Peer-to-Peer Networking and Applications
https://doi.org/10.1007/s12083-021-01146-x

Deep reinforcement learning-based incentive mechanism design for short video sharing through D2D communication

Zhuo Li[1] · **Wentao Dong**[1] · **Xin Chen**[1]

Received: 30 November 2020 / Accepted: 30 March 2021

Abstract

With the development of 5th generation (5G) wireless communication networks and the popularity of short video applications, there has been a rapid increase in short video traffic in cellular networks. Device-to-device (D2D) communication-based short video sharing is considered to be an effective way to offload traffic from cellular networks. Due to the selfish nature of mobile user equipment (MUEs), how to dynamically motivate MUEs to engage in short video sharing while ensuring the Quality of Service, which makes it critical to design an appropriate incentive mechanism. In this paper, we firstly analyze the rationale for dynamically setting rewards and penalties and then define the rewards and penalties setting dynamically for maximizing the utility of the mobile edge computing server (RPSDMU) problem. The problem is proved NP-hard. Furthermore, we formulate the dynamic incentive process as the Markov Decision Process problem. Considering the complexity and dynamics of the problem, we design a Dynamic Incentive Mechanism algorithm of D2D-based Short Video Sharing based on Asynchronous Advantage Actor-Critic (DIM-A3C) to solve the problem. Simulation results show that the proposed dynamic incentive mechanism can increase the utility of mobile edge computing server by an average of 22% and 16% compared with the existing proportional incentive mechanism (PIM) and scoring-based incentive mechanism (SIM). Meanwhile, DIM-A3C achieves a higher degree of satisfaction than PIM and SIM.

Keywords Short video sharing · D2D communications · Dynamic incentive mechanism · Deep reinforcement learning

1 Introduction

Recently, with the rapid development of communication technology and the continuous updating of mobile terminal equipment, many short video applications (such as Douyin [1], Youtube Go [2], KuaiShou [3], and so on) are developing at an astonishing speed. Short video applications have become the mainstream applications on the Internet today, which occupy a very large network bandwidth [4] and the trend is expected to continue [5]. The above-mentioned applications have resorted to employ edge caching servers to store and deliver the massive short videos, so as to avoid that all requests have to be fetched from the backend/origin server, which usually introduces extra user-perceived latency [6]. Edge caching not only improves users' Quality of Experience (QoE), it also effectively offloads the backhaul traffic [7]. However, there is still a large amount of redundant traffic on the wireless side, which brings a heavy burden on the cellular network. Against such a background, wireless traffic in cellular networks can be effectively offloaded and the energy efficiency and spectral efficiency of the network can be improved by using D2D communication technology and short video sharing between mobile devices [8–11].

Most previous research on D2D communication has focused on resource allocation and interference management, an implicit assumption under which is that MUEs would always deliver content to others unconditionally [12]. However, the selfishness of MUEs may make them reluctant to engage in D2D-based short video sharing due to

This article is part of the Topical Collection: *Special Issue on Convergence of Edge Computing and Next Generation Networking* Guest Editors: Deze Zeng, Geyong Min, Qiang He, and Song Guo

✉ Zhuo Li
lizhuo@bistu.edu.cn

Wentao Dong
15634130256@163.com

[1] School of Computer Science, Beijing Information Science and Technology University, Beijing, China

Published online: 06 May 2021

the consumption of power and computing resources that exists when MUEs participate in content sharing [13]. The conflict among MUEs and the tension between the base station and the MUEs are formulated as a Stackelberg game and a fixed incentive scheme is derived in [14]. Due to the fact that Users have different levels of interest in shared content and the state is constantly changing as the process proceeds, it is not efficient to set a fixed incentive. Therefore, one of the main challenges faced by D2D-based short videos sharing is: When MUEs succeed or fail in D2D-based short video sharing, how to dynamically set rewards or penalties to motivate MUEs to actively participate in content sharing and at the same time constrain their behavior. Edge computing [15, 16] as an extension of the cloud is an emerging computing platform. Similar to the cloud, edge computing assists the MUEs by providing computing resources, data storage, and application services to possess location awareness, maintain low latency, support heterogeneity, and improve QoS of the applications, especially the compute-intensive and delay-sensitive ones [17]. In this paper, based on edge computing, we focus on the dynamic rewards and penalties of D2D-based short video sharing in cellular networks, with the goal of maximizing the utility on the network side. Meanwhile, we propose a dynamic incentive mechanism of D2D-based short video sharing based on deep reinforcement learning (DRL) to solve the challenge mentioned above.

Our contributions are summarized as followed:

1. We focus on setting up rewards and punishments dynamically for maximizing the utility of the MEC server. The rewards and punishments setting dynamically for maximizing the utility of the MEC server (RPSDMU) problem is formulated and proved NP-hard.
2. We analyze the interference model of the network during D2D-based short video sharing and the definition of the type of short video content provider (CP), and use the channel conditions of D2D-based short video sharing, the number of CPs in that cellular network at a certain time and the busyness of MUE, etc., as the basis for setting rewards and penalties. For example, as the number of CPs in a cellular network gradually increases, the process of short videos sharing will be easier and therefore the rewards should be appropriately reduced on the network side.
3. Based on the analysis of interference models and type of CPs in a cellular network, we formulate the incentive process as a Markov Decision Process framework and propose the Dynamic Incentive Mechanism algorithm of D2D-based Short Video Sharing based on Asynchronous Advantage Actor-Critic (DIM-A3C) to optimize it dynamically with the objective of maximizing the utility of the MEC server.
4. Thorough experiments are taken to verify the effectiveness of the DIM-A3C. Simulation results reveal that the proportional DIM-A3C can increase the utility of the MEC server by an average of 22% and 16% compared with the existing PIM and SIM. Meanwhile, DIM-A3C achieves a higher degree of satisfaction than PIM and SIM.

The structure of this paper is as follows. We review the related work in Section 2. Section 3 describes the system model. The Section 4 gives the dynamic incentive mechanism. We formulate the problem and apply a DRL algorithm to solve the problems in Section 5, then we provide simulation experiments and analysis in Section 6. Finally, we draw the conclusion in Section 7.

2 Related work

Due to the incredible increase in the number of mobile terminal devices, the lightning diffusion of cloud-based services, and the rapid development of a large number of short video applications, the network traffic in mobile communication systems has increased exponentially. In addition, the simultaneous access of a mass of mobile terminal devices and the emerging new services make the current communication system face many challenges. As a new generation communication standard, 5G [18] has received extensive attention from many researchers. D2D-based communication technology, as one of the key technologies for 5G communication systems, can not only solve the problems caused by the extremely large traffic in 5G communication systems but also effectively reduce latency and meet the needs of computing marginalization [19]. For D2D content sharing, it is usually controlled by the cellular network to offload peak-time backhaul traffic and effectively improve quality of service, while mobile user devices' willingness of content sharing makes a difference to the system performance [20]. Therefore, in D2D-based content sharing, incentive mechanisms have been extensively studied, which play an important role [21].

