Edge-Assisted Spectrum Sharing for Freshness-Aware Industrial Wireless Networks: A Learning-Based Approach

Mingyan Li, *Student Member, IEEE*, Cailian Chen, *Member, IEEE*, Huaqing Wu, *Member, IEEE*, Xinping Guan, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*— Information freshness is essential to industrial wireless networks (IWNs) and can be quantified by the age-of-information (AoI) metric. This paper addresses an AoI-aware spectrum sharing (AgeS) problem in IWNs, where multiple device-to-device (D2D) links opportunistically access the spectrum to satisfy their AoI constraints while maximizing primal links' throughput. Particularly, we orchestrate the access of D2D links in a distributed manner. Since distributed scheduling results in incomplete observation, D2D links share the spectrum with uncertainty on the transmission environment. Therefore, we propose a distributed scheduling scheme, called D-age, to deal with the transmission uncertainty in the AgeS problem, where an adaptation of actor-critic method is adopted with AoI constraints tackled in the dual domain. To address the non-stationary environment and multi-agent credit assignment issue, cooperative multi-agent reinforcement learning (MARL) approach is developed, where multiple local actors are designed to guide D2D links to make real-time decisions via distributed scheduling policies, which are evaluated by an edge-assisted global critic with action-aware advantage functions. Integrated with graph attention networks (GATs), the critic selectively learns contextual information by assigning different importances to neighboring links, which enables the evaluation of scheduling policies in a scalable and computation-efficient manner. Theoretical guarantee of the time-averaged AoI constraints is provided and the effectiveness of D-age in terms of both AoI violation ratio and the capacity of primal links is demonstrated by simulation.**

Manuscript received 26 April 2021; revised 17 September 2021 and 11 January 2022; accepted 7 March 2022. Date of publication 8 April 2022; date of current version 12 September 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1702100; in part by the National Natural Science Foundation of China under Grant 61933009, Grant 62025305, Grant 62172066, and Grant 62002042; in part by the Natural Sciences and Engineering Research Council of Canada; and in part by the Chongqing Natural Science Foundation (Postdoctoral) under Grant cstc2021jcyj-bsh0141. The associate editor coordinating the review of this article and approving it for publication was H. Zhang. *(Corresponding author: Cailian Chen.)*

Mingyan Li is with the College of Computer Science, Chongqing University, Chongqing 400030, China (e-mail: limy2021@cqu.edu.cn).

Cailian Chen and Xinping Guan are with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: cailianchen@sjtu.edu.cn; xpguan@sjtu.edu.cn).

Huaqing Wu is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: wu482@mcmaster.ca).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3160857.

Digital Object Identifier 10.1109/TWC.2022.3160857

***Index Terms*— Age of information, spectrum sharing, edge-assisted IWNs, multi-agent reinforcement learning.**

I. Introduction

INDUSTRIAL wireless networks (IWNs) enable ubiquitous connection and automated information transmission for tremendous devices, which have sensing, processing, and communication capabilities to support various monitoring applications. The performance of monitoring systems depends heavily on the freshness of state update packets, which can be captured by age of information (AoI) [1], also referred to as age. Age measures the time elapsed since the generation of the last packet delivered to the destination and is a paramount timeliness metric in industrial monitoring systems [2].

Moreover, as a promising solution for IWNs, the beyond fifth-generation (5G) networks are expected to support ubiquitous state monitoring, where the spectrum sharing paradigm can be utilized to improve both spectrum and energy efficiency. As an indispensable technique, the device-to-device (D2D) communication can enable multiple sensor-actuator device pairs to communicate directly via the LTE-A interface over cellular spectrum [3] and thus can provide real-time transmission with reduced communication hops due to the proximity of devices. However, each D2D link selfishly competes for the restricted resources in a fully distributed manner, which may lead to adverse effects on the transmissions of not only other D2D links but also traditional cellular uplinks.

As an emerging technology, edge intelligence (EI) has received lots of interest. The physical proximity between computing servers and information-generation sources promises several benefits, e.g., comprehensive data processing and resource management [4]. By pushing the artificial intelligence frontier to the edge, EI is also envisioned to boost a wide range of flexible scheduling strategies by hierarchical learning across end devices, edge nodes, and cloud datacenters [5]. Therefore, the potential of edge-assisted IWNs can be excavated, where the base station (BS) is connected with an edge server to enable intelligent link scheduling. Under this architecture, we consider a spectrum sharing problem for age-aware monitoring systems, which comprise both D2D links for timely state update and throughput-oriented cellular uplinks as shown in Fig. 1. Cellular uplinks, also named primal links, preoccupy the spectrum to deliver the traffic going outside a field network e.g., to a remote controller or edge server, whereas D2D links

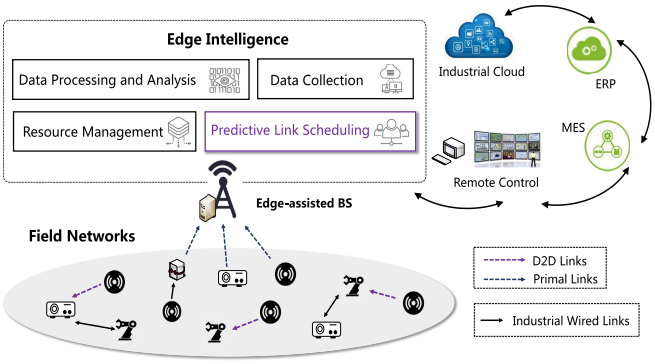


Fig. 1. A framework of IWNs based on edge intelligence.

communicate directly within the field network. To improve spectrum efficiency, D2D links attempt to share the spectrum occupied by primal links, where the co-channel interference should be minimized and thus a trade-off between the capacity of primal links and age guarantee needs to be further explored by D2D links with the assistance of the BS in the edge.

A. Related Work

Since age was first proposed [6], significant progress on the centralized age-aware link scheduling has been achieved [7]–[16]. The sum-age minimization problem was considered for broadcast networks in [7], [8], where the BS delivered updates over a single wireless link and could update at most one user for each time slot. Based on packet successful delivery probabilities, the age-aware scheduling problem under unreliable channels was firstly studied in [8]. To improve spectrum efficiency, the problem of constructing compatible link sets was investigated to enable the concurrent delivery of several links over a wireless channel under given interference models [9]–[11] or channel state information (CSI) [12]. Considering heterogeneous transmission demands, there are some works dealing with the joint optimization of age and throughput [11], [14]–[16]. Specifically, the works [14], [15] addressed the age minimization with throughput constraints, where scheduling policies were proposed in [14] based on dual or Lyapunov optimization. The authors in [15] defined a virtual queue as the difference between the required and actual throughput. Then, the long-term average constraint was turned into a queue stability requirement and the dual decomposition was utilized to develop an iterative method. Analogously, a centralized scheduler was proposed [11] to construct compatible link sets with throughput captured by the virtual queue. Then, this scheduler optimized the linear combination of the virtual-queue length and weighted-average age. The work [16] solved a multi-criteria optimization problem to minimize the time-averaged age and maximize throughput, where a linearized algorithm was proposed to transform this problem into mixed-integer linear programming so that a weakly Pareto-optimal point can be found by commercial software.

The above-mentioned works assume the network setup is known a priori. In practice, complete information on network state may not always be available, especially in emerging networking paradigms such as cognitive radio networks and ad-hoc networks. These networks are inherently distributed, where myopic devices autonomously access the networks with uncertainty on transmission environment (i.e., unknown link quality and interference from others). To address it, age-aware distributed solutions have gained attention recently due to the advantages of latency and overhead reduction [17]. Among them, many works adopted the method of orthogonal multiple access [17]–[23]. With this contention-based random access, most works were dedicated to enhance the CSMA [18]–[20] or ALOHA [21] framework, where the transmission probabilities of devices in each contention slot were designed based on local age information. Moreover, the authors in [22] proposed to reject some devices' access based on Whittle's index, which was identified as an important enhancement for current random access channel with massive connectivity. In cellular systems, a stochastic crowd avoidance algorithm was proposed [17] to schedule the uplink communication so as to avoid duplicated usage of resource blocks (RBs). The work [23] implemented the distributed semi-persistent scheduling scheme with a collaboration method to reduce packet collisions. To coordinate the multiple access of devices, the studies [17], [20], [23] allowed exchanging some age-related information through e.g., broadcasting [17], piggyback-based messages [23] and a common transmission tax signal [20].

The above age-aware distributed solutions [17]–[22] are well-suited to data collection tasks based on star topology. Different from these works as well as the previously discussed centralized solutions, we deal with the spectrum sharing problem under non-orthogonal multiple access [24], where adjacent D2D device pairs can reuse RBs within the cell to improve spectrum efficiency and thus potentially interfere with each other. A similar scenario was considered in [13] where multiple vehicle-to-vehicle links share the spectrum under the guidance of a centralized policy with global CSI. Since this age-aware spectrum sharing problem is correlated in successive time slots due to the age evolution of D2D links, distributed scheduling is more advantageous than centralized allocation since D2D links can instantly make scheduling decisions based on their current age states. Therefore, the distributed inband mode of D2D communication is adopted, where D2D packets are not delivered across the BS and thus conventional pilot signals for channel estimation cannot be utilized, resulting in no prior knowledge of transmission environment. In this respect, the problem of age-aware distributed spectrum sharing with transmission uncertainty needs further investigation.

B. Challenges and Contributions

As shown in Fig. 1, we consider a monitoring system with multiple D2D links in cellular-based IWNs, where local devices can transmit over multiple non-overlapping RBs. Due to the diversity of age requirements of D2D links, avoiding the violation of age constraints is more beneficial than minimizing the average age [25]. Moreover, we consider a cooperative age-aware optimizing problem so that competitive D2D links are encouraged to sacrifice for the greater good, which is

different from the previous distributed solutions. Specifically, *the global objective is to design an age-aware and distributed spectrum sharing scheme for D2D links with the assistance of EI so that diverse age constraints of D2D links can be guaranteed while the capacity of primal links can be maximized.*

To achieve it, each D2D link should autonomously sample and transmit state packets based on its current age state without prior knowledge of global CSI and scheduling strategies of other D2D links. Hence, we opt for model-free reinforcement learning (RL) approaches to deal with this transmission uncertainty [25]–[27]. One way to tackle the distributed spectrum sharing problem is to apply traditional RL methods for each agent of D2D links to learn transmission environment. This is the idea behind independent Q-learning (IQL), where multiple agents learn distributed policies from their own action-observation history. However, the environment keeps changing from the perspective of each agent as other co-learners update their policies, which is evidenced by the non-stationary environment issue. To circumvent it, deep multi-agent reinforcement learning (Deep-MARL) has gained lots of attention recently [20], [28]–[34], where deep neural networks (DNNs) are introduced for complex learning tasks. These works mainly focus on competitive learning problems.

Inspired by the development of Deep-MARL, we aim to design a learning-based spectrum sharing scheme so that the transmission uncertainty of D2D links can be efficiently addressed and the above cooperative objective can be satisfied. To this end, the following challenges need to be solved. *Challenge (i)*: Gradient-based learning methods are not directly applicable to our problem with age constraints. *Challenge (ii)*: Different from the previous MARL-related works, our problem aims to solve a cooperative MARL task, where the optimization of individual D2D links' scheduling policies is performed based on a global objective. As a result, the learning gradient for each partial-observable D2D link does not explicitly reflect how its action contributes to the global reward, which is known as the multi-agent credit assignment problem. *Challenge (iii)*: Learning capacity is another problem of large-scale scheduling for D2D links, where key environment information can be extracted to enable the efficient learning of distributed policies.

To solve these challenges, an age-aware spectrum sharing scheme, namely, D-age, is proposed for edge-assisted IWNs and our major contributions can be summarized as follows.

