Intelligent Network Selection Algorithm for Multiservice Users in 5G Heterogeneous Network System: Nash *Q*-Learning Method

Mingfang Ma, Anqi Zhu, Songtao Guo, *Senior Member, IEEE*, and Yuanyuan Yang, *Fellow, IEEE*

***Abstract*—The 5G heterogeneous network architecture integrates different radio access technologies (RATs), which will support the large-scale communication connection of massive Internet-of-Things (IoT) devices. However, as the rapid growth of IoT connections, personalized requirements of services requested and heterogeneity deepening of the network system, how to design an intelligent network selection scheme for user devices (UDs) is becoming a crucial challenge in the 5G heterogeneous network system. Most of the existing network selection methods only optimize the selection strategies from the user side or network side, which results in heavy network congestion, poor user experience, and system performance degradation. Accordingly, we propose a multiagent *Q*-learning network selection (MAQNS) algorithm based on Nash *Q*-learning, which can learn a joint optimal selection strategy to improve system throughput and reduce user blocking on the premise of ensuring the requirements of IoT services. In particular, we apply the discrete-time Markov chains to model the network selection, and the analytic hierarchy process (AHP) and gray relation analysis (GRA) are jointly utilized to obtain user preferences for each network. Finally, performance evaluation demonstrates that comparing to the existing schemes, MAQNS proposed cannot only improve system throughput and reduce user blocking but also promote user experience on average energy efficiency and delay.**

***Index Terms*—Analytic hierarchy process (AHP), grey relation analysis (GRA), heterogeneous wireless networks, Nash equilibrium, Nash *Q*-learning, network selection.**

I. Introduction

THE PARADIGM of the Internet of Things (IoT) has rapidly matured nowadays [1]. As the driving force of IoT, emerging services, such as smart health service, virtual reality and augmented reality (VR&AR) service, and industrial machinery service have been widely applied and promoted hundreds of millions of devices connected to the networks [2], [3]. It is predicted that by 2025, the total number of global IoT connections will reach 24.6 billion [4]. However, existing networks such as the 4th generation wireless systems (4G) cannot withstand huge IoT connections. In addition, massive machine-type communication (mMTC) defined by 3GPP requires the network to efficiently support the large-scale communication connection of massive IoT devices while providing high service quality [5], which has caused great pressure on the application scenarios of future wireless networks (e.g., 5G).

In order to meet these challenges in 5G, the heterogeneous network, as a novel network structure, is particularly critical in 5G technology [1]. Heterogeneous networks are composed of different radio access technologies (RATs) containing 3GPP and IEEE families, and use the unique and complementary characteristics of each RAT to provide massive user devices (UDs) with multiple access networks and multifarious network services [6], [7]. Utilizing heterogeneous network technology, 5G with backward compatibility can integrate different RATs, such as LTE-advanced (LTE-A) and 802.11.AX standard-based Wi-Fi 6, providing greater coverage and capacity for massive IoT connections, as well as supporting personalized services requested by UDs.

As the deepening of network heterogeneity, the complex environment of multiple wireless networks in 5G has become a major challenge for network selection of IoT UDs. From the perspective of users, rational users will selfishly select the network that has the optimal performance or can maximize their own utility [8]. However, they ignored the network information, such as load status, resulting in most IoT UDs accessing the same network base stations (BSs) or access points (APs), while other BSs or APs available may be in the idle mode. This situation will further aggravate load imbalance, network congestion, and resource wasting. From the perspective of the network operator, the network operator will maximize an important system indicator, mostly throughput [9], because, to a great extent, the profit captured counts on the number of bytes transmitted. Nevertheless, the diverse requirements of different services are neglected, bringing about the increase of user blocking and the decline in user experience. These issues make the traditional network selection methods no longer effective and demand reliable ones to be beneficial for both users and network operators. Thus,

Manuscript received December 6, 2020; revised March 20, 2021; accepted April 9, 2021. Date of publication April 13, 2021; date of current version July 23, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61772432 and Grant 61772433; in part by the Natural Science Key Foundation of Chongqing under Grant cstc2020jcyj-zdxmX0026; in part by the Fundamental Research Funds for the Central Universities under Grant 2020CDCGJSJ071, Grant 2020CDCGJSJ038, and Grant 2019CDYGZD004; and in part by Zhejiang Lab under Grant 2021LC0AB01. *(Corresponding author: Songtao Guo.)*

Mingfang Ma and Songtao Guo are with the Key Laboratory of Dependable Service Computing in Cyber-Physical-Society (Ministry of Education) and the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: songtao_guo@163.com).

Anqi Zhu is with the Robotics Research Center, College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China.

Yuanyuan Yang is with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA.

Digital Object Identifier 10.1109/JIOT.2021.3073027

how to design a valid network selection algorithm, particularly in the 5G heterogeneous network, is a crucial and immediate topic.

As a result, a win–win network selection approach should jointly take the benefits of the user and network side into account. Based on the above consideration, we are mainly committed to design a network selection algorithm for IoT UDs in the 5G heterogeneous network architecture that considers the role of ensuring user experience and optimizing the long-term network performance. The algorithm is dedicated to improving system throughput and reducing user blocking on the premise of ensuring the personalized requirements of UDs who request different IoT services. More precisely, on the basis of Nash *Q*-learning, we use the agent to represent each network in our model and, thus, different types of networks form a multiagent structure. Then, we propose the intelligent network selection algorithm based on the Nash *Q*-learning method, named multiagent *Q*-learning network selection (MAQNS) algorithm. For the sake of maximizing the long-term performance of the multiagent structure, access of UDs is not only determined by the current network status but also depends on the expected future demand.

The contributions of this work can be summarized as follows.

1) In contrast to the existing network selection methods, in order to ensure user experience, reduce user blocking, and improve system throughput, we perform the optimization of network selection for IoT UDs from the perspective of both user and network.
2) Aiming at capturing the dynamic access of UDs in the network selection, we build a network selection model based on the discrete-time Markov chains. Then, we propose to utilize Nash *Q*-learning to design an MAQNS algorithm, which can learn the joint selection strategies of each network by trial and error. In this way, the Nash equilibrium can be achieved.
3) In order to better satisfy the differentiated service requirements and ensure user experience efficiently, we utilize the discrete-time Markov chains to formulate the network selection, and jointly adopt analytic hierarchy process (AHP) and gray relation analysis (GRA) to achieve user preferences for each network, which will be appropriately modeled into the network reward function.
4) Evaluation results indicate that as the number of UDs increases, compared with the existing schemes [10]–[13], our proposed MAQNS cannot only effectively improve system throughput and reduce user blocking but also significantly boost user experience on average energy efficiency and delay. In addition, the MAQNS algorithm can efficiently maintain load balancing for heterogeneous networks and achieve an appropriate access for IoT UDs.

The remainder of this article is organized as follows. In Section II, the related work on network selection is briefly reviewed. Section III introduces the system model. Section IV describes the requirements of each service, and Section V presents the Markov model of network selection. Next, Section VI discusses the MAQNS network selection algorithm. Then, in Section VII, we analyze the complexity and convergence of the proposed MAQNS. Finally, we verify the evaluation results in Section VIII and conclude this article in Section IX.

II. Related Work

Generally, the mechanisms that study access selection in a heterogeneous network scenario can be classified into the user-centric mechanism [11]–[19] and the network-centric mechanism [7], [9], [10], [20]–[22].

In [8] and [14], the user-centric network selection decisions mainly depend on the received signal strength (RSS), UDs will access the network with the optimal RSS. Similarly, other specific network attributes, such as the peak rate [15] and signal-to-noise ratio [16] are also used to determine network access. These above methods based on the single network attribute criterion are simple and easy to implement, which can enable UDs to access the network with good performance in one respect. However, it cannot accurately reflect the individual service requirements. Besides, the network selection method based on the single decision attribute has poor applicability in the complex heterogeneous network environment [23].

Therefore, for the sake of eliminating the limitation of single-attribute network selection methods and improving service quality for users, the multiattribute decision-making (MADM) theory has been widely studied. The typical MADM approaches include the technique for order preference by similarity to ideal solution (TOPSIS) [11], [17], simple additive weighting (SAW) [12], [18], and multiplicative exponent weighting (MEW) [13], [19]. In [11], an access selection solution based on the TOPSIS is proposed in the multiple wireless networks, which considers service requirements, and prioritizes the network order for access. El Helou *et al.* [12] proposed a hybrid approach according to the SAW for network selection, which takes the cost and network parameters into account, and individual UDs choose the optimal network for the sake of maximizing their own utility selfishly. Araniti *et al.* [13] analyzed the relationship between the energy consumption and transmission quality of streaming service, and constructed a mixed network selection algorithm. These MADM approaches comprehensively consider multiple indicators (e.g., cost, delay and packet loss) of candidate networks, and UDs will collect the information of networks, compute access selection decision metrics based on their own preferences. However, the performance of these indicators can compensate each other, which may enable UDs to access the network with the best comprehensive performance but overloaded [24].

Few network-centric access selection mechanisms are studied in [9], [10], and [7], [20]–[22]. In [20], the proposed algorithm can make UDs prefer to access WLAN when it is available; otherwise, UDs will access LTE. In [21], the voice users prefer to access WLAN and UMTS is the first choice for data users. However, these methods ignore the joint optimization between different heterogeneous networks, which may hinder effective utilization for the overall network

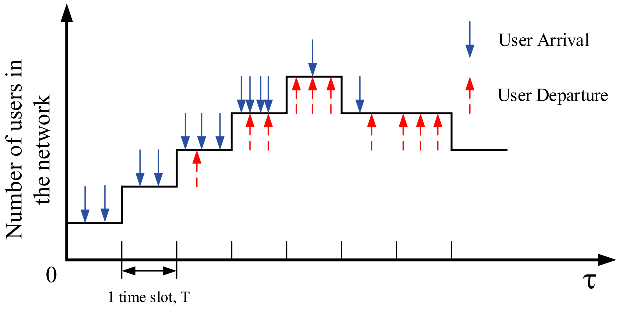

Fig. 1. System time model.

