Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive Internet of Things

Yang, Helin; Zhong, Wen-De; Chen, Chen; Alphones, Arokiaswami; Xie, Xianzhong

2020

Yang, H., Zhong, W.-D., Chen, C., Alphones, A., & Xie, X. (2020).

Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive Internet of Things. IEEE Internet of Things Journal, 7(6), 5677-5689.

doi:10.1109/JIOT.2020.2980586

<https://hdl.handle.net/10356/142885>

<https://doi.org/10.1109/JIOT.2020.2980586>

© 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at:
<https://doi.org/10.1109/JIOT.2020.2980586>

Downloaded on 26 Apr 2023 21:03:49 SGT

Deep Reinforcement Learning Based Energy-Efficient Resource Management for Social and Cognitive Internet of Things

Helin Yang, *Student Member, IEEE*, Wen-De Zhong, *Senior Member, IEEE*, Chen Chen, *Member, IEEE*, Arokiaswami Alphones, *Senior Member, IEEE*, and Xianzhong Xie, *Member, IEEE*

Abstract—Internet of things (IoT) has attracted much interest due to its wide applications such as smart city, manufacturing, transportation, and healthcare. Social and cognitive IoT is capable of exploiting the social networking characteristics to optimize the network performance. Considering the fact that IoT devices have different quality of service (QoS) requirements (ranging from ultra-reliable and low-latency communications (URLLC) to minimum data rate), this paper presents a QoS-driven social-aware enhanced device-to-device (D2D) communication network model for social and cognitive IoT by utilizing social orientation information. We model the optimization problem as a multi-agent reinforcement learning formulation, and a novel coordinated multi-agent deep reinforcement learning based resource management approach is proposed to optimize the joint radio block assignment and transmission power control strategy. Meanwhile, prioritized experience replay (PER) and coordinated learning mechanisms are employed to enable communication links to work cooperatively in a distributed manner, which enhances the network performance and access success probability. Simulation results corroborate the superiority in the performance of the presented resource management approach, and it outperforms other existing approaches in terms of meeting the energy efficiency and the QoS requirements.

Index Terms—Internet of Things, social-awareness, D2D communication, resource management, deep reinforcement learning, quality of service, energy efficiency.

I. INTRODUCTION

With the widespread development of the wireless communication technologies, internet of things (IoT) has attracted much attention due to its wide range of applications, such as smart city, smart grid, smart industry, transportation, healthcare and so on [1]. In IoT, the key is not only simply to connect things with things, but also to connect a large

number of smart connected devices with intelligent sensing, communication and computing capabilities [2]. With the advanced wireless communication, more and more smart devices are involved in IoT, resulting in the tenfold increased amount of raw data, the spectrum radio scarcity and burden resource management in wireless IoT networks.

The references [3]–[8] adopted cognitive or mobile computing technologies (e.g., fog computing, edge computing and cloud computing) to support the data processing, knowledge discovery, experience learning and efficient data exchanging for IoT in smart or cognitive ways. In [4]–[7], the massive raw data generated from smart things were processed to be social guiding data, which are used to support smart decision making in IoT or industrial IoT (IIoT) environments. In addition, Aazam *et al.* in [6] proposed a quality of service (QoS)-aware approach for resource allocation in IoT based on feedback intrinsic properties. Apart from cognitive computing technologies, a new research paradigm called social IoT has attracted much attention [9]–[13], where real-world physical things can form their own social networks like human beings to implement self-cooperation and self-management. The authors in [9] and [10] combined the real-world social information into IoT networks to improve the communication connectivity. Moreover, in order to satisfy the strict delay requirements of IoT devices, the social content sharing-oriented matching schemes were presented in social IoT [11], [12]. Under cognitive mobile computing and social information orientation, IoT has the ability to intelligently exploit the devices' behaviors or experience and smartly learn to adjust the strategy to effectively realize the intelligent management, data transmission and smart control [9]–[13].

Recently, cognitive IoT based cognitive radio (CR) technology has been investigated to improve the utilization of scarce radio spectrum [8], [14]–[17], where secondary devices can sense the state of licensed spectrum of priority devices and reuse the spectrum under interference constraints. The literatures [14] and [15] adopted CR to realize the cognitive spectrum sensing, spectrum access and interference management in wireless sensor networks. Considering that energy efficiency (EE) plays an important role in IoT, since most of devices (e.g., sensors, actuators and controllers) are power-constrained, the energy efficient resource allocation approaches were proposed to maximize the network EE with considering the minimum data rate and interference constraints in CR-based IoT networks [16], [17]. Recently, considering

This work is support by Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc. and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme, and National Natural Science Foundation of China under Grant 61901065 and 61502067, the Key Research Project of Chongqing Education Commission under Grant No. KJZD-K201800603, the Key Project of Science and Technology Research of Chongqing Education Commission under Grant No. KJZD-M201900602, the Chongqing Nature Science Foundation under Grant No. CSTC2018jcyjAX0432 and CSTC2016jcyjA0455.

H. Yang, A. Alphones and W. Zhong are with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), 50 Nanyang Avenue, Singapore 639798. (e-mail: hyang013@e.ntu.edu.sg)

C. Chen is with School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, 400044, China.

X. Xie is with Chongqing Key Lab of Computer Network and Communication Technology, Chongqing University of Posts and Telecommunications, Chongqing, China.

the fact that ultra-reliable and low latency communications (URLLC) in 5G are closely related to wireless networks, where some services (i.e., industrial automation and vehicular communications) may require the low end-to-end latency in the range of 1-5 ms with the transmission reliability of 99.999 % or higher [1], [18], [19]. Even the studies [11] and [12] considered strict delay requirements of IoT devices, it is more piratical to take both the stringent high reliability and low latency requirements into account.

In addition to the CR technology, device-to-device (D2D) communication has been widely employed in IoT networks to reduce transmission delay, power consumption and improve spectrum efficiency [20]-[28], where two devices can directly communicate with each other by reusing the same radio resource of cellular devices under interference constraints. Considering the fact that URLLC in 5G is closely related to industrial networks, the works in [20] and [21] proposed resource management approaches to guarantee the stringent low latency and high reliability constraints of IoT devices in underlying D2D networks. In [9], [11], [22], [23], the authors investigated how to integrate social networking characteristics into D2D-based IoT or IIoT networks for the resource optimization, where the resource allocation approaches were also proposed to enhance the network performance and ensure the QoS requirements of devices. In addition, the references [26]-[28] provided their investigations to verify that D2D communication is also considered as a promising technology to achieve high data rates, ultra-low latency communication and high spectral efficiency. Moreover, the works in [24], [25], [29] integrated CR capabilities with D2D communications for IoT devices (called cognitive D2D), where cognitive D2D devices have the ability to perform spectrum sensing to opportunistically assign the channels of cellular devices, which enhances the spectrum utilization and reduce the power consumption. The above reported works [20], [21], [24]-[25] mainly focused on the physical layer information in their presented models, and the mentioned studies [9], [11], [22], [23] integrated social networking characteristics into D2D-based IoT networks, but these works did not consider different QoS requirements (e.g., high reliability, low latency, minimum data rate and D2D-to-cellular interference) into their proposed social-aware models, where these QoS requirements can be contained into social orientation information to perform the resource optimization.

Recently, many studies have investigated application of reinforcement learning (RL) in the field of resource management for wireless social networks [30]-[35]. Qian *et al.* in [31] proposed an optimal relay selection approach based on Q-learning to maximize the sum rate of D2D links in social networks. In [32], a novel big data deep reinforcement learning (DRL) approach was studied to optimize resource management in mobile trust-based social networks, where the intrinsic nature of social networks was exploited to generate effective resource management rules. Tang *et al.* in [33] leveraged the social tie structure among devices to improve beneficial computation offloading decision making, and a distributed RL algorithm was presented to deal with network uncertainty. The references [21], [34], and [35] proposed RL based mobile resource management mechanism to guarantee diverse vehicular social

requirements under dynamic environments. However, most of the works [31]-[35] did not consider the stringent reliability and latency constraints into the optimization problem.

