a) = Q_{(t-1)}(s,a) + \delta\left[Re\_t(s,a) + \varepsilon \max Q_{(t-1)}(s',a') - Q_{(t-1)}(s,a)\right] \tag{18}$$

where $0 < \delta < 1$, it represents the learning rate. The specific algorithm is shown in Algorithm 1. Fig. 2 illustrates the flowchart of the presented algorithm to better indicates the stepwise process.

Algorithm 1: Proposed algorithm

1: **Initialize:** For state $s \in S$ and action $\alpha \in A$
   $Q(s,a) = 0$
   Initialize the reward value $R(s,a)$
2: End
3: **Learning:**
   Generate a binary random number of rand(.) for all users
4: **if** rand(.) $< \varepsilon$
5: Calculate the action a by Eq. (15) and (16)

6: Choose a channel with a higher probability for access
7: else
8: Choose the action to make Q reach the maximum
9: End **if**
10: Perform action a according to Eq. (14)
11: Calculate the total return value $R(s, a)$
12: Observe the next state s', according to Eq. (18)
13: Update the Q list
14: End

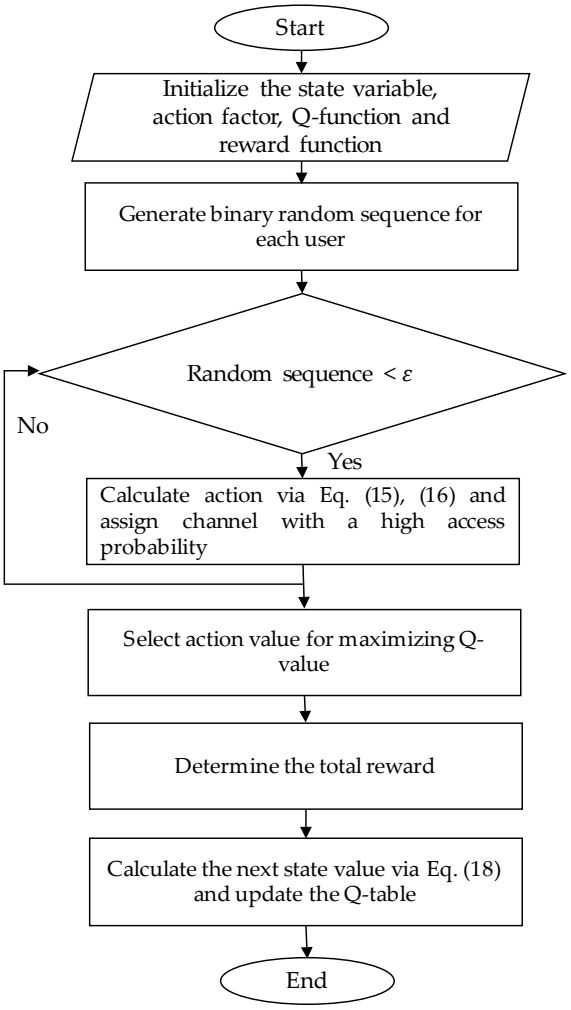

Figure 2: Algorithm flowchart

4 Simulation results and performance analysis

In this section, a computer simulation method is used to evaluate the Q-based power control and channel selection joint optimization algorithm proposed in this paper for D2D communication. The simulation platform used OPNET's LTE-TDD network simulation package [Ding, Lei, Karagiannidis et al. (2017)]. The simulation parameters are shown in Tab. 1. Compare the performance of our schemes with the following three schemes: The first type uses D2D communication with random access and is called Random; the second type uses macro base station communication and is called All to BS. The third one is the greedy algorithm for channel selection [Cao, Li, Zhao et al. (2017)], which is denoted as Y-greed, and the parameter Y=0.5. First, we compare the trend of the average utility (i.e., the average throughput of D2D user pairs) with Q-learning convergence time at different Boltzmann [Cao, Li, Zhao et al. (2017)] temperatures. The results are shown in Fig. 3. The Boltzmann temperature τ affects the convergence rate of Q-Learning. When the value of τ is small, that is, the Boltzmann temperature is relatively low, Q-Learning has a fast convergence speed and a large channel selection probability. On the other hand, as τ increases, the probability distribution of channel selection is almost the same. At any value of τ, the average utility is less than the maximum possible throughput $Th_{max} = 120$ kbps.