- According to the cooperative objective, an age-aware spectrum sharing (AgeS) problem is firstly formulated for diverse QoS provisioning, where transmitting spectrum and power are co-designed for individual D2D links. Then, this problem is addressed in the dual domain such that age constraints are handled in a distributed and adaptable manner and an analytical global objective for the learning-based policy optimization can be obtained.
- To tackle transmission uncertainty, D-age is designed by utilizing the framework of centralized training with decentralized execution. In specific, D-age contains multiple local actors to learn distributed scheduling policies for D2D links and a shared critic. The latter locates in the edge-assisted BS and takes the responsibility for information integration of primal links as well as policy evaluation via action-aware advantage functions. This collaborative hierarchy is efficiently integrated into the design of EI solutions for the AgeS problem. As a result, the non-stationary environment issue and multi-agent credit assignment problem are addressed.
- To enhance scalability, we further integrate the edge-assisted critic with graph attention networks (GATs). In this way, contextual information (e.g., the location and co-channel interference information of D2D links) can be efficiently utilized and extracted by the attention and graph mechanisms. As a result, the critic can intelligently select relevant knowledge when estimating the policies of local actors. This makes D-age computationally efficient and well amenable to large-scale graph-structured IWNs.

The remainder of this paper is organized as follows The system model and problem formulation are proposed in Section II. Then, primal-dual learning is introduced in Section III to address *Challenge (i)*. The detail of D-age is proposed in Section IV, which elaborates on the learning framework in terms of solving *Challenge (ii)*. To address *Challenge (iii)*, a novel GAT-based critic is designed in Section V. Finally, the performance analysis of D-age is provided in Section VI and Section VII concludes this paper.

II. System Model and Problem Formulation

In this section, we firstly describe the system model and then elaborate on the AgeS problem formulation and its dual counterpart. The key notations are summarized in Table I.

A. System Model

Consider a monitoring system in IWNs, which contains a set of cellular primal links $\mathcal{K} = \{1, \ldots, K\}$ for throughput-oriented uplink transmission and a set of transmitter-receiver D2D links $\mathcal{N} = \{1, \ldots, N\}$ for timely status update. A D2D link is denoted either by just device $n \in \mathcal{N}$ or by pair (n, n'). The devices of both primal and D2D links are equipped with the LTE-A interface under the coverage of an edge-assisted BS b, where several consecutive subcarriers are grouped to form orthogonal spectrum sub-bands, referred to as RBs. Primal links have been preassigned RBs with a fixed transmission power p_0, i.e., the kth primal link occupies the kth RB, and thus $\mathcal{K}$ denotes the set of RBs as well. Moreover, a time-slotted system is considered, which is indexed by t, and time slots are synchronized across all the devices by the BS.

For D2D links, state update packets usually contain a small amount of information, however, requiring stringently fresh data, which can be captured by the metric of age. Formally, the age of D2D link $n \in \mathcal{N}$ at time slot t is defined as $q_{n,t}$. It measures the time elapsed since the generation of the packet that was most recently delivered to the destination. We let D2D link n be characterized by an age constraint with its age limit

A_n, given by,

$$q_{n,t} \leq A_n, \quad \forall n \in \mathcal{N}. \tag{1}$$

In this work, we consider that D2D links are non-buffer systems, which focus on the freshest information. Thus, the generate-at-will policy [9] is employed such that whenever n is chosen to transmit a state packet at a certain time slot, it samples the packet at the beginning of that time slot. Based on the generate-at-will policy, the packet sampling and transmission can be jointly designed for D2D links and thus the age evolution of $q_{n,t}$ can be written as,

$$q_{n,t} = \begin{cases} q_{n,t-1} + 1, & \text{if } r_{n,t} = 0, \\ 1, & \text{if } r_{n,t} \neq 0, \end{cases} \tag{2}$$

where $r_{n,t}$ denotes the data rate of D2D link n at time slot t, which will be introduced in the next. Note that $q_{n,t}$ will reduce to 1 if a state packet is sampled and successfully received in the receiver of D2D link n at time slot t. Otherwise, $q_{n,t}$ will increase linearly with t.

B. Communication Model

The spectrum-power joint scheduling action is defined for D2D link $n \in \mathcal{N}$ as the tuple $\phi_{n,t} := \{c_{n,t}, p_{n,t}\} \in \Phi$ composed of the operating RB variable $c_{n,t}$ and transmit power variable $p_{n,t}$, where Φ is the action space and $\Phi = \{\mathcal{K} \times \mathcal{P}\}$. For the sake of practical circuit restriction, the power control option is discretized and limited to four levels,[1] i.e., $\mathcal{P} = [23, 10, 5, -200]$dBm, where the choice of -200dBm denotes no packet sampling and delivery [3]. To improve spectrum efficiency, D2D links are allowed to reuse the RBs, where colliding links transmit together resulting in interference. Hence, the signal to interference plus noise ratio (SINR) of D2D link n or primal link k at time slot t is given by

$$\begin{aligned} \zeta_{n,t} &= \frac{p_{n,t} g_{nn'}[c_{n,t}]}{\sigma^2 + \sum_{m \in \{C_{c_{n,t}} \setminus n\}} p_{m,t} g_{mn'}[c_{n,t}]}, \\ \zeta_{k,t} &= \frac{p_0 g_{kb}[k]}{\sigma^2 + \sum_{m \in \{C_k \setminus k\}} p_{m,t} g_{mb}[k]}, \end{aligned} \tag{3}$$

where σ^2 is the power of the additive white Gaussian noise. $C_k \subset \mathcal{N} \cup \{k\}$ is the link set including both primal link k and the D2D links operating over the kth RB. Similarly, $C_{c_{n,t}}$ includes the links operating over the $c_{n,t}$th RB. We consider channel fading is the same within an RB and independent across different RBs. During one time slot, the stochastic process of channel gain, $g_{nm}[k]$, of the link between transmitter n and receiver m over the kth RB, is defined as $\alpha^f_{nm} h_{nm}[k]$. Specifically, $h_{nm}[k]$ is the frequency dependent small-scale fading power component whereas α^f_{nm} denotes the frequency dependent large-scale fading effect and consists of path loss and shadowing. To ensure the successful decoding of packets, we set the SINR thresholds for both primal and D2D links as $\zeta_{p,min}$ and $\zeta_{d,min}$. ω is the bandwidth of each RB and the data rates of D2D links and primal links are defined as,

$$\begin{aligned} r_{n,t} &= \begin{cases} \omega \log(1 + \zeta_{n,t}), & \text{if } \zeta_{n,t} \geq \zeta_{d,min}, \\ 0, & \text{if } \zeta_{n,t} < \zeta_{d,min}, \end{cases} \\ r_{k,t} &= \begin{cases} \omega \log(1 + \zeta_{k,t}), & \text{if } \zeta_{k,t} \geq \zeta_{p,min}, \\ 0, & \text{if } \zeta_{k,t} < \zeta_{p,min}. \end{cases} \end{aligned} \tag{4}$$

C. AgeS Problem Formulation

Recall from the cooperative objective that we aim to design a spectrum sharing scheme for D2D links such that the data rates of primal links are maximized while guaranteeing the age constraints for D2D links. Thus, the primal AgeS problem, P, can be formulated with respect to the probability distribution of channel gain $\mathbf{g}$ as,

$$\begin{aligned} P: \max_{\boldsymbol{\phi}_t} \quad & \mathbb{E}_{\mathbf{g}} \left[\sum_{k=1}^{K} r_k(\boldsymbol{\phi}_t, \mathbf{g}) \right] \\ \text{s.t.} \quad & \mathbb{E}_{\mathbf{g}} \left[q_n(\boldsymbol{\phi}_t, \mathbf{g} | \mathbf{q}_{t-1}) \right] \leq A_n, \quad \forall n \in \mathcal{N}, \quad \forall t = 1, \ldots, T, \end{aligned} \tag{5}$$

where $\boldsymbol{\phi}_t = \{\phi_{1,t}, \ldots, \phi_{N,t}\}$ is a vector of joint scheduling decisions of D2D links at time slot t. Intuitively, the data rate of primal link k is determined by global CSI $\mathbf{g}$ and link scheduling decision $\boldsymbol{\phi}_t$, re-written as the function $r_k(\boldsymbol{\phi}_t, \mathbf{g})$. Similarly, the age of D2D link n is the function of $\mathbf{g}$ and $\boldsymbol{\phi}_t$ as well as the previous age state $\mathbf{q}_{t-1}$, denoted as $q_n(\boldsymbol{\phi}_t, \mathbf{g} | \mathbf{q}_{t-1})$. Both $r_k(\boldsymbol{\phi}_t, \mathbf{g})$ and $q_n(\boldsymbol{\phi}_t, \mathbf{g} | \mathbf{q}_{t-1})$ can be inferred by Eqn. (2)-(4).

To tackle the constrained problem P, we opt for dual optimization by firstly forming its Lagrangian dual as,

$$\mathcal{L}(\boldsymbol{\phi}_t, \boldsymbol{\lambda}_t) = \mathbb{E}_{\mathbf{g}} \left[\sum_{k=1}^{K} r_k(\boldsymbol{\phi}_t, \mathbf{g}) + \sum_{n=1}^{N} \lambda_{n,t} \left(A_n - q_n(\boldsymbol{\phi}_t, \mathbf{g} | \mathbf{q}_{t-1}) \right) \right]$$

where $\boldsymbol{\lambda}_t = (\lambda_{1,t}, \ldots, \lambda_{N,t})^{\mathrm{T}}$ is a vector of Lagrangian dual variables at time slot t. Define the dual objective $D(\boldsymbol{\lambda}_t)$ as an unconstrained maximization of $\mathcal{L}(\boldsymbol{\phi}_t, \boldsymbol{\lambda}_t)$. Then, the dual problem formulation D can be defined as,

$$\begin{aligned} D: \min_{\boldsymbol{\lambda}_t} \quad & D(\boldsymbol{\lambda}_t) \quad \text{where} \quad D(\boldsymbol{\lambda}_t) = \max_{\boldsymbol{\phi}_t} \mathcal{L}(\boldsymbol{\phi}_t, \boldsymbol{\lambda}_t) \\ \text{s.t.} \quad & \boldsymbol{\lambda}_t \geq \mathbf{0}, \quad \forall t = 1, \ldots, T. \end{aligned} \tag{6}$$

When $r_k(\boldsymbol{\phi}_t, \mathbf{g})$ is concave and $q_n(\boldsymbol{\phi}_t, \mathbf{g} | \mathbf{q}_{t-1})$ is convex, standard convex optimization results can guarantee that P and D have the same optimal value (i.e., $P^* = D^*$). In such a case, there is zero duality gap, also named strong duality.

D. Strong Duality Analysis of the AgeS Problem

In the non-convex optimization of multi-user multi-carrier systems as in Eqn. (5), the duality gap may exist (i.e., $P^* \leq D^*$). However, the duality gap can be zero under some conditions despite of non-convexity. To show this, the concept of time-sharing condition is firstly introduced.

Definition 1 Time-Sharing Condition [35]: Let $\boldsymbol{x}_t^*$ and $\boldsymbol{y}_t^*$ be the optimal scheduling decisions to the primal AgeS problem in Eqn. (5) at time slot t with corresponding $\boldsymbol{A} = \boldsymbol{A}^x$ and $\boldsymbol{A} = \boldsymbol{A}^y$, where $\boldsymbol{A} = (A_1, \ldots, A_N)^{\mathrm{T}}$ is a vector of age limits. The AgeS problem is said to satisfy the time-sharing condition if

[1]The partition granularity of power level can be increased for more fine-grained power control. However, it will take more time for the critic to explore the joint action space of multiple D2D users, which has the cardinality of $|\Phi|^N$.

