Joint Computation Offloading and Resource Allocation Optimization in Heterogeneous Networks with Mobile Edge Computing

Jing Zhang, Weiwei Xia, Feng Yan, Lianfeng Shen
National Mobile Communications Research Laboratory, Southeast University
Nanjing, Jiangsu, 210096, China
Email: {jingzhang, wwxia, feng.yan, lfshen}@seu.edu.cn

***Abstract*—In this paper, we propose a distributed joint computation offloading and resource allocation optimization (JCORAO) scheme in heterogeneous networks (HetNets) with mobile edge computing (MEC). An optimization problem is formulated to provide the optimal computation offloading strategy policy, uplink subchannel allocation, uplink transmission power allocation and computation resource scheduling. The optimization problem is decomposed into two sub-problems due to the NP-hard property. In order to analyse the offloading strategy, a sub-algorithm named distributed potential game is built. The existence of Nash equilibrium (NE) is proved. To jointly allocate uplink subchannel, uplink transmission power and computation resource for the offloading MTs, a sub-algorithm named cloud and wireless resource allocation algorithm (CWRAA) is designed. The solutions for subchannel allocation consist of uniform zero frequency reuse (UZFR) method without interference and fractional frequency reuse method based on Hungarian and graph coloring (FFR-HGC) with interference. A distributed JCORAO scheme is proposed to solve the optimization problem by the mutual iteration of the two sub-algorithms. Simulation results show that the distributed JCORAO scheme can effectively decrease the energy consumption and task completion time with lower complexity.**

***Index Terms*—Mobile edge computing; heterogeneous networks; offloading strategy; resource allocation; game theory.**

I. Introduction

As the popularity of smart phones, laptops and tablets is increasing dramatically, more novel sophisticated applications are emerging, such as face recognition, interactive gaming and augmented reality [1]. However, running computationally demanding applications at the mobile terminals (MTs) is constrained by the limited battery power and scarce computing capabilities [2]. Suitable solution impeding the performance of service qualities of the MTs is to offload the complicated applications as the tasks to a cloud server [3]. Computation offloading has given rise to an exponential growth of demand for not only high data rate in wireless networks but also high computational capability in cloud server.

One recently proposed solution for tackling the data rate issue is the use of heterogeneous networks (HetNets). HetNets often indicate the use of multiple types of access nodes in a wireless network. Multiple small cells and the traditional macro cells constitute HetNets [4], which meet MTs' high-rate requirements. Small cells with small coverage area and low transmission power usually include microcells, picocells, femtocells and relays [5]. The previous signal processing and transmission techniques applied in the conventional cellular networks may not be efficient to meet MTs' requirements of high throughput. The deployment of low-cost small cells is a very significant way to improve spectrum and energy efficiency.

In addition, to solve the computational capability issue, mobile edge computing (MEC) system has been a typical paradigm that combines wireless network service and cloud computing to enable MTs to enjoy the abundant wireless resources and vast computation power ubiquitously [6]. MEC is an IT service environment and has cloud-computing capability located at the edge of the mobile networks, within the radio access networks and in close proximity to MTs [7]. MEC server is a data center typically collocated with a base station in a network cell, and accessible by nearby MTs via one-hop wireless connection [8]. MEC allows MTs to perform computation offloading by uploading their computational tasks to the MEC server via HetNets [9]. In terms of network topology, the computation resources of MEC are supposed to be in proximity of the MTs so as to decrease transmission delay. Besides, MTs can save energy consumption by trading off heavy computational load for lightweight communication [10].

In the previous researches, many works investigated the computation offloading and resource allocation strategies in the scenario of MEC [11]-[19]. The authors of [11]-[13] studied the computation offloading strategy. The works in [14]-[15] mainly laid emphasis on joint radio and cloud resource allocation algorithms. Some researches [16]-[19] focused on joint computation offloading and resource allocation. There were also many works studying the resource allocation algorithm in the HetNets [20]-[24]. The HetNets are confronted with many challenges due to the limited radio communications capabilities, such as interference management and wireless resource allocation. However, only the authors of [25]-[28] considered the heterogeneity of networks in the context of MEC. Nevertheless, the authors of [25] did not consider the offloading strategy. The wireless resource allocation was not involved in [26]. The work in [27] did not consider the

This work is supported in part by the National Natural Science Foundation of China (No.61741102, No.61471164, No.61601122). The corresponding author is Weiwei Xia, Email: wwxia@seu.edu.cn

impact factor of monetary cost that MTs paid for wireless and computation resources. The authors of [28] only concentrated on single MT in the coverage of small base station rather than multi-MTs.

Different from the previous works, this paper jointly optimizes the offloading strategy, subchannel allocation, uplink power allocation and CPU-cycle assignment in the HetNet with MEC. When solving the resource allocation problem, monetary cost is considered including wireless and computation resource. In addition, there is competition among numerous MTs over both constrained communication resources in HetNets and limited computation resources in the MEC server. This paper proposes a distributed joint computation offloading and resource allocation optimization (JCORAO) scheme in HetNets with MEC. The main contributions of this paper are listed as follows.

1) An optimization problem is formulated to provide the optimal computation offloading strategy policy, uplink subchannel allocation, uplink transmission power allocation and computation resource scheduling. The objective of the optimization problem is to minimize all MTs' cost while satisfying offloading latency constraints.
2) The optimization problem is decomposed into two subproblems due to the NP-hard property. On one hand, a sub-algorithm named distributed potential game is built to model and analyse the offloading strategy. The existence of Nash equilibrium (NE) is proved. On the other hand, to jointly allocate uplink subchannel, uplink transmission power and computation resource for the offloading MTs, a sub-algorithm named cloud and wireless resource allocation algorithm (CWRAA) is designed. A distributed JCORAO scheme is proposed to solve the optimization problem by the mutual iteration of the two sub-algorithms. In the CWRAA, interference management is taken into consideration for uplink subchannel allocation. CWRAA focuses on two situations. One is the subchannel allocation using uniform zero frequency reuse (UZFR) method where no interference exists among MTs. Another is the subchannel allocation using fractional frequency reuse based on Hungarian method and graph coloring (FFR-HGC) method that pays attention to interference migration among MTs.
3) Simulation results show that the distributed JCORAO scheme outperforms other algorithms by making tradeoff between the total cost and algorithm complexity. In addition, the distributed JCORAO scheme can effectively decrease the energy consumption and task completion time. Furthermore, FFR-HCG method is an effective way to mitigate the interference among neighboring MTs.

The rest of the paper is organized as follows. Section II introduces some related work. In Section III, system model and optimization problem are presented. Section IV introduces the distributed JCORAO scheme. In Section V, the simulation results are shown. Finally, conclusion is given in Section VI.

II. Related Work

A. Offloading strategy in MEC environment

Computation offloading and resource allocation for MEC systems have attracted significant attention in recent years. Some previous researches investigated the computation offloading mechanism design. Chen et al. in [11] formulated the computation offloading strategy making problem among multiple MTs for MEC as a distributed game. The authors of [12] established a socially aware computation offloading game considering the social tie structure among mobile users. Zhang et al. in [13] utilized auction theory to model the matching relationship between MEC server and MTs so as to offload tasks to the optimal MEC server. Works on resource allocation have also acquired some achievements. The study of [14] jointly allocated communication and computation resources to minimize the total MTs energy consumption under latency constraints by successive convex approximation. The authors of [15] concentrated on how to tackle the allocation of the communication and computational resources among the MTs to achieve low latency.

B. Joint offloading and resource allocation in MEC environment

There are many excellent works on offloading strategy and resource allocation respectively. There are also some literatures jointly considering offloading strategy and resource allocation. The author of [16] jointly decided the offloading strategy, the CPU-cycle frequencies for mobile execution, and the transmit power for computation offloading. However, energy consumption was not involved in [16] since the energy used by MTs was assumed to be renewable resources. Wang et al. in [17] studied offloading strategy, subcarrier allocation for task offloading and CPU time allocation for task execution in the MEC server. The work in [18] jointly optimized the offloading selection, radio resource allocation, and computational resource allocation coordinately to make the energy consumption minimum. In [19], a power consumption minimization problem with task buffer stability constraints was formulated and an online computation offloading algorithm was studied based on Lyapunov optimization.

C. Resource allocation in HetNets

Many works concentrate on the resource allocation in HetNets. The authors in [20] studied the tradeoff between energy efficiency and spectral efficiency in multicell HetNets. User association and power allocation in mmWave-based ultra dense networks were modeled as a mixed-integer programming problem in [21]. The work in [22] used the Lyapunov optimization method to explore the dynamic subchannel and power allocation in spectrum sharing heterogeneous small cell networks. A heuristic, joint QoE-aware resource allocation and dynamic pricing algorithm was proposed to maximize the mobile network operators profit while providing high users QoE in [23]. The work in [24] investigated interference management and power allocation problem in two-tier HetNets with massive MIMO by appropriate approximation.

D. Offloading strategy and resource allocation in HetNets with MEC

All researches above did not combine the HetNets and MEC. However, there have been some research works considering the scenario of HetNets with MEC. The authors of [25] jointly allocated the transmit precoding matrices of the MTs and the CPU cycles of MEC server to minimize the overall MTs energy consumption, while meeting latency constraints based on a novel successive convex approximation technique, but this paper did not consider the offloading strategy. The work in [26] jointly optimized the computation offloading and content caching strategy considering the total revenue of the network. However, the wireless resource allocation was not involved in it. In [27], the authors jointly optimized offloading and radio resource allocation to minimize energy consumption under the latency constraints, but they did not consider the monetary cost that MTs paid for wireless and computation resources. The study in [28] took the computation offloading, physical resource block and MEC computation resource allocation into consideration. However, it only concentrated on single MT in the coverage of small base station rather than multi-MTs.

III. System Model and JCORAO Problem Formulation

In this section, system model including network model, communication model and computation model are described firstly, then the optimization problem is formulated.