According to the above analysis, in this paper, a QoS-driven social-aware network model is presented to integrate the social information into cognitive D2D-based IoT networks, in order to optimize the network performance and guarantee different QoS requirements (e.g., high reliability, low latency and minimum data rate). Then, the social-aware energy-efficient resource management optimization problem is modelled as a multi-agent RL formulation. A novel coordinated multi-agent DRL-prioritized experience replay (PER) based resource management approach is proposed to optimize the joint radio block (RB) assignment and transmission power control strategy. The key contributions can be summarized as follows:

- We present a QoS-driven social-aware network model for cognitive D2D-based IoT networks, where the social orientation properties (including QoS requirements) are utilized to perform resource optimization for IoT devices. To the best of our knowledge, this is the first time that the QoS-driven social-awareness is introduced in IoT.
- We formulate a joint social-aware RB assignment and power allocation problem with the goal of maximizing the network EE while considering the URLLC, minimum data rate and interference constraints. In addition, we transform the optimization problem into a multi-agent RL formulation. Hence, each device has the ability to intelligently make its resource management decision according to its own instantaneous observations.
- A distributed coordinated RB assignment and transmission power control approach based on multi-agent DRL-PER is proposed to guarantee both the strict reliability and latency requirements on URLLC services in social-aware scenario. Specifically, a proper QoS-aware reward function is built to cover both the network EE and devices' QoS requirements into the learning process. PER and coordinated learning are adopted to efficiently improve the learning efficiency and convergence speed.
- The effectiveness of the proposed coordinated multi-agent DRL-PER based resource management approach is evaluated by the comprehensive simulations, and we also compare the performance with other existing approaches.

The rest of this paper is organized as follows. The social-aware network model is presented in Section II. Section III formulates the resource management problem and the problem is solved in Section IV. Simulation results are presented in Section V and Section VI concludes the paper.

II. NETWORK MODEL

We consider a cognitive D2D-based IoT network, as shown in Fig. 1, which consists of a base station (BS), a number of cellular user equipments (denoted CUEs) and cognitive IoT devices which will establish a number of D2D communication pairs. CUEs communicate with the BS over orthogonal spectrum RBs, while two IoT devices can directly communicate with each other if they are close enough to each other

Fig. 1. Physical and social-aware cognitive D2D-based IoT networks.

when they require a communication demand. Cognitive IoT devices in D2D pairs have the ability to sense the spectrum working status (busy or idle) and the surrounding interference level of CUEs' RBs. With these knowledge, D2D pairs can intelligently make decisions on assigning the RBs while guaranteeing that the generated interference from D2D pairs to CUEs is below a tolerant interference threshold.

As seen from Fig. 1, the network consists of the physical domain and social domain. In the physical layer domain, the wireless link establishment of one D2D pair depends on the communication distance between two IoT devices, while in the social layer domain, the devices establish a mobile social network according to the social relationships of different devices and communities. In the social layer, the devices are divided into a number of social communities based on their traffic interests, behaviors similarities, information sharing and so on. In addition, the network is well aware of the specified QoS requirements of devices when devices report their services in their communities in the social layer.

A. Physical Layer Model

The network has K CUEs and M cognitive IoT devices, where each device is equipped with one single antenna. The CUE set and cognitive IoT device set are denoted by $\mathcal{K} = \{1, 2, \dots, K\}$ and $\mathcal{M} = \{1, 2, \dots, M\}$, respectively. These cognitive IoT devices may form N D2D communication pairs and the D2D-pair set is denoted by $\mathcal{N} = \{1, 2, \dots, N\}$. The network uses the orthogonal frequency division multiple access (OFDMA) technique to provide communication services in the uplink scenario, and there are K orthogonal RBs being allocated to the CUEs with each CUE being allocated with one RB. In addition, D2D pairs can reuse RBs of CUEs to build communication links.

Physical graph: The physical graph for the D2D communication graph is introduced as $\mathcal{G}^P = \{\mathcal{M}, \Upsilon^P\}$, where $\mathcal{M}$ is the vertex set and $\Upsilon^P = \{(i, j) : e_{i,j}^P = 1, \forall i, j \in \mathcal{M}\}$ is the edge set. $e_{i,j}^P = 1$ if and only if the transmission distance (denoted by $d_{i,j}$) between the i -th IoT device and the j -th IoT device is smaller than a given distance threshold (denoted by

$d_{\max}$); otherwise, $e_{i,j}^P = 0$, which can be expressed as

$$e_{i,j}^P = \begin{cases} 1, & d_{i,j} \leq d_{\max} \\ 0, & d_{i,j} > d_{\max} \end{cases} \quad (1)$$

Let h_k denotes the channel gain from the k -th CUE to the BS, and let h_n denotes the channel gain of the n -th D2D pair. Let $g_{n,k,B}$ denotes the interference channel gain from the n -th D2D pair's transmitter to the BS on the k -th CUE's RB, and let $g_{k,n}$ denotes the interference channel gain from the k -th CUE to the n -th D2D pair's receiver. Let $g_{n',n}$ denotes the interference channel gain from the n' -th D2D pair's transmitter to the n -th D2D pair's receiver.

In the spectrum reusing scenario, the k -th CUE suffers the co-channel interference from the transmitters of the D2D pairs when it shares the RB with D2D pairs. Hence, the received uplink signal-to-interference-plus-noise ratio (SINR) at the BS for the k -th CUE is expressed as

$$\gamma_k^c = \frac{P_k^c h_k}{\sum_{n \in \mathcal{N}} \rho_{n,k} P_{n,k}^d g_{n,k,B} + \sigma^2} \quad (2)$$

where P_k^c and $P_{n,k}^d$ are the transmit power levels of the k -th CUE and the n -th D2D pair's transmitter on the k -th RB, respectively. $\rho_{n,k}$ is a binary variable, $\rho_{n,k} \in \{0, 1\}$; $\rho_{n,k} = 1$ indicates that the n -th D2D pair assigns the RB of the k -th CUE; otherwise, $\rho_{n,k} = 0$. σ^2 is the additive white Gaussian noise power.

At the same time, RB resource sharing also causes the interference to the D2D pairs, which is generated from the co-channel CUE and co-channel D2D pairs on the same RB. Therefore, the received uplink SINR at n -th D2D pair's receiver sharing the RB of the k -th CUE can be given by

$$\gamma_{n,k}^d = \frac{P_{n,k}^d h_n}{\sum_{n' \in \mathcal{N}, n' \neq n} \rho_{n',k} P_{n',k}^d g_{n',n} + P_k^c g_{k,n} + \sigma^2} \quad (3)$$

In (3), we would like to mention that there should be no D2D to D2D interference when these two D2D pairs reuse the RBs of two different CUEs.

B. Social Layer Model

In the social layer domain, the social characteristics depends on the real activities and behaviors of the corresponding human beings, which has two main relationship characteristics: social community and social tie. Social community is formed based on human beings' social relations, such as social interests and behaviors. Hence, the devices in the mobile network can be divided into a number of communities according to their social relationships, i.e., partners, interests and background. The term social tie in the social network is the relationship or trustiness levels among human beings. For example, the human beings in the same community have the high social tie while human beings from different communities have weak social tie.

As mentioned before, social information can be used for the better resource management. For example, we consider a social industry scenario (community Z) as shown in Fig. 1, where the workers, excavator controller, truck driver, etc., can

form a social community. These human beings are more likely to exchange the real-time messages to control the machines or devices, which requires the low packet transmission delay and high transmission reliability. In this case, trusted D2D links can be established among them to support the communication services and the delay constraints need to be carefully considered in the resource management.

Social graph: Since the social ties also determine the D2D communication link establishment, we introduce the social graph as $\mathcal{G}^S = \{\mathcal{M}, \Upsilon^S\}$, where $\Upsilon^S = \{(i, j) : e_{i,j}^S, \forall i, j \in \mathcal{M}\}$ is the edge set. $e_{i,j}^S$ is the communication demand between the i -th IoT device and the j -th IoT device, and $e_{i,j}^S$ indicates the social tie strength of these two IoT devices in the social layer domain.