The current incentive mechanism design schemes could be generally classified into two categories, one is the price-based, the other is the reputation-based. In the price-based mechanisms [22–24], MUEs will view the D2D-based content sharing process as a market transaction according to economic related theories when they participate in D2D content sharing. In [24], the incentive mechanism puts

a price on each D2D process and pays the MUEs who successfully share the content, thereby motivating more MUEs to actively participate in the D2D-based sharing process. In the reputation-based mechanisms [25, 26], incentive mechanisms evaluate MUEs according to their historical behavior records. In the past period, the more successful times of participating in D2D-based content sharing, the higher the reputation of the MUEs. However, none of the incentive mechanisms mentioned in the above studies can provide continuous and dynamic incentives for the D2D network to maximize the economic utility of the network side.

Deep reinforcement learning combines the perceptual abilities of deep learning with the decision-making abilities of reinforcement learning [27], which has been widely used in incentive mechanism design [28–30]. In [27], Zeng et al. introduce a model-free DRL approach to efficiently manage the resources at the network edge. In [29], to maximize the overall utility of vehicle drivers, Zhao et al. propose a social-aware incentive mechanism by deep reinforcement learning, to derive the optimal long-term sensing strategy for all vehicles. In [30], Zhan et al. study the deep reinforcement learning-based incentive mechanism for federated learning to motivate edge nodes to contribute model training. DRL has also been used in D2D content caching and sharing. In [31], Zhang et al. build the model of cache sharing and transaction execution consensus, and they further formulate cache placement and scene selection as Markov Decision Process problems. DRL could be classified into two basic methods based on value function and strategy gradient. Since value-based algorithms are generally applicable for problems with a discrete action-space [32], Chen et al. [33] apply a policy-based algorithm, namely proximal policy optimization (PPO) to obtain the continuous incentive in the D2D content sharing. Due to the uncertainty of coming system states, the problem of setting the appropriate rewards or penalties is more complicated. Exiting works also consider designing dynamic incentives to maximize the system gains [34–36]. However, all of the above work only considered the incentive mechanism of general content in D2D sharing, but does not consider the specific scenario of a short video application, and does not consider the impact of user interest on the incentive mechanism.

3 System model

In this paper, we consider a cellular cell scenario where there are multiple MUEs within the coverage of a single MEC server. The D2D-based short video sharing model is shown in Fig. 1. We only consider the sharing process of one short video on the network. A certain number of traditional cellular mobile user equipment (CMUEs) and D2D mobile user equipment (DMUEs) are randomly distributed in the cellular cell. The short video sharing process of DMUEs multiplexes the uplink of CMUEs. DMUEs can be classified into short video providers (CPs) and receiver (CRs). CPs have obtained the short video, and CRs are DMUEs who have not received content. Formally, let $\mathbb{M} = \{CP_1, \cdots, CP_m, \cdots, CP_M\}$ be a set of CPs that describes M CPs. $\mathbb{N} = \{CR_1, \cdots, CR_n, \cdots, CR_N\}$ denotes a set of CRs that describes N CRs.

3.1 Channel interference model

In this paper, in order to analyse the impact of the channel interference model on the incentive mechanism in more detail, we mainly consider the Path Loss fading and independent-flat Rayleigh fading in the channel model. We assume that all wireless links experience independent-flat Rayleigh fading at the same frequency and the Path Loss [37]. The Path Loss is inversely proportional to the communication distance with a loss factor of α. Let $h_{m,n}$ be Rayleigh fading coefficient between CP_m and CR_n, $d_{m,n}$ be the distances from CP_m to short video receiver CR_n, p_m^T be the transmitting power of CP_m. Consequently, the received power of CR_n receives the short video sent by CP_m is $p_m^T d_{m,n}^{-\alpha} h_{m,n}$. As DMUEs multiplex the uplink transmission channels of CMUEs, a DMUE receives interference from one CMUE and other DMUEs [38]. When CP_m shares a short video with CR_n via D2D, the total interference of CR_n can be expressed as:

$$I_{m,n} = p_c^T d_{c,n}^{-\alpha} h_{c,n} + \sum_{CP_k \in \mathbb{M}, k \neq m} p_k^T d_{k,n}^{-\alpha} h_{k,n}, \tag{1}$$

where p_c^T and p_k^T are the transmitting power of CMUE c and CP_k, respectively. In cellular communication networks, Qos can accurately reflect the quality of communication of MUEs, and Signal to Interference plus Noise Ratio (SINR) can be an important indicator to ensure Qos. SINR is given by:

$$r_{m,n} = \frac{p_m d_{m,n}^{-\alpha} h_{m,n}}{\mathrm{I}_{m,n} + \sigma^2}, \tag{2}$$

where σ^2 denotes the power of additive white Gaussian noise. According to the Shannon Theory, the short video content transmission rate between CP_m and CR_n can be expressed as:

$$R_{m,n} = B\log_2(1 + r_{m,n}), \tag{3}$$

where B is channel bandwidth of the cellular network.

3.2 Definition of CP types

Considering that different short video CPs or the same CP have different content sharing capabilities in different

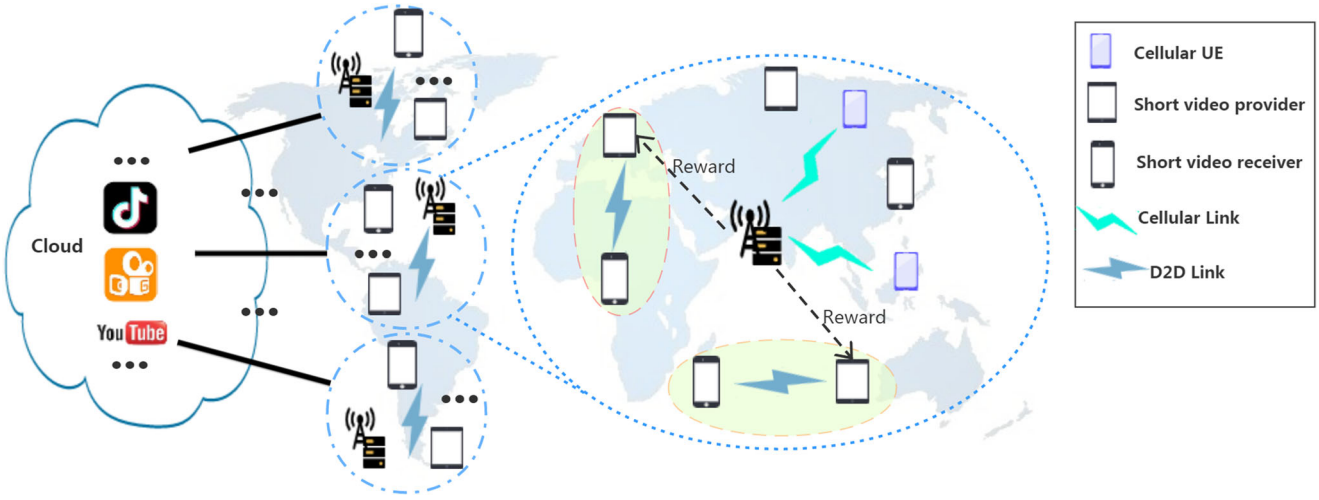


Fig. 1 The D2D-based short video sharing model

periods, we should dynamically set different rewards for CPs that complete the D2D-based short video sharing task. For CP_m and CR_n that are about to a share short video, the quality of completing the D2D-based sharing task depends on the D2D channel conditions, the remaining power E_m of the CP_m, and the busyness l_m of the CP_m. For example, the higher CP_m's remaining power E_m and the lower CP_m's busyness l_m, the less cost paid by the CP_m to complete short video sharing. The type $\vartheta_{m,n}$ of CP_m refer to CR_n is defined as:

$$\vartheta_{m,n} = \frac{d_{m,n}^{-\alpha} h_{m,n}}{I_{m,n} \psi_m l_m}, \tag{4}$$

where ψ_m is the price paid by the CP_m when it consumes a unit of electricity and ψ_m is expressed as:

$$\psi_m = \omega {E_m}^{-1}, \tag{5}$$

where ω is defined as the conversion coefficient that converts the current power value into the cost of the unit power consumption of the device.