A. Network Model

In HetNets, each MT has complicated tasks to be dealt with and needs to decide local computing or cloud computing. Local computing will occupy MTs' local computation resources and consume large quantities energy. In addition, the task completion delay may be very high due to the limited computation capabilities. To cope with these problems, edge cloud computing allows MTs to offload their computational tasks to the MEC server via HetNets. Then each MT is associated with a clone in MEC server, which executes the compute-intensive tasks on behalf of that MT. Computation offloading may save energy consumption and time delay. As shown in Fig. 1, an example of MEC system includes MEC server and HetNets. The MEC server can be a small data center deployed on the edge of HetNets by telecom operators. It connects to the macro base station (MBS) and provides computation resources (e.g. CPU cycles per second) for MTs by the HetNets. It can serve for surrounding MTs to extend their computation capability and can deal with tasks parallelly. In a particular cell of a two-tier HetNets, J small base stations (SBSs) and one MBS provide communication resources (e.g. subchannels) to K MTs. The set of MBS and SBSs is denoted by $\mathcal{J} = \{0, 1, 2, ..., J\}$ in which 0 represents the MBS and $\{1, 2, ..., J\}$ denote the SBSs in a cell. Let the set of MTs served by BS j denote as $\mathcal{V}_j$ ($j \in \mathcal{J}$) and the set of all MTs as $\mathcal{K} = \{1, 2, ..., K\}$. The total number of MTs is K. Furthermore, the set of offloading MTs is denoted by $\mathcal{K}^c$, and the set of MTs for local computing is denoted by $\mathcal{K}^l$. Besides,

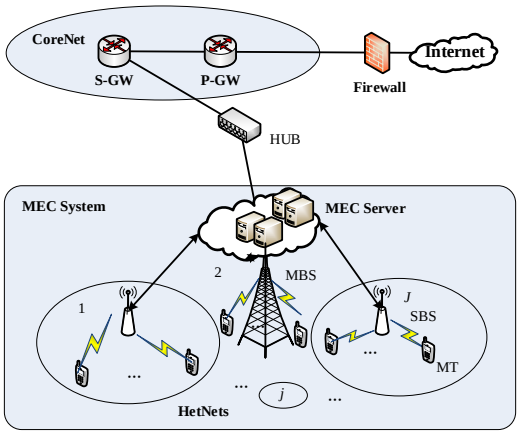


Fig. 1. An example of heterogeneous networks with mobile edge computing.

$|\mathcal{K}^c| = K^c = \sum_{k=1}^{K} a_k$ and $|\mathcal{K}^l| = K^l = K - K^c$. There are N available orthogonal OFDM subchannels that can be assigned for uplink communication in a cell of the HetNets. Let $\mathcal{N} = \{1, 2, ..., N\}$ denote the set of subchannels. MBS and SBSs can reuse subchannels in set $\mathcal{N}$. The bandwidth of each subchannel is w_0. The MTs subscribed to one MBS (SBS) are allocated orthogonal OFDM subchannel while the MTs subscribed to different BSs can share the same subchannels. Therefore, there exists intra-cell interference [27] among MTs. Moreover, for simplicity, we only consider MTs and BSs with single-antenna in this paper.

It is assumed that each MT has computationally intensive and delay sensitive tasks to be completed at present moment. Typical tasks offloading from MTs usually include two aspects: CPU cycles to be used to execute the tasks and the amount of data to be transmitted to MEC server. Each MT could offload the tasks to the MEC server through the BS with which it is associated, or execute the computation tasks locally. For MT k ($k \in \mathcal{K}$), the tasks are characterized by s_k the number of instructions to be executed and by b_k the size of input data necessary to be transferred. The tasks of MT k are supposed to be completed within $\widetilde{T}_k$ which is the task completion time threshold that does not affect MTs' experience. The offloading strategy set of MTs is defined as $\mathcal{A} = \{a_1, a_2, ..., a_k, ..., a_K\}$, $k \in \mathcal{K}\}$. $a_k = 1$ implies that MT k offloads its tasks to MEC server. $a_k = 0$ indicates that MT k executes its tasks locally. For MEC, the server can deal with tasks from all MTs due to multi-tasking capability. MEC server is capable of handling f_S instructions per unit time and the tasks of MT k are allocated the number of f_k instructions per unit time under the constraint of $\sum_k f_k \leq f_S$. Similar to previous work in MEC [23], a quasi-static scenario is considered where the set of MTs $\mathcal{K}$ remains unchanged during a computation offloading period.

The notations mainly used in this paper are summarized in Table I.

B. Communication Model

Each MT needs to obtain full channel state information (CSI) of all uplink subchannels. The signal-to-interference-plus-noise ratio (SINR) for MT k in BS j using subchannel

TABLE I
PARAMETER NOTATIONS

Symbol	Definition
J	The number of BSs.
K	The number of MTs.
N	The number of subchannels.
$\mathcal{J}$	The set of base stations, $\mathcal{J}=\{0,1,...,J\}$.
$\mathcal{N}$	The set of subchannels, $\mathcal{N}=\{1,...,N\}$.
$\mathcal{A}$	The offloading strategy set of MTs.
$\mathcal{C}$	Subchannel association table, $K \times N$.
$\mathcal{K}^c$	The MTs set of offloading.
$\mathcal{K}^l$	The MTs set of local computation.
j	The base station index $j \in \mathcal{J}$.
k	The MT index $k \in \mathcal{K}$.
n	The subchannel index $n \in \mathcal{N}$.
V_j	The set of MTs subscribing to BS j.
w_0	The bandwidth of one subchannel.
σ^2	The power of the additive white Gaussian noise.
I''_{kn}	The interference coming from adjacent cells.
I'_{kn}	The intra-cell interference from other BSs to MT k of BS j.
p_{kn}	The transmission power of MT k allocated to subchannel n.
h_{kn}	The channel gain of MT k in subchannel n.
R_{kn}	The transmission rate of MT k in subchannel n.
r_k	The transmission rate of MT k.
b_k	The input data size of MT k.
s_k	The tasks load of MT k.
f_k	The CPU cycles allocated to MT k by MEC server.
F_S	The total computational resources of MEC server.
f_k^l	The local computational capability of MT k.
β	Transmission rate price coefficient of BSs.
q	Computation resources price coefficient of MEC server.
γ_k^T	The weight of local execution time cost for MT k.
γ_k^E	The weight of energy consumption in offloading for MT k.
γ_k^M	The weight of monetary cost in offloading for MT k.

n can be expressed as

$$\text{SINR}_{kn}^{(j)} = \frac{p_{kn} h_{kn}^{(j)}}{\sigma_{kn}^2 + I''_{kn} + I'^{(j)}_{kn}} \tag{1}$$

where σ_{kn}^2 is defined as the power of the additive white Gaussian noise at subchannel n ($n \in \mathcal{N}$), p_{kn} is the transmission power of MT k at subchannel n, h_{kn}^j is the channel gain between MT k and BS j at subchannel n. The interference coming from MBSs and SBSs in adjacent cells is denoted by I''_{kn}. For sake of simplicity, we regard I''_{kn} as constants. The intra-cell interference from other BSs to MT k of BS j in current cell is denoted by $I'^{(j)}_{kn}$. In particular, we define $I'^{(j)}_{kn} = \sum_{j' \neq j} \sum_{k' \in V_{j'}} p_{k'n} h_{kn}^{(k')}$ where $h_{kn}^{(k')}$ represents the channel gain between MT k in BS j and MT k' in BS j' on subchannel n. In heterogeneous cellular cell, we introduce a subchannel association table $\mathcal{C}$, which is an $K^c \times N$ matrix with binary variable c_{kn}. The binary variable means whether subchannel n is assigned to the uplink communication of MT k. c_{kn}=1 represents that subchannel n is assigned to the uplink of MT k and c_{kn}=0 otherwise. The throughput of the uplink communication for MT k in BS j can be given by $R_{kn}^{(j)} = W_0 \log(1 + c_{kn} \cdot \text{SINR}_{kn}^{(j)})$. In the subsequent context, the superscript j is omitted in $\text{SINR}_{kn}^{(j)}, h_{kn}^{(j)}$ and $R_{kn}^{(j)}$ when MT k is attributed to BS j.

The uplink transmission rate r_k of MT k is given as

$$r_k = \sum_{n \in \mathcal{N}} R_{kn} \tag{2}$$

For subchannel allocation, we utilize two categories of solutions according to the number of MTs and subchannels:

1) If $N \geq K^c$, UZFR method is applied. There is no interference among MTs in this approach since sufficient orthogonal subchannels are available. Each MT is assigned with an equal number of orthogonal subchannels expressed as $n_k = \lfloor \frac{N}{K^c} \rfloor$.

2) If $N < K^c$, FFR-HGC method is used. In the scenario of frequency reuse for uplink channels, the interference among MTs is inevitable. For mitigating the interference, we take advantage of Hungarian method initially put forward in [29] and graph coloring originally proposed in [30] synthetically to complete the fractional frequency reuse (FFR). For reducing the complexity, we assume one MT can only use one subchannel in FFR-HGC. The detailed description is shown in Section IV.

C. Computation Model

The offloading latency T_k^c consists of four parts [31], the uplink communication delay Δ^{ul}, backhaul link delay Δ^{bh}, downlink delay Δ^{dl} and cloud task processing delay Δ^{exe}. The backhaul link rate between BS and MEC server is much higher than wireless link so that we can neglect Δ^{bh}. Compared with the size of input bits b_k, the size of output bits from MEC server is less, so the downlink delay Δ^{dl} is regarded as a constant ε. We use $T_k = \tilde{T}_k - \varepsilon$ to represent the delay that comprises the uplink communication delay Δ^{ul} and cloud task processing delay Δ^{exe} when MT decides to offload its tasks.