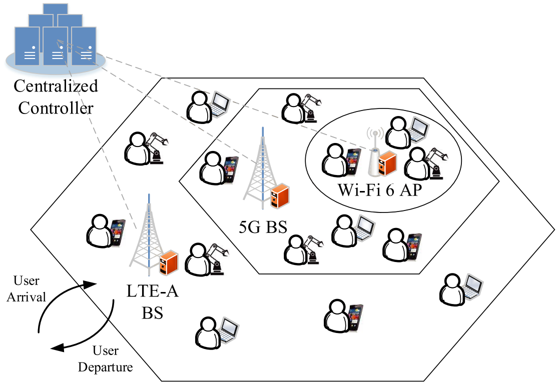

Fig. 2. 5G heterogeneous wireless access networks.

resources. Roy *et al.* [7] built a Markov RAT selection model tackled from an operator's perspective and designed an optimal scheme based on a threshold structure, which intended to promote the total throughput under the constraint of blocking probability of voice users in an LTE-WiFi heterogeneous networks. Nevertheless, it is hard to determine the optimal threshold accurately to improve throughput.

Recently, machine learning has also been used as a tool to study the network-centric selection mechanism [9], [10], [22], these schemes can enable UDs to grasp the global information of the heterogeneous network system, such as the history which BSs or APs have been accessed by other UDs. Du *et al.* [9] introduced a Q-learning algorithm based on knowledge transfer, so as to optimize the access selection result and improve the system throughput. Wang *et al.* [10] designed a network selection algorithm on the basis of random forest and Q-learning to optimize experience of users with single service request and reduce the average delay. Nonetheless, most current access selection approaches based on machine learning only focus on optimizing a certain performance of the network system in the scenario of a single service type, ignoring that the requirements of UDs requesting different services are not the same. Thus, these approaches are not feasible in 5G where more and more attention is paid to differentiated service requests. It is inevitable that in the future wireless network system, which provides more IoT services, both the user side and network side need to be well considered to achieve a win–win situation.

III. System Model

For the purpose of representing the dynamic changes in the number of UDs under the heterogeneous networks, it is necessary to formalize the time into a discrete-time model. Therefore, as illustrated in Fig. 1, the continuous time can be divided into equispaced time intervals, named time slot τ, where T denotes the duration time of one time slot. Besides, the beginning of each slot is considered as the decision-making epoch. In this way, we assume that the changes in variables, such as user arrivals, accesses, and departures occur at the decision epoch of each time slot, the variables will maintain unchanged during time slot τ and the variables will change again with a probability at the next decision epoch. It is worth noting that there are only user arrivals in the network at the first time slot, and no user departures from the network.

The 5G system will integrate different RATs and allow the coexistence of different RATs, aiming at improving the capabilities of the 5G system in multiple aspects [25]. It is known that the 5G system cannot only achieve high data rate and low latency but also enhance the performance of large-scale device connections and satisfy the requirements of mass services [26]. The locality of edge computing makes the BSs equipped with a certain amount of computing resources, which provide computing and storage capabilities at the edge of network to collect users' information (e.g., service requirements) in real time and reduce delay for UDs [27]. Furthermore, the 5G system enables UDs to obtain services from different available access networks at anytime or anywhere.

In this article, as shown in Fig. 2, a 5G heterogeneous wireless access network model is considered, which is composed of Wi-Fi 6 AP, LTE-A base station BS, and 5G BS, where the centralized controller is used to gather the global network information. At the same time, UDs located in the 5G heterogeneous network system can request three types of IoT services, including smart health service, VR&AR service, and industrial machinery service. As a rising medical application of smart health service, providing patients with remote medical treatment, especially remote surgeries, requires precise implementation of remote operations in a stable environment. Therefore, smart health service has a high demand for delay and jitter generally [28]. Typical application scenarios of VR&AR service include holographic navigation and smart VR game, which requires high bandwidth to provide users with a good experience [29]. The application scenarios of industrial machinery service include real-time monitoring of production equipment and remote control of construction machinery, which have relatively high demand on delay [30].

Without loss of generality, it is assumed that the users' arrival and departure in the network meet two independent stochastic poisson processes. More precisely, we can use $\lambda_k^m[\tau]$ and $\mu_k^m[\tau]$, respectively, to express the user arrival rate and departure rate in time slot τ, where k and m represent the type of service and type of network, respectively. Therefore, in time slot τ, the probability of x users with service request k arriving at network m is shown in the following formula:

$$P\{x \text{ arrivals}\} = \frac{\left(\lambda_k^m[\tau]T\right)^x \cdot e^{-\lambda_k^m[\tau]T}}{x!}. \quad (1)$$

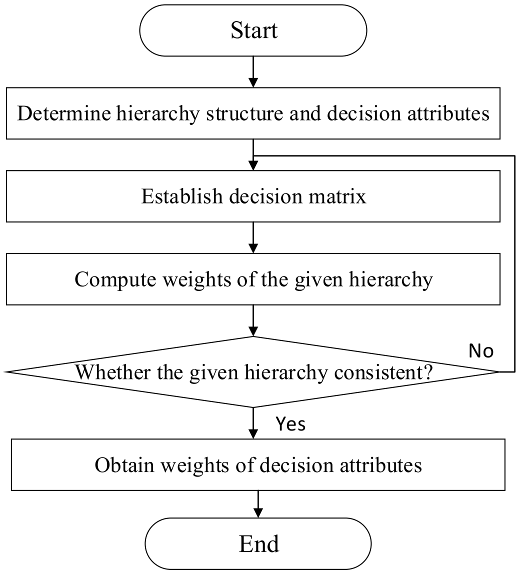
Start
Determine hierarchy structure and decision attributes
Establish decision matrix
Compute weights of the given hierarchy
Whether the given hierarchy consistent?
No
Yes
Obtain weights of decision attributes
End

Fig. 3. Flowchart of AHP in MAQNS.

Similarly, in time slot τ, the probability of y users with service request k departing from the network m can be represented by

$$P\{y \text{ departures}\} = \frac{\left(\mu_k^m[\tau]T\right)^y \cdot e^{-\mu_k^m[\tau]T}}{y!}. \tag{2}$$

IV. Measurement of Service Preference

In view of the 5G services having differentiated requirements on the network attributes, in order to provide good service experience for users, it is reasonable to analyze the service preference for different network attributes. In this section, we intend to use AHP to obtain service preferences on the attributes of network bandwidth, energy efficiency, delay, jitter, packet loss rate (PLR), and price. Furthermore, we apply AHP and GRA comprehensively to get the weighted gray correlation coefficient (GCC), which can be modeled as the preference of users for each candidate network.

A. Analytic Hierarchy Process

In the network selection, AHP is usually used to measure the preference weights of various services on network attributes by expert experience, and it is regarded that the obtained weights belong to subjective weight [31]. Since there is only one criterion layer to be considered in our AHP algorithm, the AHP proposed can be executed according to the flowchart presented in Fig. 3.

The specific details to measure the preferences weight by AHP are illustrated as follows.

1) Structuring the hierarchical model as shown in Fig. 4. The hierarchical model of AHP consists of three layer: a) the target layer denotes the optimum networks that UDs requesting specific service expect to access; b) the criteria layer indicates the demands of services for network attributes; and c) the solution layer expresses the candidate networks in the system model.

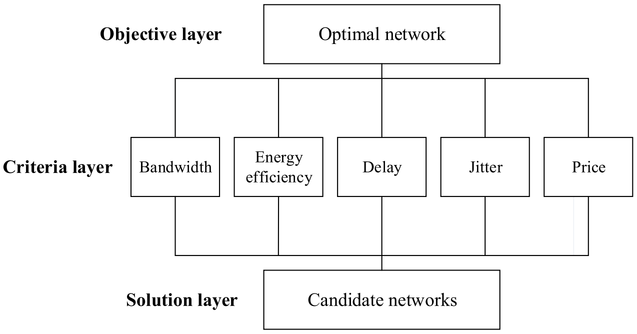

Fig. 4. Hierarchical model in AHP.

2) Normalizing the network attributes needed by each service to compare the relative importance between different attributes needed by a certain service in the next step. The benefit attributes can be normalized by

$$f'_{kl} = \frac{f_{kl}}{f_l^{\max}}, f_l^{\max} = \max\{f_{1l}, \ldots, f_{kl}, \ldots, f_{Kl}\}. \tag{3}$$

The cost attributes can be normalized by

$$f'_{kl} = \frac{f_l^{\min}}{f_{kl}}, f_h^{\min} = \min\{f_{1l}, \ldots, f_{kl}, \ldots, f_{Kl}\} \tag{4}$$

where f_{kl} expresses the demand of service k for network attribute l, and there are totaly L network attributes considered in the criteria layer; therefore, $l = 1, 2, \ldots, L$.

3) Constructing judgment matrix $G = (g_{uv})_{H*H}$ by

$$G = (g_{uv})_{H*H} = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1v} \\ g_{21} & g_{22} & \cdots & g_{2v} \\ \vdots & \vdots & \vdots & \vdots \\ g_{u1} & g_{u2} & \cdots & g_{uv} \end{bmatrix} \tag{5}$$

where matrix G is the judgment of the relative importance to the network attributes needed by a specific service. In detail, we can use 1–9 levels to compare the normalized value of attribute u and v to get g_{uv} [32]. The smaller the difference value between attribute u and v, the smaller the weight assigned to g_{uv}. Similarly, g_{vu} is the measurement of the relative importance of the attribute v to u and $g_{vu} = (1/g_{uv})$.

4) Calculating the preference weight by

$$w_v = \frac{\sum_{u=1}^{L} g_{uv}}{L}, \quad \sum_{v=1}^{L} w_v = 1. \tag{6}$$

Therefore, we can get the preference weight of service k on different network attributes proposed in the criterion layer, i.e., $W^k = \{w_b^k, w_e^k, w_d^k, w_j^k, w_p^k, w_c^k\}$.