3.3 Constructing the interest vector

When a user browses the short video content, the MEC server will record the specific details of the user's access. The historical records of users' access to different types of short video content can be used as a basis for discovering users' interests. In the process of using short video applications, if a user likes or forwards a short video he is watching, we think that the user is interested in the short video content. We uniformly refer to the user's operations such as likes and forwards as being approved by the user and define the user's interest in the short video content of type j as:

$$v_j = \frac{f_j}{F_j} \tag{6}$$

where f_j represents the number of user approvals for short videos of type j, F_j denotes the total number of short videos of type j browsed by the user. According to formula (6), we can get the interest vector of user u for various types of short videos:

$$\mathbb{V}^u = [v_1^u, \cdots, v_j^u, \ldots, v_J^u] \tag{7}$$

3.4 Process of D2D-based short video sharing

According to Section 3.3, we can obtain the vector $\mathbb{V}^{CR_n} = [v_1^{CR_n}, \cdots, v_j^{CR_n}, \ldots, v_J^{CR_n}]$ of CR_n's interest in various types of short video content, where $v_j^{CR_n}$ denotes the degree of CR_n's interest in the short video of type j. The process of implementing D2D-based short video sharing is generally as follows.

1. The CR_n sends its set of interest vector $\mathbb{V}^{CR_n}$ for each type of short video to the base station.
2. The base station records and broadcasts the information from the CR_n to the cell. When the distance between a CP and the CR_n is within the maximum range allowed for D2D communication and the CP is willing to participate in the D2D-based short video sharing process, the CP sends a signal to the base station. The signal sent includes the transmission rate between the CP and the CR_n, the information about the short video to be shared, and the device type. If multiple CPs preparing to share a short video to the CR_n at the same time, the base station selects a certain CP to

participate in sharing process according to the received signal, thereby maximizing the benefit of the system.
3. The base station sends the reward and penalty information to the CP. The MEC server on the base station side will calculate the reward for successful transmission and the penalty for failure based on the specific information of the CP.
4. After the CP receives the reward and penalty information, it calculates the self-utility this short video sharing will bring to itself. If the utility is greater than 0, the CP starts sharing the short video to the CR_n.

4 Dynamic incentive mechanism design

4.1 Dynamic reward model

In a cellular network, when CPs share short videos for other MUEs, they need to consume their power, storage, and computing resources. Therefore, the network side should consider giving the CPs corresponding rewards in a certain way. When a CP successfully shares a short video via D2D, the size of the reward given to the CP by the network side is related to the number of CPs in the cellular network. The state of the cellular network is defined as S_M when there are M CPs in the cellular network. Assuming that the type of short video to be shared is j, when the CP_m successfully shares the short video to the CR_n via D2D, the reward to the CP_m can be expressed as follows:

$$C_{m,n}^{s,M} = \frac{\bar{w}\eta\vartheta_{m,n}v_j^{CR_n}\mathrm{L(N+M)}}{\mathrm{L}_{\max}M} = \frac{\bar{w}\eta d_{m,n}^{-\alpha}h_{m,n}\mathrm{L}v_j^{CR_n}\mathrm{(N+M)}}{I_{m,n}\psi_m l_m \mathrm{L}_{\max}M} \tag{8}$$

where η is the conversion coefficient for converting user equipment type metrics into economic utility. L is the length of the short video successfully shared, and $\mathrm{L}_{\max}$ represents the maximum length of the short video that can be shared through D2D. β is a constant, it is the conversion factor that converts the proportion of the number of CPs in the cellular network into economic utility, and $\bar{w}$ denotes the weight to be optimized. It can be seen that the higher the CP type and the longer the length of the short video successfully shared, the higher the reward for the CP, which we call high-score CP. As the number of CPs in the cellular network increases, the process of D2D-based short video sharing becomes relatively easy, and therefore the rewards that CPs receive for successfully sharing a short video should decrease accordingly. Since the reward setting depends on the signal information sent by the CP to the MEC server-side base station, to prevent the CP from sending false signal information to obtain a higher reward, the MEC server should punish the dishonest CP accordingly [39], and the size of the punishment should be proportional to the reward size, i.e.,

$$C_{m,n}^{f,M} \propto C_{m,n}^{s,M}. \tag{9}$$

4.2 Utility model of MEC server

In a cycle S_M of incentive renewal, the utility of the MEC server is the difference between its benefit and the incentive cost paid.In the state S_M, the utility of MEC server is calculated as follows:

$$\mathcal{R}'(S_M) = C_p N_M^s + \sum_{CP_m\in\mathbb{M},CR_n\in\mathbb{N}} C_{m,n}^{f,M} - \sum_{CP_m\in\mathbb{M},CR_n\in\mathbb{N}} C_{m,n}^{s,M}, \tag{10}$$

where N_M^f denotes the times of successful sharing of a short video in a scheduling cycle of state S_M, C_p denotes the benefit that the MEC server would receive from a successful short video sharing.

4.3 Utility model of CP

The CPs are individually rational. Only when their utility is greater than 0, will the CPs share a short video based on D2D. Before the short video sharing in each scheduling period, the CP will evaluate the utility of this content sharing based on its specific situation and the size of the incentive that the MEC server will give.

Let P_H and P_L be the failure probability of high-score CP and low-score CP in D2D communication,respectively. $C_{m,n}^{f,M}$ and $C_{m',n'}^{f,M}$ respectively denote the penalties for high-score CP and low-score CP after D2D Communication failure,i.e. $C_{m',n'}^{f,M} < C_{m,n}^{f,M}$. In the state S_M, when CP_m shares a short video with CR_n, the cost to CP_m can be expressed as:

$$g_{m,n}^M = \frac{\sigma}{2}p_m^T. \tag{11}$$

Therefore, the utility of CP_m is the difference between its benefits and cost, as follows:

$$u_{m,n}^M = C_{m,n}^{s,M} - g_{m,n}^M - P_H C_{m,n}^{f,M}. \tag{12}$$

To effectively incentivize CPs to actively participate in the D2D-based short video content sharing process while ensuring the utility of the MEC server, the incentive mechanism should be designed to satisfy both IC and IR conditions.