1) Edge cloud computing: The energy consumption of MT k including uplink and downlink energy consumption is given by

$$E_k^c = p_k \Delta_k^{ul} + p_k^r \Delta_k^{dl} \tag{3}$$

where $p_k = \sum_{n \in \mathcal{N}} c_{kn} p_{kn}$. p_k^r denotes the received power of MT. If one MT is assigned multi-subchannel, the transmit power p_k is averagely distributed with $p_{kn} = \frac{p_k}{\sum_{n \in \mathcal{N}} c_{kn}}$.

Monetary cost of MTs can be expressed as

$$M_k^c = \beta r_k + q f_k \tag{4}$$

The first item is the communication cost and the second item is the computation cost. For BS, the unit price of transmission rate is β. For MEC server, the unit price of computation resources is q.

The offloading latency of MT k by MEC server computing is defined as

$$T_k^c = \Delta^{ul} + \Delta^{dl} + \Delta^{bh} + \Delta^{exe} \tag{5}$$

Δ^{ul} is given as $\Delta^{ul} = b_k / r_k$ and Δ^{exe} is defined as $\Delta^{exe} = s_k / f_k$. According to (3) and (4), the overhead of the edge cloud computing approach in terms of energy consumption and monetary cost can be computed as

$$z_k^c = \gamma_k^E E_k^c + \gamma_k^M M_k^c \tag{6}$$

where $\gamma_k^E \in R^+$ means the impact factor of energy consumption on the overhead of MT k and keeps the energy

consumption as the same order of magnitude. $\gamma_k^M \in R^+$ is defined as the impact factor of monetary cost. It should be noticed that $f_k = 0$ if the tasks are executed locally.

2) Local Computing: Let f_k^l denote the computation capacity of MT k. Different MTs have different computation capacity. According to [32], the energy consumption is given by

$$E_k^l = \kappa s_k (f_k^l)^2 \tag{7}$$

where κ is the effective switched capacitance relying on the chip architecture [24].

The local execution latency of MT k by local computing is denoted as

$$T_k^l = \frac{s_k}{f_k^l} \tag{8}$$

According to (7) and (8), the overhead of the local computing approach in terms of energy consumption and local execution time cost can be computed as

$$z_k^l = \gamma_k^E E_k^l + \gamma_k^T (T_k^l - \widetilde{T}_k) \tag{9}$$

where $T_k^l - \widetilde{T}_k$ denotes the local execution time cost and γ_k^T represents the impact factor of the local execution time cost. If $T_k^l > \tilde{T}_k$, the second term makes the overhead of the local computing increase, vice versa.

Task computation time is equal to T_k^l if MT k decide local computing. Otherwise, task computation time is equal to offloading latency T_k^c.

D. JCORAO Problem Formulation

The MEC server makes the offloading strategy for MT k on comparison of its local and offloading computation overhead, i.e., comparison of

$$\begin{cases} a_k = 1, z_k^l > z_k^c \\ a_k = 0, z_k^l \leq z_k^c \end{cases} \tag{10}$$

The cost for MT k can be computed as

$$z_k = (1 - a_k) z_k^l + a_k z_k^c \tag{11}$$

The aim of JCORAO is to provide the optimal computation offloading strategy policy $\mathcal{A}^*$, uplink subchannel allocation $\mathcal{C}^*$, uplink transmission power allocation $\mathbf{P}^*$ and computation resource scheduling $\mathbf{F}^*$ for all MTs such that the total cost is minimized. Therefore, the optimization problem can be formulated as

$$\min_{\mathcal{A},\mathcal{C},\mathbf{P},\mathbf{F}} Z(\mathcal{A},\mathcal{C},\mathbf{P},\mathbf{F}) = \sum_{i=1}^{K} (1 - a_k) z_k^l + a_k z_k^c \tag{12}$$

$$\begin{aligned} s.t. C1 &: T_k^c \leq \widetilde{T}_k, \forall k \\ C2 &: \sum_k f_k \leq f_S \\ C3 &: f_k \geq 0, \forall i \\ C4 &: 0 \leq p_k \leq {p_k}^T \\ C5 &: a_k \in \{0,1\}, \forall k \in \mathcal{K} \\ C6 &: c_{kn} \in \{0,1\}, \forall n \in \mathcal{N}, k \in \mathcal{K} \\ C7 &: \sum_k c_{kn} \in \{0,1\}, \forall k \in \mathcal{V}_j \end{aligned}$$

where $\mathcal{A} = (a_1, a_2, ..., a_K)$, $\mathbf{P} = \{p_k | 0 \leq p_k \leq {p_k}^T, k \in \mathcal{K}\}$ and $\mathbf{F} = \{f_k | 0 \leq f_k, \sum_k f_k \leq f_S, k \in \mathcal{K}\}$. C1 is the offloading latency constraint that does not affect MTs' experience. The maximum processing capability constraint of MEC server is indicated by Constraint C2. Constraints C3 means the non-negativity of computation resources. Constraint C4 manifests the change range of uplink transmission power. Constraint C7 ensures that one subchannel in the same BS can be used by only one MT or no use.

The key challenge in (12) is that the integer constraint from the above optimization objective. $a_k \in \{0,1\}$ and $c_{kn} \in \{0,1\}$ make (12) become a mixed integer programming problem. Problem (12) is non-convex and NP-hard, thus it is extremely urgent to design an efficient and simplified mechanism. Next, the distributed JCORAO scheme is proposed to allow the MTs to determine the offloading strategy $\mathcal{A}$, the subchannel selection $\mathcal{C}$, power control $\mathbf{P}$ and computation resource requirements $\mathbf{F}$ by themselves.

IV. The Distributed JCORAO Scheme

In this section, the distributed JCORAO scheme is proposed to solve the optimization problem. The scheme consists of two sub-algorithms. One is the distributed potential game. Another is the CWRAA. Driven by the finite improvement property (FIP) [11] and the existence of NE of potential game, offloading strategy $\mathcal{A}$ is formulated as a distributed potential game. When tasks of MTs are offloaded to the MEC server, the CWRAA is designed to acquire the subchannel selection $\mathcal{C}$, power control $\mathbf{P}$ and computation resource requirements $\mathbf{F}$ for these MTs. A distributed JCORAO scheme solves the optimization problem by the mutual iteration of the two sub-algorithms.

A. Game Formulation and CWRAA

$\mathbf{a}_{-k} = \{a_1, ..., a_{k-1}, a_{k+1}, ..., a_K\}$ is denote as the computation offloading strategy profile by all other MTs except MT k. Given strategy profile $\mathbf{a}_{-k}$, MT k would like to select a proper decision a_k, by using either the local computing ($a_k = 0$) or the edge cloud computing ($a_k = 1$) to minimize its own computation overhead in the competitive environment. Mathematically, the distributed computation offloading strategy making problem is formulated as:

$$\min_{a_k \in \{0,1\}} u_k(a_k, \mathbf{a}_{-k}) = (1 - a_k) z_k^l + a_k z_k^c, \forall k \in \mathcal{K} \tag{13}$$

According to (6) and (9), we can obtain the overhead function of MTs as

$$u_k(a_k, \mathbf{a}_{-k}) = \begin{cases} z_k^c, a_k = 1 \\ z_k^l, a_k = 0 \end{cases} \tag{14}$$

We then formulate the distributed computation offloading strategy making problem as a distributed potential game $G = \{\mathcal{K}, (a_k)_{k \in \mathcal{K}}, (u_k)_{k \in \mathcal{K}}\}$ which is described as follows:

Players. Each MT is one player and there are K participants selecting local computing or edge cloud computing.

Strategies. The offloading strategy $a_k \in \{0,1\}$ is the strategy for MT k. $\mathcal{A}$ is the offloading strategy profile for all MTs .

Cost function. The overhead function $u_k(a_k, \mathbf{a}_{-k})$ in (14) is denoted as the cost function for MT k. The cost function for offloading MT k is z^c. If MT chooses local computing, the cost function will be z^l.

The solution for the game model is NE, the definition is denoted as:

Definition 1. *A strategy profile $\mathcal{A}^* = (a_1^*, a_2^*, ..., a_K^*)$ is a NE of the distributed potential game model. At the equilibrium $\mathcal{A}^*$, no player can further reduce its cost by unilaterally altering its strategy, i.e.,*

$$u_k(a_k^*, \mathbf{a}_{-k}^*) \leq u_k(a_k, \mathbf{a}_{-k}^*), \forall a_k \in \{0,1\}, k \in \mathcal{K}$$

The NE has significant self-stability property such that the MTs at the equilibrium can derive a mutually satisfactory solution and no MT has the incentive to deviate. This property is very important to the non-cooperative computation offloading problem, since the MTs are selfish to act in their own interests.

From the objective function (12), we can see that the offloading strategies $\mathcal{A}$ are associated with $\mathcal{C}, \mathbf{P}, \mathbf{F}$. The solving process of these variables requires mutual iteration.