5) Checking the consistency of judgment matrix G. Because the consistency means that the judgment on the importance of attributes is reasonable, we use consistency index to check the judgment matrix G. First, we calculate the consistency index CI by

$$\text{CI} = \frac{\lambda_{\max} - L}{L - 1} \tag{7}$$

Algorithm 1 AHP in MAQNS

```
Input: Bandwidth $R_b$, energy efficiency $R_e$, delay $R_d$, jitter $R_j$, PLR $R_p$, price $R_c$ required by service $k$;
Output: Preference weight of service $k$ on different network attributes: $W^k = \{w_b^k, w_e^k, w_d^k, w_j^k, w_p^k, w_c^k\}$;
1: Establish the hierarchical model $O - C - P$ (Fig. 4) based on the requirements of service $k$;
2: Normalizing network attributes needed by service $k$ by formula (3) and (4);
3: Construct a judgment matrix $G$ in the form of formula (5);
4: Obtain the weight of bandwidth requirement $w_b^k$, energy efficiency requirement $w_e^k$, delay requirement $w_d^k$, jitter requirement $w_j^k$, PLR requirement $w_p^k$, price requirement $w_c^k$ according to formula (6);
5: if the matrix $G$ is not consistent then
6:     Return to step 3;
7: else
8:     Output $W^k = \{w_b^k, w_e^k, w_d^k, w_j^k, w_p^k, w_c^k\}$;
9: end if
```

in which $\lambda_{\max}$ indicates the largest eigenvalue of matrix G, and we can get it by the following formula:

$$\lambda_{\max} = \frac{\sum_{v=1}^{L} \sum_{u=1}^{L} w_v g_{uv}}{w_v}. \tag{8}$$

Then, the consistency ratio is denoted by

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \tag{9}$$

where RI represents the average random consistency index, which is the arithmetic mean value of the eigenvalue of the random judgement matrix multiple times [33], and the value of RI corresponding to the matrix of each order can be estimated according to Satty's research [34]. For consistency ratio CR, if $\mathrm{CR} \leq 0.1$, the matrix G is consistent adequately; thus, the weight vector $W^k = \{w_b^k, w_e^k, w_d^k, w_j^k, w_p^k, w_c^k\}$ is considered to be reasonable; otherwise, it is inconsistent and the judgment matrix G needs to be modified to meet the consistent condition.

The processes to generate weights by AHP can be summarized in Algorithm 1.

B. Gray Relational Analysis Algorithm

By calculating the GCC between the discrete sequences and the ideal sequence defined, respectively, GRA can measure the correlation grade of these sequences, and the sequence with the largest GCC is the best alternative [35]–[37]. In the network selection, GRA is usually used to measure the correlation between the attributes of the candidate networks and that of the ideal network we set, and combined with AHP to compute the weighted GCC to indicate the correlation between candidate networks and the ideal network. The greater the value of weighted GCC, the closer the correlation between the corresponding candidate network and the ideal network.

In MAQNS, The specific steps of the GRA proposed are given in the following.

Algorithm 2 GRA in MAQNS

```
Input: Parameters of network attributes: network bandwidth $B_m$, energy efficiency $E_m$, delay $D_m$, jitter $J_m$, PLR $Pl_m$, price $P_m$;
    Weight of service k on network attributes: bandwidth $w_b^k$, energy efficiency $w_e^k$, delay $w_d^k$, jitter $w_j^k$, PLR $w_p^k$, price $w_c^k$ obtained by Algorithm 1 in MAQNS;
Output: Weighted grey correlation coefficient $\Xi_k^m$;
1: Get original decision matrix $E$;
2: Normalize original decision matrix $E$ into matrix $D$;
3: Determine the ideal sequence $d^*$ using formula (10);
4: Calculate the grey correlation coefficient $\xi_{mh}$ according to formula (11);
5: Compute the weighted grey correlation coefficient $\Xi_k^m$ by formula (12);
6: Output the weighted grey correlation coefficient $\Xi_k^m$
```

1) Constructing network attribute matrix $E = (e_{ml})_{M*L}$, where M expresses the number of candidate networks in our model, while L indicates the total number of decision attribute types of each network, then e_{ml} characterizes the attribute l of network m.
2) Normalizing the network attribute according to (3) and (4) to compare the importance among different types of attributes.
3) Determining the ideal network d^*, which is the optimal value of each attributes selected from candidate networks and can be expressed as

$$d^* = \left\{ d_l^* | d_l^* = \left(\max_m \{d_{ml}\} | d_{ml} \in d_b \right) \left(\min_m \{d_{ml}\} | d_{ml} \in d_c \right) \right\}, \quad m \in M, l \in L \tag{10}$$

where d_b and d_c, respectively, indicate the set of benefit and cost network attributes.

4) Acquiring GCC ξ_{ml}

$$\xi_{ml} = \frac{\min_m \min_l \left| d_{ml} - d_l^* \right|}{\left| d_{ml} - d_l^* \right| + \rho \max_m \max_l \left| d_{ml} - d_l^* \right|} + \frac{\phi \max_m \max_l \left| d_{ml} - d_l^* \right|}{\left| d_{ml} - d_l^* \right| + \phi \max_m \max_l \left| d_{ml} - d_l^* \right|} \tag{11}$$

where $\phi \in [0, 1]$ expresses the resolution coefficient. ξ_{ml} indicates the gray correlation degree between the attribute l of candidate network m and that of the ideal network, the greater the value of ξ_{ml}, the closer the attribute l of network m is to that of the ideal network.

5) Measuring the weighted GCC Ξ_k^m

$$\Xi_k^m = \Sigma_{l=1}^{L} w_l^k \xi_{ml} \tag{12}$$

where w_l^k indicates the preference weight of service k to network attribute l obtained by the AHP algorithm, and UDs with service k request will tend to prefers the network with a lager Ξ_k^m.

In summary, the GRA algorithm can be described in Algorithm 2.

V. Discrete Markov Model

Aiming at capturing the dynamic access of UDs in the network selection and optimize the long-term performance for the 5G heterogeneous wireless networks, the discrete MDP model is employed to formulate the network selection problem. The MDP model contains four ingredients, including network states, actions taken by networks, state transition probability and network reward.

A. Network States

Due to the arrival and departure of users in the network will active the dynamic changes of each network in our model correspondingly, we consider that the network state space can be denoted as $S = \{s | s = n_k^m, m \in M, k \in K\}$, in which n_k^m means the number of UDs with service k request provided by network m. The network state will keep unchanged until the users' arrival or departure in the network happens.

B. Network Actions

Networks will take actions at each state, thus the network action a in our model can be regarded as the selection decision of each network, i.e., the network determines to provide a certain type of service requested by UDs. Therefore, the set of action $A = \{a | a = 1, \ldots, k, \ldots, K\}$ represents the available service types that the network selects to provide at a specific state.

C. State Transition Probability

When the network determines to perform action a^τ in state s^τ, the state of the network will transmit with a probability at the decision epoch of the next time slot $s^{\tau+1}$, which can be viewed as the state transition probability $p(s^{\tau+1}|s^\tau, a^\tau)$, where $s^\tau \in S$, $s^{\tau+1} \in S$, and $a \in A$. The state transition probability here is determined by users' arrival and departure, as well as the actions taken by networks. However, when the state transition probability is difficult to obtain, we will take advantage of the Nash Q-learning algorithm to learn what actions can be taken through trial and error.

D. Network Reward

The network reward $r(s, a)$ means that the reward gained by the network after the network chooses action $a \in A$ in state s. In our model, in order to optimize the system performance and user experience, the reward function $r(s, a)$ is considered as the sum of the network utility and blocking cost

$$r(s, a) = F(s, a) + O(s, a) \tag{13}$$

in which $F(s, a)$ indicates the network utility. When network m choose UDs who request service k, $F(s, a)$ is expressed as follows:

$$F(s, a) = \sum_{k=1}^{K} N_k^m \Xi_k^m \pi_k^m, \;\; m \in M, \;\; k \in K \tag{14}$$

where N_k^m is the number of UDs with service k request accessing network m, Ξ_k^m defines the weighted GCC computed through the GRA algorithm, and π_k^m means the throughput acquired by UD who request service k accessing network m. The calculation method of throughput can be given by the following formula:

$$\pi_k^m = N_{ru}^{m,k} N_{\text{sym}}^m N_{\text{sub}}^m \log_2\big[s_{ize}\big(\text{mod}^m\big)\big] R\big(\text{cod}^m\big)(1 - \text{BER}) \tag{15}$$

where $N_{ru}^{m,k}$ denotes the number of resource units occupied by the UD with service k request in the network m, N_{sym}^m and N_{sub}^m, respectively, indicate the number of symbols and subcarriers in a resource unit, $s_{ize}(\text{mod}^m)$ and $R(\text{cod}^m)$, respectively, denote the sizes of constellation and coding rate of the network m, besides, BER is the error rate of network m with corresponding modulation.

Blocking cost $O(s, a)$ that appeared in $r(s, a)$ is defined as follows:

$$O(s, a) = -c_{\text{cost}}^m \sum_{k=1}^{K} \lambda_k^m \left(1 - \sum_{m=1}^{M} P_k^m\right) \tag{16}$$

in which c_{cost}^m means the blocking coefficient of a certain network, while P_k^m indicates the probability that the network m choose UDs with service k request to access.

VI. Nash Q-Learning-Based Network Selection

It is necessary for network operators to select an optimal access strategy, so as to keep a higher system performance. As a result, we formulate a multiagent structure network selection model on the basis of noncooperative stochastic game. Furthermore, an intelligent network selection algorithm, named MAQNS, is proposed by adapting Nash Q-learning [38] into the multiagent structure.

A. Stochastic Game Model

Depending on the concept of noncooperative stochastic game, each network in the system is considered as an agent; thus, the 5G heterogeneous network system will form a multiagent structure. The definition on the stochastic game model is illustrated in Definition 1.

Definition 1: For a multiagent structure, we can use a tuple to define the stochastic game

$$\text{MASG}(S, A_1, \ldots, A_M, r_1, \ldots, r_M, P)$$

in which S indicates the state space of multiagent, M refers to the number of agents, which also characterize the number of networks in our model; thus, $A_1 \ldots, A_M$ actually expresses each agent's action space. In addition, the action space of agent m can be denoted as A_m, then r_m is the reward of agent m for taking an action. Finally, P means the state transition probability of the agent.