Definition 1 IC: The CPs must send their real information to the base station on the MEC server-side instead of sending false information to get higher rewards,i.e.,

$$C_{m',n'}^{s,M} - g_{m',n'}^M - P_L C_{m',n'}^{f,M} \geq C_{m,n}^{s,M} - g_{m',n'}^M - P_L C_{m,n}^{f,M}, \tag{13}$$

$$C_{m,n}^{s,M} - g_{m,n}^{M} - P_H C_{m,n}^{f,M} \geq C_{m',n'}^{s,M} - g_{m,n}^{M} - P_H C_{m',n'}^{f,M}. \tag{14}$$

The left side of inequality (13) shows the utility obtained by the low-score CP after sending true signal information to the MEC server, and the right side of the inequality (13) represents the utility obtained by the low-score CP impersonating the high-score CP and sending false signal information to the MEC server. In other words, inequality (13) denotes that if a low-score CP pretends to be a high-score CP and sends false signal information to the MEC server-side base station, its utility will be reduced, so the low-score CP will not imitate the signal information sent by the high-score CP. Similarly, inequality (14) denotes that high-score CP will not pretend to be a low-score CP to send signal information to the MEC server-side base station. After satisfying the IC constraint, the CP will send its real signal to the MEC server-side base station according to its situation, instead of sending false signal information, which ensures the authenticity and validity of the signal.

Definition 2 IR: The utility of CPs in a cellular network after participating in the D2D-based short video sharing process is no less than 0,i,e.,

$$C_{m,n}^{s,M} - g_{m,n}^{M} - P_H C_{m,n}^{f,M} \geq 0. \tag{15}$$

Every individual engaged in economic activities is self-interested. Inequality (15) denotes that the CPs can get non-negative rewards if they actively participate in the D2D-based short video sharing process.

5 Dynamic incentives based on deep reinforcement learning

5.1 Modeling based on Markov decision process

State space Assuming that there are M CPs in state S_t, and at this time, K CPs intend to share short video content with different CRs in the cellular network. These CPs have sent their information to the MEC server-side base station and the MEC server agrees to their short video sharing.we denote:

$$S_t = \{\{R_1, \vartheta_1, \mathbb{V}^{CR_1}, M\}, \cdots, \{R_k, \vartheta_k, \mathbb{V}^{CR_k}, M\}, \cdots \{R_K, \vartheta_K, \mathbb{V}^{CR_K}, M\}\}, \tag{16}$$

where R_k indicate the short video content transmission rate $R_{m,n}$ of D2D-pair (CP_m, CR_n) whose index is k. θ_k is the device type $\vartheta_{m,n}$ of CP_m relative to CR_n. In the state S_t, we assume that the number of CPs in the cellular network is always M.

Action space

$$a_t = \{\{C_1^{s,M}, C_1^{f,M}\}, \cdots, \{C_k^{s,M}, C_k^{f,M}\}, \cdots, \{C_K^{s,M}, C_K^{f,M}\}\}, \tag{17}$$

where $C_k^{s,M}$ denotes the reward that CP_m can receive when a certain D2D-pair (CP_m, CR_n) whose index is k have successfully shared a short video, otherwise the reward is $C_k^{f,M}$.

Reward In state S_t, after taking action a_t, the agent of DRL system will feedback the immediate reward $\mathcal{R}_t$. We use the utility of the MEC server in an incentive update period S_M as the instant reward of deep reinforcement learning, i.e.

$$\mathcal{R}_t = \mathcal{R}'(S_M). \tag{18}$$

5.2 Optimization objective

From the perspective of the MEC server-side, the MEC server hopes to maximize its long-term utility by adopting different actions a_t according to an optimal control strategy π for different states S_t of CPs. Therefore, our optimization goal is to maximize the cumulative discount utility on the MEC server side of the entire short video sharing process, i.e.,

$$\max \sum_{t=0}^{T} \gamma \mathcal{R}_t, \tag{19}$$

where T is the end time of the short video content sharing process. When all DMUEs have obtained a certain short video content, the D2D-based sharing process ends. γ is the discount rate. The smaller the value of γ, the more important the network's recent revenue is than the future revenue.

5.3 Complexity of RPSDMU

Theorem 1 RPSDMU is an NP-hard problem.

Proof First we introduce the subset-sum problem (SUBSET-SUM), which is defined as follows: Given a finite set H of positive integers and an integer target $k > 0$, finding out whether there is a subset $\mathrm{H}' \subseteq \mathrm{H}$ whose sum of all elements is k. For example $\mathrm{H} = \{1, 2, 3, 14, 49, 98, 343, 686, 2409\}$ and $k = 3115$, so the subset $\mathrm{H}' = \{1, 2, 3, 14, 686, 2409\}$ is a solution of the problem.

$$\text{SUBSET-SUM} = \{< \mathrm{H}, k >: \text{there is a subset} \mathrm{H}' \subseteq \mathrm{H}, \text{and } k = \sum_{h \in \mathrm{H}'} h\} \tag{20}$$

□

Next, we reduce SUBSET-SUM to RPSDMU problem.

We construct an instance for the RPSDMU problem. Suppose that the rewards set by the MEC server for CP are all positive integers, and in the state S_t, after CP gets a reward of size C, the system must go to the next identical state S_{t+1}. In the RPSDMU problem, our goal is to maximize the utility of the network side, i.e., to minimize the cost of rewards expended on the network side. Assuming that the minimum cost is k, that is, we need to find a subset H′ in all possible reward sets H, so that the sum of all elements in the subset H′ is k.

Since the SUBSET-SUM problem is a special case of the RPSDMU problem and the SUBSET-SUM problem has been proved to be NP-hard, the RPSDMU problem is NP-hard. This completes the proof.

5.4 Deep reinforcement learning

In DRL, the agent interacts with the environment and guides the behavior with the obtained reward, so as to find an optimal strategy π. In the future interaction with the environment, the agent can obtain the maximum reward by making decisions according to the strategy π [40]. Algorithms for solving reinforcement learning problems can be classified as policy-based algorithms and value-based algorithms. The output of value-based algorithms is the values of all actions, and the agent will choose the action with the highest value. However, it only applies to the problem of discrete actions. Policy-based algorithms directly output the probabilities of various actions to be taken in the next step, and then take corresponding actions according to the probabilities [41], which makes it possible to select the appropriate action in continuous actions. However, policy-based algorithms typically converge to a local rather than the global optimum. In addition, the DRL algorithms mentioned above, such as DDQN [42], DDPG [43],and etc, all use the experience replay mechanism to eliminate the relevance of training data, but the experience replay mechanism has the following problems:

1) The acquisition of experience data relies on the interaction between the agent and the environment, and each interaction consumes a lot of memory and computing resources.

2) The experience replay mechanism requires the agent to learn using the off-policy approach, which can only update model parameters based on the data generated by the old policy.