For IoT devices i and j , the social tie is calculated as [9]

$$e_{i,j}^S = ((\text{corr}(p_i, p_j) + 1))/2 \quad (4)$$

where p_i and p_j denote the probability of sharing/selecting similar contents for IoT devices i and j , which are the variables with the probability density function (PDF) $\mathcal{P}_i(p_i)$ and $\mathcal{P}_j(p_j)$ over the space $[0,1]$, respectively. In (4), the social tie $e_{i,j}^S \in [0,1]$. Naturally, the social tie between two IoT devices in the same community is usually strong, while it is generally weak when the two IoT devices come from different communities.

C. Social Aware Model

In the cognitive D2D-based IoT network, both the physical constraints and social constraints should be considered in the resource management. By combining both the physical graph and social graph, we can introduce the QoS-driven physical-social graph as $\mathcal{G}^{PS} = \{\mathcal{M}, \Upsilon^{PS}, \Phi^{QoS}\}$, where $\Upsilon^{PS} = \{(i, j) : e_{i,j}^{PS} = e_{i,j}^P e_{i,j}^S, \forall i, j \in \mathcal{M}\}$ is the edge set. $e_{i,j}^{PS}$ is the physical-social score which is determined by both the transmission distance and the social tie strength of the D2D pair (i, j) . $e_{i,j}^{PS}$ varies from 0 to 1, i.e., $e_{i,j}^{PS} \in [0,1]$. $e_{i,j}^{PS} = 0$ indicates that IoT devices i and j cannot form a D2D link. Φ^{QoS} is the QoS requirement, including the high reliability, low latency and minimum data rate. Different social communities have different service requirements. For example, in the social industry scenario (community Z) as shown in Fig. 1, the human beings are more likely to exchange the real-time messages to control the machines or devices, which require the low packet transmission delay and high transmission reliability (called specified QoS requirements). In this case, trusted D2D links can be established among them to support the communication services and the delay constraints need to be carefully considered in the resource management. Hence, the social aware model needs to be well designed after knowing the different specific services requirements, as it determines the flexibility for meeting the specified QoS requirements.

Under the physical and social constraints, the data rate of the n -th D2D pair on the k -th RB can be expressed as

$$R_{n,k}^d = e_{i,j}^{PS} \log_2(1 + \gamma_{n,k}^d) \quad (5)$$

And the data rate of n -th D2D pair over the selected RBs is $R_n^d = \sum_{k \in \mathcal{K}} \rho_{n,k} R_{n,k}^d$.

In addition, the data rate of the k -th CUE is given by

$$R_k^c = \log_2(1 + \gamma_k^c) \quad (6)$$

We would like to mention that the IoT devices in the same community contract with each other frequently so they have the good trusted relationship of each other. In this case, the D2D pairs can directly reuse the RBs resource of the CUEs in the same community. By contract, CUEs only share their RBs resource with the D2D pairs in the cooperated communities, otherwise, they reject to share their radio source with other communities due to the trustless relationship and selfishness.

III. NETWORK REQUIREMENTS AND PROBLEM FORMULATION

In practical IoT networks, different CUEs or IoT devices may have different QoS requirements, i.e., some devices have stringent latency and transmission reliability demands but they are less interested in the high data rate (called **URLLC services**), while other devices require the normal communication services with the minimum data rate constraints but they have no stringent latency and reliability requirements (called **normal services**). Hence, it is natural to take these practical requirements into account as constraints in the mathematical way. This section formulates a theoretical optimization problem for the resource management in the cognitive D2D-based IoT network with the objective of maximizing the network EE while guaranteeing the different QoS requirements of devices.

A. Network Requirements

1) URLLC requirements: The real-time communication services (URLLC services) have strict latency and transmission reliability requirements, but they are less interested in the high data rate, such as hazard monitoring, real-time safety information exchange, and so on. Generally, the requirements of URLLC can be captured by end-to-end delay of every packet and the outage probability, respectively.

In the URLLC scenario, we assume that each D2D pair n follows the independent and identically distributed (i.i.d.) Poisson distribution with the arrival rate λ_n and the data packet size L_n^{Pac} in bits. Generally, the total delay (T_d) of one transmission packet of the n -th D2D pair consists of the queue waiting time (T_w), the transmission time (T_t), the channel access delay (T_a), the delay of packet reception and processing operation (T_{rp}), which can be given by [19], [36]-[38]

$$T_d = T_w + T_t + T_a + T_{rp} \quad (7)$$

In (7), the transmission time of the n -th D2D pair' packet is expressed as $T_t = L_n^{\text{Pac}}/R_n^d$. The latency constraint in the URLLC scenario can be satisfied by controlling the probability of the total delay T_d exceeding the maximum delay threshold value T_{max} . And the probability should be less than the

maximum tolerable violation probability $p_{\max}^{\text{Lat}}$, which can be written as

$$p_n^{\text{Lat}} = \Pr \{T_d \geq T_{\max}\} \leq p_{\max}^{\text{Lat}} \quad (8)$$

In addition, the reliability requirement can be captured by guaranteeing the received SINR (denoted by $\gamma_n^{\text{d,URLLC}}$) at the receiver, which should be bigger than the target threshold $\gamma_{n,\min}^{\text{d,URLLC}}$. Then, the reliability constraint is given by

$$p_n^{\text{SINR}} = \Pr \left\{ \gamma_n^{\text{d,URLLC}} < \gamma_{n,\min}^{\text{d,URLLC}} \right\} \leq p_{\max}^{\text{SINR}} \quad (9)$$

where $p_{\max}^{\text{SINR}}$ is the maximum tolerable violation probability in terms of SINR.

2) *Minimum data rate requirements*: The CUEs and some D2D pairs usually require the minimum data rate requirements, although they have no latency constraints. Let $R_{k,\min}^c$ and $R_{n,\min}^{\text{d,Nor}}$ denote the minimum data rate requirements of CUE k and D2D pair n with the normal services, respectively.

Hence, the minimum data rate requirements of CUE k and D2D pair n should be considered in the resource management, and the constraints are respectively expressed as

$$\begin{aligned} R_k^c &\geq R_{k,\min}^c, \quad \forall k \\ R_n^{\text{d}} &\geq R_{n,\min}^{\text{d,Nor}}, \quad \forall n \end{aligned} \quad (10)$$

3) *Interference constraint*: In the spectrum sharing scenario, even though D2D pairs are allowed to share the licensed RBs of CUEs to improve the spectral efficiency, the transmission of D2D links should be carefully monitored and managed to control the generated interference to CUEs within a tolerable level. Thus, the interference constraint of each RB of its correspond CUE k can be given by

$$\sum_{n \in \mathcal{N}} \rho_{n,k} P_{n,k}^{\text{d}} g_{n,k,B} \leq I_{\max} \quad (11)$$

where $I_{\max}$ is the tolerant interference threshold of each RB.

B. Problem Formulation

The sum achievable data rate and the sum power consumption can be expressed as

$$R_{\text{sum}} = \sum_{k \in \mathcal{K}} R_k^c + \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} \rho_{n,k} R_{n,k}^{\text{d}} \quad (12)$$

$$P_{\text{sum}} = \sum_{k \in \mathcal{K}} P_k^c + \sum_{k \in \mathcal{K}} \sum_{n \in \mathcal{N}} \rho_{n,k} P_{n,k}^{\text{d}} + K P_{\text{fix}}^c + M P_{\text{fix}}^{\text{d}} \quad (13)$$

where P_{fix}^c and $P_{\text{fix}}^{\text{d}}$ denote the circuit power consumption of each CUE and each IoT device.

