IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 28, 2019, accepted June 6, 2019, date of publication June 19, 2019, date of current version July 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923630

Energy-Aware Joint User Association and Resource Allocation for Coded Cache-Enabled HetNets

FANGFANG YIN[1], ANYUE WANG[2], DANPU LIU[1], AND ZHILONG ZHANG[1]

[1]Beijing Laboratory of Advanced Information Network, Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, L-1855, Luxembourg

Corresponding author: Zhilong Zhang (zhilong.zhang@outlook.com)

This work was supported in part by the Beijing Natural Science Foundation under Grant L172032, in part by the National Natural Science Foundation of China under Grant 61801051, in part by the Key Research and Development Program Projects in Shanxi Province under Grant 2019ZDLGY07-10, and in part by the Open Project of Information Security Laboratory of National Defense Scientific Research and Test under Grant 2016XXAQ09.

ABSTRACT Cache-enabled heterogeneous networks (HetNets) have recently emerged as an attractive solution to meet the exponentially increasing demand on mobile data traffic. However, the power consumption and the backhaul limitation of small base stations (SBSs) have become bottlenecks to deploy HetNets. How to relieve the burden of backhauls via wireless caching and enable the HetNets to operate in an energy-efficient way are still open issues. Aiming to minimize the power consumption while guaranteeing QoS requirements of users, in this paper, we address the problem of joint user association (UA) and resource allocation (RA) for coded cache-enabled HetNets. First, based on the many-to-many matching game between the virtual SBSs and users (VSU), we propose a low-complexity joint UA and power allocation (PA) algorithm (JUPVA). Then, considering the unequal BA, we design a three-phase optimization algorithm (JURVA), which makes a joint decision on UA, PA, and BA iteratively. The simulation results demonstrate that the proposed algorithms yield significant performance improvement in terms of power consumption.

INDEX TERMS Coded caching, HetNets, many-to-many matching, resource allocation, user association.

I. INTRODUCTION

With the proliferation of mobile devices and the prosperity of content providers, recent years have witnessed a dramatic increase in mobile multimedia services. The latest Cisco VNI report predicts that mobile multimedia traffic will account for 80% of overall mobile data traffic in 2021 [1]. Deploying heterogeneous networks (HetNets), which are composed of small cell base stations (SBSs) and macro base stations (MBSs), is a thriving solution to cope with this challenge [2]. However, due to the dense deployment of SBSs and the unprecedented data traffic growth, the base stations (BSs) consume 60%-80% of the total power in cellular networks, which will directly cause both serious environment problems and sharp rising energy costs for network operators [3]. Hence, both industry and academia advocate ‘‘green communications'' that work towards reducing the power consumption of HetNets [4].

In HetNets, even if users are uniformly distributed in geography, traditional UA schemes based on signal-to-interference-plus noise ratio (SINR) may still lead to an extreme load imbalance. In this case, the resources of underloaded BSs cannot be fully utilized, while the users associated to the overloaded BSs may not be well served due to the insufficient resources [5]. Furthermore, the download rate of a user is generally proportional to the wireless resources that it occupies [6]. Thus, to fully utilize the system resources and improve network performance, the UA and RA strategies have to be jointly considered as a central problem, and directly affect the transmission power consumption of BSs.

Accompanied with HetNets, caching contents at BSs is regarded as a promising innovation to improve the user experience and significantly alleviate the backhaul congestion of SBSs [7]. Uncoded caching systems are usually designed to maximize the file hit rate or the proportion of files received by the users in [8], [9]. Compared to the uncoded caching, the coded caching provides higher probability of

The associate editor coordinating the review of this manuscript and approving it for publication was Chunlong He.

re-constructing the content at the desired receiver [10]–[12], which results in a decrease of data transmission and power consumption of backhauls. Actually, coded caching is well suited to emerging HetNets which consist of a dense deployment of local-coverage SBSs with high data rates, along with sparsely distributed, large-coverage MBS. For example, authors in [13], [14] put forward a maximum-distance separable (MDS) encoded caching scheme to achieve energy-saving edge computing in HetNets.