In this paper, we apply an actor-critic algorithm, namely Asynchronous Advantage Actor-Critic (A3C) [44], as this algorithm uses an asynchronous training mechanism to make full use of computing resources and improve the training speed of the model. In A3C, the gradient update method of the actor network's parameter θ can be expressed as:

$$d\theta = d\theta + \nabla_\theta \log \pi_\theta(s_t, a_t) A(s,t) + c\nabla_\theta H(\pi(s_t, \theta)), \quad (21)$$

where H is the entropy, and the hyperparameter c controls the strength of the entropy regularization term. The advantage function $A(s,t)$ in Eq. 21 denotes the advantage

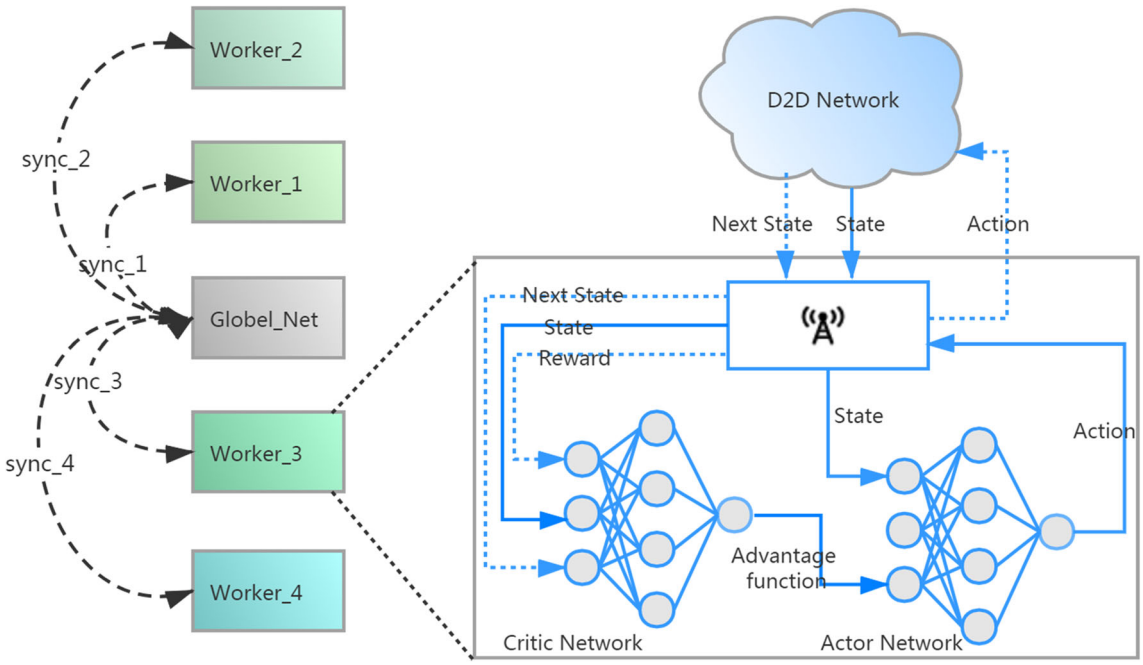


Fig. 2 Incentive mechanism of D2D-based short video sharing by DRL

of taking action a compared to the action suggested by policy π. To reduce the variance of sampled trajectories, $A(s,t)$ is usually defined as:

$$A(s,t) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}) - V(s_t), \tag{22}$$

where r_t represents the instant reward obtained in state s_t. $V(s_{t+k})$ and $V(s_t)$ denote the state value of state s_{t+k} and state s_t obtained by the critic network, respectively.

5.5 Dynamic incentive mechanism based on deep reinforcement learning

In order to effectively motivate CPs in the cellular networks to actively participate in D2D-based short video sharing, we design a Dynamic Incentive Mechanism algorithm of D2D-based Short Video Sharing based on Asynchronous Advantage Actor-Critic (DIM-A3C). As shown in Fig. 2, the DIM-A3C uses an asynchronous training framework. The global network is a public neural network model that includes both actor and critic components. Each worker runs in its own thread, and its neural network structure is the same as the global network. Each thread independently interacts with the environment to obtain experience data, and these threads do not interfere with each other. After the threads interact with the environments a certain amount of data, the loss function' ural network is calculated, and the calculated gradient is used to update the global neural network. Every once in a while, each worker replaces its parameters with global neural network parameters. The idea of asynchronous training solves the problem of slow convergence of traditional actor-critic algorithms. In addition, the DIM-A3C algorithm executes multiple agents asynchronously, which eliminates the correlation between training data through different states experienced by the agents in parallel, without the need for experience replay to store historical data, which can adapt to the optimization problems with a large state space. The combination of asynchronous training and actor-critic methods can not only reduce the calculation time, but also accommodate to the dynamic changes of the cellular network. In the D2D-based short video sharing process, after the CP sends a signal which includes the transmission rate, the information about the short video to be shared, and the device type to the base station, the base station would determine the reward and penalty according to the DIM-A3C algorithm. Consequently, the DIM-A3C algorithm is described in Algorithm1.

Algorithm 1 DIM-A3C algorithm.

Input: global shared parameter vectors θ and w, thread-specific parameter vectors θ' and w', global shared counter $T = 0$, maximal global shared counter $T_{\max}$, maximal thread step counter T_{local}
Output: global shared parameter vectors θ and w
Initialize thread step counter $t \leftarrow 1$;
while $T <= T_{\max}$ **do**
  Reset gradients: $d\theta \leftarrow 0$ and $dw \leftarrow 0$;
  Synchronize current thread parameter $\theta' = \theta$ and $w' = w$;
  Get state S_t;
  while *terminal is not* S_t *and* $t - t_{start} < t_{\max}$ **do**
    Perform a_t according to policy $\pi(a_t|S_t;\theta')$;
    Receive reward $\mathcal{R}_t$ and new state S_{t+1};
    $t \leftarrow t+1$;
    $T \leftarrow T+1$;
  end
  if S_t *is terminal state* **then**
    $Q = 0$;
  else
    $Q = V(S_t, w')$;
  end
  for $i \in \{t-1, \cdots, t_{start}\}$ **do**
    $Q \leftarrow \mathcal{R}_i + \gamma Q$;
    Accumulate gradients wrt θ':
    $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|S_i;\theta')(Q - V(S_i;w')) + \nabla_{\theta'} H(\pi(S_i,\theta'))$;
    Accumulate gradients wrt w':
    $dw \leftarrow dw + \partial(Q - V(S_i;w'))^2 \Big/ \partial w'$;
  end
  Perform asynchronous update of θ using $d\theta$ and of w using dw ;
end

In Algorithm 1, we define the loss function as:

$$L_{tol} = L_\pi + c_v L_v + c_{reg} L_{reg}, \tag{23}$$

where L_π is the loss of the policy, L_v is the value error and L_{reg} is a regularization term, c_v and c_{reg} are constants. The objective function is defined as:

$$J(\pi) = E[A(S,a) \cdot \log \pi(a|S)]. \tag{24}$$

We take $L_\pi = -J(\pi)$. Therefore, for all n samples in the minibatch, L_π can be expressed as follows:

$$L_\pi = -\frac{1}{n} \sum_{i=1}^{n} A(S_i, a_i) \cdot \log \pi(a_i, S_i). \tag{25}$$

