Intelligent Network Selection Algorithm for Multi-Service 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 networks 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 network system, how to design an intelligent network selection scheme for user devices (UDs) is becoming a crucial challenge in 5G heterogeneous network system. Most of the existing network selection methods only optimize the selection strategies from user side or network side, which resulting in heavy network congestion, poor user experience and system performance degradation. Accordingly, we propose a Multi-Agent 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 Grey Relation Analysis (GRA) are jointly utilised to obtain user preferences for each network. Finally, performance evaluation demonstrates that comparing to existing schemes, MAQNS proposed can not only improve system throughput and reduce user blocking, but also promote user experience on average energy efficiency and delay.**

***Index Terms*—Nash Q-learning, Nash equilibrium, AHP, GRA, Heterogeneous wireless networks, Network selection.**

I. Introduction

The paradigm of 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 4th generation wireless systems (4G) cannot withstand huge IoT connections. In addition, Massive MachineType 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, 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 idle mode. This situation will further aggravate load imbalance, network congestion and resource wasting. From the perspective of 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, 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 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 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 multi-agent structure. Then we propose the intelligent network selection

M. Ma and S. 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, P. R. China.

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

Y. Yang is with the Department of Electrical and Computer Engineering, Stony Brook University, New York, USA, 11794.

Corresponding author: S. Guo, e-mail:songtao_guo@163.com.

algorithm based on the Nash Q-learning method, named Multi-Agent Q-learning Network Selection (MAQNS) algorithm. For the sake of maximizing the long-term performance of multi-agent structure, access of UDs is not only determined by the current network status, but also depends on the expected future demand.

The detailed contributions of our paper are reflected in the following aspects.

- 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.
- Aiming at capturing the dynamic access of UDs in the network selection, we build a network selection model based on discrete time Markov chains. Then we propose to utilize Nash Q-learning to design a multi-agent Q-learning network selection algorithm, which can learn the joint selection strategies of each network by trial and error. In this way, the Nash equilibrium can be achieved.
- 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 Grey Relation Analysis (GRA) to achieve user preferences for each network, which will be appropriately modeled into network reward function.
- Evaluation results indicate that as the number of UDs increases, comparing with existing schemes [10]–[13], our proposed MAQNS can not 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 rest of our paper is organized as follows: In Section II, 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 paper in Section IX.

II. Related Work

Generally, the mechanisms which study access selection in a heterogeneous network scenario can be classified into user-centric mechanism [11]–[19] and network-centric mechanism [9] [10] [20]–[23].

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 peak rate [15] and signal-to-noise ratio [16] are also used to determine network access. These above methods based on 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 can not accurately reflect the individual service requirements. Besides, the network selection method based on single decision attribute has poor applicability in complex heterogeneous network environment [24].

Therefore, for the sake of eliminating the limitation of single-attribute network selection methods, and improving service quality for users, multi-attribute decision-making (MADM) theory has been widely studied. The typical MADM approaches include technique for order preference by similarity to ideal solution (TOPSIS) [11] [17], simple additive weighting (SAW) [12] [18], 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. 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 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 over-loaded [25].

Few network-centric access selection mechanisms are studied in [9] [10] [20]–[23]. 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 resources. Roy et al. [22] 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] [23], 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 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3073027, IEEE Internet of Things Journal

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 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 is illustrated in the Fig. 1, the continuous time can be divided into equi-spaced time intervals named as 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.

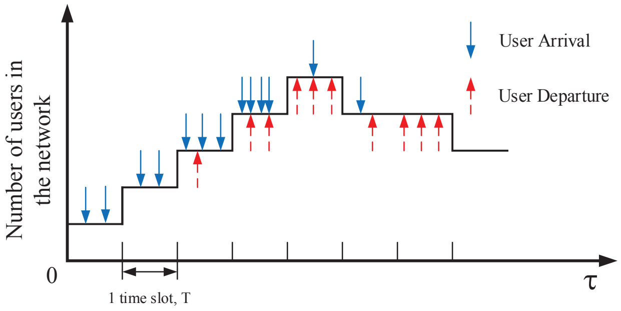

Fig. 1. System time model

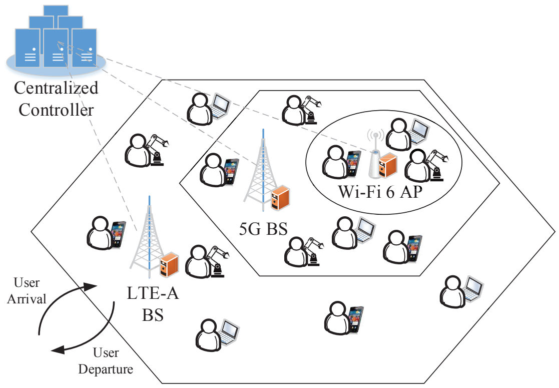

Fig. 2. 5G heterogeneous wireless access networks

5G system will integrate different RATs and allow the coexistence of different RATs, aiming at improving the capabilities of 5G system in multiple aspects [26]. It is known that 5G system can not 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 [27]. 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 [28]. Furthermore, 5G system enables UDs to obtain services from different available access networks at anytime or anywhere.