Although many works have been devoted to the optimization of UA, RA and caching in HetNets, these aspects are usually studied separately and rarely jointly considered. Realizing the great potentials and open issues, we address the joint optimization problem of UA and RA in cache enabled HetNets, which involves the following challenges:

- To improve the utilization efficiency of wireless resource, the UA and RA strategies which are coupled together should be prudently studied.
- Owing to the dense deployment of SBSs, it is significant to minimize the power consumption of BSs and backhauls between SBSs and MBSs, while ensuring the QoS of users.
- Many previous works indicate that coded caching provides a more feasible solution than uncoded scheme. Hence, such a caching strategy should be jointly considered in future cache-enabled HetNets.

To address these challenges, we build a comprehensive model which captures the key components of edge caching in HetNets and the total power costs.

In this paper, we consider a downlink MDS encoded cache enabled HetNets in which the MBS both assigns the SBSs and allocates wireless resources to a set of users. To the best of our knowledge, this is the first work that takes into account the multi-association, bandwidth and power allocation, to minimize the power consumption of HetNets by the aid of many-to-many matching game. The main contributions of this paper are summarized as follows:

- We model the system power consumption of HetNets into four parts: transmission power, static power, cache power and backhaul power. Then, we formulate a joint problem to optimize UA and RA, and aim at minimizing the total power consumption.
- *Joint UA and PA with Virtual SBSs and users (VSU)-based matching Algorithm (JUPVA)*: Based on the assumption that the available bandwidth of each SBS is subdivided equally among its associated users, the original problem is transformed into UA and PA subproblems. Then, we develop the virtual SBSs and users (VSU)-based matching game to reformulate the UA subproblem. Finally, given the SBSs-users mapping information, the PA subproblem is amenable to Linear programming (LP) techniques to allocate power for users.
- *Joint UA and RA with VSU-based matching algorithm (JURVA)*: To solve the original power consumption minimization (PCM) problem, considering the unequal BA, we introduce a simplified BA strategy and thus obtain the joint UA and RA optimization with VSU-based many to many matching. We design a centralized three-phase iterative algorithm (JURVA) that determines the UA, BA and PA successively.
- Simulation results show that the proposed JUPVA and JURVA yield significant performance improvement in terms of power consumption after a small number of iterations. It is also shown that the proposed algorithms converge to a *two-sided swap stable* matching.

The rest of this paper is organized as follows. Section II introduces the related work. Section III describes system model and problem formulation. In section IV, a joint optimization method is proposed to obtain the solution by dividing the PCM problem into two subproblems, where the VSU-based many to many matching and LP are also discussed. Section V formulates the PCM problem as a three-phase optimization and presents the proposed JURVA. Simulation results are presented in Section VI and conclusions are drawn in Section VII.

II. RELATED WORK

Since SBSs are closer to users and usually provide larger bandwidth than the MBS, caching at the SBSs has attracted significant attention in HetNets. Compared to the uncoded caching, the careful placement of coded content in caches leads to a more significant decrease of the end-to-end delay and backhaul pressure for the HetNets [11]–[14]. For example, [11] and [12] have found the optimal content placement in all the SBSs by restructuring the contents with MDS codes to reduce the backhaul rate. Focusing on the content placement phase in HetNets, [13] investigates the tradeoff between expected backhaul rate and energy consumption. Authors in [14] jointly consider the coded caching and cooperative caching issues, and propose a cooperative coded caching scheme to optimize energy consumption.

The joint optimization of UA and RA policy in cache-enabled wireless networks has also been widely studied [15]–[18]. Due to the limited radio resource of BSs, user's downlink data rate is seriously affected by the channel condition and the allocated radio resource. Thus, UA and RA policies need to be well designed to improve the network performance in cache enabled HetNets. In [15], a system utility optimize problem is introduced which jointly considers UA, RA, caching and computing offloading policies. In [16], a green content caching and mobile UA mechanism is proposed to maximize the number of users requests served by the SBSs for energy-harvesting enabled HetNets. With given caching policy, authors [17], [18] jointly optimize UA and RA for a general HetNet, aiming to improve the network throughput and the total utility, respectively. However, previous works [16]–[18] which have jointly designed cache, UA and RA policies, have not considered the energy consumed by the dense deployment of SBSs.