The value function satisfies the Bellman Equation, i.e.,

$$V(S_0) = \mathcal{R}_0 + \gamma\mathcal{R}_1 + \gamma^2\mathcal{R}_2 + \cdots + \gamma^{n-1}\mathcal{R}_{n-1} + \gamma^n V(S_n). \tag{26}$$

According to the Eq. 26, we can get the value error:

$$e = \mathcal{R}_0 + \gamma\mathcal{R}_1 + \gamma^2\mathcal{R}_2 + \cdots + \gamma^{n-1}\mathcal{R}_{n-1} + \gamma^n V(S_n) - V(S_0). \tag{27}$$

Thus, for n samples in the minibatch, we can get the following formula:

$$L_v = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{28}$$

In order to balance exploration and exploitation during the interaction between agent and environment, we introduced entropy H, to make the distribution of output actions more balanced. Averaging over n samples in a batch, the regularization term can be expressed as:

$$L_{reg} = -\frac{1}{n}\sum_{i=1}^{n} H(\pi(S_i)). \tag{29}$$

According to the above formulas, we can obtain the loss function as follows:

$$L_{tol} = \frac{1}{n}[\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} H(\pi(S_i)) - \sum_{i=1}^{n} A(S_i, a_i) \cdot \log\pi(a_i, S_i)]. \tag{30}$$

The above loss function can be optimized by the optimization method of the deep neural network to achieve the purpose of optimizing network parameters [45]. Therefore, the system can dynamically set the reward according to the dynamic changes of the environment.

6 Evaluation

In this section, we use Tensorflow to evaluate the performance of the proposed DIM-A3C in comparison with two other methods:

Proportional Incentive Mechanism (PIM): In this incentive mechanism,

when a CP successfully shares short video content with a CR, the MEC server will give the CP a reward that is proportional to the CP's device type. In other words, a higher type of CP will receive a larger reward.

Scoring-based Incentive Mechanism (SIM): In this incentive mechanism, the size of the reward given to the CP by the MEC server is a weighted sum of the score of historical behavior and the score of expected instant price.

We consider a single-cell network with 200 CPs and 200 CRs, and we only focus on the sharing process of one short video content in this cellular network. The size of the short video content is 50 M Bytes. Some other important parameters in the simulation process are given in Table 1.

Figure 3 shows the utility of the MEC server during training, it reveals that the MEC server's utility of SIM and DIM-A3C convergence after about 1000 training episodes, which means the systems can choose a proper reward to CPs a under given state. Obviously, the utility of the MEC server of DIM-A3C is relatively higher than PIM and SIM. This is because DIM-A3C can set a lower reward for CPs based on the current system state and interest vector of CRs while ensuring that CPs actively participate in D2D-based short video sharing. It also reveals that with the SIM, the utility of the MEC server is very low at first and keeps increasing as training progresses until it is better than PIM after about 800 training episodes because SIM can set appropriate reward based on the score of historical behavior and the score of expected instant price.

Figure 4 shows the total utility of all CPs in the D2D-based sharing process of a short video. Apparently, PIM obtains a relatively stable total utility of all CPs. As the training process progresses, the total utility of all CPs of SIM continues to decrease until it reaches a level approximately equivalent to PIM. It is also noteworthy that compared with PIM and SIM, DIM-A3C performs better in terms of the total utility of CPs because the MEC server can always encourage CPs to share short videos at a relatively small cost, so that the CPs' total utility is always kept at a low level.

Figure 5 depicts the degree of satisfaction (denoted by Ds) of three incentive mechanisms, which is defined as follows:

$$Ds = \frac{\Lambda}{\Gamma}, \tag{31}$$

Table 1 Parameters setting

Parameter	Specifications
Training episodes	1000
Number of Sharing process	200
Actor learning rate	1e-4
Critic learning rate	1e-3
Entropy_bate	1e-2
Gamma	0.9
Number of workers	8
Iterators of update global net	10
D2D bandwidth	20MHz
Noise spectral density	-170dBm/Hz
Path loss exponent	4
Transmitting power	20dBm