In this paper, as shown in Fig. 2, a 5G heterogeneous wireless access networks 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 [29]. 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 [30]. 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 [31].

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 the $\lambda_k^m[\tau]$ and $\mu_k^m[\tau]$ respectively to expresses the user arrival rate and departure rate in time slot τ, where k and m represents 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:

$$P\{x\ arrivals\} = \frac{(\lambda_k^m[\tau]T)^x \cdot e^{-\lambda_k^m[\tau]T}}{x!} \tag{1}$$

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

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

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 grey correlation coefficient, which can be modeled as the preference of users for each candidate network.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3073027, IEEE Internet of Things Journal

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 [32]. Since there is only one criterion layer to be considered in our AHP algorithm, the AHP proposed can be executed according to the flow chart presented in Fig. 3.

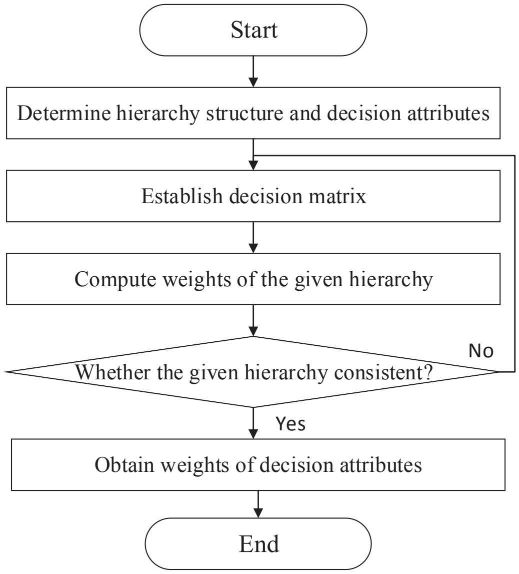
Start
Determine hierarchy structure and decision attributes
Establish decision matrix
Compute weights of the given hierarchy
Whether the given hierarchy consistent?
No
Yes
Obtain weights of decision attributes
End

Fig. 3. The flow chart of AHP in MAQNS

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: the target layer denotes the optimum networks that UDs requesting specific service expect to access, the criteria layer indicates the demands of services for network attributes, and the solution layer expresses the candidate networks in the system model.

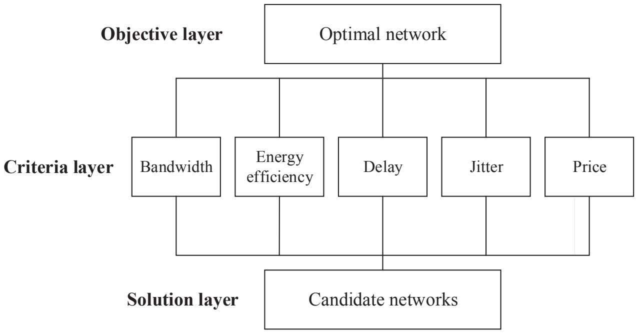

Fig. 4. Hierarchical model in AHP

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

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

The cost attributes can be normalized by

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

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 $l = 1, 2, ..., L$.

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 $1-9$ level to compare the normalized value of attribute u and v to get g_{uv} [33]. And 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 $g_{vu} = \frac{1}{g_{uv}}$.

4) Calculating the preference weight by

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

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. At first, we calculate the consistency index CI by:

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

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

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

where RI represents the average random consistency index, which is the arithmetic mean value of the eigenvalue of the random judgement matrix multiple times [34], and the value of RI corresponding to the matrix of each order can be estimated according to Satty's research [35]. For consistency ratio CR, if $CR \leq 0.1$, 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.

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
```

B. Grey Relational Analysis Algorithm

By calculating the grey correlation coefficient (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 [36]–[38]. 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 grey correlation coefficient to indicate the correlation between candidate networks and the ideal network. The greater the value of weighted grey correlation coefficient, 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:

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 the formula (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

$$\begin{aligned} d^* =& \{d_l^* | d_l^* = (\max_m\{d_{ml}\} \mid d_{ml} \in d_b), \\ & (\min_m\{d_{ml}\} \mid d_{ml} \in d_c)\},\ m \in M, l \in L \end{aligned} \tag{10}$$

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

4) Acquiring the grey correlation coefficient ξ_{ml}:

$$\begin{aligned} \xi_{ml} =& \frac{\min\limits_m \min\limits_l |d_{ml} - d_l^*|}{|d_{ml} - d_l^*| + \rho \max\limits_m \max\limits_l |d_{ml} - d_l^*|} \\ &+ \frac{\phi \max\limits_m \max\limits_l |d_{ml} - d_l^*|}{|d_{ml} - d_l^*| + \phi \max\limits_m \max\limits_l |d_{ml} - d_l^*|} \end{aligned} \tag{11}$$

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

5) Measuring the weighted grey correlation coefficient 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 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.

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 $\Xi_k^m$
```

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. And 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. And 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 $s^{\tau} \in S$, $s^{\tau+1} \in S$, $a \in A$. 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 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 the network m choose UDs who request service k, $F(s,a)$ is expressed as below:

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

where N_k^m is the number of UDs with service k request accessing network m, Ξ_k^m defines the weighted grey correlation coefficient computed through 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:

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

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 indicates the number of symbols and subcarriers in a resource unit, $s_{ize}(mod^m)$ and $R(cod^m)$ respectively denotes the size of constellation and coding rate of the network m, besides, BER is the error rate of network m with corresponding modulation.