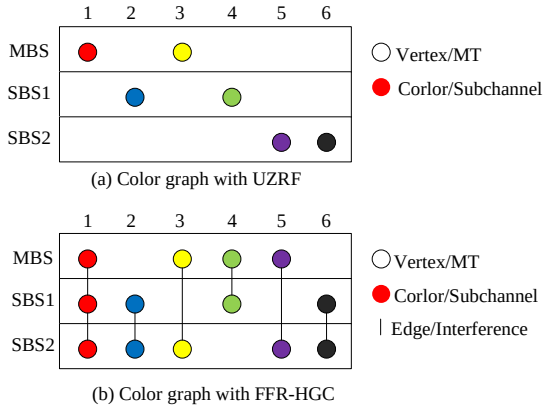
In the potential game, initially, offloading strategy profile $\mathcal{A}$ is set as $\mathcal{A}_0$ of which the elements are all 1 representing all MTs choose offloading tasks to MEC server. Given the strategies $\mathcal{A}$ of all MTs, the CWRAA is proposed to allocate the cloud and wireless resources for the MTs that prepare to offload tasks to MEC server. Given the resource allocation, the offloading strategy $\mathcal{A}$ is updated by potential game until achieving NE. The purpose of the CWRAA is to minimize the total cost of all offloading MTs. According to (3), (4), (5) and (6), the objective function of the CWRAA is defined as:

$$\begin{aligned}\min_{\mathcal{C},\mathbf{P},\mathbf{F}} Z^c(\mathcal{C},\mathbf{P},\mathbf{F}) &= \sum_{k=1}^{K} a_k z_k^c \qquad (15)\\ &= \sum_{k=1}^{K} a_k \Big(\gamma_k^E \frac{b_k \sum_{n=1}^{N} c_{kn} p_{kn}}{\sum_{n=1}^{N} c_{kn} w_0 \log_2(1+\alpha_{kn} p_{kn})} \\ &+ \gamma_k^M (\beta \sum_{n=1}^{N} c_{kn} w_0 \log_2(1+\alpha_{kn} p_{kn}) + q f_k)\Big)\end{aligned}$$

subject to constrains C1-C7 except C5. The first term means energy consumption while the second term represents monetary cost. $\alpha_{kn} = \text{SINR}_{kn}/p_{kn}$ represents the channel parameter.

For subchannel allocation $\mathcal{C}$, two categories of solutions are utilized according to the number of offloading MTs and subchannels: 1) If $N \geq K^c$, UZFR method is applied without considering interference among MTs. 2) If $N < K^c$, FFR-HGC method is used by considering interference among MTs.

Before introducing the CWRAA without interference and with interference, the color graph is described. As Fig. 2 shows, one color represents one subchannel and one vertex refers to a MT. The vertexes subscribed to the same BS are supposed to be assigned different color. Thus, the maximum capability of one BS for MTs is N. For example, there is one MBS, two SBSs and six subchannels in Fig. 2. Fig. 2 (a) describes the color graph with UZRF and there are six offloading MTs in total. The number of offloading MTs is equal to the number of subchannels and one MT is allocated one orthogonal subchannel. Thus there is no interference among MTs. The case of FFR-HGC is shown as Fig. 2 (b). The number of offloading MTs, $4+4+5=13$, is more than the number of subchannels so that subchannels must be reused and there exists interference among MTs. The edge between two MTs denotes the interference intensity in Fig. 2 (b). With the color graph described above, the subchannel assignment problem is formulated as a graph coloring problem.



(a) Color graph with UZRF

(b) Color graph with FFR-HGC

Fig. 2. Color graph.

B. CWRAA without Interference

Due to the identical channel gain among N subcahnnels, the uplink transmission rate r_k of MT k for UZRF method can be transformed as according to (2):

$$r_k = w_0 n_k \log_2(1+\alpha_{kn} p_{kn}) \qquad (16)$$

where $n_k = \lfloor \frac{N}{K^c} \rfloor$ denotes the number of subchannels of MT k assigned by BS and p_{kn} represents transmission power of one subchannel scheduled by MT. The total transmission power of MT k is defined as $p_k = n_k p_{kn}$. $\lfloor x \rfloor$ represents that fractions are rounded down. $\alpha_{kn} = h_{kn}/\sigma_{kn}^2$ is the channel parameter for MT k on subchannel n. It should be noticed that $p_k = 0$ if the tasks are executed locally. According to (16), (15) can be transformed as:

$$\begin{aligned}\min_{\mathbf{P},\mathbf{F}} Z^c(\mathbf{P},\mathbf{F}) = \sum_{i=1}^{K} a_k \Big(&\gamma_k^E \frac{b_k p_{kn}}{w_0 \log_2(1+\alpha_{kn} p_{kn})} \qquad (17)\\ &+ \gamma_k^M (\beta w_0 n_k \log_2(1+\alpha_{kn} p_{kn}) + q f_k)\Big)\end{aligned}$$

The partial derivative of (17) is shown as below:

$$\begin{aligned}\frac{\partial Z^c}{\partial p_{kn}} = &\frac{b_k \gamma_k^E}{w_0 n_k \log_2(1+\alpha_{kn} p_{kn})} \Big(1 - \frac{1}{\log_2^2(1+\alpha_{kn})} \\ &\cdot \frac{\alpha_{kn} p_{kn}}{(1+\alpha_{kn} p_{kn}) \ln 2}\Big) + \gamma_k^M \beta w_0 n_k \frac{\alpha_{kn}}{1+\alpha_{kn} p_{kn} \ln 2} \qquad (18)\end{aligned}$$

where $\alpha_{kn} p_{kn}$ is denoted as SINR and the value is larger than 1, thus $\frac{\partial Z^c}{\partial p_{kn}} > 0$ and it implies that the function is an increasing function with respect to p_{kn}. To minimize the cost

function Z^c, it is more beneficial when the value of p_{kn} and f_k is smaller. However, the offloading latency T_k^c increases with p_{kn} and f_k decreasing and offloading latency is less than $\tilde{T}_k$. Therefore, according to C1, the relation between p_{kn} and f_k can be described as:

$$\frac{b_k}{w_0 n_k \log_2(1+\alpha_{kn}p_{kn})} + \frac{s_k}{f_k} = T_k \tag{19}$$

From (19), we have $1+\alpha_{kn}p_{kn} = 2^{\xi_k/\tau_k}$ and $p_{kn} = (2^{\xi_k/\tau_k}-1)/\alpha_{kn}$, where $\xi_k = b_k/(w_0 n_k)$ and $\tau_k = T_k - s_k/f_k$. Variable τ_k denotes the uplink transmission time from MT k to BSs. To minimize the cost function, the problem in (17) can be simplified as:

$$\min_{\boldsymbol{\tau}} y(\boldsymbol{\tau}) = \sum_k \gamma_k^E \tau_k n_k \frac{2^{\xi_k/\tau_k}-1}{\alpha_k} + \gamma_k^M(\beta\frac{b_k}{\tau_k} + q\frac{s_k}{T_k-\tau_k}) \tag{20}$$

subject to $\forall k \in \mathcal{K}$

$$C8: \frac{b_k}{w_0 n_k \log_2(1+\alpha_k p_k^T/n_k)} \le \tau_k \le T_k$$
$$C9: \sum_k \frac{s_k}{T_k-\tau_k} \le f_S$$

C8 denotes the changing range of τ_k. The minimum value of τ_k is the ratio of b_k and maximum uplink transmission rate r_k^m, $r_k^m = w_0 n_k \log_2(1+\alpha_k \frac{p_k^T}{n_k})$. C9 is the transformation of C2.

$$\frac{\partial^2 y}{\partial \tau_k^2} = \frac{\gamma_k^E n_k (a_k \ln 2)^2 \cdot 2^{\xi_k/\tau_k}}{\alpha_k \tau_k^3} + 2\gamma_k^M \beta \frac{b_k}{\tau_k^3} + 2\gamma_k^M q \frac{s_k}{(T_k-\tau_k)^3} \tag{21}$$

Derived from (21), $\frac{\partial^2 y}{\partial \tau_k^2} > 0$. The convex of the function is proved. The Lagrange function can be attained based on KKT (Karush Kuhn Tucker) conditions as below (22).

$$\begin{aligned} L(\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\nu},\theta) &= \sum_k \gamma_k^E \tau_k n_k \frac{2^{\xi_k/\tau_k}-1}{\alpha_k} + \gamma_k^M(q\frac{s_k}{T_k-\tau_k}) + \\ &\gamma_k^M \beta \frac{b_k}{\tau_k} + \sum_k \mu_k\Big(\frac{b_k}{w_0 n_k \log_2(1+\alpha_k p_k^T/n_k)} - \tau_k\Big) \\ &+ \sum_k \nu_k(\tau_k - T_k) + \theta(\sum_k \frac{s_k}{T_k-\tau_k} - f_S) \end{aligned} \tag{22}$$

where the variables μ_k, ν_k, θ are all nonnegative coefficients representing the Lagrange multipliers. The KKT conditions are as follows, for $\forall k$.

$$\begin{aligned} &\frac{\partial L}{\partial \tau_k} = \gamma_k^E n_k \frac{2^{\xi_k/\tau_k}-1}{\alpha_k}(1-\frac{\xi_k \ln 2}{\tau_k}) - \gamma_k^E n_k \xi_k \ln 2 \frac{1}{\alpha_k \tau_k} \\ &+ (\gamma_k^M q + \theta)\frac{s_k}{(T_k-\tau_k)^2} - \gamma_k^M \beta \frac{b_k}{\tau_k^2} - \mu_k + \nu_k = 0 \end{aligned} \tag{23}$$

$$\mu_k\Big(\frac{b_k}{w_0 n_k \log_2(1+\alpha_k p_k^T/n_k)} - \tau_k\Big) = 0 \tag{24}$$

$$\nu_k(\tau_k - T_k) = 0 \tag{25}$$

$$\theta(\sum_k \frac{s_k}{T_k-\tau_k} - f_S) = 0 \tag{26}$$

The optimal τ_k^* can be obtained from the KKT condition. Then p_{kn}^* and f_k^* can be derived by (27) and (28).

$$p_k^* = n_k \frac{2^{\xi_k/\tau_k^*}-1}{\alpha_k} \tag{27}$$

$$f_k^* = \frac{s_k}{T_k - \tau_k^*} \tag{28}$$

Lagrange multipliers update as below.

$$\mu_k(t+1) = [\mu_k(t) + \delta(t)(tmin_k - \tau_k)]^+ \tag{29}$$

$$\nu_k(t+1) = [\nu_k(t) + \delta(t)(\tau_k - tmax_k)]^+ \tag{30}$$

$$\theta(t+1) = [\theta(t) + \delta(t)(\sum_k \frac{s_k}{T_k-\tau_k} - f_S)]^+ \tag{31}$$

where variable t represents the t_{th} iteration, $\delta(t)$ implies the step of the iteration and $[z]^+ = max\{z, 0\}$. $tmin_k$ represents the left side of C8 and $tmax_k$ denotes the right side of C8. The optimal resource allocation can be iteratively derived by utilizing the KKT condition.