It should be noted that the definition of state, action, reward, and transition probability here is completely consistent with the definition in Section V. In state s^τ, the agents select actions $a_1, \ldots, a_M$ independently and get rewards $r_m(s, a_1, \ldots, a_M)$, while its state s^τ will change with a probability at the start of the next state with the constraint as follows:

$$\sum_{s^{\tau+1} \in S} P\Big(s^{\tau+1}|s^\tau, a_1, \ldots, a_M\Big) = 1. \tag{17}$$

For the discounted stochastic game, each agent tries to maximize total discounted rewards V_m under a given initial state s^0

$$V_m(s, \delta_1, \dots, \delta_M) = \sum_{\tau=0}^{\infty} \alpha^\tau E\Big(r_1^\tau | \delta_1, \dots, \delta_M, s^0 = s\Big) \quad (18)$$

in which δ_m is assumed as the strategy taken by agent m, and $\alpha \in [0, 1)$ expresses the discount factor.

As for the Nash equilibrium in the stochastic game, which usually can be denoted as a tuple of $(\delta_1^*, \dots, \delta_M^*)$ with M strategies, and it will satisfy the following condition, i.e., for $s \in S$ and $\delta_m \in \Delta_m$

$$V_m\big(s, \delta_1^*, \dots, \delta_M^*\big) \geq V_m\big(s, \delta_1^*, \dots, \delta_{m-1}^*, \delta_m, \delta_{m+1}^*, \dots, \delta_M^*\big) \quad (19)$$

where Δ_m is used to express the set of the available strategies for agent m. Therefore, in a Nash equilibrium, the strategy of each agent should be the best response to that of other agents.

B. Nash Q-Learning in Multiagent Structure

In Nash Q-learning, the Nash Q-function claims that the joint actions of all the agents in the multiagent structure should be determined, and the Nash Q-function of agent m is considered as $(s, a_1, \dots, a_M)$. As all agents in the multiagent structure follow the joint Nash equilibrium strategies, the Nash Q-value of agent m is the sum of the current reward and its future rewards. Therefore, the Nash Q-function of agent m is denoted by

$$\begin{aligned} &Q_m^*\big(s^\tau, a_1, \dots, a_M\big) \\ &\quad = r_m\big(s^\tau, a_1, \dots, a_M\big) \\ &\qquad + \alpha \sum_{s^{\tau+1} \in s} P\Big(s^{\tau+1} | s^\tau, a_1, \dots, a_M\Big) \cdot V_m\Big(s^{\tau+1}, \delta_1^*, \dots, \delta_M^*\Big) \end{aligned} \quad (20)$$

in which $(\delta_1^*, \dots, \delta_M^*)$ denotes the joint Nash equilibrium strategy, and $r_m(s^\tau, a_1, \dots, a_M)$ expresses the reward earned by the agent m for taking a_m in state s^τ under the joint actions taken by other agents. $V_m(s^{\tau+1}, \delta_1^*, \dots, \delta_M^*)$ means the total discounted rewards earned by agent m under the condition of all agents adopt the Nash equilibrium strategies at the start of the next state.

Based on the above mentioned, we can achieve the purpose of maximizing long-term system performance. Formula (21) represents the revenue of agent m in state $s^{\tau+1}$ for adopting the Nash equilibrium

$$\text{Nash } Q_m^\tau\Big(s^{\tau+1}\Big)\delta_1\Big(s^{\tau+1}\Big) \dots \delta_M\Big(s^{\tau+1}\Big)Q_m^\tau\Big(s^{\tau+1}\Big). \quad (21)$$

For the purpose of learning Nash equilibrium revenues, each agent should observe not only its own reward but also the rewards of other agents. Before the indexed learning agent m starts to learn its Q-values, the Q-values should be initialized at first. After that, agent m observes the current state and takes action with the ε-greedy strategy, in each time slot. Furthermore, agent m would observe its reward, behaviors, and rewards of other agents, as well as its new state. Then, the learning agent m computes Nash equilibrium and updates Q-values by

$$\begin{aligned} Q_m^{\tau+1}\big(s^\tau, a_1, \dots, a_M\big) = \big(1 - \beta^\tau\big)Q_m^\tau\big(s^\tau, a_1, \dots, a_M\big) \\ + \beta^\tau\Big[r_m^\tau + \alpha \text{Nash } Q_m^\tau\Big(s^{\tau+1}\Big)\Big] \end{aligned} \quad (22)$$

where $\beta \in [0, 1)$ indicates the learning rate.

Our MAQNS algorithm applies the ε-greedy strategy to enable the agent to explore and exploit available actions during the process of learning. More precisely, the agent explores in probability $\varepsilon(s)$ and exploits Nash Q-values in probability $1 - \varepsilon(s)$, and $\varepsilon(s)$ is defined by

$$\varepsilon(s) = \frac{1}{\ln\big(\sum h(s, a) + 3\big)}, \varepsilon(s) \in [0, 1] \quad (23)$$

where $h(s, a)$ is the number of state–action pairs that the learning agent traverses up to the current learning. In addition, it is assumed that the learning rate β will vary with $h(s, a)$ by

$$\beta = \frac{i}{h(s, a)}, \;\; i \in (0, 1). \quad (24)$$

C. MAQNS Network Selection Algorithm

In the following, we will mainly give the prescription on the proposed MAQNS algorithm.

At the start of the algorithm, MAQNS initializes the input items, including the user arrival rate, departure rate, available network resource units, discount factor, exploration probability, learning rate, as well as the Q-values and the state for each agent. Then, based on the user arrival rate, the learning agent m would like to observe its state s^τ at the time slot $\tau = 1$ (line 4), and the algorithm proceeds to perform the while loop (line 5).

Agent m takes action with the $\varepsilon(s)$-greedy strategy according to the other agents' actions observed, as well as computes its reward and other rewards by (13) (line 12). Next, the MAQNS will update the state–action pairs $h(s, a)$, $\varepsilon(s)$ and the available resource units of candidate networks (lines 13–15). Afterward, the state of the agent m changes into a new state (line 18), and the Q-values are updated to compute Nash equilibrium by (21). This procedure will be repeated unless the learning time slot Γ reaches, then the algorithm comes to end (line 20).

Algorithm 3 makes a detailed summary of the proposed MAQNS algorithm.

VII. Complexity and Convergence Analysis

In this section, we intend to give analysis on the complexity and display a theoretical proof of the convergence for our MAQNS algorithm.

A. Complexity of the MAQNS

In our multiagent structure model, the state space and action space of each agent are, respectively, assumed as $|S|$ and $|A|$. If the action set $|A_1| = \cdots = |A_M| = |\Lambda|$, which means that each agent owns a space capacity with $|S||A|^M$ to maintain the Q-tables. Therefore, space complexity of MAQNS is $M|S||A|^M$, which refers to the space complexity of MAQNS is linearly

Algorithm 3 Nash Q-Learning-Based Network Selection

```
Input: User arrival rate $\lambda_k^m$, user departure rate $\mu_k^m$, network
    available resource units $C_m$, discounted factor $\alpha$, exploration
    probability $\varepsilon(s)$, the number of state-action pair
    $n(s,a) \longleftarrow 0$, learning rate $\beta$;
Output: Q-values of $M$ agents;
When $\tau = 0$, initialize $s^0$ by $\lambda_k^m$;
Assign a learning agent $m$ and get its Q-values
    $Q_m^0(s, a_1, \dots, a_M) = 0$ as well as the other agents'
    Q-values $Q_{m'}^0(s, a_1, \dots, a_M) = 0$, where $s \in S$;
for $\tau = 1$ to $\Gamma$ do
    Observe $s^\tau$ based on $\lambda_k^m[\tau]$;
    while the state of agent $m$ is $s^\tau$ do
        for $m = 1$ to $M$ do
            if Exploration then
                Randomly take $a_m^\tau$;
            else
                Adopt $a_m^\tau$ corresponding to Nash $Q$;
            end if
            Observe $a_1^\tau, \dots, a_M^\tau$ to calculate $r_1^\tau, \dots, r_M^\tau$
                through formula (13);
            $h(s,a) = h(s,a) + 1$;
            Update $\varepsilon(s)$ using formula (23);
            Update $C_m$;
        end for
    end while
    Get the next state $s^{\tau+1}$;
    Agent $m$ updates its own Q-values and that of others
        $Q^{\tau+1}(s^\tau, a_1, \dots, a_M)$ by formula (22);
end for
```

with the state space, polynomial with the action space, and exponential with the number of agents.

Actually, with regard to the time complexity of Q-learning, it is hard to give a theoretical analysis due to its iteration nature. Therefore, it is necessary to use the qualitative analysis to clarify the complexity. In Algorithm 3, in each learning round, MAQNS will perform if-else judgment once in M cycles. Besides, the heterogeneity in the 5G heterogeneous system is not large in reality, that is, the types of networks M are limited, thus there are more than M Q-values need to be stored in each server by using MAQNS algorithm. Beyond that, the ε-greedy strategy adopted effectively enhances overall efficiency of the algorithm.

B. Convergence Analysis

The proof of the convergence needs two essential assumptions over the learning rate.

Assumption 1: In the MAQNS algorithm, each agent requires to visit its state s and action $a_m \in A_m$ repeatedly.

Assumption 2: The learning rate β^τ should meet the following two conditions.

1)

$$0 \le \beta^\tau(s, a_1, \dots, a_M) < 1$$

$$\sum_{\tau=0}^{\infty} \beta^\tau(s, a_1, \dots, a_M) = \infty$$
$$\sum_{\tau=0}^{\infty} \left[\beta^\tau(s, a_1, \dots, a_M)\right]^2 < \infty. \tag{25}$$

2) If $(s^\tau, a_1^\tau, \dots, a_M^\tau) \neq (s, a_1, \dots, a_M)$, then $\beta^\tau(s, a_1, \dots, a_M) = 0$.