TABLE I
THE LIST OF KEY NOTATIONS

Notation	Definition	Notation	Definition
n, m	D2D links in the D2D link set $\mathcal{N}$	k	Primal links in $\mathcal{K}$ or their operating RBs
$q_{n,t}, A_n$	Age of D2D link n and its age limit	$r_{n,t}, r_{k,t}$	Data rate of D2D link n and primal link k
$\phi_{n,t}, z_{n,t}, x_{n,t}, y_{n,t}$	Joint spectrum-power scheduling action of n	$\pi\langle\cdot\rangle$	Scheduling policy for D2D links
$\lambda_{n,t}, \beta$	Dual variable of n and its learning rate	$\boldsymbol{\theta}_{n,t}, \alpha$	Primal parameters of n and its learning rate
d_t	Dual value at each time slot t	$\xi_{n,t}, o_{n,t}$	Age margin and local observation of D2D link n
s_t, a_t	Global state and action of D2D links	$Q^c(s_t, a_t), \eta$	Action-state function and the learning rate
$\mathcal{A}_n^c(s_t, a_t)$	Advantage function for each local actor n	$\mathcal{S}_n$	Neighbor set of a primal or D2D link
$\rho_{n,t}, \kappa_{n,t}$	Input and output features of GAT-based critic	a_{nm}	Attention coefficients of GAT-based critic

for any $\boldsymbol{A}^x$ and $\boldsymbol{A}^y$, and for any $0 \le v \le 1$, there always exists a feasible decision $\boldsymbol{z}_t$, such that

$$\mathbb{E}_{\mathbf{g}}[q_n(\boldsymbol{z}_t, \mathbf{g}|\boldsymbol{q}_{t-1})] \le vA_n^x + (1-v)A_n^y, \quad \forall n \in \mathcal{N}$$
$$\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(\boldsymbol{z}_t, \mathbf{g})] \ge v\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(\boldsymbol{x}_t^*, \mathbf{g})] + (1-v)\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(\boldsymbol{y}_t^*, \mathbf{g})].$$

The multi-user multi-carrier spectrum optimization, e.g., the AgeS problem, has theoretically been proved to satisfy the time-sharing condition and the detail can be seen in [35]. The intuition behind this is that the maximum value P^* is an increasing function of the age limit $\mathbf{A}$, which implies P^* is a concave function of $\mathbf{A}$. Based on the characteristic of time-sharing condition, the AgeS problem is then proved to have the property of strong duality if feasible solutions exist as proposed in the following proposition.

Proposition 1: Define $A_{n,t}^{min} := min_{z_t} \mathbb{E}_{\mathbf{g}}[q_n(\boldsymbol{z}_t, \mathbf{g}|\boldsymbol{q}_{t-1})]$ as the minimum achievable age limit A_n for D2D link n. For the non-convex AgeS problem at time slot t as shown in Eqn. (5), if the age limit satisfies $\boldsymbol{A} > \boldsymbol{A}_t^{min} = (A_{1,t}^{min}, \ldots, A_{N,t}^{min})^{\mathrm{T}}$, it fulfills the characteristic of strong duality.

Proof: The proof's details are provided in the Appendix. ∎

Therefore, the AgeS problem can be solved in the dual domain with additional variables $\lambda_{n,t}$ introduced for each age constraint to form an unconstrained formulation.

III. PRIMAL-DUAL LEARNING FOR THE AGES PROBLEM

Although age constraints can be effectively addressed in the dual domain, finding a solution directly is not feasible since 1) the one-slot optimization problem should be solved in time sequence, where age evolution is correlated in successive time slots; 2) D2D links have no prior knowledge of transmission environment. In this section, we propose to harness the model-free primal-dual RL method to address the AgeS problem from the centralized and long-term perspective. Based on it, the distributed deployment will be proposed in the next section.

A. Primal-Dual Learning With DNN Parameterization

Taking age evolution into consideration, we firstly define a centralized policy $\pi^c\langle\boldsymbol{\phi}_t|\mathbf{q}_{t-1}, \mathbf{g}\rangle$, which outputs the scheduling solution $\boldsymbol{\phi}_t$ for D2D links based on global CSI and previous age state. As a result, the dual AgeS problem as shown in Eqn. (6) becomes a functional optimization problem, which entails a large computational complexity. Hence, a surrogate with the use of parameterization is utilized, accomplished by introducing a parametrization of $\pi^c\langle\boldsymbol{\phi}_t|\mathbf{q}_{t-1}, \mathbf{g}, \boldsymbol{\theta}\rangle$. $\boldsymbol{\theta} \in \Theta$ is defined as the primal parameter vector, where $\Theta \in \mathbb{R}^w$ is the space of primal parameters with a finite dimension w. For simplicity, we denote the centralized scheduling policy as $\pi^c\langle\boldsymbol{\theta}\rangle$. Accordingly, the dual problem is re-formulated as,

$$D_\theta: \min_{\boldsymbol{\lambda}_t} \max_{\boldsymbol{\theta}} \mathcal{L}(\pi^c\langle\boldsymbol{\theta}\rangle, \boldsymbol{\lambda}_t), \quad \text{s.t.} \quad \boldsymbol{\lambda}_t \ge \mathbf{0}, \quad \boldsymbol{\theta} \in \Theta, \tag{7}$$

where

$$\mathcal{L}(\pi^c\langle\boldsymbol{\theta}\rangle, \boldsymbol{\lambda}_t) = \mathbb{E}_{\mathbf{g}}\left[\sum_{k=1}^{K} r_k(\pi^c\langle\boldsymbol{\theta}\rangle, \mathbf{g}) + \sum_{n=1}^{N} \lambda_n(A_n - q_n(\pi^c\langle\boldsymbol{\theta}\rangle, \mathbf{g}|\mathbf{q}_{t-1}))\right].$$

Consequently, this optimization problem is performed over $\boldsymbol{\theta}$ rather than the policy $\pi^c\langle\boldsymbol{\theta}\rangle$ directly.

Solving such a surrogate incurs some inevitable loss of optimality. Nevertheless, this issue can be mitigated by exploiting well-known parametric functions with near-universal approximation property, e.g., deep neural networks (DNNs) [36], [37]. Moreover, Theorem 1 in [37] implies that the duality gap satisfies $P^* - D_\theta^* \le \left\|\boldsymbol{\lambda}^*\right\|_1 L\epsilon$. $\left\|\boldsymbol{\lambda}^*\right\|_1$ is the multiplier norm, which is related to the assumption stating that service demands can be provisioned with some slack. The positive constant L is introduced to guarantee Lipschitz continuity of scheduling policies. ϵ is defined as the error of approximating the scheduling policy with DNN parameterization. According to the universal approximation theorem of DNNs, a large class of functions can be approximated with an arbitrarily small ϵ using a DNN with only a single layer of arbitrarily large size. Therefore, the duality gap theoretically can be very small when using near-universal DNN parametrization.[2]

To address the dual problem in Eqn. (7), the method of stochastic gradient descent (SGD) is utilized to iteratively update both the primal parameters of DNN and dual variables

[2]DNNs cannot be arbitrarily large due to computation limitation, especially for large-scale D2D links, which may lead to unsatisfactory learning capability, i.e., large ϵ. This scalability issue will be addressed in Section V.

at each time slot t as,

$$\boldsymbol{\theta}_{i+1} = \Pi_{\Theta}\{\boldsymbol{\theta}_i + \alpha \nabla_{\boldsymbol{\theta}_i} \mathbb{E}_{\mathbf{g}} \underbrace{[\sum_k r_k(\pi^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g}) + \sum_n \lambda_{n,t,i}(A_n - q_n(\pi^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g}|\mathbf{q}_{t-1}))]}_{d(\pi^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g}|\mathbf{q}_{t-1})}\}, \tag{8}$$

$$\lambda_{n,t,i+1} = [\lambda_{n,t,i} - \beta(A_n - q_n(\pi^c\langle\boldsymbol{\theta}_{i+1}\rangle, \mathbf{g}|\mathbf{q}_{t-1}))]^+ \quad \forall n, \tag{9}$$

where i denotes the iteration number, α, β are the update stepsizes for primal parameters and dual variables, and Π_Θ is the projection operator to Θ. The gradient primal-dual updates in Eqn. (8)-(9) successively move primal and dual variables towards the maximum and minimum points of the Lagrangian function, respectively. $\lambda_{n,t,i+1}$ is updated using a *zeroth-order* representation, which is guaranteed to converge to the optimal as long as β is sufficiently small [38].

As for primal update, the gradient of $d(\pi^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g}|\mathbf{q}_{t-1})$ is needed, which can be considered as a function to output the dual value $d_t := \sum_k r_{k,t} + \sum_n \lambda_{n,t}(A_n - q_{n,t})$.[3] However, due to the unknown CSI $\mathbf{g}$, the straightforward evaluation of Eqn. (8) is intractable since neither the functions $r_k(\pi^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g})$ and $q_n(\boldsymbol{\pi}^c\langle\boldsymbol{\theta}_i\rangle, \mathbf{g}|\mathbf{q}_{t-1})$ nor their gradients are explicitly known a priori. Hence, the above primal-dual learning method is considered as a baseline method upon which we can develop a model-free algorithm.

B. Model-Free RL for the AgeS Problem

For primal update, the policy gradient method of RL is adopted with the aim to maximize the following expected long-term discounted reward[4] at time slot t_0:

$$\mathbb{E}_{\mathbf{g}}[\sum_{t=t_0}^{T} \gamma^{t-t_0} d_t],$$

where $\pi^c\langle\boldsymbol{\theta}_t\rangle$ is learned by continuous interactions with the environment and is used to make scheduling decisions in time sequences according to the age evolution of D2D links. Specifically, the primal parameters of $\pi^c\langle\boldsymbol{\theta}_t\rangle$ are updated at each time slot based on Policy Gradient Theorem [37], where the gradient of $d(\pi^c\langle\boldsymbol{\theta}_t\rangle, \mathbf{g}|\mathbf{q}_{t-1})$ can be inferred as,

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{g}}[\sum_t \gamma^{t-t_0} d(\pi^c\langle\boldsymbol{\theta}_t\rangle, \mathbf{g}|\mathbf{q}_{t-1})] = \mathbb{E}_{\mathbf{g}}[\sum_t \gamma^{t-t_0} d_t \nabla_{\boldsymbol{\theta}_t} \log \pi^c\langle\boldsymbol{\theta}_t\rangle].$$

d_t is the feedback of dual value and can be considered as the scheduling reward for spectrum sharing at time slot t. Then, the primal update can be performed based on SGD similar to Eqn. (8) to maximize the long-term discounted reward.

[3]To extend this work to the case where the age constraints of primary links are considdered, extra dual variables $\lambda_{k,t}$ can be introduced with dual value $d_t := \sum_k r_{k,t} + \sum_k \lambda_{k,t}(A_k - q_{k,t}) + \sum_n \lambda_{n,t}(A_n - q_{n,t})$.

[4]Depending on the value of T, there are three cases: (1) $T = t_0$, which is the degenerating case. In this case, D2D users will not sample and transmit state packets if their age constraints are not violated without considering the long-term influence. (2) $1 < T < \infty$ and (3) $T \to \infty$, which are the finite and infinite-horizon cases, respectively. Planning algorithms with finite horizon are forced to maintain different plans for different time slots, leading to undesired complexity. For example, the optimal policy near the end of the time horizon, might differ substantially from the optimal choice earlier in time, even under identical conditions (e.g., same state). Moreover, our problem is a continuous scheduling task without obvious ending points, and thus the RL-based AgeS problem aims to maximize the long-term discounted reward with infinite-horizon, that is $T \to \infty$.

IV. D-Age: An Edge-Assisted Distributed Scheduling Orchestrator for the AgeS Problem

There are some challenges that limit the implementation of the centralized scheduling policy in large-scale D2D-enabled IWNs, e.g., 1) the exploration of large action space is time-consuming; and 2) high latency from waiting for the orders of the BS may lead to outdated decisions. Hence, a distributed scheduling orchestrator, D-age, is proposed. Under the guidance of D-age, D2D links independently learn their scheduling policies, which forms an MARL problem. However, traditional RL approaches such as Q-Learning or policy gradient are poorly suited to this multi-agent learning environment because 1) the dual value is not accessible at D2D links since only the BS has the information on the capacity of primal links; and 2) the non-stationarity and multi-agent credit assignment problem may cause instability during the distributed learning of the cooperative AgeS problem. D-age deals with these problems by adopting the edge-assisted learning framework of centralized training and decentralized execution (CTDE) [30].

A. Edge-Assisted Distributed Scheduling for D2D Links

1) Partially-Observable Markov Decision Process (POMDP) of D2D Links: Under MARL problem, each D2D link n makes decisions based on the distributed scheduling policy defined as $\pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle$, which is parametrized with its own primal parameters $\boldsymbol{\theta}_{n,t}$ and also referred to as $\pi_n^d\langle\boldsymbol{\theta}_{n,t}\rangle$ for simplicity. It inputs local observations and takes decisions without the information on other D2D links known a prior, which can be described as a POMDP, given by:

- The observation space O, where $o_{n,t} = \{\lambda_{n,t-1}, \xi_{n,t-1}\} \in O$ is defined as the local observation of D2D link n at time slot t. $\xi_{n,t-1}$ is calculated as the age margin to capture the trend of age violation and $\xi_{n,t-1} := A_n - q_{n,t-1}$.
- The action space Φ, where $\phi_{n,t} = \{c_{n,t}, p_{n,t}\} \in \Phi$ is the scheduling decision at time slot t.
- The transition probability $\Pr(o_{t+1}|o_{n,t}, \phi_{n,t})$ denotes the probability of transiting to o_{t+1} at the next time slot by performing the decision $\phi_{n,t}$ under the current state $o_{n,t}$.
- The instantaneous reward d_t, which is the global dual value shared by all D2D links.
- The discount factor $\gamma \in (0, 1)$.