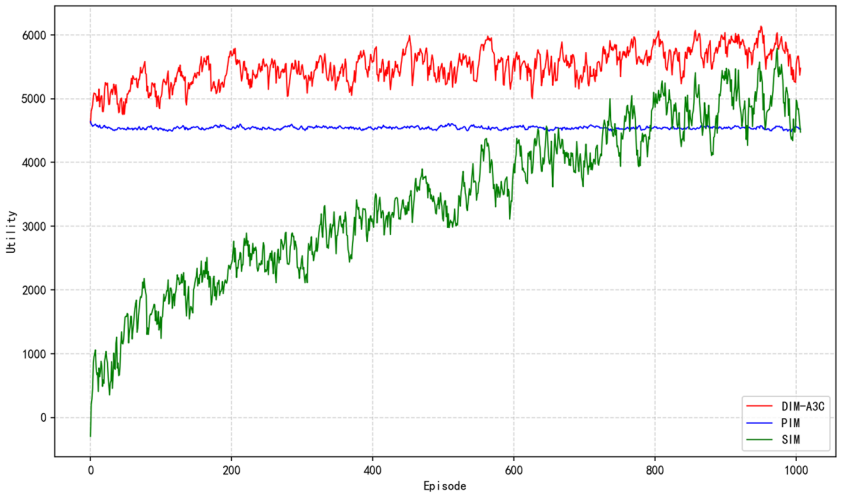


Fig. 3 Utility of MEC server during training

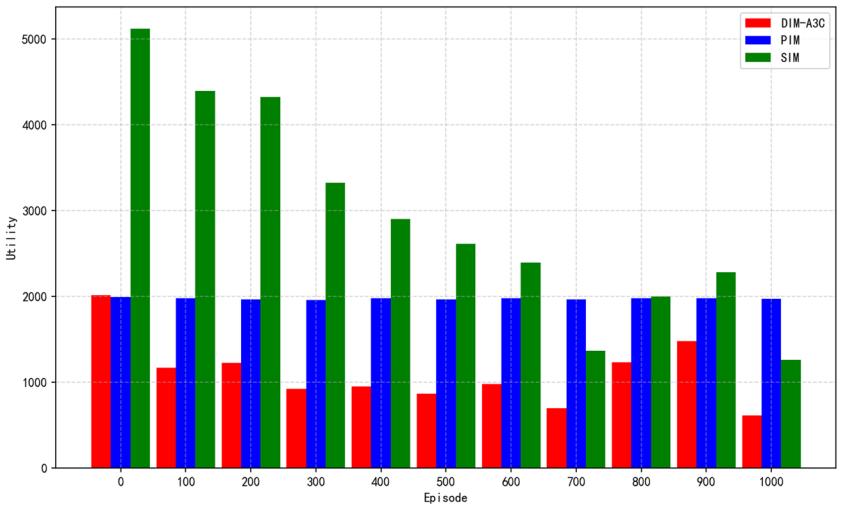


Fig. 4 Total utility of all CPs

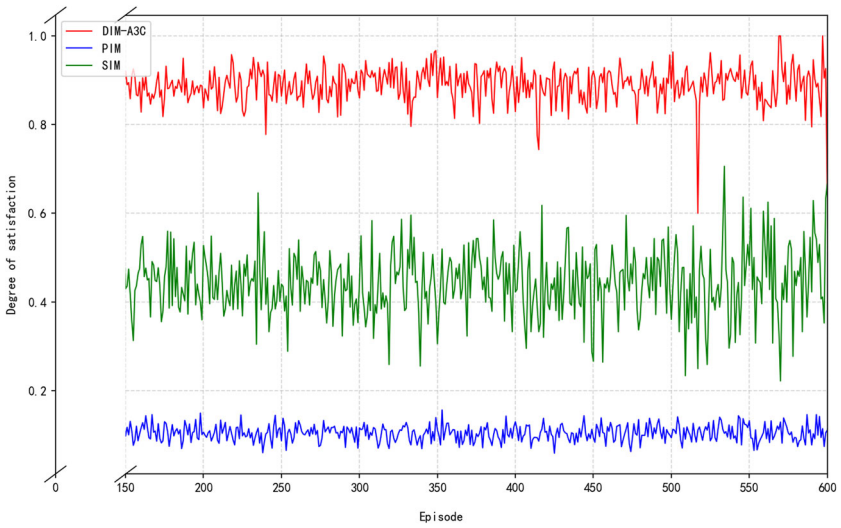
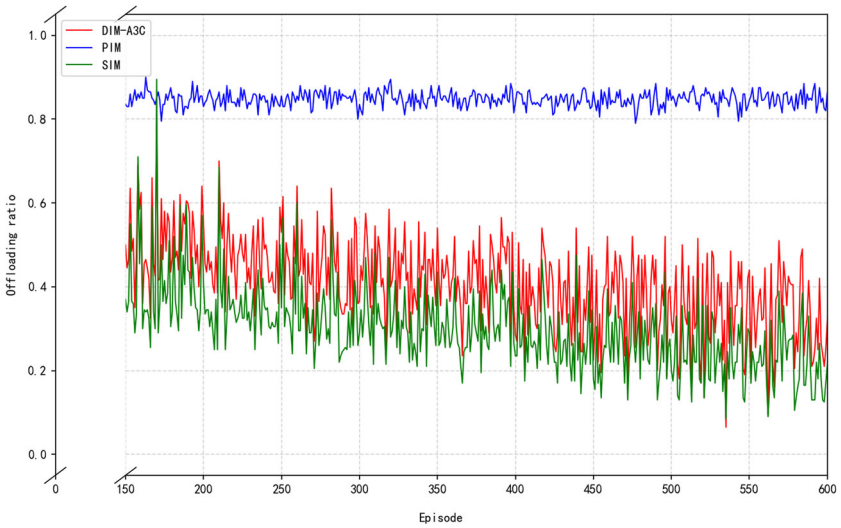


Fig. 5 Degree of CRs's satisfaction

Fig. 6 Offloading ratio



where Γ represents the number of times a short video has been shared via D2D, and Λ denotes the number of times the short video was approved by CRs after being shared via D2D. Predictably, DIM-A3C gains the highest degree of satisfaction because it considers CR's interest in the short video to be shared during the reward setting process. In the D2D-based short video sharing process, if the CR is not interested in the short video content to be shared, the MEC server will set a smaller reward for the CP sharing the short video, and then if the utility obtained by the CP is a negative value, the D2D sharing process of the short video will be terminated. Neither PIM nor SIM takes the CR's interest in short videos into account during the sharing process, and thus CR's degree of satisfaction was relatively low. Note that since SIM takes into account historical behavior, the degree of satisfaction obtained by SIM is better than PIM.

In Fig. 6, we compare the offloading ratio of DIM-A3C with PIM and SIM. As shown in Fig. 6, the offloading ratio of PIM is relatively higher by DIM-A3C and SIM because PIM always sets larger rewards for CPs to enable more devices to participate in the D2D sharing process of short videos. We can find that with DIM-A3C, rewards are dynamically set according to the network state, and the CR's interest vector is also taken into account when makes decisions. Therefore, DIM-A3C has a lower offloading ratio than PIM, but in most cases higher than SIM.

In this paper, we maximize the utility of the MEC server from the perspective of the operator, while ensuring that the utility of the users who own the short videos, namely the CPs, is non-negative. Of course, as the system effectively rejects the short video sharing of little significance in order to increase the utility of the MEC server, it results in a relatively low offloading ratio. Choosing the appropriate CPs would be an effective way to increase the offloading ratio. Therefore, in future work, how to choose the proper CPs will be our focus.

7 Conclusion

In order to enable more UEs to participate in the D2D sharing process of short videos and maximize the utility of the MEC server, we define a dynamic rewards and punishments setting for maximizing the utility of the MEC server problem RPSDMU, which has been proven NP-hard. What's more, we model it as a Markov Decision Process and propose a dynamic incentive mechanism algorithm DIM-A3C based on DRL. Simulation results prove that the proposed improved incentive mechanism is effective in motivating UEs to participate in D2D sharing process of short videos while maximizing the utility of the MEC server. In the future, we will further investigate how to choose appropriate CPs in a cellular network.

Acknowledgements This research is partly supported by the National Natural Science Foundation of China (Nos.61872044), Beijing Municipal Program for Top Talent, Beijing Municipal Program for Top Talent Cultivation (CIT & TCD201804055), Open Program of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDDXN001).

References

1. DouYin, https://www.douyin.com/. Accessed 1 Nov 2020
2. Youtube Go, https://www.youtube.com/. Accessed 1 Nov 2020
3. KuaiShou, https://www.kuaishou.com/. Accessed 1 Nov 2020
4. Oughton E, Frias Z, Russell T et al (2018) Towards 5g: scenario-based assessment of the future supply and demand for mobile telecommunications infrastructure. Technol Forecast Soc Chang 133:141–155
5. Cisco Annual Internet Report (2018–2023) White Paper, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html. Accessed 1 Nov 2020
6. Zhang Y et al (2020) Autosight: distributed edge caching in short video network. IEEE Netw 34:194–199