C. CWRAA with Interference

A fractional frequency reuse based on hungrian and graph coloring methods (FFR-HGC) is applied to allocate subchannels for MTs when $K^c > N$. The frequency reuse among MTs results in intra-cell interference. Therefore, the purpose of FFR-HGC method is to mitigate the interference received at the MTs from the MTs of other BSs and achieve fractional frequency reuse at the same time. In order to execute graph coloring, the constructed interference graph in Fig 2(b) is modified into a weighted interference graph, where the weight of every directed edge is calculated as

$$\rho_{km}|_{k\in\mathcal{V}_j, m\in V_{j'}} = \begin{cases} 0, j = j' \\ p_k h_{kn}^m, j \ne j' \end{cases} \tag{32}$$

where h_{kn}^m represents the channel gain between the MT k of BS j and the MT m of BS j'. p_k denotes the transmission power of MT k associated to BS j and is set as fixed value in the process of FFR-HGC. The weight ρ_{km} means the intensity of interference at MT m associated to BS j'.

The steps of the FFR-HGC are described below.

1) Initialization: In this step, the MEC server sets the subchannel association table $\mathcal{C}(K^c \times N)$ mentioned above to zeros, and initializes the interference table $\mathcal{O}$, which is also an $K^c \times N$ table. Table $\mathcal{O}$ has real-valued variable o_{kn} representing the sum interference from all other offloading MTs experienced by MT on color n. So, o_{kn} is given by

$$o_{kn}|k \in \mathcal{V}_j = \sum_{m\in V_{j'}|j'\ne j} c_{mn}\rho_{km} \tag{33}$$

The interference table $\mathcal{O}$ is set as zeros in the initialization step, too. We set the uncolored vertices as U. Its initial value U_0 is set as all offloading MTs $\mathcal{K}^c$.

2) Orthogonal Subchannel Allocation: Since the number of MTs is more than the amount of subchannels and the number of orthogonal subchannels is N, we should take measures to select N MTs from U_0 to take up the N orthogonal subchannels. Hence, to maximize the throughput of the N MTs, we apply a method besed on Hungarian method [33] to allocate the subchannels. Once the N subchannels allocated, the N MTs will be selected. The method is denoted as:

$$c_{kn} = \arg\max \sum_{k=0}^{K} r_k, 1 \le k \le K, 1 \le n \le N \tag{34}$$

There are $K^c - N$ MTs left needing allocated subchannels on which there exists interference from other MTs. The set of uncolored vertices U is updated and the size of U is $K^c - N$.

3) Finding the Color with the Smallest Interference: In order to mitigate the interference on MT $k \in U$, the subchannel with smallest interference in current time should be assigned to MT k. So it is necessary to find the color with the smallest interference. We search for the color by searching for the color on which MT k can achieve the highest transmission rates. Assuming color n is assigned to MT k, we calculate the estimated transmission rate of MT k as follows:

$$r_{kn}|k \in \mathcal{V}_j = w_0 \log_2(1 + \frac{p_k h_{kn}}{\sigma_{kn}^2 + o_{kn}}) \tag{35}$$

Therefore the expected $\bar{n}$ is derived by:

$$\bar{n} = \arg_{n \in \mathcal{N}} \max\{r_{kn}\} \tag{36}$$

Then $\bar{n}$ is allocated to MT k.

4) Update Tables: Both the subchannel association table $\mathcal{C}$ and the interference table $\mathcal{O}$ are updated in this step. According to the subchannel allocation to vertex k in the previous step, the corresponding variables of the assigned colors in table $\mathcal{C}$ are set to 1, and the interference caused by this new assignment is calculated and updated in table $\mathcal{O}$.

5) Update the Set of Uncolored Vertices: The vertex k got colored will be excluded from the uncolored vertices set U and U is updated.

6) Check Whether all Vertices are Colored: The uncolored vertices set U will be checked. If the set U is not empty, steps 3) to 5) will be repeated. If set U is empty, we will go to the next step.

7) Color Assignment: The set of colors will be allocated to the corresponding vertices according to the subchannel association table $\mathcal{C}$.

After assigning the subchannels, the transmission power and CPU cycles are allocated according to (15).

The partial derivative of (15) is shown as (37) when $c_{kn} = 1$. If $c_{kn} = 0$, $\dfrac{\partial Z^c}{\partial p_{kn}}$ is equal to 0.

$$\frac{\partial Z^c}{\partial p_k} = \frac{b_k \gamma_k^E}{w_0 \log_2(1 + \alpha_{kn} p_k)} (1 - \frac{1}{\log_2^2(1 + \alpha_{kn})} \cdot \frac{\alpha_{kn} p_k}{1 + \alpha_{kn} p_k \ln 2}) + \gamma_k^M \beta w c_{kn} \frac{\alpha_{kn}}{1 + \alpha_{kn} p_k \ln 2} \tag{37}$$

where $\alpha_{kn} p_k$ is denoted as SINR and the value is larger than 1, thus $\frac{\partial Z^c}{\partial p_k} > 0$ implies that the function is an increasing function with respect to p_k.

The solution for p_k^* and f_k^* is similar with CWRAA without interference and does not be repeated it here.

D. The Existence of NE

We then study the existence of NE of the distributed potential game model. To proceed, we first introduce an important concept of potential game [17].

Definition 2. *A game is called an exact potential game if it admits a potential function $\phi(\mathcal{A})$ such that for every $k \in \mathcal{K}, \mathbf{a}_{-k}$, and $a_k, a_k' \in \mathcal{A}_k$, if*

$$u_k(a_k, \mathbf{a}_{-k}) - u_k(a_k', \mathbf{a}_{-k}) = \phi(a_k, \mathbf{a}_{-k}) - \phi(a_k', \mathbf{a}_{-k})$$

Theorem 1. *Every ordinal potential game with finite strategy sets owns as least one pure-strategy NE and has the FIP.*

Ordinal potential game includes exact potential game [34]. A nice property of ordinal potential game is that it always admits a NE.

Theorem 2. *The potential game model using UZFR subchannel allocation method is an exact potential game with the potential function as given in (38), and hence always has a NE and the finite improvement property.*

$$\phi(\mathcal{A}) = (1 - a_k)\Big(\sum_{k' \ne k} (\frac{\gamma_{k'}^E p_{k'n} b_{k'}}{w_0 \lfloor \frac{N}{(1 + \sum_{j \ne k}^K a_j)} \rfloor \log_2(1 + \alpha_{k'} p_{k'n})} + \gamma_{k'}^M M_{k'}^c) + z_k^l \Big) + a_k \sum_{k=1}^{K} z_k^c \tag{38}$$

Proof: Based on (13), we have that

$$u_k(1, \mathbf{a}_{-k}) - u_k(0, \mathbf{a}_{-k}) = z_k^c - z_k^l \tag{39}$$

Based on (38), $\phi(1, \mathbf{a}_{-k})$ and $\phi(0, \mathbf{a}_{-k})$ can be written as follows respectively,

$$\phi(1, \mathbf{a}_{-k}) = \sum_{k=1}^{K} z_k^c = z_k^c + \sum_{k' \ne i}^{K} z_{k'}^c = z_k^c + \sum_{k' \ne k} (\gamma_{k'}^M M_{k'}^c + \frac{\gamma_{k'}^E p_{k'n} b_{k'}}{w_0 \lfloor N/(1 + \sum_{j \ne k}^K a_j) \rfloor \log_2(1 + \alpha_{k'} p_{k'n})}) \tag{40}$$

$$\phi(0, \mathbf{a}_{-k}) = \sum_{k' \ne k} (\frac{\gamma_{k'}^E p_{k'n} b_{k'}}{w_0 \lfloor N/(1 + \sum_{j \ne k}^K a_j) \rfloor \log_2(1 + \alpha_{k'} p_{k'n})} + \gamma_{k'}^M M_{k'}^c) + z_k^l \tag{41}$$

From (40) and (41), we can achieve that

$$\phi(1, \mathbf{a}_{-k}) - \phi(0, \mathbf{a}_{-k}) = z_k^c - z_k^l \tag{42}$$

From (39) and (42), we obtain that $\phi(1, \mathbf{a}_{-k}) - \phi(0, \mathbf{a}_{-k}) = u_k(1, \mathbf{a}_{-k}) - u_k(0, \mathbf{a}_{-k})$. Similarly, we can derive that $\phi(0, \mathbf{a}_{-k}) - \phi(1, \mathbf{a}_{-k}) = u_k(0, \mathbf{a}_{-k}) - u_k(1, \mathbf{a}_{-k})$ as well. Therefore, the game model utilizing UZFR method is an exact potential game and there is at least one pure-strategy NE and has the FIP.