The condition 2) expresses that each agent only needs to update the Q-values for the current state s^τ and actions $a_1^\tau \dots, a_M^\tau$.

Next, the following two lemmas provide a significant support for our proof of the convergence.

Lemma 1: If the learning rate β^τ meets Assumption 2, and mapping $P^\tau : \mathbb{Q} \to \mathbb{Q}$ fits the condition: assuming that there are a number $\eta \in (0, 1)$ and a sequence $\vartheta^\tau \ge 0$ converges to 0 at the probability of 1; thus, the following formula holds for $Q \to \mathbb{Q}$ and $Q^* = E[P^\tau Q^*]$:

$$\left\|P^\tau Q - P^\tau Q^*\right\| \le \eta \left\|Q - Q^*\right\| + \vartheta^\tau \tag{26}$$

in which $\mathbb{Q}$ denotes the space of Q tables. Then, the iteration rule $Q^{\tau+1} = (1 - \beta^\tau)Q^\tau + \beta^\tau(P^\tau Q^\tau)$ converges to Nash equilibrium Q^* at the probability of 1.

In Lemma 1, P^τ is a pseudo-contraction operator. The definition of P^τ in the M-player stochastic game is omitted. Then, Lemma 2 demonstrates that $Q^* = E[P^\tau Q^*]$.

Lemma 2:

1) In a discounted stochastic game, there exists a Nash equilibrium revenue $(V_1(\delta_1^*, \dots, \delta_M^*), \dots, V_M\ (\delta_1^*, \dots, \delta_M^*))$, and $(\delta_1^*,\ \dots,\ \delta_M^*)$ denotes a Nash equilibrium strategies in the discounted stochastic game. In addition, $V_m(\delta_1^*, \dots, \delta_M^*) = (V_m(s_1, \delta_1^*, \dots, \delta_M^*), \dots, V_m(s_M, \delta_1^*, \dots, \delta_M^*))$.
2) In the stage game $(Q_1^*(s), \dots, Q_M^*(s))$, there exist Nash equilibrium revenues $(V_1(s, \delta_1^*, \dots, \delta_M^*), \dots, V_M(s, \delta_1^*, \dots, \delta_M^*))$. $(\delta_1^*(s), \dots, \delta_M^*(s))$, which is a Nash equilibrium point of the stage game

$$\begin{aligned} Q_m^*\left(s^\tau, a_1, \dots, a_M\right) &= r_m\left(s^\tau, a_1, \dots, a_M\right) \\ &\quad + \alpha \sum_{s^{\tau+1} \in S} P\left(s^{\tau+1} | s^\tau, a_1, \dots, a_M\right) \\ &\quad \times V_m\left(s^{\tau+1}, \delta_1^*, \dots, \delta_M^*\right). \end{aligned} \tag{27}$$

According to 1) and 2) stated in Lemma 2, $V_m(s) = \delta_1(s) \dots \delta_M(s) Q_m^*(s)$. Furthermore, by Lemma 2, we can derive the following lemma (Lemma 3).

Lemma 3: Based on the stochastic game, in our multiagent model, $Q^* = E[P^\tau Q^*]$, where $Q^* = (Q_1^*, \dots, Q_M^*)$.

Proof: According to (27) in Lemma 2, it can be derived that

$$\begin{aligned} Q_m^*\left(s^\tau, a_1, \dots, a_M\right) &= r_m\left(s^\tau, a_1, \dots, a_M\right) \\ &\quad + \alpha \sum_{s^{\tau+1} \in S} P\left(s^{\tau+1} | s^\tau, a_1, \dots, a_M\right) \\ &\quad \times \delta_1^*\left(s^{\tau+1}\right), \dots, \delta_M^*\left(s^{\tau+1}\right) Q_m^*\left(s^{\tau+1}\right) \\ &= \sum_{s^{\tau+1} \in S} P\left(s^{\tau+1} | s^\tau, a_1, \dots, a_M\right) \end{aligned}$$

$$
\begin{aligned}
&\times \Big(r_m\big(s^\tau, a_1, \ldots, a_M\big) + \alpha\delta_1^*\big(s^{\tau+1}\big) \\
&\qquad \ldots, \delta_M^*\big(s^{\tau+1}\big)Q_m^*\big(s^{\tau+1}\big)\Big) \\
&= E\big[P_m^\tau Q_m^*(s, a_1, \ldots, a_M)\big]. \qquad (28)
\end{aligned}
$$

The above formula holds for all $s, a_1, \ldots, a_M$; thus, we can obtain that $Q^* = E(P^\tau Q^*)$. The following verifies the correctness of (26) in Lemma 1. According to the definition on the distance between two points, it can be deduced that

$$
\begin{aligned}
\big\|Q - Q^*\big\| &= \max_{m\in M}\max_{s\in S}\big\|Q_m(s) - Q_m^*(s)\big\| \\
&= \max_{m\in M}\max_{s\in S}\max_{a^1,\ldots,a^M} \\
&\qquad \big|Q_m(s, a_1, \ldots, a_M) - Q_m^*(s, a_1, \ldots, a_M)\big|. \quad (29)
\end{aligned}
$$

Based on (29), we can derive that

$$
\begin{aligned}
&\big\|P^\tau Q_m - P^\tau Q_m^*\big\| \\
&\quad = \max_{m\in M}\max_{s\in S}\max_{a^1,\ldots,a^M} \\
&\qquad \big|P^\tau Q_m(s, a_1, \ldots, a_M) - P^\tau Q_m^*(s, a_1, \ldots, a_M)\big|. \quad (30)
\end{aligned}
$$

Then, by (22) and the iteration rule $Q^{\tau+1} = (1 - \beta^\tau)Q^\tau + \beta^\tau(P^\tau Q^\tau)$, we can get $P^\tau Q_m = \beta^\tau[r_m^\tau(s, a_1, \ldots, a_M) + \alpha\text{Nash } Q_m^\tau(s^{\tau+1})]$.

Finally, (30) can be transformed into the following:

$$
\begin{aligned}
&\big\|P^\tau Q_m - P^\tau Q_m^*\big\| \\
&\quad = \max_{m\in M}\max_{s\in S} \\
&\qquad \times \big|\alpha\delta_1(s)\ldots\delta_M(s)Q_m(s) - \alpha\delta_1^*(s)\ldots\delta_M^*(s)Q_m^*(s)\big| \\
&\quad = \alpha\max_{m\in M} \\
&\qquad \times \big|\delta_1(s)\ldots\delta_M(s)Q_m(s) - \delta_1^*(s)\ldots\delta_M^*(s)Q_m^*(s)\big|. \quad (31)
\end{aligned}
$$

Based on the definition of the Nash equilibrium, the learning agent takes an action rationally by conjecturing on the behaviors of other agents. Therefore, it can be deduced that

$$\delta_m(s)\delta_{-m}(s)Q_m(s) \geq \hat{\delta}_m(s)\delta_{-m}(s)Q_m(s) \qquad (32)$$

$$\delta_m(s)\delta_{-m}(s)Q_m(s) \leq \delta_m(s)\hat{\delta}_{-m}(s)Q_m(s). \qquad (33)$$

For all $\hat{\delta}_m \in \delta(A_m)$, $\hat{\delta}_{-m} \in \delta(A_{-m})$, in which

$$\delta_{-m}(s) = \delta_1(s)\ldots\delta_{m-1}(s)\delta_{m+1}(s)\ldots\delta_M(s). \qquad (34)$$

For all Q and $\hat{Q} \in \mathbb{Q}$, (35) can be obtained as follows:

$$
\begin{aligned}
&\Big\|P_\tau Q - P_\tau\hat{Q}\Big\| \\
&\quad = \max_m\Big\|P_\tau Q_m - P_\tau\hat{Q}_m\Big\| \\
&\quad = \max_m\max_s\Big\|P_\tau Q_m(s) - P_\tau\hat{Q}_m(s)\Big\| \\
&\quad = \max_m\max_s \\
&\qquad \Big|\alpha\delta_1(s)\ldots\delta_M(s)Q_m(s) - \alpha\hat{\delta}_1(s)\ldots\hat{\delta}_M\hat{Q}_m(s)\Big| \\
&\quad = \max_m\alpha \\
&\qquad \Big|\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s)\Big|. \quad (35)
\end{aligned}
$$

According to the Nash equilibrium strategy, we have the following.

TABLE I
NETWORK ATTRIBUTES REQUIRED BY 5G SERVICES

Network Attributes	5G Service Types		
	Smart Health	***VR&AR***	***Indus. Mach.***
BW(Mbps)	10	200	10
EE($1e-6$ J/bit)	30	50	1
Delay(ms)	1	20	1
Jitter(ms)	1	2	3
PLR(per 10^6)	10	20	30
Price	10	10	7
Resource units required	6	15	2

1) If $\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \geq 0$, then

$$
\begin{aligned}
&\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\
&\quad \leq \delta_m(s)\delta_{-m}(s)Q_m(s) - \delta_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\
&\quad \leq \delta_m(s)\hat{\delta}_{-m}(s)Q_m(s) - \delta_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\
&\quad \leq \Big\|Q_m(s) - \hat{Q}_m(s)\Big\|. \qquad (36)
\end{aligned}
$$

2) If $\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \leq 0$, similarly, we can obtain that

$$
\begin{aligned}
&\Big|\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s)\Big| \\
&\quad \leq \Big\|Q_m(s) - \hat{Q}_m(s)\Big\|. \qquad (37)
\end{aligned}
$$

Therefore, based on (36) and (37), the following formula holds:

$$\Big\|P_\tau Q - P_\tau\hat{Q}\Big\| \leq \max_m\max_s\alpha\Big\|Q_m(s) - \hat{Q}_m(s)\Big\| = \alpha\Big\|Q - \hat{Q}\Big\|. \qquad (38)$$

Because $\hat{Q} \in \mathbb{Q}$ and $Q^* \in \mathbb{Q}$, (26) in Lemma 1 is proved, and the MAQNS algorithm satisfies the convergence condition. ■

VIII. NUMERICAL RESULTS AND EVALUATION

This section conducts a series of experiments to evaluate the efficacy of the proposed network selection scheme. First, we study differentiated access selection of each service in our model. Then, we examine the convergence by setting a fixed user arrival rates, as well as compare the influence of different discount factors on the performance of the MAQNS algorithm. Last but not least, we compare the performance of the MAQNS with other several alternative selection schemes [10]–[13].