7. Din IU, Hassan S, Khan MK, Guizani M, Ghazali O, Habbal A (2018) Caching in information-centric networking: strategies, challenges, and future research directions. IEEE Communications Surveys & Tutorials 20:1443–1474
8. Bastug E, Bennis M, Debbah M (2014) Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun Mag 52(8):82–89
9. Chen Z, Kountouris M (2016) D2D caching vs. small cell caching: where to cache content in a wireless network? In: 2016 IEEE 17th international workshop on signal processing advances in wireless communications (SPAWC), pp 1–6
10. Wu D, Zhou L, Cai Y, Qian Y (2018) Collaborative caching and matching for D2D content sharing. In: IEEE wireless communications, vol 25, pp 43–49
11. Yu X, Tan C, Ma L, Zheng M, Bu Z (2017) Maximized traffic offloading by content sharing in D2D communication. In: 2017 IEEE 86th vehicular technology conference (VTC-Fall), pp 1–5
12. Asadi A, Wang Q, Mancuso V (2014) A survey on device-to-device communication in cellular networks. IEEE Commun 16(4):1801–1819
13. Wu D, Zhou L, Cai Y (2017) Social-aware rate based content sharing mode selection for D2D content sharing scenarios. IEEE Transactions on Multimedia 19(11):2571–2582
14. Chen Z, Liu Y, Zhou B, Tao M (2016) Caching incentive design in wireless D2D networks: a Stackelberg game approach. In: 2016 IEEE international conference on communications (ICC), Kuala Lumpur, pp 1–6
15. Satyanarayanan M (2017) The emergence of edge computing, vol 50, pp 30–39
16. Zhao M, Wang W, Wang Y et al (2019) Load scheduling for distributed edge computing: a communication-computation tradeoff. Peer-to-Peer Netw Appl 12:1418–1432
17. Yang R, Yu FR, Si P, Yang Z, Zhang Y (2019) Integrated blockchain and edge computing systems: a survey, some research issues and challenges. IEEE Communications Surveys & Tutorials 21(2):1508–1532
18. Imran A, Zoha A (2014) Challenges in 5G: how to empower SON with big data for enabling 5G. Network IEEE 28(6)
19. Al-Habashna A, Wainer G (2020) QoE awareness in progressive caching and DASH-based D2D video streaming in cellular networks[J]. Wireless Networks: The Journal of Mobile Communication, Computation and Information 26(3)
20. Pan Y, Pan C, Zhu H, Ahmed QZ, Chen M, Wang J (2017) On consideration of content preference and sharing willingness in D2D assisted offloading. IEEE Journal on Selected Areas in Communications 35(4):978–993
21. Sun Q, Tian L, Zhou Y, Shi J, Zhang Z (2020) Incentive scheme for slice cooperation based on D2D communication in 5G networks. China Communications 17(1):28–41
22. Zhang Y, Song L, Saad W et al (2015) Contract-based incentive mechanisms for device-to-device communications in cellular networks. IEEE Journal on Selected Areas in Communications 33(10):1–1
23. Yang L, Zhu H, Wang H et al (2019) Incentive propagation mechanism of computation offloading in fog-enabled D2D networks. In: 2018 IEEE 23rd international conference on digital signal processing (DSP)
24. Zhou Z, Liu P, Feng J et al (2019) Computation resource allocation and task assignment optimization in vehicular fog computing: a contract-matching approach. IEEE Transactions on Vehicular Technology 68(4):3113–3125
25. Jiang J, Zhang S, Li B et al (2015) Maximized cellular traffic offloading via device-to-device content sharing, vol 34, pp 82–91
26. Jiang J, Zhang S, Li B et al (2015) Maximized cellular traffic offloading via device-to-device content sharing. IEEE Journal on Selected Areas in Communications 34(1):82–91
27. Zeng D, Gu L, Pan S, Cai J, Guo S (2019) Resource management at the network edge: a deep reinforcement learning approach. IEEE Network 33(3):26–33
28. Zhan Y, Liu CH, Zhao Y, Zhang J, Tang J (2020) Free market of multi-leader multi-follower mobile crowdsensing: an incentive mechanism design by deep reinforcement learning. IEEE Transactions on Mobile Computing 19(10):2316–2329
29. Zhao Y, Liu CH (2021) Social-aware incentive mechanism for vehicular crowdsensing by deep reinforcement learning. IEEE Trans Intell Transp Syst 22(4):2314–2325
30. Zhan Y, Li P, Qu Z, Zeng D, Guo S (2020) A learning-based incentive mechanism for federated learning. IEEE Internet of Things Journal 7(7):6360–6368
31. Zhang R, Yu FR, Liu J et al (2020) Blockchain-incentivized D2D and mobile edge caching: a deep reinforcement learning approach. IEEE Network 34(4):150–157
32. Volodymyr M, Koray K, David S et al (2019) Human-level control through deep reinforcement learning. Nature 518(7540):529–33
33. Chen M, Wang H, Chu X (2020) Continuous incentive mechanism for d2d content sharing: a deep reinforcement learning approach. In: 2020 IEEE international conference on communications workshops (ICC Workshops), pp 1–6
34. Zhao N, Liang Y, Pei Y (2018) Dynamic contract incentive mechanism for cooperative wireless networks. In: IEEE transactions on vehicular technology, vol 67, pp 10970–10982
35. Wang H, Yang Y, Wang E, Wang L, Li Q, Yu Z (2020) Incentive mechanism for mobile devices in dynamic crowd sensing system. In: IEEE transactions on human-machine systems
36. Wang H, Guo S, Cao J, Guo M (2018) Melody: a long-term dynamic quality-aware incentive mechanism for crowd sourcing. In: IEEE transactions on parallel and distributed systems, vol 29, pp 901–914
37. Zhang T, Wang H, He J (2016) An incentive mechanism under Hidden-Action for device-to-device content sharing. In: 2016 IEEE 13th international conference on signal processing (ICSP), pp 1288–1292
38. Asadi A, Wang Q, Mancuso V (2014) A survey on device-to-device communication in cellular networks. IEEE Communications Surveys & Tutorials 16(4):1801–1819
39. La QD, Quek TQS, Lee J et al (2016) Deceptive attack and defense game in honeypot-enabled networks for the internet of things, vol 3, pp 1025–1035
40. Gu L, Zeng D, Li W, Guo S, Zomaya AY, Jin H (2020) Intelligent VNF orchestration and flow scheduling via model-assisted deep reinforcement learning. IEEE Journal on Selected Areas in Communications 38(2):279–291
41. Sutton RS, Mcallester D, Singh S et al (1999) Policy gradient methods for reinforcement learning with function approximation. Submitted to Advances in Neural Information Processing Systems
42. Van Hasselt H, Guez A, Silver D (2015) Deep reinforcement learning with double q-learning. arXiv:1509.06461
43. Lillicrap TP, Hunt JJ, Pritzel A et al (2015) Continuous control with deep reinforcement learning. Computerence 8(6):A187
44. Mnih V, Badia AP, Mirza M et al (2016) Asynchronous methods for deep reinforcement learning. In: International conference on machine learning, pp 1928–1937
45. Li Z, Chen H, Lin K et al (2021) From edge data to recommendation: a double attention-based deformable convolutional network. Peer-to-Peer Netw Appl

Zhuo Li received the Ph.D. degree from Nanjing University in 2012. He is an Associate Professor in Beijing Information Science and Technology University. His research interests include wireless network, distributed computing and network security, etc.

Wentao Dong received the BS degree in 2019 from Shandong Jianzhu University. He is currently a master student majored in Computer Science with the school of Beijing Information Science and Technology University. His main research interest is mobile edge computing.

Xin Chen received the Ph.D. degree in computer science from the Beijing Institute of Technology, Beijing, China. He is currently a Professor with the Computer School, Beijing Information Science and Technology University. His current research interests include performance evaluation of wireless networks. Dr. Chen received the Postdoctoral Fellowship in Computer Architecture from Tsinghua University in 2006. He is a Senior Member of the China Computer Federation (CCF), a member of the CCF Technical Committee of Theoretical Computer Science, and the CCF Technical Committee of Petri Nets.