Theorem 3. *The distributed game model using FFR-HGC subchannel allocation method is an exact potential game with*

the potential function as given in (43), and hence always has a NE and the finite improvement property.

$$\phi(\mathcal{A}) =(1-a_k)(\sum_{k'\neq k}(\frac{\gamma_{k'}^E p_{k'} b_{k'}}{w_0 \log_2(1+\frac{h_{k'n}p_{k'n}}{\sigma_{k'n}^2+I_{k'n}''+I_{k'n}'^j+h_{kn}p_{kn}})} + \gamma_{k'}^M M_k^c) + z_k^l) + a_k \sum_{k=1}^{K} z_k^c \tag{43}$$

Proof: Based on (43), $\phi(1, \mathbf{a}_{-k})$ and $\phi(0, \mathbf{a}_{-k})$ can be written as follows respectively,

$$\phi(1, \mathbf{a}_{-k}) = \sum_{k=1}^{K} z_k^c = z_k^c + \sum_{k'\neq k}\Big(\frac{\gamma_{k'}^E p_{k'} b_{k'}}{w_0 \log_2(1+\frac{h_{k'n}p_{k'n}}{\sigma_{k'n}^2+I_{k'n}''+I_{k'n}'^j+h_{kn}p_{kn}})} + \gamma_{k'}^M M_{k'}^c\Big) \tag{44}$$

$$\phi(0, \mathbf{a}_{-k}) = \sum_{k'\neq k}\Big(\frac{\gamma_{k'}^E p_{k'} b_{k'}}{w_0 \log_2(1+\frac{h_{k'n}p_{k'n}}{\sigma_{k'n}^2+I_{k'n}''+I_{k'n}'^j+h_{kn}p_{kn}})} + \gamma_{k'}^M M_{k'}^c\Big) + z_k^l \tag{45}$$

From (44) and (45) ,we can achieve that

$$\phi(1, \mathbf{a}_{-k}) - \phi(0, \mathbf{a}_{-k}) = z_k^c - z_k^l \tag{46}$$

From (39) and (46), we obtain that $\phi(1, \mathbf{a}_{-k}) - \phi(0, \mathbf{a}_{-k}) = u_k(1, \mathbf{a}_{-k}) - u_k(0, \mathbf{a}_{-k})$. Similarly, we can derive that $\phi(0, \mathbf{a}_{-k}) - \phi(1, \mathbf{a}_{-k}) = u_k(0, \mathbf{a}_{-k}) - u_k(1, \mathbf{a}_{-k})$ as well. Therefore, the distributed game model using FFR-HGC method is an exact potential game and there is at least one pure-strategy NE and has the FIP.

E. Algorithm Description

In this section, we describe the process of the distributed JCORAO scheme. Due to the decentralized mechanism, each MT makes the computation offloading strategy locally and it is beneficial for reducing the controlling and signaling overhead in the system. The NE is achieved by Algorithm 1 and Algorithm 2. Algorithm 2 is the sub-algorithm of Algorithm 1. When the NE is attained and the optimal offloading strategy profile $\mathcal{A}^*$ and resource allocation $\mathcal{C}^*, \mathbf{P}^*, \mathbf{F}^*$ are obtained, all MTs will follow the optimal offloading strategies without deviation because of the property of NE.

For Algorithm 1, in the initial phase, all MTs choose to offload their tasks into MEC server. Then we compute the local execution cost by (8) and obtain offloading execution cost by Algorithm 2. By comparing the size of the two costs, the offloading strategy profile $\mathcal{A}$ is updated. In the cycle phase, each MT does not update their offloading strategy until all MTs have no motivations to change their strategy. In each episode, MTs intend to decrease respective cost and have no incentive to decrease the total cost of all MTs such that each MT makes its decision by comparing own local execution cost with offloading execution cost. In addition, the optimal resource allocation $\mathcal{C}^*, \mathbf{P}^*, \mathbf{F}^*$ is recalculated in this episode when the offloading strategy profile $\mathcal{A}$ is updated.

Algorithm 1 Process of the distributed JCORAO scheme

Input: K: number of MTs;
l: the index of iteration times;
$b_k, s_k, \boldsymbol{\gamma}^E, \boldsymbol{\gamma}^M, \boldsymbol{\gamma}^T, \beta_k, q_k, h_k, \sigma_k^2, T_k, p_k^T, \kappa, f_k^l$.
Output: $\{\mathcal{A}^*, \mathcal{C}^*, \mathbf{P}^*, \mathbf{F}^*\}$:optimal resource allocation

initialize: $\mathcal{A}_0$
for $k = 1$ to K **do**
  compute the local execution cost z_k^l by (8).
  use Algorithm 2 to get optimal resources $\mathcal{C}^*, p_k^*, f_k^*$ and corresponding offloading cost z_k^c.
  if $z_k^l > z_k^c$ **then**
    $a_k = 1$
  else
    $a_k = 0$
  end if
end for
update $\mathcal{A}$.
while $\mathcal{A} \neq \mathcal{A}_0$ **do**
  $\mathcal{A}_0 = \mathcal{A}$ and $l = l + 1$
  for $k = 1$ to K **do**
    $a_k = 1$ and update $\mathcal{A}$
    utilize Algorithm 2 to get corresponding offloading cost z_k^c.
    if $z_k^c > z_k^l$ **then**
      $a_k = 0$
    else
      $a_k = 1$ and update the offloading strategy $\mathcal{A}$.
    end if
  end for
end while
the offloading strategy profile $\mathcal{A}^*$ and optimal resource allocation $\mathcal{C}^*, \mathbf{P}^*, \mathbf{F}^*$ are obtained.

For Algorithm 2, in initial phase, current offloading strategy profile $\mathcal{A}$ determines to utilize UZFR or FFR-HGC to allocate subchannels. In cycle phase, the optimal communication and computation resources are attained iteratively based on KKT condition. The offloading execution cost of offloading MTs z_k^c is computed in last phase.

By executing Algorithm 1 and Algorithm 2, we achieve NE such that the optimal offloading strategy $\mathcal{A}^*$ and optimal resource allocation $\mathcal{C}^*, \mathbf{P}^*, \mathbf{F}^*$ are obtained.

V. SIMULATION RESULTS

In this section, we use computer simulations to evaluate the performance of the distributed JCORAO scheme.

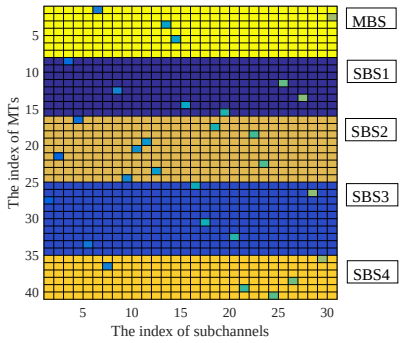
A. Parameter Settings

In the simulation, one MBS and 4 SBSs are deployed in a $100 \times 100m^2$ area. The MBS is located in the center of the area and SBSs are placed in the four corners of the world. The number of MTs associating to BS j is a randomly integer. There are $\sum_{j=0}^{j=4} V_j$ MTs conducting joint computation offloading and resource allocation optimization. The initial cost function weights are set as $\gamma_k^E = \gamma_k^T = \gamma_k^M = 0.5$. The transmission power of single MT, p_{nk} is set to 10 dBm

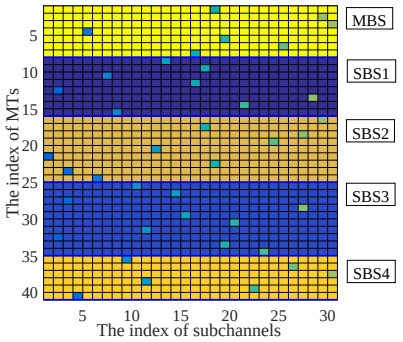
Algorithm 2 Process of CWRAA

Input: K^c:number of offloading MTs;
$\mathcal{A}$:current offloading strategy profile;
$max\_iteration$:maximum number of iterations;
Output: $\mathcal{C}^*$:optimal subchannel allocation table;
p_k^*:optimal communication resources;
f_k^*: optimal computation resources.

initialize:set initial Lagrange multiplier $\boldsymbol{\mu}_0, \boldsymbol{\nu}_0, \theta_0$.
if $(\sum_{k=1}^{K} a_k) \leq N$ **then**
  use UZFR to derive subchannels $\mathcal{C}^*$.
else
  use FFR-HGC to derive subchannels $\mathcal{C}^*$.
end if
for $n = 1$ to $max\_iteration$ **do**
  set $\delta = 1/(50 + n)$
  for $k = 1$ to K^c **do**
    compute τ by (24) based on KKT condition.
  end for
  update Lagrange multiplier $\mu_k(t+1), \nu_k(t+1)$ and $\theta(t+1)$ by (29)(30)(31)
  k=k+1
end for
compute p_k^* by (27).
compute f_k^* by (28).
compute offloading execution cost of offloading MTs z_k^c.



(a) FFR-HCG in initial phase.



(b) FFR-HCG in final phase.

Fig. 3. The subchannel distribution among MTs. (a)Subchannel allocation with FFR-HCG in initial phase. (b)Subchannel allocation with FFR-HCG in final phase.

at the beginning. The channel gain models presented in 3GPP standardization [35] are adopted here.

For MTs, the maximum transmission power and offloading latency threshold are respectively set as the 20 dBm and 3 s. The local computation capability of MTs follows the Gaussian distribution $CN(\mu_1, \sigma_1^2)$, where the mean $\mu_1 = 1000$ Mega/s, and the standard deviation $\sigma_1 = 50$. The data size of the tasks and computing load follows the Gaussian distribution $CN(\mu_2, \sigma_2^2)$ and $CN(\mu_3, \sigma_3^2)$, where $\mu_2 = \mu_3 = 1000$ KB and $\sigma_2 = \sigma_3 = 50$ [32]. According to realistic measurements, κ is set as 10^{-11} [24].

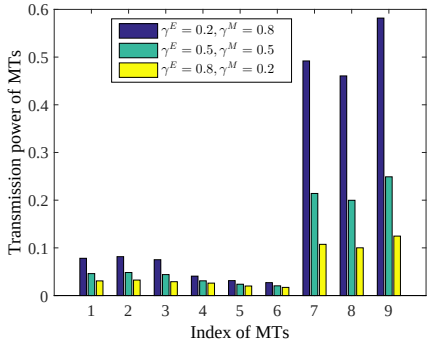
For the wireless access, we set the channel bandwidth of each subchannel $w_0 = 5$ MHz and the channel power gain of the MTs follows the Gaussian distribution $CN(\mu_4, \sigma_4^2)$, where $\mu_4 = 10, \sigma_4 = 1$. There are in all 30 subchannels. In addition, thermal noise power of the MTs follows the Gaussian distribution $CN(\mu_5, \sigma_5^2)$, where $\mu_5 = 5, \sigma_5 = 1$. For the MEC server, we set the maximum computation capability f_S as 40000 Mega cycles. The price for communication rate is 0.05 \$/Mbit. The charge for computation resources is 0.1 \$/Mega.