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

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

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 multi-agent structure network selection model on the basis of non-cooperative stochastic game. Furthermore, an intelligent network selection algorithm named MAQNS is proposed by adapting Nash Q-learning [39] into multi-agent structure.

A. Stochastic Game Model

Depending on the concept of non-cooperative stochastic game, each network in the system is considered as an agent, thus the 5G heterogeneous networks system will form a multi-agent structure. And the definition on the stochastic game model is illustrated in Definition 1.

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

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

in which S indicates the state space of multi-agent, M refers to the number of agents, which also characterize the 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 the A_m, then r_m is the reward of agent m for taking an action. Finally, P means the state transition probability of 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:

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

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

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

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

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

$$\begin{aligned} & V_m(s, \delta_1^*, \ldots, \delta_M^*) \\ & \geq V_m(s, \delta_1^*, \ldots, \delta_{m-1}^*, \delta_m, \delta_{m+1}^*, \ldots, \delta_M^*) \end{aligned} \tag{19}$$

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 Multi-agent Structure

In Nash Q-learning, the Nash Q-function claims that the joint actions of all the agent in the multi-agent structure should be determined, and the Nash Q-function of agent m is considered as $(s, a_1, ..., a_M)$. As all agents in the multi-agent 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

$$Q^*_m(s^\tau, a_1, \dots, a_M) = r_m(s^\tau, a_1, \dots, a_M) \\ + \alpha \sum_{s^{\tau+1} \in s} P(s^{\tau+1}|s^\tau, a_1, \dots, a_M) \cdot V_m(s^{\tau+1}, \delta^*_1, \dots, \delta^*_M) \tag{20}$$

in which $(\delta^*_1, \dots, \delta^*_M)$ denotes the joint Nash equilibrium strategy, $r_m(s^\tau, a_1, \dots, a_M)$ 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 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 Nash equilibrium.

$$Nash\ Q^\tau_m(s^{\tau+1})\delta_1(s^{\tau+1}) \dots \delta_M(s^{\tau+1}) Q^\tau_m(s^{\tau+1}) \tag{21}$$

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 ε-greedy strategy, in each time slot. Furthermore, the 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

$$Q^{\tau+1}_m(s^\tau, a_1, \dots, a_M) = (1 - \beta^\tau) Q^\tau_m(s^\tau, a_1, \dots, a_M) \\ + \beta^\tau [r^\tau_m + \alpha Nash\ Q^\tau_m(s^{\tau+1})] \tag{22}$$

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

Our MAQNS algorithm applies 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 the $\varepsilon(s)$ is defined by

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

where the $h(s, a)$ is the number of state-action pair 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

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

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 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 $\varepsilon(s)$-greedy strategy according to the other agents' actions observed, as well as computes its reward and other rewards by formula (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 (line 13-15). Afterwards, the state of the agent m changes into a new state (line 18), and the Q-values are updated to compute Nash equilibrium by formula (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.

Algorithm 3 Nash Q-learning based Network Selection

```
Input: User arrival rate $\lambda^m_k$, user departure rate $\mu^m_k$, 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;
 1: When $\tau = 0$, initialize $s^0$ by $\lambda^m_k$;
 2: Assign a learning agent $m$ and get its Q-values
    $Q^0_m(s, a_1, \dots, a_M) = 0$ as well as the other agents' Q-
    values $Q^0_{m'}(s, a_1, \dots, a_M) = 0$, where $s \in S$;
 3: for $\tau = 1$ to $\Gamma$ do
 4:   Observe $s^\tau$ based on $\lambda^m_k[\tau]$;
 5:   while the state of agent $m$ is $s^\tau$ do
 6:     for $m = 1$ to $M$ do
 7:       if Exploration then
 8:         Randomly take $a^\tau_m$;
 9:       else
10:         Adopt $a^\tau_m$ corresponding to $Nash\ Q$;
11:       end if
12:       Observe $a^\tau_1, \dots, a^\tau_M$ to calculate $r^\tau_1, \dots, r^\tau_M$
          through formula (13);
13:       $h(s, a) = h(s, a) + 1$;
14:       Update $\varepsilon(s)$ using formula (23);
15:       Update $C_m$;
16:     end for
17:   end while
18:   Get the next state $s^{\tau+1}$;
19:   Agent $m$ updates its own Q-values and that of others
      $Q^{\tau+1}(s^\tau, a_1, \dots, a_M)$ by formula (22);
20: end for
```