Since the global state, defined as s_t, is partially-observed by each D2D link n with local observation $o_{n,t}$, the transition probability $\Pr(o_{n,t+1}|o_{n,t}, \phi_{n,t})$ may not be stationary as other D2D links update their policies. As a result, the transmission environment is non-stationary from the perspective of any individual D2D link. Moreover, all scheduling policies of D2D links $\pi_n^d\langle\boldsymbol{\theta}_{n,t}\rangle$ are trained with the global reward d_t, which encourages D2D links to sacrifice for greater common benefits. However, the multi-agent credit assignment problem exists. As a result, each D2D link's policy is learned in a way that is not explainable by the global feedback.

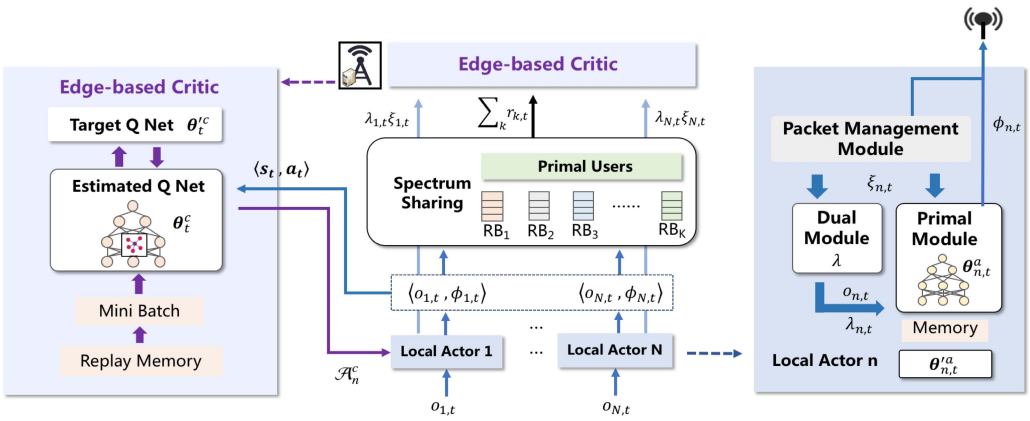


Fig. 2. Learning framework overview.

2) Actor-Critic Learning Based on the CTDE Framework: To deal with non-stationary environment, we propose to train a single meta-critic that can criticise each scheduling policy under the CTDE framework. In specific, each D2D link has a local actor and all the actors share a global critic which estimates the action-state function at each time slot t,

$$Q^c(s_t, a_t) = \mathbb{E}_{\mathbf{g}}\left[\sum_t \gamma^{t-t_0}\Big(\sum_k r_{k,t} + \sum_n \lambda_{n,t}\xi_{n,t}\Big)|s_0 = s_t, a_0 = a_t\right],$$

where $s_t = (o_{1,t}, \ldots, o_{N,t})$ and $a_t = (\phi_{1,t}, \ldots, \phi_{N,t})$ are the global state and action of D2D links at t, respectively. Based on $Q^c(s_t, a_t)$, the critic outputs the Q value as a prediction of long-term discounted reward $\sum_t \gamma^{t-t_0} d_t$ with the beginning state and action $s_0 = s_t$ and $a_0 = a_t$. Since only the BS has the access to the data rates of primal links, the critic is set to be located in the BS, where computationally complex works can be done with the assistance of edge computing.

The action-state function is trained with the deep Q network (DQN) algorithm, where the skills of off-line training and experience replay can be performed due to the stable environment brought by the global critic [39]. Formally, the Q network of $Q^c(s_t, a_t)$ is trained by the temporal-difference (TD) error with the learning rate η,

$$Loss_t^c = \mathbb{E}_{\mathbf{g}}\left[Q^c(s_t, a_t) - \Gamma_t\right], \tag{10}$$

where

$$\Gamma_t = d_t + \gamma Q^{c\prime}(s_{t+1}, \boldsymbol{\pi}^{d\prime}\langle s_{t+1}\rangle). \tag{11}$$

$Q^{c\prime}$ is the target Q network and $\boldsymbol{\pi}^{d\prime}\langle s_{t+1}\rangle$ denotes the joint action in the next scheduling outputted by the target policy networks $\pi_n^{d\prime}$ of local actors. This off-policy learning is able to improve the training stability of both actors and the critic [40].

3) Advantage Functions of Critic: The distributed scheduling policy of local actor n is trained with the feedback Q value based on policy gradient methods and thus sometimes suffers from high variance of gradient estimates. Therefore, a state-dependent bias $b^c(s_t)$ is often adopted so that the distributed policy is updated with the advantage function $\mathcal{A}^c(s_t, a_t)$ rather than just $Q^c(s_t, a_t)$, where $\mathcal{A}^c(s_t, a_t) = Q^c(s_t, a_t) - b^c(s_t)$.

Algorithm 1 The Workflow of Local Actors

Input: Set $\mathcal{D}^a, \alpha, \beta, \tau$; Initialize π_n^d, $\pi_n^{d\prime}$;
for $t = 1, 2, 3 \ldots$ **do**
1. Get the local observation $o_{n,t}$ and output the scheduling decision $\phi_{n,t}$ via $\pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}^a\rangle$;
2. Compute age margin $\xi_{n,t}$ locally and update dual parameter $\lambda_{n,t}$ in an adaptive manner based on

$$\lambda_{n,t} = \begin{cases} \lambda_{n,t-1} - \beta\xi_{n,t}\ , & \text{if } \xi_{n,t} < 0 \\ 0 \qquad\quad , & \text{if } \xi_{n,t} \geq 0 \end{cases}$$

3. Deliver $o_{n,t}$, $\phi_{n,t}$, $\xi_{n,t}$, $\lambda_{n,t}$, and $\hat{\pi}_{n,t}$ to the critic;
4. Receive the advantage function from the critic and then generate a sample $\{o_{n,t}, \mathcal{A}_n^c\}$;
if Get a batch of $\mathcal{D}^a$ samples for an episode **then**
Update primal parameters to renew the estimated policy network π_n^d based on

$$2\boldsymbol{\theta}_{n,t+1} = \Pi_\Theta\{\ \boldsymbol{\theta}_{n,t} + \alpha\mathbb{E}_{\mathbf{g},\pi_{-n}^d}[\mathcal{A}_n^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\boldsymbol{\theta}_{n,t}\rangle]\}; \tag{12}$$

end
5. Update the target policy network $\pi_n^{d\prime}$ based on

$$\boldsymbol{\theta}_{n,t}'^a \leftarrow \tau\boldsymbol{\theta}_{n,t}^a + (1-\tau)\boldsymbol{\theta}_{n,t}'^a.$$

end

However, the multi-agent credit assignment problem in our fully cooperative task will cause exacerbated high-variance issues. For example, a local actor cannot distinguish the performance of an action when receiving a low reward $Q^c(s_t, a_t)$ since it may be caused by other links' unsatisfactory actions. To mitigate this issue, an action-dependent advantage function is proposed for each D2D link, defined as,

$$\mathcal{A}_n^c(s_t, a_t) = Q^c(s_t, a_t) - b_n^c(s_t, a_t^{-n}). \tag{13}$$

Algorithm 2 The Workflow of the Edge-Assisted Critic

Input: Set $\gamma, \mathcal{D}^c, M, \eta, \tau$; Initialize Q^c, $Q^{c\prime}$;
for $t = 1, 2, 3 \ldots$ **do**
1. Keep collecting local information and get a mini-batch $\{e_i\}_{i\in\{1\ldots M\}}$;
2. Output corresponding advantage functions for each actor $\mathcal{A}_n^c$ according to Eqn. (13)-(14);
3. Observe the primal links' data rates and get the instantaneous dual value d_t;
4. Train the estimated Q network Q^c based on $\{e_i\}_{i\in\{1\ldots M\}}$ according to Eqn. (10)-(11);
5. Update the target Q network $Q^{c\prime}$ based on

$$\boldsymbol{\theta}_t^{\prime c} \leftarrow \tau\boldsymbol{\theta}_t^c + (1-\tau)\boldsymbol{\theta}_t^{\prime c}.$$

end

$b_n^c(s_t, a_t^{-n})$ is the counterfactual multi-agent baseline [31], which is calculated as,

$$b_n^c(s_t, a_t^{-n}) = \sum_{\phi_n} \pi_n^d\langle\phi_n|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle Q^c(s_t, a_t^{-n}, \phi_n). \tag{14}$$

where a_t^{-n} and $\boldsymbol{\pi}_{-n}^d$ denote the joint action and scheduling policy of actors except n, respectively. The impact of other D2D links is excluded in $\mathcal{A}_n^c(s_t, a_t)$ so that only the actions that directly influence the rewards are encouraged. Accordingly, the gradient of local actor n is re-written as,

$$\mathbb{E}_{\mathbf{g},\boldsymbol{\pi}_{-n}^d}\left[\mathcal{A}_n^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\boldsymbol{\theta}_{n,t}\rangle\right].$$

Although the global $Q^c(s_t, a_t)$ is shared among all D2D links, $\mathcal{A}_n^c(s_t, a_t)$ is customized to each D2D link. Moreover, this action-dependent baseline is theoretically shown to be better than the state-only baseline in terms of variance reduction [31].

To learn $Q^c(s_t, a_t)$ and output $\mathcal{A}_n^c(s_t, a_t)$, the following information is required by the critic: 1) the local observation $o_{n,t}$ and scheduling action $\phi_{n,t}$; and 2) the action distribution $\hat{\pi}_{n,t} = (\pi_n^d\langle\phi_1|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle, \pi_n^d\langle\phi_2|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle, \ldots, \pi_n^d\langle\phi_{|\Theta|}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle)$. The critic collects this information to generate an experience $e_t = \{s_t, a_t, d_t, s_{t+1}, a_{t+1}\}$, and accumulates a dataset of experiences. At each time slot, the critic randomly selects a mini-batch of experiences with the batchsize M to train $Q^c(s_t, a_t)$. Then, $\mathcal{A}_n^c(s_t, a_t)$ should be outputted by the critic to D2D links so that distributed policies can be trained. All the information can be delivered after scheduling decision making since the training of DNNs is off-line.

B. Framework Overview and Workflow of D-Age

The learning framework overview of D-age is delineated in Fig. 2. For local actors, the packet management module is in charge of monitoring state sampling and data encapsulation. Each local actor also has a memory module, which stores training samples of an episode. The sizes of memory for both the critic and actors are set to $\mathcal{D}^c$ and $\mathcal{D}^a$, respectively. Moreover, multiple DNNs are constructed for both policy networks in local actors and Q networks in the critic with the parameters $\boldsymbol{\theta}_{n,t}^a$ and $\boldsymbol{\theta}_t^c$ as well as $\boldsymbol{\theta}_{n,t}^{\prime a}$ and $\boldsymbol{\theta}_t^{\prime c}$ for their target networks. Formally, the workflows of actors and the critic are summarized in Alg. 1 and Alg. 2, respectively. Each actor iteratively updates its primal parameters of DNNs in the primal module and dual variables in the dual module. Dual variables are trained locally while primal parameters are learned based on advantage functions. To improve learning stability, target networks for both actors and the critic (i.e., $\pi_n^{d\prime}$ and $Q^{c\prime}$) are renewed using the soft target update as in DDPG [40]. Thus, target networks are updated by slowly tracking their estimated networks (i.e., π_n^d and Q^c) with $\tau \ll 1$ according to Step 5 in Alg. 1 and Alg. 2. Based on the updated policy networks, actors autonomously make scheduling decisions for D2D links with local observations at the beginning of each time slot so that reactions can be made in a timely and distributed manner.