The goal is to maximize the network EE (EE is the ratio of the sum data rate and the sum energy consumption: $\eta_{EE} = R_{\text{sum}}/P_{\text{sum}}$) while satisfying the network requirements and constraints shown in Section III.A. Then, the resource management problem (joint RB assignment and power allocation)

is formulated as

$$\begin{aligned} \max_{\rho, \mathbf{P}} \quad & \eta_{EE} \\ \text{s.t.} \quad & \text{(a) : (8), (9), (10), (11); } \rho_{n,k} \in \{0, 1\}, \quad \forall n, \forall k; \\ & \text{(b) : } \sum_{k \in \mathcal{K}} \rho_{n,k} \leq L, \quad \forall n \in \mathcal{N}; \\ & \text{(c) : } P_k^c \leq P_{\max}^c, \quad \forall k \in \mathcal{K}; \\ & \text{(d) : } \sum_{k \in \mathcal{K}} \rho_{n,k} P_{n,k}^{\text{d}} \leq P_{\max}^{\text{d}}, \quad \forall n \in \mathcal{N}. \end{aligned} \quad (14)$$

where ρ and $\mathbf{P}$ denote the RB assignment and power allocation strategies, respectively. $P_{\max}^c$ and $P_{\max}^{\text{d}}$ are the maximum transmit power levels of each CUE and each D2D IoT device. Constraint (14b) is imposed to ensure that each D2D pair can only assign at most L RBs at each time slot. Constraints (14c) and (14d) specify the transmit power constraint of each CUE and each D2D pair.

IV. COORDINATED MULTI-AGENT DRL-PER BASED RESOURCE MANAGEMENT

Clearly, the optimization problem shown in (9) is difficult to be solved as it is a non-convex combination and NP-hard problem. In addition, in realistic social networks, the capabilities of devices, the channel quality and service applications of devices will change dynamically. Hence, it is generally infeasible to applied traditional optimization solutions to deliver a resource management policy.

Model-free RL is a dynamic programming technique which can be adopted to solve the optimization problem by learning the optimal solution over dynamic social networks. Hence, considering the fact that a number of devices attempt to access spectrum resource, this paper formulate the optimization problem as a multi-agent formulation, and a multi-agent DRL-PER based resource management approach is proposed to learn the optimization policy under different QoS requirements.

A. Related Definition of Multi-Agent DRL

In social cognitive IoT networks, each communication link acts as a learning agent and the network constitutes the environment. In order to enable each agent for the transmission power control and RB assignment optimization, we need to define state, action and reward function in the multi-agent DRL-PER based resource management approach.

State Space: In each time slot, each communication link acts as a learning agent to observe the network state from the unknown environment, and then uses it for decision-making. Let $\mathcal{S}$ denote the state space. The current environment state $s \in \mathcal{S}$ includes channel information and all agents' behaviors. As a result, the state s is defined by the following parameters:

$$s = \{ \{ \text{RB}_k \}_{k \in \mathcal{K}}, \{ \gamma_k^c \}_{k \in \mathcal{K}}, \{ \gamma_{n,k}^{\text{d}} \}_{n \in \mathcal{N}, k \in \mathcal{K}}, \{ L_n^{\text{pac}} \}_{n \in \mathcal{N}}, \{ \text{QoS}_k \}_{k \in \mathcal{K}}, \{ \text{QoS}_n \}_{n \in \mathcal{N}} \} \quad (15)$$

where RB_k denotes the RB occupy status of the k -th RB. QoS_k and QoS_n indicate the QoS satisfaction levels of the k -th CUE and the n -th D2D pair, respectively, such as the satisfaction levels of the minimum data rate, latency and

reliability. Other parameters have been defined in Section II and Section III.

Action Space: The resource management of social networks comes down to the transmission power control and RB assignment for each communication link. Let $\mathcal{A}$ denote the action space. Each agent chooses an action $a \in \mathcal{A}$ from the action space under the current observed state s . In the network, the action a can be defined as

$$a = \left\{ \{P_k^c\}_{k \in \mathcal{K}}, \{P_{n,k}^d\}_{n \in \mathcal{N}, k \in \mathcal{K}}, \{\rho_{n,k}\}_{n \in \mathcal{N}, k \in \mathcal{K}} \right\} \quad (16)$$

where P_k^c , $P_{n,k}^d$ and $\rho_{n,k}$ denote the transmission power of the k -th CUE, the transmission power of n -th D2D pair transmitter on the k -th RB, and the RB assignment indicator of the n -th D2D pair on k -th RB, respectively.

Transition Probability: In the transition model, $\mathcal{P}(s_{t+1}|s_t, a_t)$ is the probability of transferring from a current state $s_t \in \mathcal{S}$ to a new state $s_{t+1} \in \mathcal{S}$ after executing an action $a_t \in \mathcal{A}$.

Reward Function: Each agent in DRL relies on the reward to evaluate the quality of the action which it selects and then further optimizes its decision strategy. Hence, the network performance can be enhanced when the designed reward at each learning step correlates with the desired objective. In the resource management problem formulated in Section III.B, our objective (reward function) is twofold: Maximizing the overall network EE while guaranteeing QoS requirements (ranging from high reliability and low latency to minimum data rate) of devices.

Here, let $\mathcal{N}'$ denotes the set of communication links in the URLLC scenario where the devices have both the reliability and latency requirements, and $\mathcal{N}''$ denotes the set of D2D communication links in the normal scenario where the devices have minimum data requirements.

As such, a QoS-aware reward function can be presented as

$$U = \eta_{EE} - \mu_1 \sum_{n \in \mathcal{N}'} (p_n^{\text{Lat}} + p_n^{\text{SINR}}) - \mu_2 \left(\sum_{k \in \mathcal{K}} p_k^{\text{Nor}} + \sum_{n \in \mathcal{N}''} p_n^{\text{Nor}} \right) \quad (17)$$

where the part 1 is the utility (EE), the part 2 and part 3 are cost functions in terms of the unsatisfied latency and reliability requirements for URLLC services, and the unsatisfied minimum data rate requirements for normal services, respectively. The parameters μ_1 and μ_2 are the weights of the latter two parts in (17).

Each agent tries to select a policy π to maximize a discounted accumulative reward, where π is a mapping from state s with the probability distribution over actions that the agent can take: $\pi(s) : \mathcal{S} \rightarrow \mathcal{A}$. The discounted accumulative reward for starting the state s with the current policy π is defined as

$$Q^\pi(s, a) = \left\{ \sum_{t=1}^{\infty} \xi \bar{U}_t(s_t, a_t) | s_0 = s, \pi \right\} \quad (18)$$

where $\bar{U}_t(s_t, a_t)$ denotes the expected reward by selecting action a_t at state s_t , and ξ denotes a discount rate $\xi \in (0, 1)$.

The function $Q^\pi(s, a)$ in (18) is usually used to test the quality of the selected policy π when the agent selects the

Fig. 2. Architecture of the coordinated multi-agent DRL-PER based resource management.

action a at the state s . The MPD model tries to search the optimal state-value function $Q^*(s, a)$, which is expressed as

$$Q^*(s, a) = \max_{\pi} Q^\pi(s, a) \quad (19)$$

Once $Q^*(s, a)$ is achieved, the optimal policy $\pi^*(s_t)$ under the current state s_t can be determined as

$$\pi^*(s_t) = \arg \max_{a_t \in \mathcal{A}} \bar{U}_t(s_t, a_t) + \sum_{s_{t+1}} \mathcal{P}(s_{t+1}|s_t, a_t) Q^*(s_{t+1}, a_t) \quad (20)$$

To calculate $Q^*(s, a)$, the iteration algorithms can be used. However, it is difficult to get the transition probability $\mathcal{P}(s_{t+1}|s_t, a_t)$ in practical environments, but RL algorithms, such as Q-learning, policy gradient, DQN and so on, are widely employed to address the problem under environment uncertainty [30], [32], [34], [36].

B. Detailed Process of Multi-Agent DRL-PER Based Resource Management

In order to learn the optimal policy for the intelligent resource management under continuous-valued state and action spaces, actor-critic (AC) DRL algorithm is reported to optimize the policy by integrating the policy-based and value-based designs into the learning structure [30]. Hence, this paper proposes a coordinated multi-agent DRL-PER based resource management approach to optimize the joint RB assignment and power control policy in social cognitive IoT networks, as shown in Fig. 2.

In the coordinated multi-agent DRL-PER learning framework, there are two networks, namely, the actor network and the critic network, where the actor network takes charge of mapping observed states through adopting a control policy to select action, and the critic network represents the state-action pairs to the expected long-term cumulative reward. Meanwhile, since IoT devices in the same community have the good trusted relationship of each other, coordinated learning is used to enhance the global utility by choosing the actions jointly instead of independently.