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 multi-agent structure model, the state space and action space of each agent are respectively assumed as $|S|$ and $|A|$. If the action set $|A_1| = \ldots = |A_M| = |\Lambda|$, which means that each agent owns a space capacity with $|S||A|^M$ to maitain the Q-tables. Therefore, space complexity of MAQNS is $M|S||A|^M$, which refers to the space complexity of MAQNS is linearly 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 the 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 network M is 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 require 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)
$$\begin{aligned} &0 \leq \beta^\tau(s, a_1, \ldots, a_M) < 1, \\ &\sum_{\tau=0}^{\infty} \beta^\tau(s, a_1, \ldots, a_M) = \infty, \\ &\sum_{\tau=0}^{\infty} [\beta^\tau(s, a_1, \ldots, a_M)]^2 < \infty \end{aligned} \tag{25}$$

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

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

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 $\vartheta^\tau \geq 0$ converges to 0 at the probability of 1, thus the following formula holds for $Q \to \mathbb{Q}$ and $Q^* = E[P^\tau Q^*]$:

$$\parallel P^\tau Q - P^\tau Q^* \parallel \leq \eta \parallel Q - Q^* \parallel + \vartheta^\tau \tag{26}$$

in which $\mathbb{Q}$ denotes the space of Q tables. And 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 $(V_1(\delta_1^*, \ldots, \delta_M^*), \ldots, V_M(\delta_1^*, \ldots, \delta_M^*))$, and $(\delta_1^*, \ldots, \delta_M^*)$ denotes a Nash equilibrium strategies in the discounted stochastic game. In addition, $V_m(\delta_1^*, \ldots, \delta_M^*) = (V_m(s_1, \delta_1^*, \ldots, \delta_M^*), \ldots, V_m(s_M, \delta_1^*, \ldots, \delta_M^*))$.

2) In the stage game $(Q_1^*(s), \ldots, Q_M^*(s))$, there exists Nash equilibrium revenues $(V_1(s, \delta_1^*, \ldots, \delta_M^*), \ldots, V_M(s, \delta_1^*, \ldots, \delta_M^*)).(\delta_1^*(s), \ldots, \delta_M^*(s))$, which is a Nash equilibrium point of stage game.

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

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

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

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

$$\begin{aligned} &Q_m^*(s^\tau, a_1, \ldots, a_M) = r_m(s^\tau, a_1, \ldots, a_M) \\ &+ \alpha \sum_{s^{\tau+1} \in S} P(s^{\tau+1}|s^\tau, a_1, \ldots, a_M) \\ &\cdot \delta_1^*(s^{\tau+1}), \ldots, \delta_M^*(s^{\tau+1}) Q_m^*(s^{\tau+1}) \\ &= \sum_{s^{\tau+1} \in S} P(s^{\tau+1}|s^\tau, a_1, \ldots, a_M) \\ &\cdot (r_m(s^\tau, a_1, \ldots, a_M) + \alpha \delta_1^*(s^{\tau+1}) \ldots, \delta_M^*(s^{\tau+1}) Q_m^*(s^{\tau+1})) \\ &= E[P_m^\tau Q_m^*(s, a_1, \ldots, a_M)] \end{aligned} \tag{28}$$

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 formula (26) in Lemma 1. According to the definition on the distance between two points, it can be deduced that:

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

Based on formula (29), we can derive that:

$$\begin{aligned} &\parallel P^\tau Q_m - P^\tau Q_m^* \parallel = \max_{m \in M} \max_{s \in S} \max_{a^1, \ldots, a^M} \\ &\mid P^\tau Q_m(s, a_1, \ldots, a_M) - P^\tau Q_m^*(s, a_1, \ldots, a_M) \mid \end{aligned} \tag{30}$$

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

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

$$\begin{aligned} & \parallel P^{\tau} Q_m - P^{\tau} Q_m^* \parallel \\ = & \max_{m \in M} \max_{s \in S} |\alpha \delta_1(s) \dots \delta_M(s) Q_m(s) \\ & - \alpha \delta_1^*(s) \dots \delta_M^*(s) Q_m^*(s)| \\ = & \alpha \max_{m \in M} \mid \delta_1(s) \dots \delta_M(s) Q_m(s) \\ & - \delta_1^*(s) \dots \delta_M^*(s) Q_m^*(s)| \end{aligned} \tag{31}$$

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

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

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

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

$$\delta_{-m}(s) = \delta_1(s) \dots \delta_{m-1}(s)\delta_{m+1}(s) \dots \delta_M(s) \tag{34}$$

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

$$\begin{aligned} & \parallel P_{\tau} Q - P_{\tau} \hat{Q} \parallel \\ = & \max_m \parallel P_{\tau} Q_m - P_{\tau} \hat{Q}_m \parallel \\ = & \max_m \max_s \parallel P_{\tau} Q_m(s) - P_{\tau} \hat{Q}_m(s) \parallel \\ = & \max_m \max_s \mid \alpha \delta_1(s) \dots \delta_M(s) Q_m(s) \\ & - \alpha \hat{\delta}_1(s) \dots \hat{\delta}_M \hat{Q}_m(s) \mid \\ = & \max_m \alpha \mid \delta_m(s) \delta_{-m}(s) Q_m(s) \\ & - \hat{\delta}_m(s) \hat{\delta}_{-m}(s) \hat{Q}_m(s) \mid \end{aligned} \tag{35}$$