Table 1: Simulation parameters

Parameter	Value
Packet Length	1200 byte
BS cell radius	500 m
BS Spectrum Bandwidth	20 MHz
Frame structure	Type 2
Channel Bandwidth B	200 MHz
BS power	46 dBm
Noise power δ^2	-128 dBm
Cellular user average power p_c	23 dBm
BS antenna gain	16 dBi
User antenna gains	4 dBi

Fig. 4 shows the trend of the average convergence time of Q-Learning for different numbers of D2D users. It can be seen that when the Boltzmann temperature τ is low, the convergence time of the proposed algorithm is equivalent to the greedy algorithm, and the temperature rises and the convergence speed decreases, and it takes more time to achieve convergence. Fig. 5 shows the trend of total user throughput for different numbers of D2D users. It can be seen from Fig. 5 that our proposed Q-Learning solution has the largest user throughput and is a greedy algorithm that is followed, and the All to BS scheme has the lowest throughput. This is because All to BS does not take advantage of the performance gain brought by D2D communication. All users communicate through the base station, which will inevitably lead to an increase in network load and congestion and a decrease in throughput. In random mode, the performance gain is not obvious due

to the randomness of the user's choice of D2D communication. It can also be seen from Fig. 5 that our proposed Q-Learning scheme brings an increase in system throughput, which is close to the greedy algorithm, but the algorithm complexity is much lower than the greedy algorithm. Also, as the Boltzmann temperature τ decreases, the number of users choosing D2D links increases, so the average throughput of users also increases.

Fig. 6 shows the trend of user throughput of different power control algorithms in comparison with our proposed Q-based resource optimization algorithm under different minimum signal-to-interference and noise ratios (SINRs) (that is, different channel states). When the user selects channel j for $(U_m, U_{m'})$ at time t, if channel j is selected, $c_j^m = 1$, otherwise $c_j^m = 0$, so the SINR can be expressed as

$$SINR_{\min} = \min\left(SINR_{j,m}\right) = \min\left(\frac{p^m c_j^m G_{m,m'}^j}{\sum_{n\in M} p_d^n c_j^n G_{n,m'}^j + \delta^2}\right), \forall\, m \in M, j \in \kappa \tag{19}$$

It can be seen from Fig. 6 that user throughput is a convex function of $SINR_{\min}$. At the same $SINR_{\min}$, the All to BS algorithm has the lowest system throughput. The reason is that when the $SINR_{\min}$ is low, the channel state is poor and the user throughput is small. However, when $SINR_{\min}$ is too high, the number of channels that meet the user's quality of service requirements will also decrease the throughput. Therefore, under $SINR_{\min}$, the user throughput function is convex.

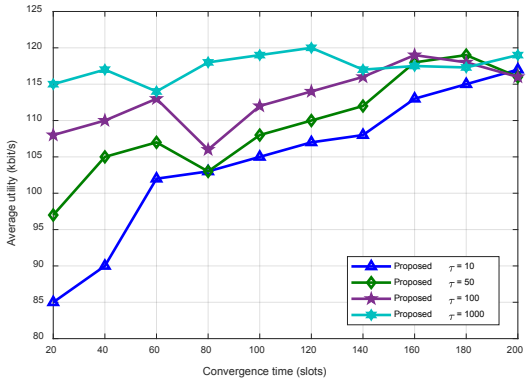

Figure 3: Comparison of the average utility of the proposed algorithm versus convergence time under different τ

Fig. 7 is a comparison of the average transmission power of different power control algorithms when the number of users is different. Since different D2D links work in different frequency bands, they do not interfere with each other. As can be seen from Fig. 7, the average transmission power of the proposed algorithm is significantly smaller than other algorithms and very close to the greedy algorithm, but the complexity is much lower than the greedy algorithm.