C. Convergence Analysis of D-Age

Comparing Eqn. (12) with Eqn. (8), the distributed primal update for D2D link n includes the extra stochastic process $\boldsymbol{\pi}_{-n}^d$, which denotes the transmission interference from other D2D links. However, the global training of the critic guarantees the stationary environment for learning regardless of local scheduling policy changes. As a result, the following theorem can be introduced to validate the convergence of distributed scheduling policies in an ergodic manner considering the stochastic processes of channel and other D2D links' policies.

Theorem 1: For the long-term AgeS problem, where D2D links follow the local actors and the edge-assisted critic of D-age with $\gamma \to 1$, if the learning rates of actors and the critic satisfy,

$$\sum_{t=0}^{\infty}\alpha_t = \infty, \quad \sum_{t=0}^{\infty}\alpha_t^2 < \infty, \quad \sum_{t=0}^{\infty}\eta_t = \infty,$$
$$\sum_{t=0}^{\infty}\eta_t^2 < \infty \quad \text{and} \quad \lim_{t\to\infty}\frac{\alpha_t}{\eta_t} = 0, \tag{15}$$

(i) the policy gradient of each D2D user n follows,

$$\liminf_t \mathbb{E}_{\mathbf{g},\boldsymbol{\pi}_{-n}^d}[\mathcal{A}_n^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\boldsymbol{\theta}_{n,t}\rangle] = 0$$

with probability 1; and (ii) by setting a smaller age limit $A_n' < A_n$ for D2D link n, its ergodic age constraint can be feasible with probability 1, i.e.,

$$\lim_{T\to\infty}\frac{1}{t}\sum_{t=0}^{T} q_{n,t} < A_n \quad w.p.1.$$

Proof: The proof's details are provided in the Appendix. ∎

Although optimal solutions may not be achieved due to transmission uncertainty, the SGD-based primal update can guide D2D links to find local maximum of long-term discounted reward and their scheduling policies asymptotically converge with the almost sure feasibility of time-averaged age constraints. However, respecting the age constraints in an ergodic manner does not imply that the instantaneous one will always be satisfied, which is inevitable due to multiple stochastic processes. Alternatively, enforcing D-age with $A_n' < A_n, \forall n$, will increase the probability of satisfying instantaneous age constraints. Here, we define the age limit margin as $A_n^{mar} = A_n - A_n'$ and its influence on the performance of D-age is evaluated in Section VI.

V. Architecture Design of Learning Networks for D-Age

It is still challenging to directly implement D-age for large-scale D2D links due to unsatisfactory learning capability. To improve the learning efficiency of D-age, GATs are integrated in this section to address the scalability issue brought by the centralized training of critic.

A. GAT-Based Critic With Edge Computing

The edge-assisted critic is trained through the methods of DQN to provide advantage functions for each actor where DNNs are constructed for Q networks. However, the centralized training of the critic may suffer from dimension disaster [32] and thus achieve unsatisfactory learning accuracy, making D-age not applicable to large-scale IWNs. Although this issue can be mitigated with edge computing, the learning capability of traditional multi-layer perception (MLP) layers is limited due to the insufficient action exploration of D2D links. Hence, a more computationally efficient network should be considered. In light of the graph-structured communication pattern of D2D links, GATs [41] are utilized to establish a novel network architecture of $Q^c(s_t, a_t)$. The traditional MLP is replaced with GATs as shown in Fig. 3(b). With the mechanisms of attention models and graph neural networks, GATs are well-suited to our AgeS problem since the interference from remote links is insignificant. Therefore, the key idea is to design the Q network from a link-level perspective, which enables the critic to deal with the features of neighborhoods with different attention coefficients and thus redundant parameters (i.e., connections between neurons) can be pruned.

Specifically, GATs only operate on groups of spatially close neighbors and thus there is no need to know the graph structure upfront. Therefore, a neighborhood set for each D2D link and primal link n is defined in advance as $\mathcal{S}_n \in \{\mathcal{N} \cup \mathcal{K}\}$ where $n \in \{\mathcal{N} \cup \mathcal{K}\}$. Abusing the notation a bit, we will also use n, m to denote the primal user if no confusion can be caused. Based on $\mathcal{S}_n$, the critic learns a specific attention pattern for each link n so that it can focus on the most relevant part of the feature input as shown in Fig. 3(a). Formally, the input to the GAT layer is a set of links' features, defined as $\boldsymbol{\rho}_t = \{\rho_{1,t}, \rho_{2,t}, \ldots, \rho_{N+K,t}\} \in \mathbb{R}^{(N+K)\times F}$, where F is the cardinality of features. Each link feature then passes through a shared weight matrix $\mathbf{W}^s \in \mathbb{R}^{F'\times F}$, which is an MLP layer and serves as a linear transformation of $\rho_{n,t}$ into a higher-level feature $\rho'_{n,t}$ with the cardinality F'. Accordingly, the attention coefficient a_{nm} between any two links can be obtained via an MLP-based attentional mechanism, given by $a_{nm,t} = a^c\left(\mathbf{W}^s\rho_{n,t}^T, \mathbf{W}^s\rho_{m,t}^T\right)$. The attention mechanism $a^c(\cdot): \mathbb{R}^{F'} \times \mathbb{R}^{F'} \to \mathbb{R}$ consists of multiple MLP layers to calculate the similarity between n and m, which well captures the dynamical non-linear relation between two links' performance. Then, the neighbor set $\mathcal{S}_n$ is utilized to obtain the masked and normalized attention coefficient $a'_{nm,t}$ for link n so as to drop the redundant information from remote links, given by [41]

$$a'_{nm,t} = \text{softmax}\left(a_{nm,t}\right) = \frac{\exp\left(\text{LeakyReLU}(a_{nm,t})\right)}{\sum_{i\in\mathcal{S}_n}\exp\left(\text{LeakyReLU}(a_{ni,t})\right)},$$

where softmax$(\cdot)$ is the softmax function and LeakyReLU$(\cdot)$ is the LeadkyReLU activation function used to add nonlinearity. $a'_{nm,t}$ indicates the importance of m' feature to n, based on which the output of GAT layer $\kappa_{n,t}$ for n is computed as the weighted sum of the higher-level features of its neighbors,

$$\kappa_{n,t} = \text{LeakyReLU}\left(\sum_{m\in\mathcal{S}_n} a'_{nm,t}\rho'_{m,t}\right).$$

To analyze the neighbors' features from multiple perspectives, $\kappa_{n,t}$ can also be obtained by concatenating L independent attention mechanisms, followed as,

$$\kappa_{n,t} = \|_{l=1}^{L}\,\text{LeakyReLU}\left(\sum_{m\in\mathcal{S}_n} a'^{l}_{nm,t}\rho'^{l}_{m,t}\right).$$

The overall architecture of $Q^c(s_t, a_t)$ is shown in Fig. 3(b). To obtain an advantage function, a global state-action pair (s_t, a_t) serves as an input for the embedding layer to output the feature $\rho_{n,t}$ for each link by the concatenation $\rho_{n,t} = (\text{MLP}(o_{n,t}) + \text{Lookup}(\phi_{n,t}) + \text{MLP}(loc_n))$, where loc_n is the device location of link n.[5] The embeddings $\{\rho_{n,t}\}_{n=1}^{N+K}$ of both D2D and primal links are fed into the GAT layer. Then, the output features $\{\kappa_{n,t}\}_{n=1}^{N+K}$ are generated, which are divided into two flows for the prediction of long term primal links' data rate and D2D links' dual punishment, respectively. Finally, Q values are obtained by combining these two parts,

$$Q^c(s_t, a_t) = \sum_k Q_k^c(\kappa_{k,t}) + \sum_n Q_n^c(\kappa_{n,t}). \tag{16}$$

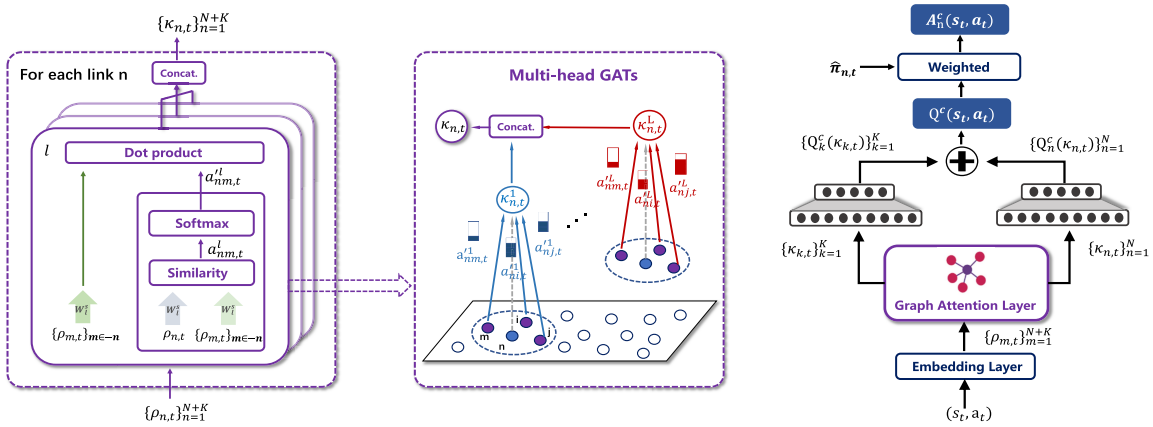
$Q_{k/n}^c(\cdot)$ denotes the 2-layer MLP for calculating the component of $Q^c(s_t, a_t)$ of each link, which is evaluated by considering the non-linear relation with other links. Based on GATs, this Q value can be obtained with the refined decomposition representation from a link-level perspective, which corresponds to the dual problem shown in Eqn. (7).

By leveraging GATs, the learning capability of $Q^c(s_t, a_t)$ can be largely improved compared with the one with traditional MLP layers. The intuition behind this is that traditional MLP layers may face the over-fitting issue due to the incomplete sampling of large joint-action space while the GAT-based critic handles this problem by changing the network architecture, which learns the hidden representations of each link through the masked attention mechanism. Moreover, the multi-head GATs allow for assigning different importances to links of the same neighborhood, enabling a leap in model capacity.

B. Complexity Analysis

Recall that scheduling decisions can be made instantly based on local information (defined as the feature with cardinality F). Hence, for local actors with the action space of $|\Phi|$, the time complexity mainly stems from both the forward process of policy networks and action distribution sampling, where the time complexity of $O(n_{DNN} + |\Phi|n_l + log|\Phi|)$ is required. n_{DNN} is the number of parameters in the DNN-based policy network except the output layer, which linearly scales with F, and n_l is the input cardinality of the output layer. Compared with that in centralized scheduling, which scales with the joint-action

[5]Both $o_{n,t}$ and loc_n are passed through a 1-layer MLP and $\phi_{n,t}$ passes through an embedding layer for feature reconstruction.



(a) Illustration of multi-head GATs (b) Overall Q network architecture

Fig. 3. Illustration of GATs and network architecture of GAT-based critic.

TABLE II
SIMULATION PARAMETERS

Parameter	Value ($d[m]$)
TTI duration	1ms
Carrier frequency f	4GHz
RB bandwidth ω	180kHz
Tx power of primal links	23dBm
Tx power of D2D links	$[23, 10, 5, -200]$dBm
Fading, shadowing st. dev.	Reyleigh, 4dB
noise power	-174dBm/Hz
LOS path loss	$18.7\log(d) + 20\log(f/5) + 46.8$
NLOS path loss	$36.8\log(d) + 20\log(f/5) + 43.8 + 5N_{wall}$

space cardinality $|\Phi|^N$, the time complexity of actors decreases from exponential with $|\Phi|$ to linear with $|\Phi|$ [3].

To guarantee the learning efficiency of local actors, most computational tasks are offloaded to the critic. This global GAT-based critic is computationally efficient and scalable to the networks that cover arbitrarily large geographical areas if the number of links per unit area remains upper bounded. Specifically, the time complexity of Q value computation comes from two parts: the GAT attention layer and 2-layer MLP. The time complexity of a single GAT attention head is expressed as $O((N+K)FF' + |\mathcal{S}|F')$, where $|\mathcal{S}|$ denotes the maximum size of $\mathcal{S}_n$ for all links [41]. For multi-head attention, the critic just needs to multiply the storage requirements by a factor of L since individual head' computation is independent and thus can be parallelized. Then, the time complexity of 2-layer MLP is $O(n_i F')$, where n_i is the output cardinality of its input layer. As shown in Eqn. (16), Q value computation can also be parallelizable from a link-level perspective and thus the total time complexity is $O((N+K)FF' + |\mathcal{S}|F' + n_i F')$. The last step is to calculate the advantage function $\mathcal{A}_n^c(s_t, a_t)$ for each link with $|\Phi|$ times the time complexity of Q value computation and still can be parallelly calculated. Therefore, edge computing can assist the parallel operation and makes the GAT-based critic work more efficiently.