1. Critic Network With Improved DNN

Since the input state and output action are high-dimensional, one of the powerful deep learning techniques, called DNN,

has been widely used to represent the function approximation. Hence, in this paper, DNN is adopted to represent the critic structure in DRL, where the Q-value function $Q^\pi(s, a)$ is represented as $Q(s, a, \omega)$ with ω being the DNN weights. The input layer of DNN is the environment state s , and the output layer provides the approximated Q-value function $Q(s, a, \omega)$. At each training step, the DNN can be trained to update the weights, ω , by minimizing the loss function

$$\mathcal{L}(\omega(t)) = [y_t - Q_t(s_t, a_t, \omega(t))]^2 \quad (21)$$

where y_t is the target value which can be estimated by:

$$y_t = U_{t+1}(s_t, a_t) + \xi \max_{a_{t+1} \in \mathcal{A}} Q_t(s_{t+1}, a_{t+1}, \omega(t)) \quad (22)$$

The error between the target value y_t and the estimated value $Q_t(s_t, a_t, \omega_t)$ is usually called temporal-difference (TD) error, which is expressed as

$$\vartheta(t) = y_t - Q_t(s_t, a_t, \omega(t)) \quad (23)$$

One important reason of adopting DQN is to sample the loss functions shown in (21) at each training step to decrease the computational cost for large-scale learning problems [30], [32], [34]. When the loss function $\mathcal{L}(\omega_t)$ is continuously differentiable with respect to the critic DNN weights ω , then ω can be updated by using the gradient descent method, which is expressed as

$$\omega(t+1) = \omega(t) + \beta_c(t) \vartheta(t) \nabla_{\omega} \mathcal{L}(\omega(t)) \quad (24)$$

where $\beta_c(t)$ is the learning rate of the critic DQN, and $\nabla(\cdot)$ is the first-order partial derivative.

2. The Actor Network

In the actor network, the policy gradient method is usually adopted to iteratively optimize the policy to find an optimal behavior policy π for each agent. We use an arbitrary function of the parameters vector θ to initially build the policy $\pi_{\theta}(s, a)$, which is denoted as $\pi_{\theta}(s, a) = \mathcal{P}(a|s, \theta)$. The policy $\pi_{\theta}(s, a)$ can be iteratively updated trying to maximize the long-term reward, where the long-term reward can be defined as the objective function [30]

$$J(\pi_{\theta}) = \{Q^\pi(s, a)\} = \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_{\theta}(s, a) Q(s, a, \omega) \quad (25)$$

where $d^\pi(s)$ denotes a stationary distribution with the current policy π_{θ} . The gradient of the policy can be achieved by taking the partial differentiation of the objective function (25) with respect to θ , which is given by

$$\nabla_{\theta} J(\pi_{\theta}) = \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_{\theta} \pi_{\theta}(s, a) Q(s, a, \omega) \quad (26)$$

Then, the policy parameters θ can be improved by

$$\theta(t+1) = \theta(t) + \beta_a(t) \nabla_{\theta} J(\pi_{\theta}) \quad (27)$$

where $\beta_a(t)$ denotes the learning rate of the actor. Instead of calculating probabilities for each action, we apply the Gaussian probability distribution [30] to compute a stochastic policy for selecting actions, which is defined by

$$\pi_{\theta} = \pi(a|s, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(a - c(s, \sigma))^2}{2\sigma^2}\right) \quad (28)$$

where $c(s, \sigma)$ is the mean of the normal distribution which is the deterministic action under the current state s . σ is the standard deviation which shows the range of exploration over all actions. The more explanations of (28) can be found in [19], [30].

3. Enhanced Coordinated Multi-Agent DRL-PER

Classical DNN use a simple uniform or random sampling mini-batch scheme for experience replay, which may ignore the importance of transition samples in experience replay buffer. Hence, a sampling scheme called PER has been reported [37] to make experience replay more efficiently, where the agent can more efficiently learn from some samples rather than other useless or redundant samples.

The probability of the j -th sampled experience on the basis of the absolute TD error is expressed as

$$\mathcal{P}_j^{\text{sam}} = p_j^{\chi_1} / \sum_j p_j^{\chi_1} \quad (29)$$

where the exponent χ_1 controls how much prioritization is used; if $\chi_1 = 0$, then it is a uniform case. $p_j = 1/\text{rank}(j) > 0$ is the priority assigned to the j -th experience where $\text{rank}(j)$ is the rank of experience j in the experience replay when the samples are ordered based on their absolute TD-error $|\vartheta_j|$.

However, since the experiences with high TD-error values are more frequently replayed, which leads to the state visitation frequency and thus causes the training process of DNN to oscillate or diverge. To handle this issue, we resort to importance-sampling weights:

$$W_j = 1 / (N_R \mathcal{P}_j^{\text{sam}})^{\chi_2} \quad (30)$$

where N_R denotes the size of the replay buffer, the parameter χ_2 controls the amount of correction which is used.

Finally, by employing the PER scheme into the target AC networks, the critic DNN loss function (21) and the actor update (26) can be rewritten as follows.

$$\hat{\mathcal{L}}(\omega) = \frac{1}{D} \sum_{j=1}^D (W_j \mathcal{L}_j(\omega)) \quad (31)$$

$$\nabla_{\theta} \hat{J}(\pi_{\theta}) = \frac{1}{D} \sum_{j=1}^D \left(\sum_{s_j \in \mathcal{S}} d^\pi(s_j) \sum_{a_j \in \mathcal{A}} \nabla_{\theta} \pi_{\theta}(s_j, a_j) Q(s_j, a_j, \omega) \right) \quad (32)$$

The above two equations are applied to update the parameters of the online critic and actor networks, respectively,

$$\omega(t+1) = \omega(t) + \beta_c(t) \vartheta(t) \nabla_{\omega} \hat{\mathcal{L}}(\omega(t)) \quad (33)$$

$$\theta(t+1) = \theta(t) + \beta_a(t) \nabla_{\theta} \hat{J}(\pi_{\theta}) \quad (34)$$

At each training time step, since the Q-function is shifted, the value estimations of target AC networks may be easy out of control if we directly use the constantly shifting set of Q-function values. We can update target critic and actor networks to gradually track the online critic and actor networks by using soft target updates with $\tau \gg 0$

$$\omega'(t+1) = \tau\omega(t) + (1-\tau)\omega'(t) \quad (35)$$

$$\theta'(t+1) = \tau\theta(t) + (1-\tau)\theta'(t) \quad (36)$$

As analyzed in Section II.C, the IoT devices in the same community contract with each other frequently so they have the good trusted relationship of each other. In this case, coordination and information sharing among the devices in the same community can decrease search time for optimized resource management. By contract, if the action is chosen independently according to the local information, each device has no information of actions selected by other devices when the actions are updated simultaneously. Consequently, the states observed by each device may fail to fully characterize the environment, and the network yields poor global performance. Hence, coordinated learning is adopted to improve the network performance.

The social network is divided into L groups (communities), where the l -th group consists of Z_l agents and the agents in the same group are also their neighboring agents. In coordinated learning manner, the expected global Q-function of each group can be approximated by the sum of local Q-functions of agent in the same group, which can be written as

$$Q^l(s^l, a^l, \omega) = \sum_{i=1}^{Z_l} Q^{l,i}(s^{l,i}, a^{l,i}, \omega) \quad (37)$$

where $s^{l,i}$, $a^{l,i}$ and $Q^{l,i}(s^{l,i}, a^{l,i}, \omega)$ are the network state, the selected action and the utility (Q-value) of the i -th agent in the l -th group, respectively. s^l and a^l denote the joint state and action of the l -th group, respectively.

The weights of the critic DNN of group l is updated as

$$\omega^l(t+1) = \omega^l(t) + \beta_c(t)\vartheta_l(t)\nabla\mathcal{L}(\omega^l(t)) \quad (38)$$

where $\vartheta^l(t)$ and $\mathcal{L}(\omega^l(t))$ are the TD error and the loss function of the l -th group, respectively.

In addition, the policy of the actor of the l -th group would be updated as

$$\theta^l(t+1) = \theta^l(t) + \beta_a(t)\nabla_{\theta}J^l(\pi_{\theta}) \quad (39)$$

where $J^l(\pi_{\theta})$ is the objective function of the l -th group.