A. Parameters Settings

In the experiments, as shown in Fig. 2, there are three types of networks considered in the 5G heterogeneous wireless networks, i.e., LTE-A, 5G, and Wi-Fi 6, and these networks are supposed to provide 5G novel services, including smart health, VR&AR, and industrial machinery service. Besides, these 5G services have diversified demands for different network attributes in Table I.

According to the analysis in [39]–[42], 5G and Wi-Fi 6 networks adopt OFDMA modulation, while LTE-A applies OFDM modulation [43], [44], in which the wireless spectrum is cut into time–frequency resource units [45]. Consequently, the available network resource units will limit the total number of UDs to access. Table II lists the network parameter setting.

TABLE II
NETWORK PARAMETERS

Parameters of Networks	**Network Types**		
	5G	***Wi-Fi 6***	***LTE-A***
Network bandwidth(MHz)	30	10	5
EE($1e-6$ J/bit)	30	10	1
Delay(ms)	1	5	30
Jitter(ms)	1	2	20
PLR(per 10^6)	0.5	1.5	15
Price	10	1	6
Available resource units	150	50	25
N^m_{sym}	28	28	14
N^m_{sub}	12	12	12
Modulation mode	256-QAM	1024-QAM	16-QAM
$size(mod^m)$	3	3	2
$R(cod^m)$	3/4	5/6	1/2
BER	10^{-6}	10^{-8}	10^{-5}

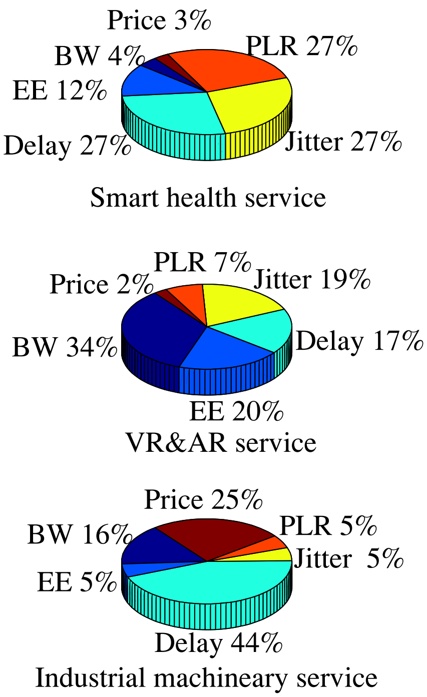

Fig. 5. Service preferences for different network attributes.

B. Service Preference Analysis

In order to consider requirements of service for access and effectively improve user experience, the AHP is applied in MAQNS to compute the weight of decision attributes, thereby reflecting the diversified service preference for different network attributes, which is represented in Fig. 5.

As can be seen from Fig. 5, industrial machinery service has the highest demand on the network delay, while we can not neglect the strict requirement of delay jitter and PLR in smart health service. As for the VR&AR service, which needs large bandwidth, higher energy efficiency and lower jitter, as well as lower price due to its high requirements on the clarity of pictures.

C. Convergence Demonstration

Convergence is an important premise for the effectiveness of the Q-learning algorithm, and the accumulated reward of the online learning algorithm can be appropriately studied to examine its convergence. The proposed MAQNS algorithm use the ε(s) greedy strategy to enable the agent to explore actions at the probability of ε(s), and select optimal actions by probability 1-ε(s). The arrival rate for users at Wi-Fi 6, 5G, and LTE-A are set to [0.2, 0.3, 0.4], respectively, and the service time is 30 s, while a time slot size is equal to 0.1 s and the discount factor α is computed to be 0.99. In addition, based on the parameters in Tables I and II, as well as the system model in Section III, Fig. 6 depicts the accumulated discounted reward of our heterogeneous wireless system versus time slot. As shown in Fig. 6 that the x-axis represents the index of each time slot which ranges from 0 to 3000, and the y-axis indicates the accumulated reward corresponding to the time slot. We can see that in the first 700 time slots, our accumulated reward is rising rapidly in the oscillation by exploration and exploitation so as to achieve the global optimum. In the next 700 time slots, it is clearly seen that the accumulated reward keeps stable in a certain range. Thereby, the convergence of the proposed MAQNS algorithm is verified in the experiment.

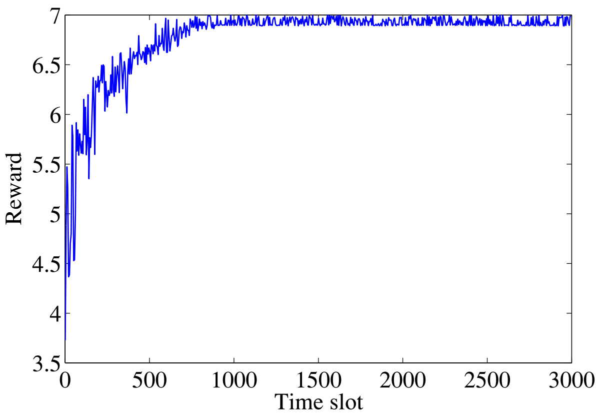

Fig. 6. Convergence of MAQNS.

D. Optimal Evaluation of Discount Factor

In order to study the influence of the discount factor on system performance, under different discount factors, we study the changes of total throughput and user blocking probability with the user arrival rate, respectively, and other parameters are set to be the same. It can be noted that in Figs. 7 and 8, comparing with other discounter factor values, when the value of discounter factor α is equal to 0.99, the total throughput and blocking probability are the best. Thus, a higher discount factor value will make more contribution to the long-term performance of the system.

E. Evaluation on the User Access

Similarly, we set the parameters as described above, and then the access ratio of UDs in different networks is evaluated to reveal the convergence of the proposed MAQNS algorithm. Fig. 9(a)–(c), respectively, describes the ratio of UDs requesting different services to access Wi-Fi 6, 5G, and LTE-A at each time slot. The access ratio of UDs refers to the ratio of the

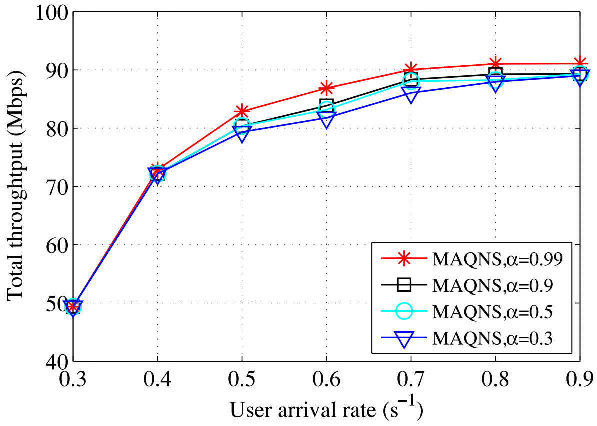

Fig. 7. Impact of the discount factor on the throughput.

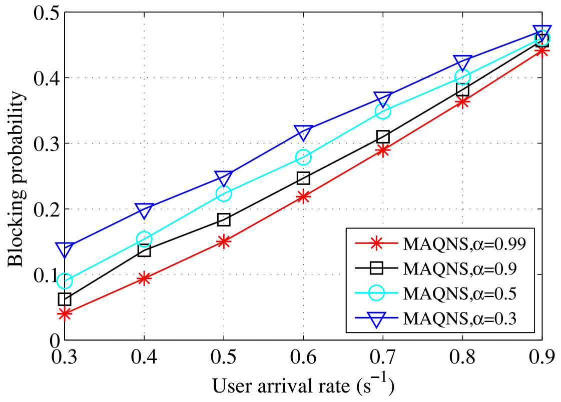

Fig. 8. Impact of the discount factor on the blocking probability.

number of UDs requesting each service accessing a specific network to the total number of UDs accessing this network.

As shown in Fig. 9(a), the *x*-axis represents the index of each time slot, and the *y*-axis indicates the access ratio of UDs requesting different services in Wi-Fi 6. From Fig. 9(a), we can observe that before the first 700 time slots, the ratio of UDs requesting industrial machinery service to access Wi-Fi 6 is on the rise, while the access ratio of UDs requesting smart health service slightly increases, and that of UDs requesting VR&AR service shows a downward trend. This is because industrial machinery service prefers Wi-Fi 6 with good delay performance and relatively low price. After 700 time slots, the access ratio of UDs tends to be stable, and the ratio of UDs requesting three types of services to access Wi-Fi 6 is 0.38:0.09:0.53.

As can be seen in Fig. 9(b), the ratio of UDs requesting smart health service to access 5G gradually increases before 700 time slots, while the access ratio of UDs requesting industrial machinery service increases marginally, and that of UDs requesting VR&AR service continues to decrease. This is because smart health service expects 5G with superior performance in terms of delay, jitter, and PLR. After 700 time slots, the access ratio of UDs requesting these three service is 0.50:0.10:0.40.

From Fig. 9(c), it can be seen that the ratio of UDs requesting VR&AR service to access LTE before 700 time slots continues to increase. On the contrary, the access ratio of UDs requesting the other two services shows a downward trend. This is because although VR&AR service has certain requirements for energy efficiency, its preferences for network delay and price are not very high compared to the other two services. When the access ratio of UDs tends to be stable after 700 time slots, the ratio of UDs requesting these three services to access LTE is 0.1:0.72:0.18.