VI. Performance Evaluation

A. Simulation Setup

In this section, the performance of D-age is evaluated with the relevant parameters for wireless settings listed in Table II [42]–[44], where d is the distance between a transmitter and a receiver, f is the operating frequency. Specifically, both D2D and primal links are randomly positioned inside a $600\times600m^2$ industrial service area, where the distance and age limit of D2D device pairs are randomly set from $[0, 30]m$ and $[4, 8]$, respectively. All devices operate based on the WINNER II indoor model, which is a link-level channel model for LTE system design [43]. In the non-line of sight (NLOS) path loss model, N_{wall} denotes the number of walls between two locations and is set to be $d/50$. Both the SINR thresholds $\zeta_{d,min}$ and $\zeta_{p,min}$ are set to be $0dB$ [3].

The related parameters of D-age are shown in Table III. To improve learning capability from the sequences of observations, actors are designed based on the long short term memory (LSTM) to address the partially-observable issue [28]. Hence, the skill of clipping gradient is used, which efficiently solves the gradient explosion issue caused by LSTM. For the GAT-based critic, $\mathcal{S}_n$ is defined as the links within the distance of 200m. Since we deal with a non-episodic task, which has no clear ending point, γ is set to be 0.5 to focus more on the recent rewards. Both actors and the critic are trained with Adam optimizer. For local actors, the learning rate is set as $\alpha = 0.0002$ while for the critic, η is set to be 0.001 for the first 400 time slots and then to 0.0005 for the remaining.

B. Simulation Results

To evaluate the performance of D-age, the following metrics are considered: 1) age violation ratio, the percentage of D2D links whose age constraints are violated; 2) capacity, calculated as the sum of primal links' data rates; 3) dual value d_t; 4) average power consumption and sleep ratio of D2D links, where the latter denotes the percentage of D2D links choosing $-200dBm$; and 5) learning capability, that includes the classification error and the binary cross entropy loss (BCELoss)

TABLE III
LIST OF LEARNING HYPERPARAMETERS

Hyperparameters for Critic	Values	Hyperparameters for Actor	Values
M, $\mathcal{D}^c$,	32, 480	$\mathcal{D}^a$	20
Activation function	LeakyReLU	Activation function	Tanh
Initialization	Normal (0,0.2)	Initialization	Normal (0,0.2)
# attention mechanisms L	8	Hidden state size of LSTM	50
Output feature cardinality of GAT, F'	32	Threshold for clipping gradient	40

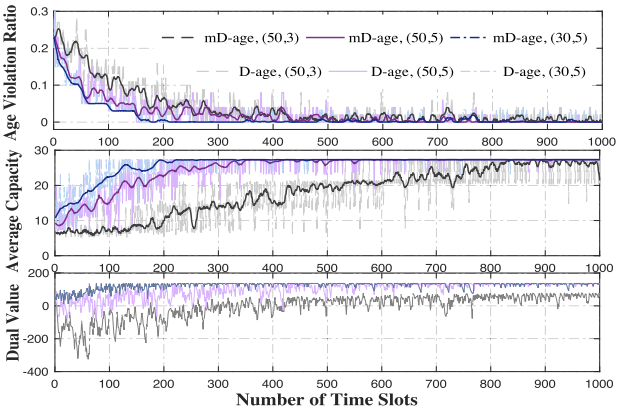


Fig. 4. Performance of D-age $\beta = 1$ with different sizes (N, K).

of the critic. Based on those metrics, D-age is compared with the following schemes,

- Shared [37], where only policy gradient and primal-dual learning are used and all D2D links share the same feedback reward d_t to train their scheduling policies;
- Individual, which is an improved version of Shared and has customized reward for each D2D link, calculated as $\sum_{k=1}^{K} r_{k,t} + \lambda_{n,t}\xi_{n,t}$ to focus on individual age constraints;
- DRQN [28], [29], where independent learning is performed via LSTM-based DQN for each D2D link to deal with partially-observed environment.

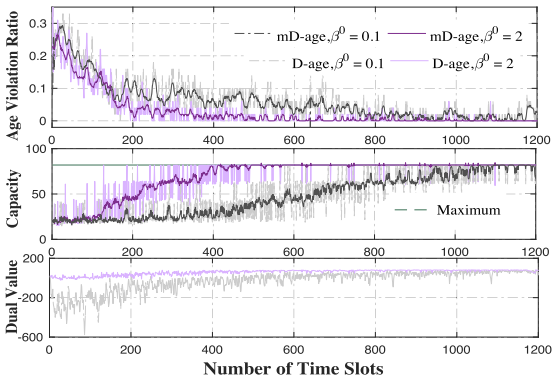
To clearly show the trend of data, we smoothen some curves, which are prefixed by *m*.

1) Performance of D-Age With Different Parameters: The performance of D-age is firstly evaluated under three network scenarios with different number of D2D links and primal links, denoted by the tuple (N, K) (i.e., (50,3), (50,5), (30,5)), shown in Fig. 4. Here, the average capacity (per RB) is evaluated due to the different numbers of RBs. It can be observed that the results of D-age with 5 RBs are much better than that with 3 RBs in terms of three metrics. This shows the importance of the resources/links density since D-age can help D2D links find suitable RBs to lessen the interference only if there are enough RBs. Moreover, D-age (30,5) achieves a greater solution in terms of data rate (up to the maximum value 27.32) and age violation ratio (down to 0.02) compared with D-age (50,5). Although D-age (50,5) has a slower convergence due to the larger number of D2D links, both D-age (50,5) and (30,5) can guide D2D links to learn stable scheduling policies. Moreover, D-age (50,3) has the highest age violation ratio and largest fluctuation due to insufficient resources. However, D-age (50,3) still can reach the maximum average capacity despite the longer learning process.

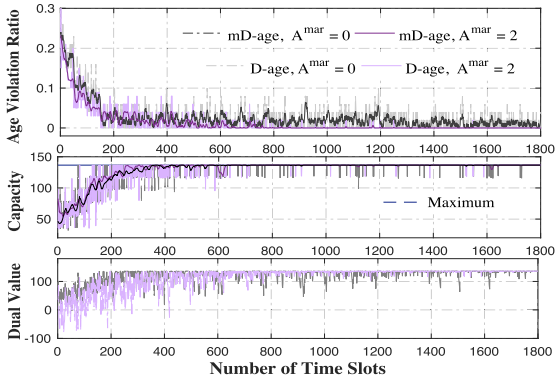
D-age with different dual learning steps β is evaluated under the network size (50,3) in Fig. 5(a). Note that the learning rate of dual variables has a great influence on the convergence rate of age violation ratio. Although β should be small enough to guarantee the optimal solution, a larger learning rate $\beta^0 = 2$ can help accelerate the age violation ratio reduction since more dual punishment is imposed. To be specific, D-age with a suitable learning rate can guide a D2D link to rank the RBs that may cause severe interference to primal links low, and also pay attention to the RBs with high interference from other D2D links so as to reduce age violation ratio. Accordingly, the dual value of D-age $\beta^0 = 2$ is larger and more stable than that with $\beta^0 = 0.1$ despite the larger degree of dual punishment. Then, the performance of D-age with different age limit margins is evaluated in Fig. 5(b). It can be clearly seen that D-age $A_n^{mar} = 2$ can largely improve the final age violation ratio compared with D-age $A_n^{mar} = 0$ since D-age $A_n^{mar} = 0$ only ensures the expected age constraints. As a result, the dual value of D-age $A_n^{mar} = 0$ keeps fluctuating due to unsatisfactory age violation ratio.

2) Comparison of D-Age With Different Scheduling Policies: D-age is compared with different scheduling policies in Fig. 6(a). In specific, Shared, the policy with the same reward for all D2D links, performs worst despite the global reward feedback since the global action space is not fully explored. As a result, useful information cannot be efficiently excavated due to the misleading credit assignment of D2D links. The curve of Shared's age violation ratio will deteriorate in the end, leading to the worst dual value. Individual functions better than Shared in three metrics as it pays more attention to age states. In other words, D2D links can be more selfish under the guidance of Individual so that age can be decreased. However, it is still confusing for D2D links because they cannot differentiate the influence of their actions on the capacity of primal links. Consequently, the age violation ratio performance of Individual is unstable in the end. As for D-age, the customized and action-dependent advantage $\mathcal{A}_{n,t}^c$ can effectively mitigate this weakness. Hence, more stable and better performance on three metrics can be achieved.

In Fig. 6(b), the performances of average power consumption and sleep ratio based on the last 800 time slots are evaluated. Compared with Individual and Shared, D-age has a lower average power consumption and higher sleep ratio in most networks since D-age gives advantage functions from the long-term and comprehensive perspective. Moreover, the

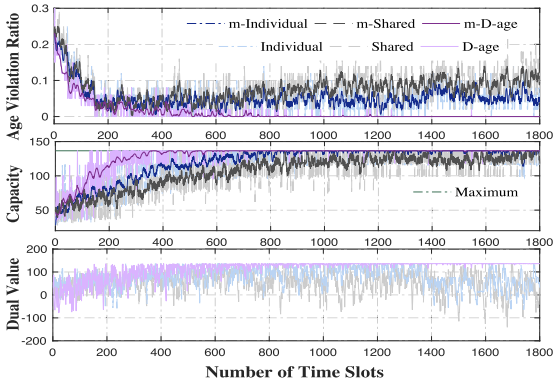


(a) Dual learning step comparison under (50,3)

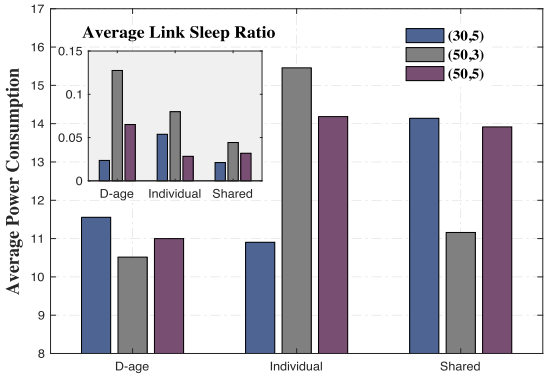


(b) Age limit margin comparison under (50,5)

Fig. 5. Performance of D-age with different parameters.



(a) Comparison of learning with different rewards



(b) Comparison of power consumption

Fig. 6. Comparison of D-age with different scheduling policies.

power consumption of D-age in the network (30,5) is higher than that in the networks (50,3) and (50,5) as D-age will encourage some D2D links that are far from primal links to increase their power and awake time in order to enhance the possibility of successful delivery. In the network (50,3) with the larger density of links, D-age tends to choose lower power levels to decrease the co-channel interference. It also prefers to guide the D2D links who are less likely to violate their age limits to shut down the sampling and delivering processes to save energy, leading to the highest sleep ratio. A similar trend is also seen in the results of Shared. Nevertheless, there is much misleading information for D2D links under the guidance of Shared, resulting in high power consumption and low sleep ratio. Individual performs the completely opposite solutions. For example, it achieves the highest power consumption in the most crowded network (50,3) as selfish D2D links only aim to increase power levels for successful delivery.