According to the Nash equilibrium strategy, we have

(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

$$\begin{aligned} & \delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\ \leq & \delta_m(s)\delta_{-m}(s)Q_m(s) - \delta_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\ \leq & \delta_m(s)\hat{\delta}_{-m}(s)Q_m(s) - \delta_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \\ \leq & \parallel Q_m(s) - \hat{Q}_m(s) \parallel \end{aligned} \tag{36}$$

(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

$$\begin{aligned} & |\delta_m(s)\delta_{-m}(s)Q_m(s) - \hat{\delta}_m(s)\hat{\delta}_{-m}(s)\hat{Q}_m(s) \mid \\ \leq & \parallel Q_m(s) - \hat{Q}_m(s) \parallel \end{aligned} \tag{37}$$

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

$$\begin{aligned} & \parallel P_{\tau} Q - P_{\tau} \hat{Q} \parallel \\ \leq & \max_m \max_s \alpha \parallel Q_m(s) - \hat{Q}_m(s) \parallel \\ = & \alpha \parallel Q - \hat{Q} \parallel \end{aligned} \tag{38}$$

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

VIII. Numerical Results and Evaluation

This section conducts a series of experiments to evaluate the efficacy of the proposed network selection scheme. Firstly, 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 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 is 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.

TABLE I
The 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

According to the analysis in [40]–[43], 5G and Wi-Fi 6 networks adopt OFDMA modulation, while LTE-A applies OFDM modulation [44] [45], in which the wireless spectrum is cut into time-frequency resource units [46]. Consequently, the available network resource units will limit the total number of UDs to access. Table II lists the network parameters 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_{sym}^m	28	28	14
N_{sub}^m	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}

B. Service Preference Analysis

In order to consider requirements of service for access and effectively improve user experience, the AHP is applied in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3073027, IEEE Internet of Things Journal

MAQNS to compute the weight of decision attributes, thereby reflecting the diversified service preference for different network attributes, which is represented in Fig. 5.

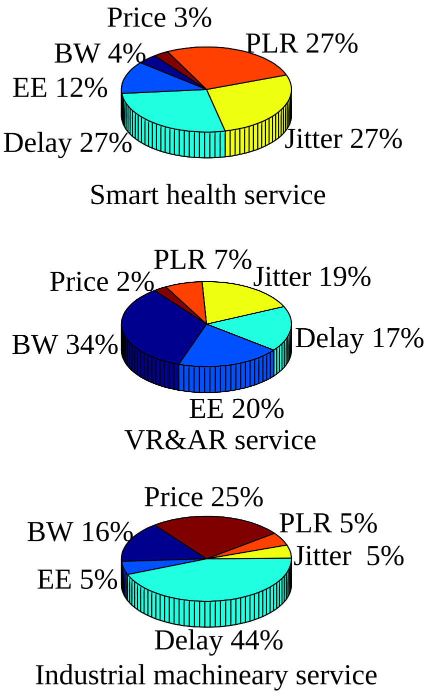

Fig. 5. Service preferences for different network attributes

As can be seen from the 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

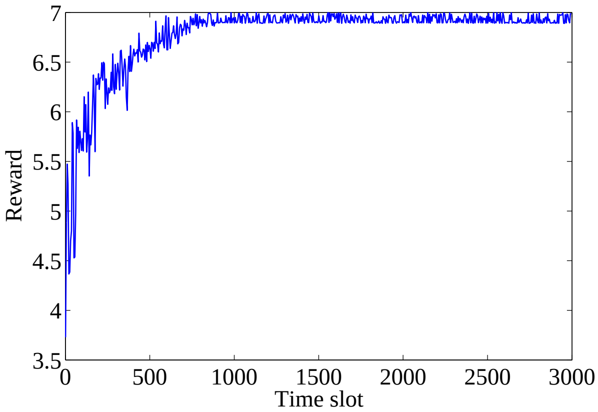

Fig. 6. Convergence of MAQNS

Convergence is an important premise for the effectiveness of Q-learning algorithm, and the accumulated reward of 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 second, while a time slot size is equal to 0.1 second and the discount factor α is computed to be 0.99. In addition, based on the parameters in Table 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 is shown in the 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 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.

D. Optimal evaluation of discount factor

In order to study the influence of discount factor on system performance, under different discount factors, we study the changes of total throughput and user blocking probability with user arrival rate respectively, and other parameters are set to be the same. It can be noticed that, in Fig. 7 and Fig. 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 system.