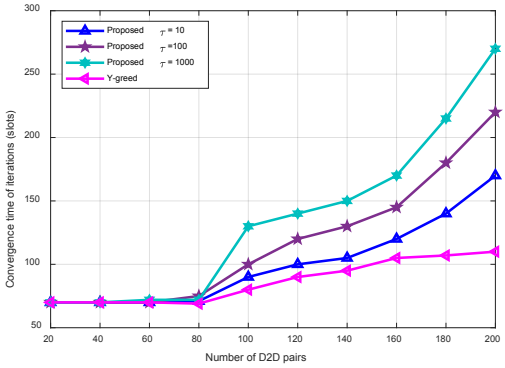

Figure 4: Comparison of the convergence time of the proposed algorithm and existing algorithm under a different number of D2D pairs

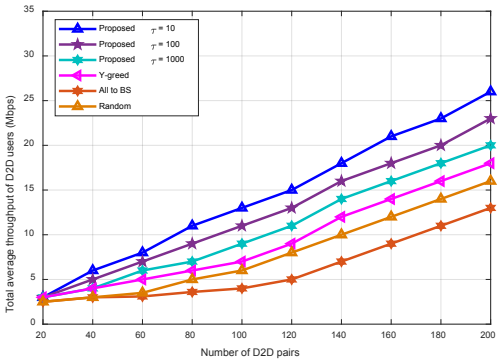

Figure 5: Comparison of the average throughput *vs.* the number of D2D user pairs of the algorithms

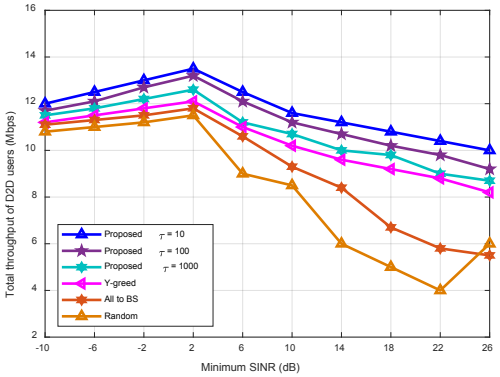

Figure 6: Comparison of the total throughput of D2D users of the algorithms versus different SINR levels

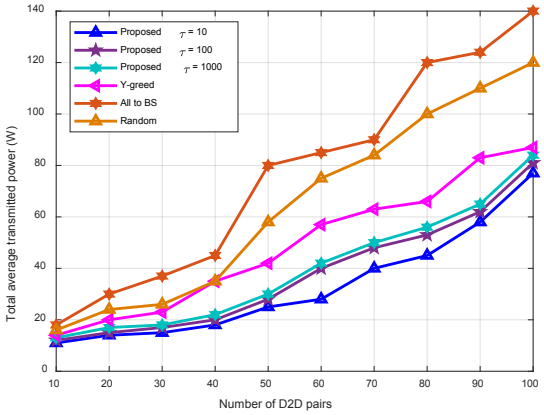

Figure 7: Comparison of the transmission power of the algorithms with an increasing number of D2D users

Tab. 2 compares the computational complexity of the proposed algorithm with conventional algorithms. It is clear from the results that the proposed algorithm has lower computational complexity as compared with the conventional algorithms. It is also clear that the proposed algorithm has also better complexity than the genetic algorithm (GA).

Table 2: Complexity comparison

Algorithm	Complexity
Random	$\mathcal{O}(M^3 + M \times N)$
All to BS	$\mathcal{O}(8M^3 \times N + N)$
Y-greed [Cao, Li, Zhao et al. (2017)]	$\mathcal{O}(M^2 \times N)$
Genetic Algorithm	$\mathcal{O}(M \times N^2 + M)$
Proposed	$\mathcal{O}(M \times N + M)$

5 Conclusions and future recommendations

With the continuous development of Internet technology, the number of users and network traffic has exploded. The era of Internet-of-things and big data is emerging. How to greatly increase the network capacity has become the biggest problem facing the development of wireless networks. D2D communication allows direct communication between user equipment at close distances, which can improve the throughput of the system, obtain high spectral efficiency and energy efficiency, and achieve a multiple of the wireless network capacity. It is regarded as one of the most promising new technologies for future wireless communication systems. This paper introduces the idea of reinforcement learning into D2D communication power control research and proposes a joint resource allocation and power control algorithm based on Q-learning in D2D and cellular heterogeneous networks. The proposed algorithm maps the power control problem into a Q-Learning problem in rate as a return function, the instantaneous change of rate can intuitively and accurately reflect the instantaneous change of throughput and

the occupation of the cellular network. Through continuous learning, adjust the power of D2D users, get the Q value table of system throughput, and choose the action with the highest Q value as the final act, and finally obtain the joint optimal strategy of channel selection and power control. Under the premise of ensuring the quality of service for cellular users, the maximum system throughput is obtained through D2D power control. Simulation results show that the proposed algorithm can maximize the system throughput and ensure the overall performance of the network. Future directions as an extension to the proposed study are to consider the interference analysis and integration of mmWave communications and evaluate the performance under different usage scenarios.