D-age is compared with DRQN in Fig. 7. Since there is no centralized critic in DRQN, the same reward as Individual is adopted and only the performances of age violation ratio and capacity are delineated. Although there is a little improvement of capacity after network training for DRQN, the result of age violation ratio gets worse, which is even worse than random scheduling (data before the start of training). This implies that DRQN is unable to find out useful information with

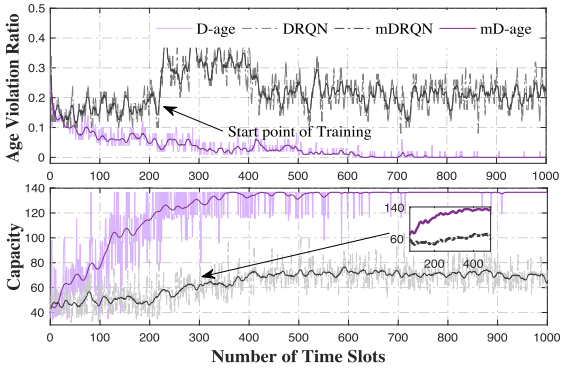


Fig. 7. Comparison of D-age with DRQN.

only LSTM-based policy networks, while D-age can tackle the non-stationary issue using the CTDE learning framework.

3) Results of GAT-Based Critic: To evaluate the learning accuracy, we compare the function approximation capability of critic with GAT (Critic-GAT, L = 1) and that with conventional MLP layers (Critic-MLP, i.e., the block of GAT in Fig. 3 is replaced by a 3-layer MLP). To be specific, the critic predicts whether D2D links will have successful delivery at each time slot with a decreasing learning rate ('lr') under the

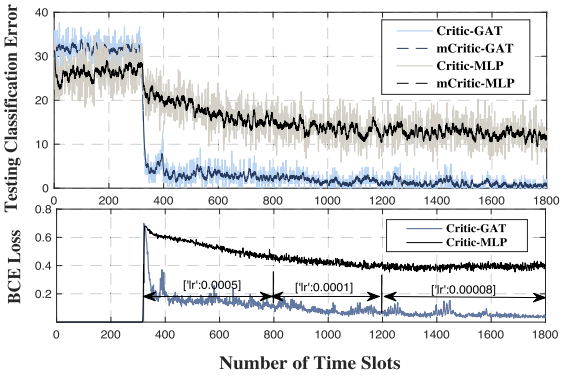


Fig. 8. Learning capability of GAT-based critic.

network size (50,5). The performance in terms of classification error and BCELoss is delineated in Fig. 8. It can be seen that the learning accuracy of Critic-GAT is much better than that of Critic-MLP as the training of Critic-MLP tends to be over-fitting due to insufficient action exploration. As a result, Critic-MLP cannot decrease the testing classification error down to 15. As for Critic-GAT, the curves of both classification error and BCELoss can rapidly converge to a lower level since Critic-GAT just pays attention to the neighbors of links rather than the whole 55 links. Therefore, useless network branches can be cut out to some degree to avoid over-fitting.

VII. Conclusion

In this paper, we have investigated the age-aware spectrum sharing problem for the edge-assisted IWNs and proposed a learning-based distributed spectrum scheduling scheme, D-age, to jointly optimize the transmitting RB and power for D2D links. The AgeS problem has been formulated as a cooperative MARL problem, where the non-stationary environment and multi-agent credit-assignment problem are tackled by the CTDE framework with local actors and an edge-assisted critic. Local actors learn individual scheduling policies for D2D links. These policies are assessed by the critic via action-aware advantage functions, obtained by GAT-based Q networks to filter out unimportant information. In general, the scalability of GAT-based critic makes D-age well-suited for distributed learning systems. Moreover, D-age provides a potential solution for future studies on the scheduling of graph-structured IoT networks with ubiquitous monitoring requirements. For our future work, we will consider the age-aware multi-hop communication for monitoring tasks, where the routing should be optimized based on the graph network topology.

Appendix

Proof of Proposition 1: We follow the proof of *Theorem 1* in [35]. Firstly, $A_{n,t}^{min} := min_{z_t} \mathbb{E}_{\mathbf{g}}\left[q_n(z_t, \mathbf{g}|\boldsymbol{q}_{t-1})\right]$ is defined as the minimum achievable age limit of D2D link n in the worst case. This guarantee that the slater's condition holds [36] i.e., for any $A_{n,t} > A_{n,t}^{min}, \forall n$, there is a strictly feasible solution z_t satisfy $\mathbb{E}_{\mathbf{g}}\left[q_n(z_t, \mathbf{g}|\mathbf{q}_{t-1})\right] < A_n, \forall n$. Since the proposed AgeS problem satisfies the time-sharing conditions, we could get that, for two distinct achievable upper bounds $\mathbf{A}^x, \mathbf{A}^y \geq \mathbf{A}_t^{min}$ with the optimal decisions $\boldsymbol{x}_t^*, \boldsymbol{y}_t^*$, we could find a feasible scheduling decision z_t with an age limit $\mathbf{A}^z = v\mathbf{A}^x + (1-v)\mathbf{A}^y$ for arbitrary constant $v \in [0,1]$ that satisfies

$$\begin{aligned}
&\mathbb{E}_{\mathbf{g}}[q_n(z_t, \mathbf{g}|\boldsymbol{q}_{t-1})] \leq vA_n^x + (1-v)A_n^y, \quad \forall n \\
&2\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(z_t^*, \mathbf{g})] \geq \mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(z_t, \mathbf{g})] \\
&\qquad \geq v\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(\boldsymbol{x}_t^*, \mathbf{g})] + (1-v)\mathbb{E}_{\mathbf{g}}[\sum_{k=1}^{K} r_k(\boldsymbol{y}_t^*, \mathbf{g})].
\end{aligned}$$

With the guarantees of slater's condition and time-sharing conditions, it can be concluded that the optimal objective $r^*(\mathbf{A})$ is a concave function of $\mathbf{A}$ with $\mathbf{A} \geq \mathbf{A}_t^{min}$. According to *Theorem 1* in [35], the primal problem and dual problem have the same optimum (i.e., $P^* = D^*$). ■

Proof of Theorem 1: The proof follows two steps. The first step is to prove the D-age with action-dependent advantages will converge under the framework of CTDE, i.e., at each iteration t, the total policy gradient of N local actors,

$$\mathcal{G}_t = \mathbb{E}_{\mathbf{g},\boldsymbol{\pi}}\left[\sum_n \mathcal{A}_n^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle\right] \tag{17}$$

will converge to 0 w.p.1.

In detail, the proof of the first step follows the convergence analysis in [31]. Formally, the gradient in Eqn. (17) is re-written as,

$$\begin{aligned}
\mathcal{G}_t = \mathbb{E}_{\mathbf{g},\boldsymbol{\pi}}[&\underbrace{\sum_n Q^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle}_{P1} \\
&\underbrace{- \sum_n b_n^c(s_t, a_t^{-n})\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle}_{P2}]
\end{aligned}$$

First, we consider the expected contribution of P2. For simplicity, we will omit the time slot superscript. Hence, we have

$$\begin{aligned}
&\mathcal{G}_{P2} \\
&= \mathbb{E}_{\mathbf{g}}[\sum_n \sum_{\boldsymbol{\phi}_{-n}} \pi_{-n}^d\langle\boldsymbol{\phi}_{-n}|\mathbf{o}_{-n}, \boldsymbol{\theta}_{-n}\rangle \sum_{\phi_n} \pi_n^d\langle\phi_n|o_n, \boldsymbol{\theta}_n\rangle \\
&\qquad \times \nabla_{\boldsymbol{\theta}_n} \log \pi_n^d\langle\phi_n|o_n, \boldsymbol{\theta}_n\rangle b_n^c(s_t, \boldsymbol{\phi}_{-n})] \\
&= \mathbb{E}_{\mathbf{g}}[\sum_n \sum_{\boldsymbol{\phi}_{-n}} \pi_{-n}^d\langle\boldsymbol{\phi}_{-n}|\mathbf{o}_{-n}, \boldsymbol{\theta}_{-n}\rangle \sum_{\phi_n} \nabla_{\boldsymbol{\theta}_n} \pi_n^d\langle\phi_n|o_n, \theta_n\rangle b_n^c(s_t, \boldsymbol{\phi}_{-n})] \\
&= \mathbb{E}_{\mathbf{g}}[\sum_n \sum_{\boldsymbol{\phi}_{-n}} \pi_{-n}^d\langle\boldsymbol{\phi}_{-n}|\mathbf{o}_{-n}, \boldsymbol{\theta}_{-n}\rangle b_n^c(s_t, \boldsymbol{\phi}_{-n,t})\nabla_{\boldsymbol{\theta}_n} 1] = 0.
\end{aligned}$$

where $\boldsymbol{\phi}_{-n}$ is the same with a^{-n}. This means that P2 does not change the expected gradient due to the bias-free baseline. Thus, the convergence of P1 is guaranteed by the traditional single-agent actor-critic policy gradient. Specifically, P1 can be inferred as,

$$\begin{aligned}
\mathcal{G}_{t,P1} &= \mathbb{E}_{\mathbf{g},\boldsymbol{\pi}}[\sum_n Q^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_{n,t}} \log \pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle] \\
&= \mathbb{E}_{\mathbf{g},\boldsymbol{\pi}}[Q^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_t} \log \boldsymbol{\Pi}_n \pi_n^d\langle\phi_{n,t}|o_{n,t}, \boldsymbol{\theta}_{n,t}\rangle] \\
&= \mathbb{E}_{\mathbf{g}}[Q^c(s_t, a_t)\nabla_{\boldsymbol{\theta}_t} \log \pi^c\langle\boldsymbol{\phi}_t|\boldsymbol{o}_t, \boldsymbol{\theta}_t\rangle]
\end{aligned}$$

where $\pi^c\langle\boldsymbol{\phi}_t|\boldsymbol{o}_t, \boldsymbol{\theta}_t\rangle$ denotes a global policy corresponding to $Q^c(s_t, a_t)$ and outputs the joint scheduling action $\boldsymbol{\phi}_t$. This yields

the standard single-agent actor-critic policy gradient, which is proved to converge to a local maximum, given that [45], 1) the policy π is differentiable; 2) Q^c is a compatible TD(μ) critic, where $\mu = 1$ and is a decay parameter denoting the impact of previous states on the current state, and Q^c uses a representation compatible with π_n^d; 3) The step size of actors should be negligible compared to that of the critic so that the actor looks stationary as far as the critic is concerned. Moreover, all learning rates of actor and critic should be non-increasing so that the learning process could slow down gradually while not stop (ensuring that stochastic processes are ergodic). To this end, the convergence conditions of learning rates should be followed as in Eqn. (15) [45].

The second step is to prove that the converged policies with a smaller age limit A_n' can guide D2D links to satisfy their time-averaged age constraints. We define $-\lambda_{n,t}\xi_{n,t}$ as a positive $c_{n,t}$ for each actor upon convergence. Let $\{c_{n,t}\}_{t\geq 0}$ be a sequence of positive real numbers. Before proceeding, we make the following mild assumptions: A1)

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} c_{n,t} = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} c_{n,t},$$

which denotes that the limit of the average of the sequence $\{c_{n,t}\}_{t\geq 0}$ exists, and A2) the discount factor γ used is such that

$$\liminf_{\nu\uparrow 1}(1-\nu)\sum_{t=0}^{\infty}\nu^t c_{n,t} \leq \sum_{t=0}^{\infty}\gamma^t c_{n,t} + M_0$$

for some $0 < M_0 < \infty$ [46]. Then, we consider the following lemma of Abel theorem [47] as,

Lemma 1 Abel theorem [47]: Let $\{c_t\}_{t\geq 0}$ be a sequence of positive real numbers, then

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} c_{n,t} \leq \liminf_{\nu\uparrow 1}(1-\nu)\sum_{t=0}^{\infty}\nu^t c_{n,t}.$$

Follows (A1)-(A2) and ***Lemma*** 1 that

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T} c_{n,t} \leq \liminf_{\nu\uparrow 1}(1-\nu)\sum_{t=0}^{\infty}\nu^t c_{n,t} \leq \sum_{t=0}^{\infty}\gamma^t c_{n,t} + M(\gamma)$$

for some $\gamma \to 1$ and a small $M(\gamma) = M_0 > 0$.