B. Performance Evaluation of distributed JCORAO Scheme

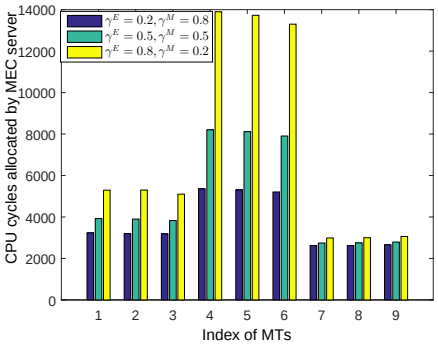
1) Subchannel Allocation with FFR-HCG Method: Fig. 3 shows the subchannel distribution among 40 MTs in the coverage of one MBS and four SBSs with 30 subchannels. Some MTs suffers interference from their neighboring MTs. Fig. 3 (a) shows the results of orthogonal subchannel allocation based on Hungarian method which is the first step in the FFR-HGC approach and Fig. 3 (b) indicates the whole results of subchannel allocation with FFR-HGC method. It can be observed that 30 orthogonal subchannels are allocated with Hungarian method at first and then the remaining MTs are allotted subchannels with color graph method. From Fig. 3 (b), we can see that the MTs in the same BS do not occupy the same subchannel and the subchannel is reused by the MTs far away, rather than the MTs near to each other. we can also observe from Fig. 3 (b) that a subchannel is used at most twice, such as subchannel 3 and 11. The results of subchannel allocation illustrate that FFR-HCG method is an effective way to mitigate the interference among neighboring MTs.

2) The effect of weighs among impact factors: In addition, we take the weights among impact factors into consideration on the transmission power and CPU cycles in Fig. 4. In order to display more clearly, we select 9 representative MTs from 40 MTs. As shown in Fig. 4 (a), the optimal transmission power to the MTs decrease with the increasing of the communication resource cost weight. It can be seen from Fig. 4 (b) that the optimal CPU cycles allocated to the MTs increase with the decreasing of the computation resource cost weight. This is reasonable since a larger γ^E will lead to the increase of cost on communication resources which in turn result in the decrease of cost on computation resources. We choose $\gamma^E = \gamma^M = 0.5$ as simulation parameters to balance the monetary cost between communication and computation cost.

3) Algorithm Comparison with Existing Algorithms: We evaluate the distributed JCORAO scheme performance compared with several baseline algorithms, such as local execution



(a) Transmission power allocation comparison.



(b) CPU cycles comparison.

Fig. 4. Comparison of transmission power and CPU cycles for different weights.(a)transmission power. (b)CPU cycles.

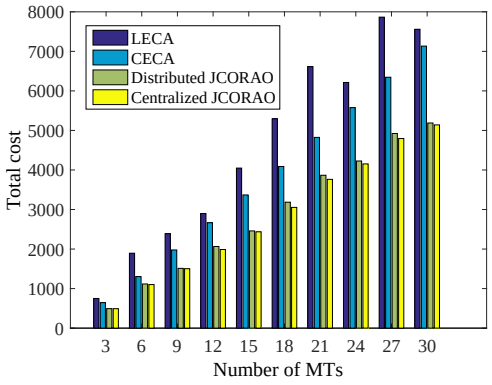
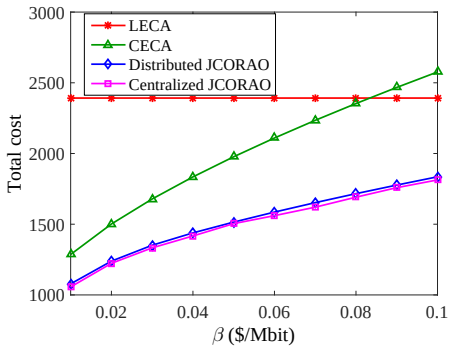
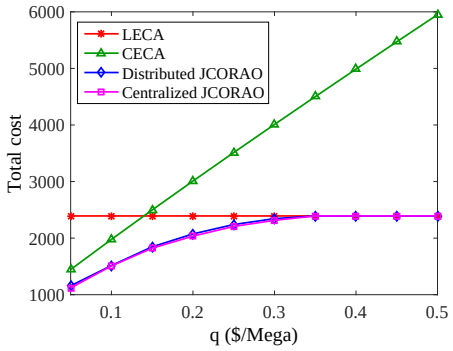


Fig. 5. The impact of number of MTs.



(a) The impact of communication resources price.



(b) The impact of computation resource price.

Fig. 6. The impact of computation resource price and computation resource price. (a)Communication resources price. (b)Computation resources price.

completely algorithm (LECA), cloud execution completely algorithm (CECA) and centralized JCORAO scheme. In LECA, all MTs decide to execute their tasks locally. On the contrary, all MTs determine to execute their tasks on the MEC server in CECA. In centralized JCORAO scheme, the method of exhaustion is utilized to solve the optimization problem of (12).

At first, the total cost comparison with the number of MTs is analyzed. As shown in Fig. 5, the total cost has a tendency to rise with the increasing of participants for all algorithms because the occupation of communication resources and computation resources is more. By comparison with LECA and CECA, the total cost of distributed JCORAO scheme is minimum. Proposed scheme's total cost is a little higher than but nearly the same as the centralized JCORAO scheme. However, the centralized JCORAO scheme has very high algorithm complexity which is NP hard problem.

Fig. 6 shows the impact of communication and computation resource prices on MTs' total cost. We can observe from Fig. 6 (a) that the total cost increases with the growth of communication price but the growth rates of total cost decreases slowly. It is due to that the cost of offloading begins to be more than the local cost and the number of offloading MTs starts descending when communication price increases. The phenomenon is more obvious in Fig. 6 (b) and the growth rate is eventually equal to zero. With the ascent of computation price, the cloud execution cost increases so that more MTs deal with tasks locally. At last, all MTs offload no task to MEC server. Therefore, the curves of LECA, distributed JCORAO and centralized JCORAO coincide when the computation price is bigger than 0.3 $/Mega.

Next, the complexity of above algorithms is analyzed. Table II describes the complexity of LECA, CECA, distributed JCORAO scheme and centralized JCORAO scheme. $max\_iteration$ is iteration times of KKT condition solution defined at Algorithm 2. We can also see the complexity difference among these algorithms in Fig. 7.

As shown in Fig. 7, the distributed JCORAO scheme and CECA spend more time to complete the tasks of MTs than LECA obviously. However, the total cost of LECA is the largest compared with the other algorithms, which can be

TABLE II
ALGORITHM COMPLEXITY COMPARISON

LECA	K
CECA	$K * max\_iteration$
Distributed JCORAO	$2 * K^c * max\_iteration + K^l$
Centralized JCORAO	$2^{K^c} * K^c * max\_iteration + K^l$

seen obviously from Fig. 5. We can also see from Fig. 7 that the running time of the distributed JCORAO scheme fluctuates a little with the number of MTs. This is because the running time is associated with the number of offloading MTs K^c which is not absolutely linear with the number of MTs K. Moreover, we can conclude from Fig. 7 that the running time of our scheme is less than CECA. Furthermore, the complexity of the distributed JCORAO scheme is much less than the centralized JCORAO scheme as shown in Table II. Therefore, the distributed JCORAO scheme outperforms other algorithms by making tradeoff between the total cost and algorithm complexity.

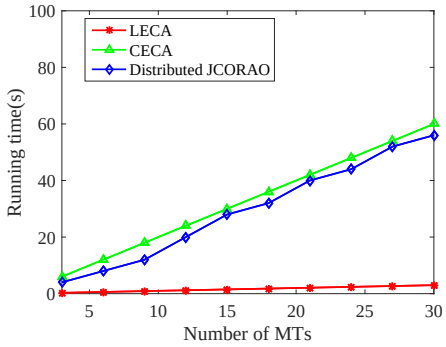


Fig. 7. Algorithm complexity comparison.

At last, we compare energy consumption and offloading latency with distributed computation offloading algorithm (DCOA) proposed in [11] and energy-efficient dynamic offloading and resource scheduling scheme (eDors) proposed in [32]. The DCOA scheme only focuses on offloading strategies in mobile cloud computing adopting a distributed potential game and is not involved in dynamic resource allocation. The eDors is a distributed algorithm consisting of offloading selection, CPU cycle control and power control. From Fig. 8, it can be seen that the energy consumption of DCOA mounts up rapidly while eDors and proposed scheme are relatively slow. When the size of input data is lower than 4 Mbit, the distributed JCORAO scheme and DCOA have approximate energy consumption but the energy consumption of distributed JCORAO scheme is slightly lower than DCOA. However, the difference becomes more obvious later. In comparison with eDors scheme, our scheme always spends less energy consumption evidently. We can observe from Fig. 9 that our proposed scheme occupies the least task completion time. In summary, proposed scheme can save more energy and complete tasks with less time than the other algorithms significantly.

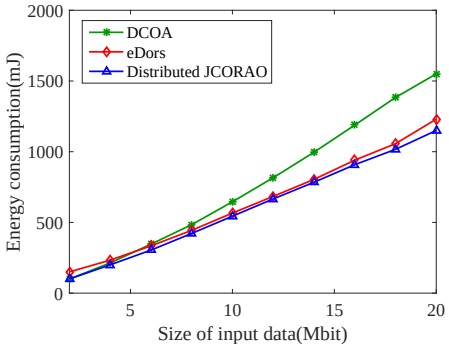


Fig. 8. Comparison of energy consumption.

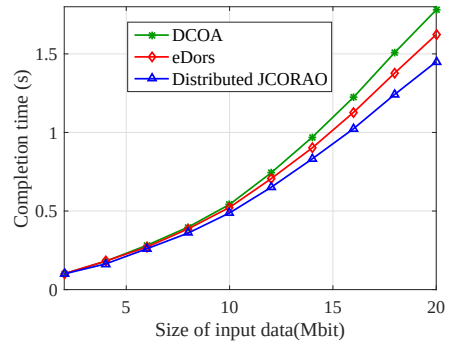


Fig. 9. Comparison of offloading latency.