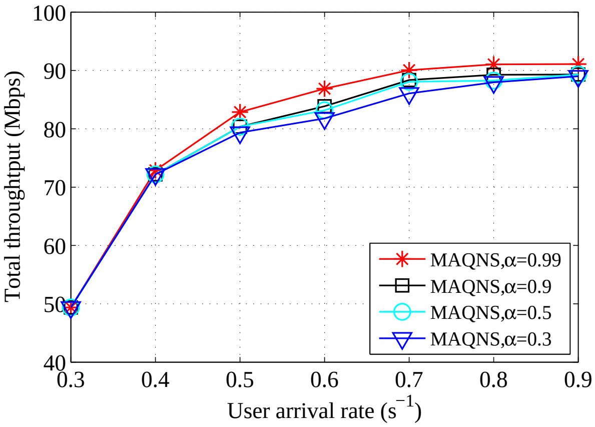

Fig. 7. Impact of discount factor on the throughput

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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3073027, IEEE Internet of Things Journal

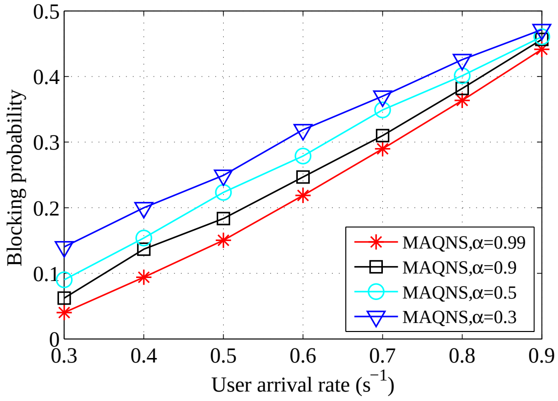

Fig. 8. Impact of discount factor on the blocking probability

Figs. 9 (a) - (c) respectively describe 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 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 the 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.

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.

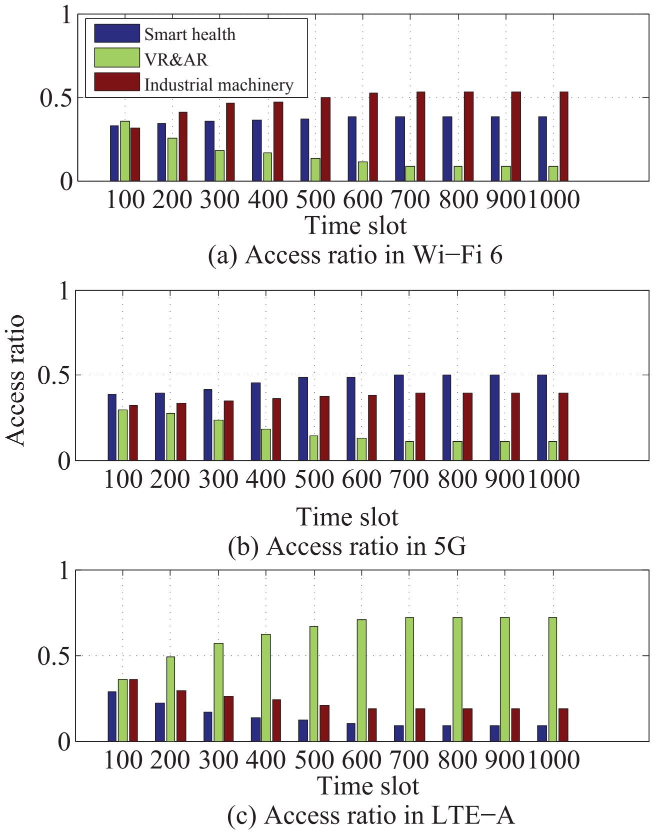

Fig. 9. User access ratio in different networks

F. Throughput Comparison

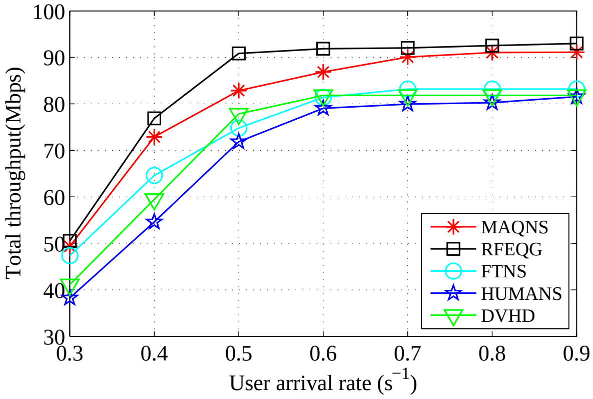

Fig. 10. Total throughput of the 5G system

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 rate 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.

As is shown in Fig. 10, the total throughput of the 5G system increases with the user arrival rate. The main resaon 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 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 RFEQG algorithm. Thus in RFEQG, the network priority of user access is 5G, Wi-Fi 6, 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, 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 other three algorithms by 9.51%.

G. Blocking Probability Comparison

As is represented in Fig. 11, the user blocking probability increases versus user arrival rate. When the network capacity can not 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 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 to maximize 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 multi-attribute decision-making and access the network with best comprehensive performance. However, when users adopt the three algorithm 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 MAQNS algorithm is reduced by at least 14.82%.