Recall that once Q^c and π_n^d converged, $\lim_t Q^c(s_t, a_t) = Q_0^c(s_t, a_t)$ $w.p.1$, denoting that Q values almost surely converge to fixed values and π_n^d will guide D2D links to choose the actions with the maximal Q value, i.e., the long-term reward $\mathbb{E}_{\mathbf{g}}[\sum_{t=0}^{\infty}\gamma^t(\sum_{k=1}^{K} r_{k,t} - \sum_{n=1}^{N} c_{n,t})]$. Based on it, we define $\lim_t \sum_{t=0}^{\infty}\gamma^t c_{n,t} = M_1$ as the small value that the learning converges to and set $M = M_1 + M_0$. Then, we have,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{\infty}\lambda_{n,t}\xi_{n,t} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{\infty}\lambda_{n,t}(A_n' - q_{n,t}) \geq -M.$$

and we can get,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{\infty} A_n' + M/\lambda_{n,t} = A_n' + M/\lambda_{n,t} \geq \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{\infty} q_{n,t}.$$

Based on it, we can conclude that by setting some smaller A_n' with $A_n' + M/\lambda_{n,t} \leq A_n$, the time-averaged age constraints can be satisfied and thus we finish the proof. ■

References

[1] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the role of age of information in the Internet of Things," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 72–77, Dec. 2019.

[2] R. V. Bhat, R. Vaze, and M. Motani, "Throughput maximization with an average age of information constraint in fading channels," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 481–494, Jan. 2021.

[3] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[4] Y. Liu, M. Peng, G. Shou, Y. Chen, and S. Chen, "Toward edge intelligence: Multiaccess edge computing for 5G and Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6722–6747, Aug. 2020.

[5] C. Zhou *et al.*, "Deep reinforcement learning for delay-oriented IoT task scheduling in SAGIN," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 911–925, Feb. 2021.

[6] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Mar. 2012, pp. 2731–2735.

[7] Y.-P. Hsu, E. Modiano, and L. Duan, "Scheduling algorithms for minimizing age of information in wireless broadcast networks with random arrivals," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2903–2915, Dec. 2020.

[8] I. Kadota, E. Uysal-Biyikoglu, R. Singh, and E. Modiano, "Minimizing the age of information in broadcast wireless networks," in *Proc. 54th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2016, pp. 844–851.

[9] Q. He, D. Yuan, and A. Ephremides, "Optimal link scheduling for age minimization in wireless systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5381–5394, Jul. 2018.

[10] S. Farazi, A. G. Klein, and D. R. Brown, "Fundamental bounds on the age of information in general multi-hop interference networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2019, pp. 96–101.

[11] N. Lu, J. Bo, and B. Li, "Age-based scheduling: Improving data freshness for wireless real-time traffic," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Jun. 2018, pp. 191–200.

[12] R. Talak, S. Karaman, and E. Modiano, "Improving age of information in wireless networks with perfect channel state information," *IEEE/ACM Trans. Netw.*, vol. 28, no. 4, pp. 1765–1778, Aug. 2020.

[13] M. K. Abdel-Aziz, S. Samarakoon, C.-F. Liu, M. Bennis, and W. Saad, "Optimized age of information tail for ultra-reliable low-latency communications in vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1911–1924, Mar. 2020.

[14] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 1844–1852.

[15] J. Sun, L. Wang, Z. Jiang, S. Zhou, and Z. Niu, "Age-optimal scheduling for heterogeneous traffic with timely throughput constraints," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1485–1498, May 2021.

[16] J. Lou, X. Yuan, S. Kompella, and N.-F. Tzeng, "Boosting or hindering: AoI and throughput interrelation in routing-aware multi-hop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 29, no. 3, pp. 1008–1021, Jun. 2021.

[17] T. Park, W. Saad, and B. Zhou, "Centralized and distributed age of information minimization with nonlinear aging functions in the Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 8437–8455, May 2021.

[18] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Miu, "Decentralized status update for age-of-information optimization in wireless multiaccess channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2276–2280.

[19] D. Yu, X. Duan, F. Li, Y. Liang, H. Yang, and J. Yu, "Distributed scheduling algorithm for optimizing age of information in wireless networks," in *Proc. IEEE 39th Int. Perform. Comput. Commun. Conf. (IPCCC)*, Nov. 2020, pp. 1–8.

[20] Z. Jiang, Y. Liu, J. Hribar, L. A. DaSilva, S. Zhou, and Z. Niu, "SMART: Situationally-aware multi-agent reinforcement learning-based transmissions," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 4, pp. 1430–1443, Dec. 2021.

[21] R. Talak, S. Karaman, and E. Modiano, "Distributed scheduling algorithms for optimizing information freshness in wireless networks," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2018, pp. 1–5.

[22] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form Whittle's index-enabled random access for timely status update," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1538–1551, Mar. 2020.
[23] F. Peng, Z. Jiang, S. Zhang, and S. Xu, "Age of information optimized MAC in V2X sidelink via piggyback-based collaboration," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 607–622, Jan. 2021.
[24] S. Maghsudi and S. Stańczak, "Channel selection for network-assisted D2D communication via no-regret bandit learning with calibrated forecasting," *IEEE Trans. Wireless Commun.*, vol. 14, no. 3, pp. 1309–1322, Mar. 2015.
[25] M. A. Abd-Elmagid, H. S. Dhillon, and N. Pappas, "A reinforcement learning framework for optimizing age of information in RF-powered communication systems," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4747–4760, Aug. 2020.
[26] E. T. Ceran, D. Gunduz, and A. Gyorgy, "Average age of information with hybrid ARQ under a resource constraint," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1900–1913, Mar. 2019.
[27] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the Internet of Things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, Nov. 2019.
[28] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Nov. 2015, pp. 1–7.
[29] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, Aug. 2017, pp. 2681–2690.
[30] R. Lowe, Y. Wu, A. Tamar, and J. Harb, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 6382–6393.
[31] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, Feb. 2018, pp. 2974–2982.
[32] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 2961–2970.
[33] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3001–3012, Jul. 2020.
[34] M. Li, C. Chen, H. Wu, X. Guan, and X. Shen, "Age-of-information aware scheduling for edge-assisted industrial wireless networks," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5562–5571, Aug. 2021.
[35] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.
[36] H. Lee, S. H. Lee, and T. Q. S. Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.
[37] M. Eisen, C. Zhang, L. F. O. Chamon, D. D. Lee, and A. Ribeiro, "Learning optimal resource allocations in wireless systems," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2775–2790, Apr. 2019.
[38] D. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1999.
[39] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
[40] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2018, pp. 1–14.
[41] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2018, pp. 1–13.
[42] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, and M. Hollick, "CBMoS: Combinatorial bandit learning for mode selection and resource allocation in D2D systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2225–2238, Oct. 2019.
[43] *WINNER II Path Loss Model*. Accessed: Jan. 6, 2012. [Online]. Available: http://www.raymaps.com/index.php/winner-ii-path-loss-model/
[44] *Guidelines for Evaluation of Radio Interface Technologies for IMT-2020*, document ITU-R M.2412-0, International Telecommunication Union, Geneva, Switzerland, 2017.
[45] V. R. Konda and J. N. Tsitsiklis, "On actor-critic algorithms," *SIAM J. Control Optim.*, vol. 42, no. 4, pp. 1143–1166, 2003.
[46] B. Demirel, A. Ramaswamy, D. E. Quevedo, and H. Karl, "DeepCAS: A deep reinforcement learning algorithm for control-aware scheduling," *IEEE Control Syst. Lett.*, vol. 2, no. 4, pp. 737–742, Oct. 2018.
[47] O. H. Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, vol. 30. New York, NY, USA: Springer, 2012.

Mingyan Li (Student Member, IEEE) received the B.E. degree in telecommunication engineering from Jilin University, Changchun, China, in 2015, and the Ph.D. degree from the Department of Electronic Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2021.

She was a Visiting Professor with the University of Waterloo, Canada, from 2019 to 2020. She joined the College of Computer Science, Chongqing University, in 2021, and currently works as a Post-Doctoral Research Associate. Her current research interests include industrial wireless networks and application in industrial automation, joint design of communication and control in industrial cyber-physical systems, software-defined networking and network slicing, ultra-reliable low-latency communication, and time-sensitive networks for industrial internet.

Cailian Chen (Member, IEEE) received the B.E. and M.E. degrees in automatic control from Yanshan University, China, in 2000 and 2002, respectively, and the Ph.D. degree in control and systems from the City University of Hong Kong, Hong Kong, SAR, in 2006.

She has been with the Department of Automation, Shanghai Jiao Tong University since 2008. She is now a Distinguished Professor. She has authored three research monographs and over 100 refereed international journal articles. She is the inventor of more than 20 patents. Her research interests include industrial wireless networks and computational intelligence and the Internet of Vehicles.

Prof. Chen has received the prestigious "IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award" in 2008 and the Best Paper Award of WCSP'17 and YAC'18. She has won the Second Prize of National Natural Science Award from the State Council of China in 2018, the First Prize of Natural Science Award from the Ministry of Education of China in 2006 and 2016, respectively, and the First Prize of Technological Invention of Shanghai Municipal, China, in 2017. She was honored "National Outstanding Young Researcher" by NSF of China in 2020 and "Changjiang Young Scholar" in 2015. She has been actively involved in various professional services. She has served as the TPC Chair for ISAS'19, the Symposium TPC Co-Chair for IEEE GLOBECOM 2016, the Track Co-Chair for VTC2016-Fall and VTC2020-Fall, and the Workshop Co-Chair for WiOpt'18. She has served as a Guest Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. She serves as an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *IET Cyber-Physical Systems: Theory and Applications*, and *Peer-to-Peer Networking and Applications* (Springer).

Huaqing Wu (Member, IEEE) received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Waterloo, ON, Canada, in 2021. She is currently a Post-Doctoral Research Fellow at McMaster University, ON, Canada. Her current research interests include vehicular networks with emphasis on edge caching, wireless resource management, space-air-ground integrated networks, and application of artificial intelligence (AI) for wireless networks. She has received the prestigious Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship Award in 2021.

Xinping Guan (Fellow, IEEE) received the B.Sc. degree in mathematics from Harbin Normal University, Harbin, China, in 1986, and the Ph.D. degree in control science and engineering from the Harbin Institute of Technology, Harbin, in 1999.

He is currently a Chair Professor with Shanghai Jiao Tong University, Shanghai, China, where he is the Dean of the School of Electronic, Information and Electrical Engineering and the Director of the Key Laboratory of Systems Control and Information Processing, Ministry of Education of China. Before that, he was the Executive Director of the Office of Research Management, Shanghai Jiao Tong University; and a Full Professor and the Dean of electrical engineering with Yanshan University, Qinhuangdao, China. He has authored and/or coauthored five research monographs, more than 200 papers in IEEE TRANSACTIONS and other peer-reviewed journals, and numerous conference papers. As a Principal Investigator, he has finished/been working on more than 20 national key projects. He is the Leader of the prestigious Innovative Research Team of the National Natural Science Foundation of China (NSFC). His current research interests cover industrial network systems, smart manufacturing, and underwater networks. He is an Executive Committee Member of the Chinese Automation Association Council and the Chinese Artificial Intelligence Association Council. He has received the Second Prize of the National Natural Science Award of China in both 2008 and 2018 and the First Prize of Natural Science Award from the Ministry of Education of China in both 2006 and 2016. He was a recipient of the "IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award" in 2008. He is the "National Outstanding Youth" honored by NSF of China, a "Changjiang Scholar" by the Ministry of Education of China, and a "State-Level Scholar" of "New Century Bai Qianwan Talent Program" of China.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on network resource management, wireless network security, the Internet of Things, 5G and beyond, and vehicular *ad-hoc* and sensor networks. He is a Registered Professional Engineer of Ontario, Canada; an Engineering Institute of Canada Fellow; a Canadian Academy of Engineering Fellow; a Royal Society of Canada Fellow; a Chinese Academy of Engineering Foreign Member; and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of IEEE Fellow Selection Committee of the ComSoc. He is the President of the IEEE Communications Society. He has received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R. A. Fessenden Award in 2019 from IEEE, Canada, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, the Joseph LoCicero Award in 2015 and the Education Award in 2017 from the IEEE Communications Society, and the Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He has also received the Excellent Graduate Supervision Award in 2006 from the University of Waterloo and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He has served as the Technical Program Committee Chair/Co-Chair for IEEE GLOBECOM'16, IEEE Infocom'14, IEEE VTC'10 Fall, and IEEE GLOBECOM'07; and the Chair for the IEEE Communications Society Technical Committee on Wireless Communications. He has served as the Editor-in-Chief for the IEEE INTERNET OF THINGS JOURNAL, *IEEE Network*, and *IET Communications*.