Acknowledgement: The authors would like to thanks the reviewers for their time and review.

Availability of Data and Materials: The data used for the findings of this study is available upon request from the corresponding authors.

Funding Statement: This work is supported by SUT research and development fund. The publication of this article was funded by the Qatar National Library. Seyed Sajad Mirjavadi also appreciates the help from the Fidar Project Qaem Company (FPQ).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Ahmad, M.; Li, Y.; Waqas, M.; Sheraz, M.; Jin, D. et al. (2018): A survey on socially aware device-to-device communications. *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2169-2197.

Alemaishat, S.; Saraereh, O. A.; Khan, I.; Choi, B. J. (2019): An efficient resource allocation algorithm for D2D communications based on NOMA. *IEEE Access*, vol. 7, pp. 120238-120247.

Alemaishat, S.; Saraereh, O. A.; Khan, I.; Affess, S. H.; Li. X. et al. (2019): An efficient precoding scheme for millimeter-wave massive MIMO systems. *Electronics*, vol. 8, no. 9, pp. 1-15.

Alfa, A. S.; Maharaj, B. T.; Lall, S.; Pal, S. (2016): Mixed-integer programming based techniques for resource allocation in underlay cognitive radio networks: a survey. *Journal of Communications and Networks*, vol. 18, no. 5, pp. 744-761.

Asheralieva, A.; Miyanaga, Y. (2016): Multi-agent Q-learning for autonomous D2D communication. *IEEE International Symposium on Intelligent Signal Processing and Communication Systems*, Phuket, Thailand, pp. 1-6.

Bakht, K.; Jameel, F.; Ali, Z.; Khan, W. U.; Khan, I. et al. (2019): Power allocation and user assignment scheme for beyond 5G heterogeneous networks. *Wireless Communications and Mobile Computing*, ID 2472783, pp. 1-11.

Chen, H.; Li, Y.; Jiang, Y; Ma, Y.; Vucetic, B. (2015): Distributed power splitting for SWIPT in relay interference channels using game theory. *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 410-420.

Chen, Q.; Yu, G.; Shan, H.; Maaref, A.; Li, G. et al. (2016): Cellular meets WiFi: traffic offloading or resource sharing? *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3354-3367.

Cao, C. H.; Li, Y.; Zhao, Y. L.; Chen, S. (2017): A two-level game theory approach for joint relay selection and resource allocation in network coding assisted D2D communications. *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2697-2711.

Ding, Z.; Lei, X.; Karagiannidis, G. K.; Schober, R.; Yuan, J. et al. (2017): A survey on non-orthogonal multiple access for 5G networks: research challenges and future trends. *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181-2195.

Ge, H.; Song, Y.; Wu, C.; Ren, J.; Tan, G. (2019**)**: Cooperative deep Q-learning with Q-value transfer for multi-intersection signal control. *IEEE Access*, vol. 7, pp. 40797-40809.

Huang, Y.; Nasir, A. A.; Durrani, S.; Zhou, X. (2016): Mode selection, resource allocation, and power control for D2D-enabled two-tier cellular network. *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3534-3547.

Jameel, F.; Hamid, Z.; Jabeen, F.; Zedally, S.; Javed, M. A. et al. (2018): A survey of device-to-device communications: research issues and challenges. *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2133-2168.

Jabeen, T.; Ali, Z.; Khan, W.; Jameel, F.; Khan, I. et al. (2019): Joint power allocation and link selection for multi-carrier buffer aided relay network. *Electronics*, vol. 8, no. 6, pp. 1-13.

Jameel, F.; Risaniemi, T.; Khan, I.; Lee, B. M. (2019): Simultaneous harvest-and-transmit ambient backscatter communications under rayleigh fading. *EURASIP Journal on Wireless Communications and Networking*, vol. 166, pp. 1-9.

Ji, M. Y.; Caire, G.; Molisch, A. F. (2016): Wireless device-to-device caching networks: basic principles and system performance. *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176-189.

Jiang, C. X.; Zhang, H. J.; Ren, Y.; Han, Z.; Chen, K. et al. (2017): Machine learning paradigms for next-generation wireless networks. *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98-105.