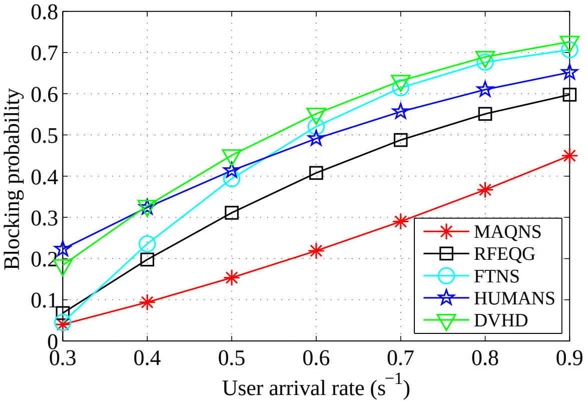

Fig. 11. Blocking probability of five algorithms

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 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 MAQNS algorithm can be improved by at least 6.65%.

I. Comparison on Delay

In this subsection, we would like to take delay as the representative of negative network attributes to evaluate the performance of five algorithms. As is 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's 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 lowest delay for the network selection. This is because MAQNS not only effectively considers service requirements of current

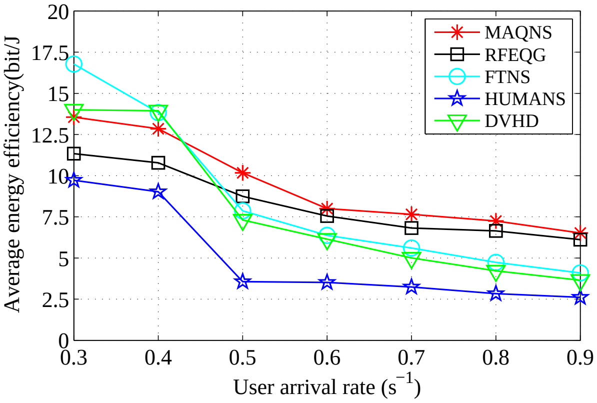

Fig. 12. Average energy efficiency experienced by users

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 subsequent users, resulting in the dramatic increase of the average delay. 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 MAQNS algorithm can be reduced by at 19.09% as the user arrival rate keeps to increase.

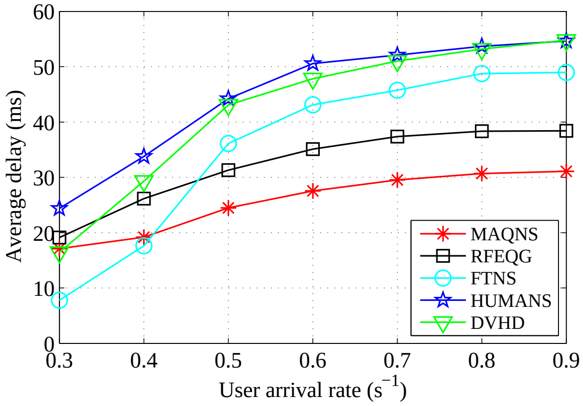

Fig. 13. Average delay experienced by users

IX. Conclusion

In this paper, under the scenario of 5G heterogeneous network system, we formulate 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 build a multi-agent structure network selection model based on the non-cooperative stochastic game, and propose an intelligent network selection algorithm named MAQNS by adapting Nash Q-learning into the multi-agent structure. Particularly, in the MAQNS, each agent update 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 is applied by MAQNS, which can be utilised 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 measure the preference weight of each services on heterogeneous networks by jointly performing AHP and GRA algorithms. Numerical results illustrate that the proposed MAQNS can not 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.

X. Acknowledgement

This work was supported partly by National Natural Science Foundation of China (No. 61772432, 61772433), the Fundamental Research Funds for the Central Universities (2019CDYGZD004), National Natural Science Foundation of China (61722105), Preresearch Fund (No. 6140449XX61001).

References

[1] L. Chettri and R. Bera, “A comprehensive survey on internet of things (iot) toward 5G wireless systems,” *IEEE Internet of Things Journal*, 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 Systems Journal*, vol. 15, no. 1, pp. 546–559, 2021.

[3] Q. Fan and N. Ansari, “Application aware workload allocation for edge computing-based iot,” *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2146–2153, 2018.

[4] GSMA, “The mobile economy,” 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 Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 426–471, 2020.

[6] A. Zhu, S. Guo, and M. Ma, “i5gaccess: Nash q-learning based multiservice edge users access in 5g heterogeneous networks,” in *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS)*, 2020, pp. 1–10.

[7] A. Roy, P. Chaporkar, and A. Karandikar, “Optimal radio access technology selection algorithm for lte-wifi network,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6446–6460, 2018.

[8] S. Ahmed and M. O. Farooq, “Analysis of access network selection ane switching metrics for lte and wifi hetnets,” in *2017 International Conference on Selected Topics in Mobile and Wireless Networking (MoWNeT)*, 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. 33 275–33 284, Jun. 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*, Jan. 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. 46 240–46 260, 2019.

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

[13] G. Araniti, P. Scopelliti, G. Muntean, and A. Iera, “A hybrid unicast-multicast network selection for video deliveries in dense heterogeneous network environments,” *IEEE Transactions on Broadcasting*, 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 *2011 IEEE Vehicular Technology Conference (VTC Fall)*, 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 P802.11-REVmd/D3.0, October 2019*, pp. 1–4647, 2019.