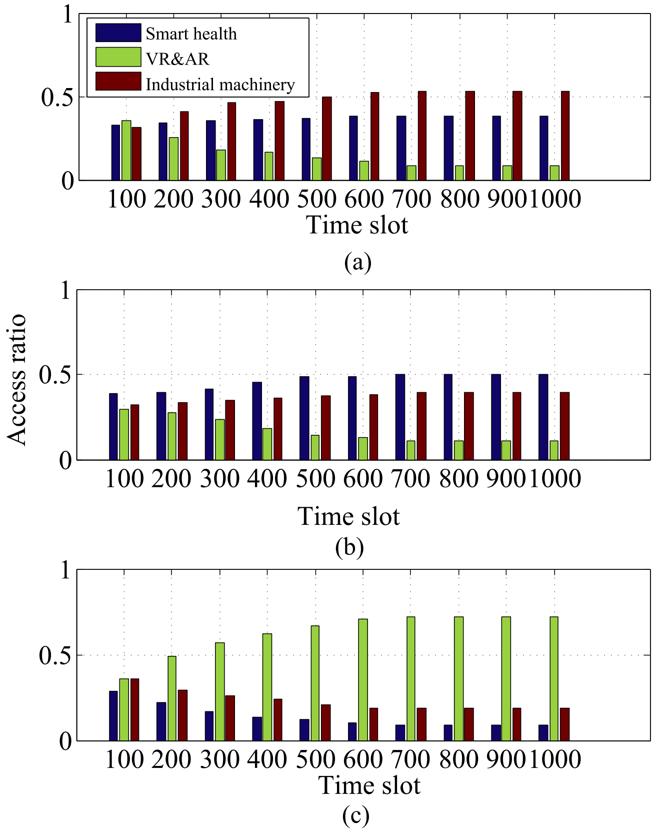

Fig. 9. Access ratio of UDs in different networks. (a) Access ratio in Wi-Fi 6. (b) Access ratio in 5G. (c) Access ratio in LTE-A.

In general, the proposed MAQNS algorithm fully considers the differentiated requirements of service and balances the access ratio of UDs accessing each network. After 700 time slots, the access ratio of UDs requesting different types of services in each network can maintain a balance. This is because MAQNS effectively makes the selection strategies of different networks follow the Nash equilibrium through exploration and exploitation, so as to enhance the system performance.

F. Throughput Comparison

We use several network selection algorithms, including FTNS [11], DVHD [12], HUMANS [13], and RFEQG [10] to compare the throughput of the proposed MAQNS. In the simulation, the *x*-axis represents the user arrival rate of LTE-A, while the user arrival rates of 5G and Wi-Fi 6 are set to be 0.1 and 0.2, respectively, which is less than that of LTE-A. In particular, the arrival rates of users requesting different services are set to be the same.

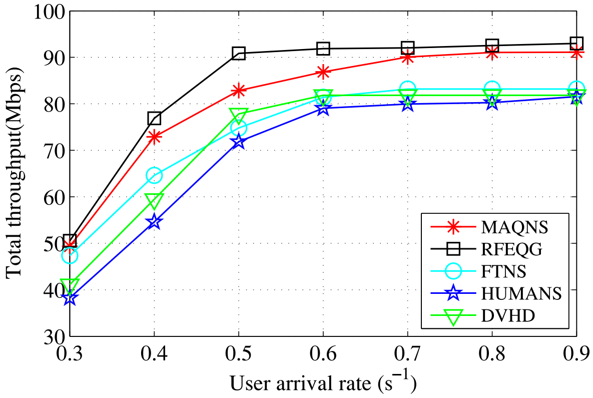

Fig. 10. Total throughput of the 5G system.

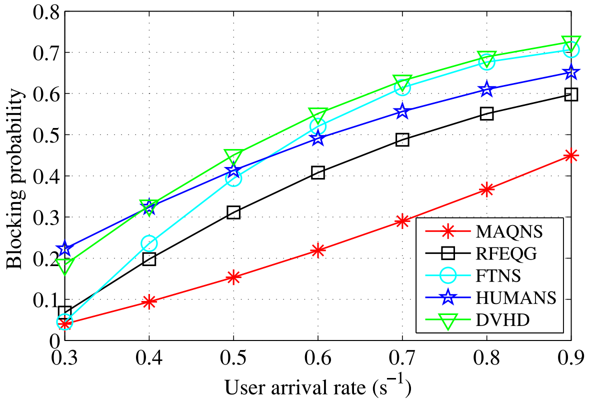

Fig. 11. Blocking probability of five algorithms.

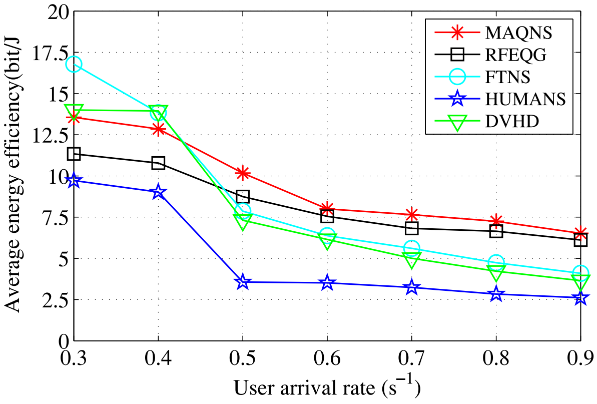

Fig. 12. Average energy efficiency experienced by users.

As shown in Fig. 10, the total throughput of the 5G system increases with the user arrival rate. The main reason for this is that the increased number of users accessing networks contributes to the system throughput. Actually, as the user arrival rate is small, the total throughput is greatly increasing. However, as the user arrival rate rises continuously, especially when the arrival rate reaches 0.5, the growth trend of the total system throughput slows down dramatically due to the resource competition among multiple users.

We can see that the RFEQG algorithm achieves the maximum throughput among these algorithms, as maximizing the total system throughput is the optimal objective in the RFEQG algorithm. Thus, in RFEQG, the network priority of user access is 5G, Wi-Fi 6, and LTE-A in descending order, and the algorithm will make the 5G BS hold as many users requesting smart health service and industrial machinery service as possible. However, the MAQNS proposed can enjoy a suboptimal throughput performance. The reason for this is that the MAQNS scheme considers not only enhancing the system performance on throughput but also reducing the influence of user blocking on throughput. In brief, the throughput obtained by MAQNS is just less than that of RFEQG by 2.03%, and more than the other three algorithms by 9.51%.

G. Blocking Probability Comparison

As represented in Fig. 11, the user blocking probability increases versus user arrival rate. When the network capacity cannot provide service requests from users, users will encounter blocking. Therefore, when user arrival rate is relatively little, the probability of users being blocked will be smaller. When the user arrival rate increases continuously, the number of users who need to access the network will increase rapidly; thus, the blocking probability of users will increase correspondingly. Particularly, MAQNS owns the lowest user blocking among the four algorithms mentioned, with the reason that MAQNS takes into account the effect of blocking cost on the network reward, which can slow the blocking probability to some extent.

RFEQG aims at maximizing the total system throughput, so it will try to access as many users requesting smart health and industrial machinery service as possible in 5G and Wi-Fi 6, which causes users requesting VR&AR service to be blocked easily. In FTNS, HUMANS, and DVHD, the users consider all the attributes of candidate networks through MADM and access the network with best comprehensive performance. However, when users adopt the three algorithms for network selection, the performance of each network attributes can compensate each other, which causes users to access the unreasonable network that has the best comprehensive performance but may be heavily loaded. This will further aggravates network congestion. In general, compared with the remaining algorithms, the blocking probability obtained by the MAQNS algorithm is reduced by at least 14.82%.

H. Comparison on Energy Efficiency

In this part, we will compare the performance on average energy efficiency of the five algorithms.

The network attributes considered in our model can be divided into positive and negative attributes. We select the average energy efficiency as a representative of positive network attributes. It can be seen in Fig. 12 that as the increase of user arrival rate, the average energy efficiency experienced by users decreases gradually.

Specifically, before the user arrival rate is greater than 0.4, the energy efficiency in the DVHD and FTNS algorithms exceeds that of MAQNS. The reason for this is that, the network resources are enough to provide for a small number of users to access. In the mean time, the proportion of users accessing 5G in DVHD and FTNS is more than that in

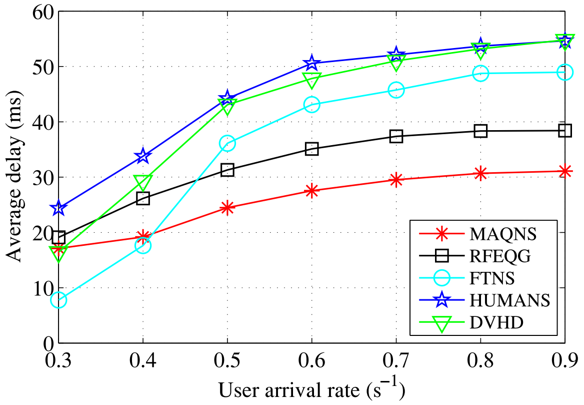

Fig. 13. Average delay experienced by users.

MAQNS, and their user blocking is especially lower. However, FTNS is affected by the 5G comprehensive performance and falls into local optimum, ignoring factors, such as network BS load; thus, the blocking probability in DVHD and FTNS rises rapidly after the user arrival rate is greater than 0.4, which results in a sharply decrease on the energy efficiency. Compared with other four algorithms, the average energy efficiency experienced by users in the MAQNS algorithm can be improved by at least 6.65%.

I. Comparison on Delay

In this section, we would like to take delay as the representative of negative network attributes to evaluate the performance of five algorithms. As shown in Fig. 13, before the user arrival rate reaches 0.4, the FTNS algorithm has a lower average delay than that of MAQNS. That is because, FTNS falls into local optimization and enables a lot of users with current service requests to access 5G, rather than reasonably consider the impact of 5G capacity on the subsequent users. After the user arrival rate exceeds 0.4, MAQNS enjoys the lowest delay for the network selection. This is because MAQNS not only effectively considers service requirements of current users but also ensures the requirements of subsequent users. HUMANS and DVHD attempt to make users select network from their own perspective, and do not consider the effect on the subsequent users, resulting in the dramatic increase of the average delay. The RFEQG algorithm does not consider service requirements so that there are many blocked users when user arrival rate increases continuously. Compared with other remaining algorithms, the average delay experienced by users in the MAQNS algorithm can be reduced by at 19.09% as the user arrival rate keeps to increase.