The coordinated multi-agent DRL-PER based resource management approach for joint RB assignment and transmission power control is shown in **Algorithm 1**.

V. SIMULATION RESULTS AND ANALYSIS

In this section, conduct simulations are provided to evaluate the performance of the proposed resource management approach. We compare the proposed coordinated multi-agent DRL-PER based resource management approach (denoted as

Algorithm 1 Resource Management with Coordinated Multi-Agent DRL-PER

1: **Input:** Social network simulator and QoS requirements of all devices.
2: Initialize DNN critic network $Q(s, a, \omega)$ and actor network π_{θ} with weights ω and θ , respectively;
3: Initialize target critic network $Q'(s, a, \omega')$ and actor network $\pi_{\theta'}$ with weights ω' and θ' , respectively;
4: Initialize replay buffer with size N_R ;
5: **for** each episode $=1, 2, \dots, E$ **do**
6: **for** each time step $t=0, 1, 2, \dots, T$ **do**
7: **for** group $l=1, 2, \dots, L$ **do**
8: In each group l , each agent shares its observation;
9: Select the joint action a_t^l using (28);
10: Execute action a_t^l , then obtain a reward U_t^l using (17), and observe a new state s_{t+1}^l ;
11: Update the global Q-function $Q^l(s^l, a^l, \omega)$ of each group l using (36);
12: Store experience $e_t^l = (s_t^l, a_t^l, U_t^l, s_{t+1}^l)$ in the replay buffer $\mathcal{D}$. If the capacity of the relay memory is full, the least used historical experience is dropped;
13: **for** $j=1, 2, \dots, D$ **do**
14: Sample experience j with the probability $\mathcal{P}_j^{\text{sam}}$ by (29);
15: Calculate corresponding IS weight W_j by (30);
16: Update the priority of transition j based on $|v_j|$;
17: **end for**
18: Update the loss function $\hat{\mathcal{L}}(\omega)$ and weights ω of the critic network using (31) and (33), respectively;
19: Update $\nabla_{\theta}\hat{J}(\pi_{\theta})$ and weights θ of the actor network using (32) and (34), respectively;
20: Update the weights ω' and θ' of the AC networks using (35) and (36), respectively;
21: **end for**
22: **end for**
23: **end for**
24: **Output:** RB assignment and transmission power control.

proposed C-DRL-PER RM) with the following approaches: **1.** The most classic AC approach, which is capable of dealing with continuous valued state and action spaces and adopted to address the resource management problem in social cognitive IoT networks (denoted by AC RM [21]); **2.** The existing QoS-driven social-aware approach which decomposes the problem (14) into two subproblems: RB assignment and power control, which can be solved iteratively in a centralized way. However, it is just a single time slot optimization solution, which may converge to the suboptimal solution due to the ignorance of the long term benefit (denoted as Baseline 1 [11]); **3.** The greedy based resource management approach (denoted as Greedy RM), where each communication link greedily chooses its RB assignment and transmission power control strategy in a competitive manner.

We consider a single cell scenario with the cell radius of 500 m, where one BS locates at the center, both the CUEs and IoT devices are randomly located in the circular area. We repeat the simulation 100 times and calculate the mean of the simulation data. The social characteristics, community groups

Fig. 3. Convergence comparisons of various approaches.

and the social relationships are formed by using the social-aware D2D communication architecture [13]. Each D2D pair can only assign at most 2 RBs at one time slot. The channel bandwidth of one RB is 180 kHz which occupies 0.5 ms in the time domain. The noise spectral density is -174 dBm/Hz. The maximum transmit power of each D2D device and each CUE are 23 dBm, respectively. The maximum transmission distance of one D2D link is 100 m. The circuit power consumption of each CUE and each D2D pair is 17 dBm. The minimum data rate requirements of each CUE and D2D link with normal services are 3 bits/s/Hz, respectively. For URLLC services [1], [18], the SINR threshold is 7 dB, the transmission reliability is 99.999% with each message size being 500 bits and the latency threshold is 1 ms with $T_a=0.1$ ms and $T_{rp}=0.2$ ms.

The DNN model consists of three connected hidden layers, containing 200, 200, and 100 neurons, respectively. The learning rate is $\alpha = 0.01$ and discount factor is set to be $\gamma = 0.98$. The values of the factor vector $[\mu_1, \mu_2]$ in (17) depend on the QoS requirements of devices, and they are used to balance the benefit (i.e., EE performance) and the costs (i.e., the punishments of the unsatisfied communications links) shown in (17), for example, if the reliability requirement is 99.99%, we then set $\mu_1 = \mu_2 = 2 \times 10^4$; if the reliability requirement is 99.9999%, we then set $\mu_1 = \mu_2 = 2 \times 10^6$.

A. Convergence Comparisons

Fig. 4 indicates the iteration process of four approaches in terms of the EE and the transmission success probability, when social tie $e^{PS}=0.6$, the number of IoT device is $M=400$ and the number of CUEs is $K=60$.

Clearly, the proposed learning approach significantly achieves the higher EE performance and probability of satisfied services than those of other approaches. Especially, compared with AC RM [21] and greedy RM, the proposed approach has the faster convergence speed and less fluctuations by adopting by PER and coordinated learning mechanisms to improve the learning efficiency and convergence speed. The greedy RM approach [37] is a simple learning approach, but it achieves poor global performance, leading to the poor EE value and transmission success probability.

Fig. 4. Performance comparisons versus different normalized social tie values.

It is worth noting that the proposed approach has comparable convergence speed with the Baseline 1 [11], but the Baseline 1 needs a central controller (i.e., at the BS) to manage resource of devices with the heavy feedback. In addition, we would like to mention that our proposed approach achieves the better EE performance than the Baseline 1, but they have the comparable transmission success probability performance. The reason is that the the Baseline 1 [11] is a QoS-driven approach which mainly focuses on the QoS guarantees without taking EE optimization into consideration. Moreover, the optimization objective in the Baseline 1 [11] is just a single time slot optimization problem, where the resource management decision is only based on the current state. The single time slot resource management decision approach may converge to the suboptimal solution and obtain the greedy-search like performance due to the lack of the historical network state and the long term benefit.

B. Performance Comparisons vs. Social Tie Levels

Fig. 4 compares the network EE and the probability of satisfied services under different normalized values of the social tie e^{PS} , when the number of IoT device is $M=400$ and the number of CUEs is $K=60$. We can see that both the network EE and the satisfied probability enhance when the social tie value increases. This is because that as the increase of the social tie value, the more number of D2D communication links are allowed to be established among IoT devices, which increases the the spectrum utilization efficiency and sum capacity of the network. Hence, both the network EE and the satisfaction level are improved.

We find that our presented learning approach outperforms the greedy RM approach and the performance advantage gradually increases when the social tie value becomes larger. Because the social orientation information is utilized to form more D2D links and improve the spectrum utilization efficiency in the proposed learning approach. In addition, our presented learning approach has the better performance than both the Baseline 1 [11] and AC RM approach [21]. The reason is that our proposed approach applies both PER and coordinated learning mechanisms to optimize the learning

Fig. 5. Performance comparisons versus the number of IoT devices.

policy, and the optimized RB assignment and power control strategy can be achieved.

C. Performance Comparisons vs. IoT Device Density

Fig. 5 provides the performance comparisons versus the number of IoT devices M when $e^{PS}=0.6$ and $K=60$. As seen in Fig. 5(a), when the number of IoT devices increases, the network EE decreases. This is because that spectrum reusing incurs co-channel interference among communication links, and the interference becomes more pervasive and the power consumption increases under the high-density scenario of devices, leading to the EE degradation.

From Fig. 5(b), for all approaches, we observe that the probability of satisfied services can be maintained at a good level when the number of IoT devices is not large, but it decreases obviously when M is large. Such the performance degradation results from the following reasons: the decrease of received SINR values under serious co-channel interference; the network cannot completely support all increased communication services due to the limited radio resource. Hence, the probability of satisfied services reduces in this case, but our proposed learning approach still achieves the higher performance than other approaches.