Jiang, J. F.; Tang, L. Y.; Gu, K.; Jia, W. J. (2020): Secure computing resource allocation framework for open fog computing. *The Computer Journal*, vol. 63, no. 4, pp. 567-592.

Khan, I.; Rodrigues J.; Al-Muhtadi, J.; Khattak, M. I.; Khan, Y. et al. (2019): A robust channel estimation scheme for 5G massive MIMO systems. *Wireless Communications and Mobile Computing*, ID 3469413, pp. 1-8.

Kiumarsi, B.; Vamvoudakis, K. G.; Modares, H.; Lewis, F. L. (2018): Optimal and autonomous control using reinforcement learning: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2042-2062.

Lee, B. M.; Patil, M.; Hunt, P.; Khan, I. (2019): An easy network onboarding scheme for the internet of things networks. *IEEE Access*, vol. 7, pp. 8763-8772.

Li, J. D.; Huang, S. (2017): Delay-aware power control for D2D communication with successive interference cancellation and hybrid energy source. *IEEE Wireless Communications Letters*, vol. 6, no. 6, pp. 806-809.

Lin, M.; Ouyang, J.; Zhu, W. P. (2016): Joint beamforming and power control for device-to-device communications underlaying cellular networks. *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 138-150.

Liu, J.; Yang, Y. H.; Lv, S. Q.; Wang, J.; Chen, H. (2019): Attention-based BiGRU-CNN for Chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, pp .1-12.

Luo, Y. J.; Qin, J. H.; Xiang, X. Y.; Tan, Y.; Liu, Q. et al. (2020): Coverless real-time image information hiding based on image block matching and dense convolutional network. *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp.125-135.

Li, W. J.; Cao, Y. X.; Chen, J.; Wang, J. X. (2017): Deeper local search for parameterized and approximation algorithms for maximum internal spanning tree. *Information and Computation*, vol. 252, pp. 187-200.

Li, H. X.; Li, W. J.; Zhang, S. G.; Wang, H. D.; Pan, Y. et al. (2019): Page-sharing-based virtual machine packing with multi-resource constraints to reduce network traffic in migration for clouds. *Future Generation Computer Systems*, vol. 96, pp. 462-471.

Maghsudi, S.; Stanczak, S. (2015): Joint selection and power control in insfrastructureless wireless networks: a multiplayer multiarmed bandit framework. *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4565-4578.

Maghsudi, S.; Stanczak, S. (2015): Channel selection for network-assisted D2D communications via no-regret bandit learning with calibrated forecasting. *IEEE Transactions on Wireless Communications*, vol. 14, no. 3, pp. 1309-1322.

Ni, W.; Collings, I. B.; Lipman, J.; Wang, X.; Tao, M. et al. (2015): Graph theory and its applications for future network planning: software-defined online small cell management. *IEEE Wireless Communications*, vol. 22, no. 1, pp. 52-60.

Orsino, A.; Ometo, A.; Fodor, G.; Moltchanov, D.; Militano, L. et al. (2017): Effects of heterogeneous mobility on D2D-and drone-assisted mission-critical MTC in 5G. *IEEE Communications Magazine*, vol. 55, no. 2, pp. 79-87.

Peng, H. X.; Li, D.; Abboud, K.; Zhou, H.; Zhao, H. et al. (2017): Performance analysis of IEEE 802. 11p DCF for multiplatooning communications with autonomous vehicles. *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2485-2498.

Ren, Y.; Liu, F. Q.; Liu, Z.; Wang, C.; Ji, Y. et al. (2015): Power control in D2D-based vehicular communications networks. *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5547-5562.

Shaikh, F. S.; Weissmuller, R. (2018): Routing in multi-hop cellular device-to-device (D2D) networks: a survey. *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2622-2657.

Saraereh, O. A.; Mohammed, S. L.; Khan, I.; Rabie, K.; Affess, S. (2019): An efficient resource allocation algorithm for device-to-device communications. *Applied Sciences*, vol. 9, no. 18, pp. 1-15.

Saraereh, O. A.; Alsaraira, A.; Khan, I.; Uthansakul, P. (2019): An efficient eesource allocation algorithm for OFDM-based NOMA in 5G systems. *Electronics*, vol. 8, no. 12, pp. 1-13.