IX. Conclusion

In this article, under the scenario of 5G heterogeneous network system, we formulated the optimal RAT selection problem into a discrete Markov model with the target of improving system throughout, reducing user blocking, and ensuring user experience. Then, we built a multiagent structure network selection model based on the noncooperative stochastic game, and proposed an intelligent network selection algorithm, named MAQNS, by adapting Nash Q-learning into the multiagent structure. Particularly, in the MAQNS, each agent updates Q-values under the joint actions taken by other agents, and the Nash equilibriums is considered as the optimal access selection decisions of UDs. In addition, the online ε-greedy strategy was applied by MAQNS, which can be utilized to promote the efficiency of the proposed algorithm and enable the algorithm to achieve convergence. Not only that, in order to guarantee various requirements of IoT services and ensure user experience efficiently, we measured the preference weight of each services on heterogeneous networks by jointly performing AHP and GRA algorithms. Numerical results illustrated that the proposed MAQNS cannot only significantly improve the total system throughput and average energy efficiency but also effectively avoid user blocking and reduce average delay user experienced to some extent.

References

[1] L. Chettri and R. Bera, "A comprehensive survey on Internet of Things (IoT) toward 5G wireless systems," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 16–32, Jan. 2020.

[2] B. Bajic, A. Rikalovic, N. Suzic, and V. Piuri, "Industry 4.0 implementation challenges and opportunities: A managerial perspective," *IEEE Syst. J.*, vol. 15, no. 1, pp. 546–559, Mar. 2021.

[3] Q. Fan and N. Ansari, "Application aware workload allocation for edge computing-based IoT," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 2146–2153, Jun. 2018.

[4] *The Mobile Economy*, GSMA, London, U.K., 2020. [Online]. Available: https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=51249388&file=2915-260220-Mobile-Economy.pdf

[5] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.

[6] A. Zhu, S. Guo, and M. Ma, "i5GAccess: Nash Q-learning based multiservice edge users access in 5G heterogeneous networks," in *Proc. IEEE/ACM 28th Int. Symp. Qual. Service. (IWQoS)*, Hang Zhou, China, 2020, pp. 1–10.

[7] A. Roy, P. Chaporkar, and A. Karandikar, "Optimal radio access technology selection algorithm for LTE-WiFi network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6446–6460, Jul. 2018.

[8] S. Ahmed and M. O. Farooq, "Analysis of access network selection ANE switching metrics for LTE and WiFi HetNets," in *Proc. Int. Conf. Sel. Topics Mobile Wireless Netw. (MoWNeT)*, Avignon, France, 2017, pp. 1–5.

[9] Z. Du, C. Wang, Y. Sun, and G. Wu, "Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer," *IEEE Access*, vol. 6, pp. 33275–33284, 2018.

[10] X. Wang, J. Li, L. Wang, C. Yang, and Z. Han, "Intelligent user-centric network selection: A model-driven reinforcement learning framework," *IEEE Access*, vol. 7, pp. 21645–21661, 2019.

[11] H. Yu, Y. Ma, and J. Yu, "Network selection algorithm for multiservice multimode terminals in heterogeneous wireless networks," *IEEE Access*, vol. 7, pp. 46240–46260, 2019.

[12] M. El Helou, S. Lahoud, M. Ibrahim, and K. Khawam, "Satisfaction-based radio access technology selection in heterogeneous wireless networks," in *Proc. IFIP Wireless Days (WD)*, Valencia, Spain, 2013, pp. 1–4.

[13] G. Araniti, P. Scopelliti, G.-M. Muntean, and A. Iera, "A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 83–93, Mar. 2019.

[14] J. Yan, L. Zhao, and J. Li, "A prediction-based handover trigger time selection strategy in varying network overlapping environment," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, 2011, pp. 1–5.

[15] *IEEE Draft Standard for Information Technology–Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks–Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Standard P802.11-REVmd/D3.0, Oct. 2019.

[16] M. Gerasimenko, N. Himayat, S.-P. Yeh, S. Talwar, S. Andreev, and Y. Koucheryavy, "Characterizing performance of load-aware network selection in multi-radio (WiFi/LTE) heterogeneous networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Atlanta, GA, USA, 2013, pp. 397–402.

[17] M. A. Senouci, S. Hoceini, and A. Mellouk, "Utility function-based topsis for network interface selection in heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, 2016, pp. 1–6.

[18] R. K. Prasad and T. Jaya, "Optimal network selection in cognitive radio network using simple additive weighting method with multiple parameters," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Tirunelveli, India, Nov. 2019, pp. 715–721.

[19] N. Zarin and A. Agarwal, "A hybrid network selection scheme for heterogeneous wireless access network," in *Proc. IEEE 28th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–6.

[20] W. Song, H. Jiang, W. Zhuang, and A. Saleh, "Call admission control for integrated voice/data services in cellular/WLAN interworking," in *Proc. IEEE Int. Conf. Commun.*, vol. 12. Istanbul, Turkey, 2006, pp. 5480–5485.

[21] W. Song, H. Jiang, and W. Zhuang, "Performance analysis of the WLAN-first scheme in cellular/WLAN interworking," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1932–1952, May 2007.

[22] K.-S. Shin, G.-H. Hwang, and O. Jo, "Distributed reinforcement learning scheme for environmentally adaptive IoT network selection," *Electron. Lett.*, vol. 56, no. 9, pp. 462–464, Apr. 2020.

[23] A. Zhu, S. Guo, B. Liu, M. Ma, J. Yao, and X. Su, "Adaptive multiservice heterogeneous network selection scheme in mobile edge computing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6862–6875, Aug. 2019.

[24] S. Pal, S. K. Das, and M. Chatterjee, "User-satisfaction based differentiated services for wireless data networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, vol. 2. Seoul, South Korea, May 2005, pp. 1174–1178.

[25] A. O. Mufutau, F. P. Guiomar, M. A. Fernandes, A. Lorences-Riesgo, A. Oliveira, and P. P. Monteiro, "Demonstration of a hybrid optical fiber–wireless 5G fronthaul coexisting with end-to-end 4G networks," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 3, pp. 72–78, Jan. 2020.

[26] C. Zhang, M. Dong, and K. Ota, "Fine-grained management in 5G: DQL based intelligent resource allocation for network function virtualization in C-RAN," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 428–435, Jun. 2020.

[27] 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.

[28] S. K. Kharroub, K. Abualsaud, and M. Guizani, "Medical IoT: A comprehensive survey of different encryption and security techniques," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Limassol, Cyprus, 2020, pp. 1891–1896.

[29] P. Zeng, W. Zhaowei, Z. Jia, L. Kong, D. Li, and X. Jin, "Time-slotted software-defined industrial Ethernet for real-time quality of service in industry 4.0," *Future Gener. Comput. Syst.*, vol. 99, pp. 1–10. Oct. 2019.

[30] J. Song, F. Yang, W. Zhang, W. Zou, Y. Fan, and P. Di, "A fast FoV-switching dash system based on tiling mechanism for practical omnidirectional video services," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2366–2381, Sep. 2020.

[31] Z. Ge and Y. Liu, "Analytic hierarchy process based fuzzy decision fusion system for model prioritization and process monitoring application," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 357–365, Jan. 2019.

[32] T. Saaty, "Decision making with the analytic hierarchy process," *Int. J. Service. Sci.*, vol. 1, pp. 83–98, Jan. 2008.

[33] M. Brunelli, *Introduction to the Analytic Hierarchy Process*. New York, NY, USA: Springer, 2015.

[34] T. Saaty, "A scaling method for priorities in hierarchical structures," *J. Math. Psychol.*, vol. 15, no. 3, pp. 234–281, Jun. 1977.

[35] Q. Xiao, M. Shan, M. Gao, and X. Xiao, "Grey information coverage interaction relational decision making and its application," *J. Syst. Eng. Electron.*, vol. 31, no. 2, pp. 359–369, Apr. 2020.

[36] L. Xu, H. Ma, and D. Ren, "Reliability analysis of tractor multi-way valves based on the improved weighted grey relational method," *J. Eng.*, vol. 2019, no. 13, pp. 86–92, May 2019.

[37] M. Han, R. Zhang, T. Qiu, M. Xu, and W. Ren, "Multivariate chaotic time series prediction based on improved grey relational analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 10, pp. 2144–2154, Oct. 2019.

[38] C. Xu, W. Zhao, L. Li, Q. Chen, D. Kuang, and J. Zhou, "A nash Q-learning based motion decision algorithm with considering interaction to traffic participants," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12621–12634, Nov. 2020.

[39] D. Xie, J. Zhang, A. Tang, and X. Wang, "Multi-dimensional busy-tone arbitration for OFDMA random access in IEEE 802.11ax," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4080–4094, Jun. 2020.

[40] K.-H. Lee, "Using OFDMA for MU-MIMO user selection in 802.11ax-based Wi-Fi networks," *IEEE Access*, vol. 7, pp. 186041–186055, 2019.

[41] S. Althunibat, R. Mesleh, and K. A. Qaraqe, "IM-OFDMA: A novel spectral efficient uplink multiple access based on index modulation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10315–10319, Oct. 2019.

[42] N.-T. Le, L.-N. Tran, Q.-D. Vu, and D. Jayalath, "Energy-efficient resource allocation for OFDMA heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7043–7057, Oct. 2019.

[43] C. B. Barneto *et al.*, "Full-duplex OFDM radar with LTE and 5G NR waveforms: Challenges, solutions, and measurements," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 10, pp. 4042–4054, Oct. 2019.

[44] T. Cui, F. Gao, A. Nallanathan, H. Lin, and C. Tellambura, "Iterative demodulation and decoding algorithm for 3GPP/LTE-A MIMO-OFDM using distribution approximation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1331–1342, Feb. 2018.

[45] W. Wu, Q. Yang, R. Liu, T. Q. S. Quek, and K. S. Kwak, "Online spectrum partitioning for LTE-U and WLAN coexistence in unlicensed spectrum," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 506–520, Jan. 2020.