D. Performance Comparisons Under Different Numbers of CUEs

Fig. 6 shows the performances of the EE and the probability of satisfied services under different numbers of CUEs K when $e^{PS}=0.6$ and $M=400$. Fig. 6 illustrates that the network performances increase with the increase number of CUEs. The reason is that with a fixed number of D2D devices, more CUEs will provide more RBs and better link conditions with higher probability (note: each CUE has one RB), and thus leading to a higher probability of successful delivering. At the same time, D2D devices have more radio resource to assign to meet their communication services. Hence, both the network EE and the probability of satisfied service enhance as the increase number of CUEs. Compared with greedy RM approach approach, other three approaches achieve the higher performance due to the resource management optimization.

Fig. 6. Performance comparisons versus the number of CUEs.

Fig. 7. Performance comparisons versus the reliability threshold.

E. Performance Comparisons Under Different Values of Reliability and Latency

Fig. 7 and Fig. 8 plot the performances of all approaches versus the reliability and latency requirements, respectively, when $K=60$, $e^{PS}=0.6$ and $M=400$. As shown in Fig. 7 and Fig. 8, for all approaches, both the network EE and the satisfied probability decrease as the reliability and latency

Fig. 8. Performance comparisons versus the maximum latency bound.

Fig. 9. Performance comparisons versus the minimum rates.

requirements increase. This indicates that the looser the reliability and latency constraints are, the larger network EE and satisfied probability we can achieve. As the reliability and latency constraints are stringent, both the power and RB resources need to be carefully managed to guarantee the high reliability and low latency requirements. Even through our proposed approach has the slight higher probability of satisfied services than that of the Baseline 1 [11], it can achieve a significant improvement of EE under the stringent constraints. This is because that our proposed approach takes the EE optimization in the learning frameworks, so the optimized RBs and power management policy is searched to improve EE performance. The proposed approach is adaptive to diverse reliability and latency-bounded constraints, thus achieving the higher probability of satisfied services.

F. Performance Comparisons vs. Different Minimum Rates

Fig. 9 shows the performances against different minimum rate requirements, when $K=60$, $e^{PS}=0.6$ and $M=400$. From Fig. 9(a), we can see that when the minimum rate requirement is less than a certain threshold, the network EE value decreases slightly, but drops obviously when the minimum rate is large. This can be explained that the approaches can still to operate at the optimal transmission rate when the minimum rate is not big, but more power needs to be transmitted to guarantee the increased minimum rates, which leads to the EE performance degradation. In addition, from Fig. 9(b), we find that the probability of satisfied services approximately remains unchanged when the minimum rate is less than 2 bit/s/Hz. However, it goes down obviously when the minimum rate is larger than 3 bit/s/Hz, because the limited radio and power resource cannot completely meet the increased minimum data rate requirements. We would like to mention that the performances of other approaches decrease quickly at the higher minimum data region, which indicates the our proposed approach is superior to other approaches.

G. Performance Comparisons vs. Interference Threshold

For the given $K=60$, $e^{PS}=0.6$ and $M=400$, the performances versus the interference threshold I_{max} are provided

Fig. 10. Performance comparisons versus the interference threshold.

Fig. 11. (a) CUEs distribute near the BS and (b) CUEs distribute near edges.

in Fig. 10. It can be observed from Fig. 10 that when the interference threshold increases, both the network EE and the probability of satisfied services improve as D2D devices become able to flexibly transit more power on RBs improve SINR values, but approximately holding constant when the threshold is beyond a certain value due to the limited transmit power budgets shown in Eq. (14). It is worth noticing that the gap between the proposed approach and other approaches is reduced with the interference threshold to some degree, because the interference threshold has less effect on the network performance since the available power budgets cannot induce the maximum interference threshold.

H. Impact of CUEs' distributions on Performance

Fig. 11. shows the two cases of the distributions of CUEs in the network, where Fig. 11 (a) indicates that CUEs distribute near the BS and Fig. 11 (b) indicates that CUEs distribute near edges, when the number of CUs, D2D devices and RBs are 6, 30 and 10, respectively. Note that we just take one simple example to demonstrate the impact of distributions of CUEs on the network performance.

The impact of distributions of CUEs of the two cases on the network performance are shown in Fig. 12. From Fig. 12, we can observe that both the EE performance and the probability of satisfied services of Case I are higher than those of Case II. The reason is that the channel gains from CUEs to the BS in Fig. 11 (a) is quit higher than those of Fig. 11 (b), leading

Fig. 12. Performance comparisons under different CUEs' distributions.

to the higher reviewed SINR values of the CUEs. In addition, in Fig. 11 (b), as CUEs are far away from the BS, more power needs to be transmitted to maintain the received SINR value due to the low channel gain. In this case, the cellular to D2D interference will increase in co-channel access, which decreases the performance of D2D pairs. Hence, the network performances are improved when CUEs distribute near the BS. This indicates that the distribution of CUEs can affect the performance of all the approaches.

VI. CONCLUSION

This paper presents a QoS-driven social-aware network architecture to integrate the social information into cognitive D2D-based IoT networks, in order to optimize the network EE performance and guarantee different QoS requirements by utilizing the social orientation characteristics. Then, we formulate resource management optimization problem as a multi-agent DRL leaning formulation under different QoS requirements. To improve the network performance and QoS satisfaction levels, the PER and coordinated learning mechanisms have been employed to enable communication links to work cooperatively in a distributed cooperative way. Simulation results verify that the effectiveness of the proposed coordinated multi-agent DRL-PER based resource management approach. Our future work will pay attention to extend the learning model in indoor environments [39]-[41].

REFERENCES

- [1] N. Saxena, A. Roy, B. J. R. Sahu, and H. Kim, "Efficient IoT gateway over 5G wireless: a new design with prototype and implementation results," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 97-105, Feb. 2017.
- [2] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for IoT networks: Computing multiple functions With antenna arrays," *IEEE Internet of Things J.*, vol. 5, no. 6, pp. 5296-5306, Dec. 2018.
- [3] M. Qin, L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Power-constrained edge computing with maximum processing capacity for IoT networks," *IEEE Internet of Things J.*, vol. 6, no. 3, pp. 4330-4343, Jun. 2019.
- [4] L. Ale, N. Zhang, H. Wu, D. Chen, and T. Han, "Online proactive caching in mobile edge computing using bidirectional deep recurrent neural network," *IEEE Internet of Things J.*, vol. 6, no. 3, pp. 5520-5530, Jun. 2019.