Saraereh, O. A.; Alsaraira, A.; Khan, I.; Choi, B. J. (2020): A hybrid energy harvesting design for on-body internet-of-things (IoT) networks. *Sensors*, vol. 20, no. 2, pp. 1-16.

Shafi, M.; Molisch, A. F.; Smith, P. J.; Haustein, T.; Zhu, P. et al. (2017): 5G: a tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201-1221.

Saraereh, O. A.; Khan, I.; Lee, B. M., Tahat, A. (2019): Efficient pilot decontamination schemes in 5G massive MIMO systems. *Electronics*, vol. 8, no. 1, pp. 1-26.

Stamou, A.; Dimitriou, N.; Kontovasilis, K.; Papavassiliou, S. (2019): Autonomic handover management for heterogeneous network in a future internet context: a survey. *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3274-3297.

Salehi, M.; Mohammadi, A.; Haenggi, M. (2017): Analysis of D2D underlaid cellular networks: SIR meta distribution and mean local delay. *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 2904-2916.

Sun, P.; Shin, K. G.; Zhang, H. L.; He, L. (2017): Transmit power control for D2D-underlaid cellular networks based on statistical features. *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4110-4119.

Sakr, A. H.; Hossain, E. (2015): Cognitive and energy harvesting-based D2D communication in cellular networks: stochastic geometry modeling and analysis. *IEEE Transactions on Communications*, vol. 63, no. 5, pp. 1867-1880.

Sun, R. X.; Shi, L. F.; Yin, C. Y.; Wang, J. (2019): An improved method in deep packet inspection based on regular expression. *The Journal of Supercomputing*, vol. 75, no. 6, pp. 3317-3333.

Wang, X. F.; Zhang, Y. H.; Leung, V. C. M.; Guizani, N.; Jiang, T. (2018): D2D big data: content deliveries over wireless device-to-device sharing in large-scale mobile. *IEEE Wireless Communications*, vol. 25, no. 1, pp. 32-38.

Wang, L.; Tang, H.; Wu, H.; Stuber, G. L. (2017): Resource allocation for D2D communications underlay in rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1159-1170.

Wang, Q.; Wang, W.; Jin, S.; Zhu, H. (2015): Quality-optimized joint source selection and power control for wireless multimedia D2D communication using stackelberg game. *IEEE Transaction on Vehicular Technology*, vol. 64, no. 8, pp. 3755-3769.

Wang, W.; Jiang, Y. B.; Luo, Y. H.; Li, J.; Wang, X. et al. (2019): An advanced deep residual dense network (DRDN) approach for image super-resolution. *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1592-1601.

Xu, H.; Huang, N.; Yang, Z.; Shi, J.; Wu, B. et al. (2017): Pilot allocation and power control in D2D underlay massive MIMO systems. *IEEE Communications Letters*, vol. 21, no. 1, pp. 112-115.

Yang, C. G.; Li, J. D.; Semasinghe, P.; He, L. (2017): Distributed interference and energy-aware power control for ultra-dense D2D networks: a mean field game. *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1205-1217.

Yuan, Y.; Yuang, T.; Feng, H.; Hu, B. (2019): Learning for matching game in cooperative D2D communication with incomplete information. *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7174-7178.

Zhang, X. M.; Zhang, Y.; Yan, F.; Vasilakos A. V. (2015): Interference-based topology control algorithm for delay-constrained mobile ad hoc networks. *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 742-754.

Zeng, D. J.; Dai, Y.; Li, F.; Wang, J.; Sangaiah, A. K. (2019): Aspect based sentiment analysis by a linguistically regularized CNN with gated mechanism. *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 3971-3980.

Zhang, J. M.; Jin, X. K.; Sun, J.; Wang, J.; Sangaiah, A. K. (2018): Spatial and semantic convolutional features for robust visual object tracking. *Multimedia Tools and Applications*, https://doi.org/10.1007/s11042-018-6562-8.

Zhang, J. M.; Wang, W.; Lu, C. Q.; Wang, J.; Sangaiah, A. K. (2019): Lightweight deep network for traffic sign classification. *Annals of Telecommunications*, https://doi.org/10.1007/s12243-019-00731-9.

Zhang, Z.; Li, Y. B.; Wang, C.; Wang, M. Y.; Tu, Y. et al. (2018): An ensemble learning method for wireless multimedia device identification. *Security and Communication Networks*, https://doi.org/10.1155/2018/5264526.