VI. CONCLUSION

In this paper, an optimization problem is formulated to acquire computation offloading strategy policy, uplink subchannel allocation, uplink transmission power allocation and computation resource scheduling at first. Then a distributed joint computation offloading and resource allocation optimization (JCORAO) scheme consisting of a potential game and CWRAA in HetNets with MEC is proposed. A distributed potential game model based on the property of FIP is established to obtain the strategy offloading policy. The existence of NE is proved in the game. For the sub-algorithm CWRAA, on one hand, we take OFDM subchannel allocation and uplink power allocation into account in HetNets. The solutions of subchannel allocation consist of UZFR and FFR-HGC according to the interference between MTs. On the other hand, the computation resource allocation in MEC is studied. The JCORAO scheme eventually solved the optimization problem by the mutual iteration of the two sub-algorithms. Finally, the simulation results is revealed. Compared with existing algorithms, the distributed JCORAO scheme can reduce the energy consumption and task completion time significantly with lower complexity.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322-2358, 4th Quart.2017.

[2] B. P. Rimal, D. Pham Van and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," in *Proc. IEEE INFOCOM WKSHPS*, San Francisco, CA, Apr. 2016, pp. 991-996.

[3] P. Mach, Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. PP, no.99, pp. 1628 - 1656, Mar.2017.
[4] H. Zhang, J. Du, J. Cheng, K. Long and V. Leung, "Incomplete CSI Based Resource Optimization in SWIPT Enabled Heterogeneous Networks: A Non-Cooperative Game Theoretic Approach," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1-1, Dec.2017.
[5] S. Barbarossa, S. Sardellitti and P. Di Lorenzo, "Communicating while computing: Distributed mobile cloud computing over 5G heterogeneous networks," *IEEE Signal Process. Mag.*, vol. 31, no. 6, pp. 45-55, Nov. 2014.
[6] J. Xu and S. Ren, "Online learning for offloading and autoscaling in renewable-powered mobile edge computing," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec.2016, pp. 1-6.
[7] ETSI, "New White Paper: ETSI's Mobile Edge Computing initiative explained," ETSI White Paper, Sept. 2015.
[8] J. Plachy, Z. Becvar and E. C. Strinati, "Dynamic resource allocation exploiting mobility prediction in mobile edge computing," in *Proc. IEEE PIMRC*, Valencia, Spain, Sept. 2016, pp. 1-6.
[9] M. Liu, Y. Liu, "Price-Based Distributed Offloading for Mobile-Edge Computing with Computation Capacity Constraints," *IEEE Wireless Commun. Lett.*, vol. PP, no.99, pp. 1-1, Dec. 2017.
[10] W. Fan, Y. Liu, B. Tang, F. Wu and Z. Wang, "Computation Offloading Based on Cooperations of Mobile Edge Computing-Enabled Base Stations," *IEEE Access*, vol. PP, no. 99, pp. 1-1.
[11] X. Chen, L. Jiao, W. Li and X. Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795-2808, Oct. 2016.
[12] L. Tang, X. Chen and S. He, "When Social Network Meets Mobile Cloud: A Social Group Utility Approach for Optimizing Computation Offloading in Cloudlet," *IEEE Access*, vol. 4, pp. 5868-5879, Sep. 2016.
[13] H. Zhang, F. Guo, H. Ji and C. Zhu, "Combinational Auction-Based Service Provider Selection in Mobile Edge Computing Networks," *IEEE Access*, vol. 5, pp. 13455-13464, Jul. 2017.
[14] A. Al-Shuwaili and O. Simeone, "Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398-401, June 2017.
[15] M. Molina, O. Muñoz, A. Pascual-Iserte and J. Vidal, "Joint scheduling of communication and computation resources in multiuser wireless application offloading," in *Proc. IEEE PIMRC*, Washington, DC, USA, Sept. 2014, pp. 1093-1098.
[16] Y. Mao, J. Zhang and K. B. Letaief, "Dynamic Computation Offloading for Mobile-Edge Computing with Energy Harvesting Devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590-3605, Dec. 2016.
[17] Y. Yu, J. Zhang and K. B. Letaief, "Joint subcarrier and CPU time allocation for mobile edge computing," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1-6.
[18] P. Zhao, H. Tian, C. Qin and G. Nie, "Energy-Saving Offloading by Jointly Allocating Radio and Computational Resources for Mobile Edge Computing," *IEEE Access*, vol. 5, pp. 11255-11268, Jun. 2017.
[19] Y. Mao, J. Zhang, S. H. Song and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec.2016, pp. 1-6.
[20] C. C. Coskun and E. Ayanoglu, "Energy- and Spectral-Efficient Resource Allocation Algorithm for Heterogeneous Networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 590-603, Jan. 2018.
[21] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung and H. V. Poor, "Energy Efficient User Association and Power Allocation in Millimeter-Wave-Based Ultra Dense Networks With Energy Harvesting Base Stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936-1947, Sept. 2017.
[22] H. Zhang, B. Wang, K. Long, J. Cheng and V. C. M. Leung, "Energy-Efficient Resource Allocation in Heterogeneous Small Cell Networks with WiFi Spectrum Sharing," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1-5.
[23] P. Trakas, F. Adelantado, N. Zorba and C. Verikoukis, "A QoE-Aware Joint Resource Allocation and Dynamic Pricing Algorithm for Heterogeneous Networks," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1-6.
[24] W. Hao and S. Yang, "Small Cell Cluster-Based Resource Allocation for Wireless Backhaul in Two-Tier Heterogeneous Networks With Massive MIMO," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 509-523, Jan. 2018.
[25] S. Sardellitti, G. Scutari and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89-103, Jun. 2015.
[26] C. Wang, C. Liang, F. R. Yu, Q. Chen and L. Tang, "Computation Offloading and Resource Allocation in Wireless Cellular Networks with Mobile Edge Computing," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4924-4938, Aug. 2017.
[27] K. Zhang et al., "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access*, vol. 4, pp. 5896-5907, Aug. 2016.
[28] C. Wang, F. R. Yu, C. Liang, Q. Chen and L. Tang, "Joint Computation Offloading and Interference Management in Wireless Cellular Networks with Mobile Edge Computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7432-7445, Aug. 2017.
[29] H. W. Kuhn, "The Hungarian method for the assignment problem, *NavalRes. Logist. Quart.*, vol. 2, no. 1/2, pp. 83C97, Mar. 1955.
[30] D.Brélaz, "New methods to color the vertices of a graph, *Communications of the ACM*, vol. 22, no. 4, pp. 251C256, Apr. 1979.
[31] A. AL-Shuwaili, O. Simeone, A. Bagheri and G. Scutari, "Joint uplink/downlink optimization for backhaul-limited mobile cloud computing with user scheduling," *IEEE Trans. Signal Inf. Process. Netw.*, vol.PP, no.99, pp.1-1, Feb. 2017.
[32] S. Guo, B. Xiao, Y. Yang and Y. Yang, "Energy-efficient dynamic offloading and resource scheduling in mobile cloud computing," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, Apr. 2016, pp. 1-9.
[33] L. Liang, G. Feng and Y. Jia, "Game-Theoretic Hierarchical Resource Allocation for Heterogeneous Relay Networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1480-1492, Apr. 2015.
[34] D. Monderer and L. S. Shaply, "Potential games," in Games and Economic Behavio, vol. 14, no. 1, pp. 124-143, Jun. 1996.
[35] Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects (Release 9), 3rd Generation Partnership Project 3GPP TS 36.814, 2012. [Online]. Available: http://www.3gpp.org/ftp/

Jing Zhang received the B.S. degree in information and telecommunication engineering from the China University of Mining and Technology, Xuzhou, China, in 2015, where she is currently pursuing the Ph.D. degree with the the Southeast University, Nanjing, China. Her current research interests include mobile edge computing, resource management and game theory.

Weiwei Xia received the M.S. and Ph.D. degree in Communications and Information System from Southeast University, Nanjing, China, in 2003 and 2011, respectively. She is currently an Associate Professor in the National Mobile Communications Research Laboratory, Southeast University. Her current research interests include mobile cloud computing and networking, resource management and performance analysis in heterogeneous wireless networks, as well as mobility management. From April 2015 to May 2016, she was a visiting scholar in the Department of Electrical and Computer Engineering, Stony Brook University, USA.

Feng Yan (M14) received the B.S. degree from Huazhong University of Science and Technology, Wuhan, China, in 2005, the M.S. degree from Southeast University, Nanjing, China, in 2008, and the Ph.D. degree from TELECOM ParisTech, Paris, France, in 2013, all in electrical engineering. From November 2013 to April 2015, he was a post-doctoral researcher in Telecom Bretagne, Rennes, France. He is currently an Associate Professor in the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His current research interests are in the areas of wireless communications and wireless networks, with emphasis on applications of homology theory and stochastic geometry in wireless networks.

Lianfeng Shen received the B.S. degree in Radio Technology and M.S. degree in Wireless Communications from Southeast University, Nanjing, China, in 1978 and 1982 respectively. In March 1982, he joined the Department of Radio Engineering of Southeast University. From 1991 to 1993, he was a visiting scholar and a consultant with the Hong Kong Productivity Council working on wireless communications for 2 years. In 1998, he was ever a Senior Consultant in the Telecom Technology Centre of Hong Kong for 1 year. Since 1997, he has been a professor at the National Mobile Communications Research Laboratory of Southeast University. His research interest has recently been focusing on the broadband mobile communications including broadband wireless access system, vehicular ad hoc network, communications protocols and so on. He is one of the editors of Journal on Communication and services as the Member of the Expert Group in Information Science of the 973 Plan in China, and the Chair of the IEEE Communication Society Nanjing Chapter.