- [5] G. Li, J. Wu, J. Li, K. Wang, and T. Ye, "Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4702-4711, Oct. 2018.
- [6] M. Aazam, K. A. Harras, and S. Zeadally, "Fog computing for 5G tactile industrial Internet of Things: QoE-aware resource allocation model," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3085-3092, May 2019.
- [7] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in SDN-based Industrial Internet of Things with edge computing," *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1351-1360, Jun. 2018.
- [8] Q. Wu *et al.*, "Cognitive Internet of Things: A new paradigm beyond connection," *IEEE Internet of Things J.*, vol. 1, no. 2, pp. 129-143, Apr. 2014.
- [9] Z. Zhou, C. Gao, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Social big-data-based content dissemination in Internet of vehicles," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 768-777, Feb. 2018.
- [10] Z. Ning, F. Xia, X. Hu, Z. Chen, and M. S. Obaidat, "Social-oriented adaptive transmission in opportunistic Internet of smartphones," *IEEE Trans. Ind. Informat.*, vol. 13, no. 2, pp. 810-820, Apr. 2017.
- [11] B. Wang, Y. Sun, S. Li, and Q. Cao, "Hierarchical matching with peer effect for low-latency and high-reliable caching in social IoT," *IEEE Internet of Things J.*, vol. 6, no. 1, pp. 1193-1209, Feb. 2019.
- [12] Y. Ai, L. Wang, Z. Han, P. Zhang, and L. Hanzo, "Social networking and caching aided collaborative computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 149-155, Dec. 2018.
- [13] Y. Li, T. Wu, P. Hui, D. Jin, and S. Chen, "Social-aware D2D communications: Qualitative insights and quantitative analysis," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 150-158, Jun. 2014.
- [14] T. M. Chiewe, C. F. Mbuya, and G. P. Hancke, "Using cognitive radio for interference-resistant industrial wireless sensor networks: An overview," *IEEE Trans. Ind. Informat.*, vol. 11, no. 6, pp. 1466-1481, Dec. 2015.
- [15] W. Li, C. Zhu, V. C. M. Leung, L. T. Yang, and Y. Ma, "Performance comparison of cognitive radio sensor networks for industrial IoT With different deployment patterns," *IEEE Systems J.*, vol. 11, no. 3, pp. 1456-1466, Sept. 2017.
- [16] A. Shahini, A. Kiani, and N. Ansari, "Energy efficient resource allocation in EH-enabled CR networks for IoT," *IEEE Internet of Things J.*, vol. 6, no. 2, pp. 3186-3193, Apr. 2019.
- [17] S. Aslam, W. Ejaz, and M. Ibnkahla, "Energy and spectral efficient cognitive radio sensor networks for Internet of Things," *IEEE Internet of Things J.*, vol. 5, no. 4, pp. 3220-3233, Aug. 2018.
- [18] 5G: Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14), document 38.913, 2017.
- [19] M. Doudou, D. Djenouri, and N. Badache, "Survey on latency issues of asynchronous MAC protocols in delay-sensitive wireless sensor networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 528-550, Sec. 2013.
- [20] A. Orsino *et al.*, "Caching-aided collaborative D2D operation for predictive data dissemination in industrial IoT," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 50-57, Jun. 2018.
- [21] H. L. Yang, X. Z. Xie, and M. Kadochi, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157-4169, May 2019.
- [22] Z. Ning, X. Wang, X. Kong, and W. Hou, "A social-aware group formation framework for information diffusion in narrowband Internet of Things," *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1527-1538, Jun. 2018.
- [23] B. Fan, H. Tian, L. Jiang, and A. V. Vasilakos, "A social-aware virtual MAC protocol for energy-efficient D2D communications underlying heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8372-8385, Sept. 2018.
- [24] X. Liu and J. L. Xie, "Priority-based spectrum access in cognitive D2D networks for IoT," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, 2017, pp. 1-6.
- [25] P. Gandotra and R. K. Jha, "Zonal-based green algorithm for augmenting the battery life in spectrum shared networks via D2D communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 405-419, Jan. 2019.
- [26] Y. Li, K. Chi, H. Chen, Z. Wang, and Y. Zhu, "Narrowband Internet of Things systems with opportunistic D2D communication," *IEEE Internet of Things J.*, vol. 5, no. 3, pp. 1474-1484, Jun. 2018.
- [27] A. Y. Awan, M. Ali, M. Naeem, F. Qamar, and M. N. Sial, "Joint network admission control, mode assignment and power allocation in energy harvesting aided D2D communication," *IEEE Trans. Ind. Informat.*, to be published. DOI: 10.1109/TII.2019.2922667.

- [28] M. Ali *et al.*, "Devices to devices (Ds2Ds) communication: towards energy efficient IoT," in *Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.
- [29] S. Baidya and M. Levorato, "Content-aware cognitive interference control for urban IoT systems," *IEEE Trans. Cognitive Commun. Netw.*, vol. 4, no. 3, pp. 500-512, Sept. 2018.
- [30] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MD: MIT press, 2017.
- [31] H. Qian, J. Yu, and L. Hua, "Relay selection algorithm based on social network combined with Q-learning for vehicle D2D communication," *IET Commun.*, vol. 13, no. 20, pp. 3582-3587, May 2019.
- [32] Y. He, C. Liang, R. Yu and Z. Han, "Trust-based social networks with computing, caching and communications: A deep reinforcement learning approach," *IEEE Trans. Network Science and Engineering.*, vol. 7, no. 1, pp. 66-79, Jan.-Mar. 2020.
- [33] L. Tang, X. Chen and S. He, "When social network meets mobile cloud: A social group utility approach for optimizing computation offloading in cloudlet," *IEEE Access*, vol. 4, pp. 5868-5879, 2016.
- [34] Y. Zhang, K. Zhang, J. Cao, H. Liu, and S. Maharjan, "Deep reinforcement learning for social-aware edge computing and caching in urban informatics," *Appear IEEE Trans. Ind. Informat.*, DOI: 10.1109/TII.2019.2953189.
- [35] M. Feng, L. Qian, and H. Xu, "Multi-robot enhanced MANET intelligent routing at uncertain and vulnerable tactical edge," in *Proc. IEEE MILCOM*, Los Angeles, CA, 2018, pp. 1-9.
- [36] H. L. Yang, X. Z. Xie and M. Kadoch, "Machine Learning Techniques and A Case Study for Intelligent Wireless Networks," *Appear IEEE Network*. DOI: 10.1109/MNET.001.1900351.
- [37] H. L. Yang A. Alphones, C. Chen, W. D. Zhong, and X. Z. Xie, "Learning-based energy-efficient resource management by heterogeneous RF/VLC for ultra-reliable low-latency industrial IoT networks," *Appear IEEE Trans. Ind. Informat.* DOI: 10.1109/TII.2019.2933867.
- [38] H. L. Yang and X. Z. Xie, "An actor-critic deep reinforcement learning approach for transmission scheduling in cognitive Internet of Things systems," *IEEE Sys. J.*, vol. 14, no. 1, pp. 51-60, Mar. 2020.
- [39] H. L. Yang, W. Zhong, C. Chen, A. Alphones and P. Du, "QoS-driven optimized design-based integrated visible light communication and positioning for indoor IoT networks," *IEEE Internet of Things J.*, vol. 7, no. 1, pp. 269-283, Jan. 2020.
- [40] P. Zhao, *et al.*, "P3-LOC: A privacy-preserving paradigm-driven framework for indoor localization," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2856-2869, Dec. 2018.
- [41] P. Zhao, J. Li, F. Zeng, F. Xiao, C. Wang and H. Jiang, "ILLIA: Enabling k -anonymity-based privacy preserving against location injection attacks in continuous LBS queries," *IEEE Internet of Things J.*, vol. 5, no. 2, pp. 1033-1042, Apr. 2018.

Chen Chen (S'13-M'19) received both the B.S. and M.Eng. degrees from University of Electronic Science and Technology of China, Chengdu, China, in 2010 and 2013, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2017. He is currently a Tenured-Track Assistant Professor with the School of Microelectronics and Communication Engineering, Chongqing University, China. His research interests include visible light communications, Li-Fi, visible light positioning, optical access networks, and digital signal processing.

Arokiaswami Alphones (M'92-SM'98) received the B.Tech. degree from Madras Institute of Technology, Chennai, India, in 1982, the M.Tech. degree from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 1984, and the Ph.D. degree in optically controlled millimeter wave circuits from Kyoto Institute of Technology, Kyoto, Japan, in 1992. Since 2001, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Singapore. His current research interests include electromagnetic

analysis on planar RF circuits and integrated optics, microwave photonics, metamaterial-based leaky wave antennas, and wireless power transfer technologies.

Helin Yang (S'15) received Ph.D. degree from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2020. He serves as a reviewer for IEEE international journals such as IEEE Communications Magazine, IEEE Transactions on Wireless Communications, IEEE Transactions on Vehicular Technology, etc. His current research interests include wireless communication, visible light communication, Internet of Things and resource management.

Xianzhong Xie (M'19) received his Ph.D. degree in communication and information systems from Xidian University, China, in 2000. He is currently Professor of School of Optoelectronic Engineering, and Director of Chongqing Key Lab of Computer Network and Communication Technology, at Chongqing University of Posts and Telecommunications (CQUPT), China. His research interests include MIMO precoding, cognitive radio networks, and cooperative communications.

Wende Zhong (SM'03) was a Senior Research Fellow with the Department of Electrical and Electronic Engineering, University of Melbourne, Australia, from 1995 to 2000. He joined Nanyang Technological University, Singapore, in 2000, as an Associate Professor and became a Full Professor in 2009, and is currently with the School of Electrical and Electronic Engineering. His current research interests include visible light communication/positioning, optical fiber communication systems and networks, optical access networks, and signal processing.