[16] M. Gerasimenko, N. Himayat, S. Yeh, S. Talwar, S. Andreev, and Y. Koucheryavy, “Characterizing performance of load-aware network selection in multi-radio (wifi/lte) heterogeneous networks,” in *2013 IEEE Globecom Workshops (GC Wkshps)*, 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 *2016 IEEE International Conference on Communications (ICC)*, 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 *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Nov. 2019, pp. 715–721.

[19] N. Zarin and A. Agarwal, “A hybrid network selection scheme for heterogeneous wireless access network,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 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 *2006 IEEE International Conference on Communications*, vol. 12, 2006.

[21] W. Song, H. Jiang, and W. Zhuang, “Performance analysis of the wlan-first scheme in cellular/wlan interworking,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1932–1952, 2007.

[22] A. Roy, P. Chaporkar, and A. Karandikar, “Optimal radio access technology selection algorithm for lte-wifi network,” *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6446–6460, 2018.

[23] K. Shin, G. Hwang, and O. Jo, “Distributed reinforcement learning scheme for environmentally adaptive IoT network selection,” *Electronics Letters*, vol. 56, no. 9, pp. 462–464, Apr. 2020.

[24] 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 of Things Journal*, vol. 6, no. 4, pp. 6862–6875, 2019.

[25] S. Pal, S. K. Das, and M. Chatterjee, “User-satisfaction based differentiated services for wireless data networks,” in *IEEE International Conference on Communications, 2005. ICC 2005. 2005*, vol. 2, May. 2005, pp. 1174–1178 Vol. 2.

[26] A. O. Mufutau, F. P. Guiomar, M. A. Fernandes, A. Lorences-Riesgo, A. Oliveira, and P. P. Monteiro, “Demonstration of a hybrid optical fibercwireless 5G fronthaul coexisting with end-to-end 4G networks,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 12, no. 3, pp. 72–78, Jan 2020.

[27] 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 Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 428–435, Jun. 2020.

[28] 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 of Things Journal*, vol. 7, no. 8, pp. 6722–6747, 2020.

[29] S. K. Kharroub, K. Abualsaud, and M. Guizani, “Medical iot: A comprehensive survey of different encryption and security techniques,” in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 1891–1896.

[30] 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 Generation Computer Systems*, vol. 99, 04 2019.

[31] 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 Transactions on Multimedia*, vol. 22, no. 9, pp. 2366–2381, 2020.

[32] Z. Ge and Y. Liu, “Analytic hierarchy process based fuzzy decision fusion system for model prioritization and process monitoring application,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 357–365, Jan 2019.

[33] T. Saaty, “Decision making with the analytic hierarchy process,” *Int. J. Services Sciences*, vol. 1, pp. 83–98, Jan. 2008.

[34] M. Brunelli, *Introduction to the analytic hierarchy process*. Springer International Publishing, 2015, p. 83.

[35] T. Saaty, “A scaling method for priorities in hierarchical structures,” *Journal of Mathematical Psychology*, vol. 15, no. 3, pp. 234–281, Jun. 1977.

[36] Q. Xiao, M. Shan, M. Gao, and X. Xiao, “Grey information coverage interaction relational decision making and its application,” *Journal of Systems Engineering and Electronics*, vol. 31, no. 2, pp. 359–369, Apr. 2020.

[37] L. Xu, H. Ma, and D. Ren, “Reliability analysis of tractor multi-way valves based on the improved weighted grey relational method,” *The Journal of Engineering*, vol. 2019, no. 13, pp. 86–92, May. 2019.

[38] M. Han, R. Zhang, T. Qiu, M. Xu, and W. Ren, “Multivariate chaotic time series prediction based on improved grey relational analysis,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 10, pp. 2144–2154, Oct. 2019.

[39] 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 Transactions on Vehicular Technology*, Sep. 2020.

[40] D. Xie, J. Zhang, A. Tang, and X. Wang, “Multi-dimensional busy-tone arbitration for ofdma random access in ieee 802.11ax,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4080–4094, Mar. 2020.

[41] K. Lee, “Using ofdma for mu-mimo user selection in 802.11ax-based wi-fi networks,” *IEEE Access*, vol. 7, pp. 186 041–186 055, Dec. 2019.

[42] S. Althunibat, R. Mesleh, and K. A. Qaraqe, “Im-ofdma: A novel spectral efficient uplink multiple access based on index modulation,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 10, pp. 10 315–10 319, Aug. 2019.

[43] N. Le, L. Tran, Q. Vu, and D. Jayalath, “Energy-efficient resource allocation for ofdma heterogeneous networks,” *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 7043–7057, Aug. 2019.

[44] C. Baquero Barneto, T. Riihonen, M. Turunen, L. Anttila, M. Fleischer, K. Stadius, J. Ryynanen, and M. Valkama, “Full-duplex ofdm radar with lte and 5g nr waveforms: Challenges, solutions, and measurements,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 10, pp. 4042–4054, Oct. 2019.

[45] 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 Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1331–1342, Dec. 2018.

[46] 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 Transactions on Communications*, vol. 68, no. 1, pp. 506–520, Oct. 2020.