Of relevance to our work are [13], [14], [19] where they focused on the optimization of energy consumption for coded cache-enabled HetNets. Authors in [13], [14] considered the caching and backhaul transport power consumption at higher layers separated from the transmission power of the associated SBSs. Due to the property of MDS coded caching, users have potential to associate with multiple SBSs for high-data-rate service. Thus, the authors in [13] mentioned the multi-UA approach to effectively exploit the advantage of MDS coded caching, however, the formulated energy optimization problem only focused on the caching and backhaul energy consumption. In another work [14], Jia *et al.* designed a cooperative content delivery scheme to share the coded file packets in the SBS cluster, i.e., each user can be served by one SBS, and other SBSs in the same cluster can assist the local SBS, by sending the coded packets to the local SBS via the link. However, the energy consumption [14] is only includes content caching, cooperative content transmission and backhaul content transmission, and the transmission power of association SBSs is also neglected. Finally, with the assumption that each popular multimedia file is coded and distributively stored in multiple energy harvesting enabled SBSs, X. Huang *et al.* investigated on-grid energy reduction problem which jointly consider UA and RA schemes [19]. However, they only focused on the transmission power of BSs, ignored the caching and backhaul power consumption. Therefore, the potential of reducing power consumption was not fully exploited in [13], [14], [19].

Last but not least, matching theory has been widely researched on resource allocation in future wireless networks [20]–[28]. A concise introduction and survey on matching theory applications was provided in [20]. Several works [21]–[24] have explored matching to improve the performance of different scenarios. Zhang *et al.* [25] derived a joint optimal carrier matching and PA scheme to minimize the end-to-end distortion in video transmission applications. Matching-based schemes for user pairing have been developed for NOMA communication scenarios in [26], aiming to improve the sum rate. In order to solve the RA problem for device-to-device (D2D) communications underlaying cellular networks, [27] proposed a many-to-many matching algorithm for obtaining a sub-optimal solution to maximize the system sum rate. In [28], the matching theory is adopted to maximize the cumulative sum of the mean opinion score (MOS). Compared with the above matching methods, we comprehensively propose a novel VSU-based matching algorithm to solve the multi-UA problem.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we firstly introduce the system model of the MDS coded cache-enabled HetNets, and then formulate a problem to minimize the power consumption.

To preserve the readability of the paper, some major notations are summarized in Table 1.

A. SYSTEM MODEL

We consider a downlink HetNets as shown in Fig. 1. The MBS also known as the control BS, provides the coverage and supports efficient radio resource control procedures, while SBSs, known as data BSs, provide high rate data transmission.

TABLE 1. Main notations.

Notation	Description
K	Number of SBSs
U	Number of users
N	File library size
B	Size of each file
C	Cache capacity of each SBS
q_{jk}	Cache fraction of file F_j at SBS k
x_{ku}	If user u associate with k, $x_{ku} = 1$
R_0	The minimum rate requirement for the user
R_{ku}	The achievable transmit rate from SBS k to user u
e_{MBS}	Backhaul energy consumption efficiency (J/bit)
q_u	Each user is allowed to be served by at most q_u SBSs
P_{ku}	The power resource that SBS k allocates to user u
W_{ku}	The bandwidth resource that SBS k allocates to user u
P_{tr}	The transmission power consumption of SBSs
P_{bh}	The power consumption of backhaul links
P_{st}	The static power consumption of SBSs
P_{ca}	The cache power consumption of SBSs
w_{ca}^{k}	Cache power efficiency of SBS k (watt/bit)
P_{st}^{k}	Static power of SBS k

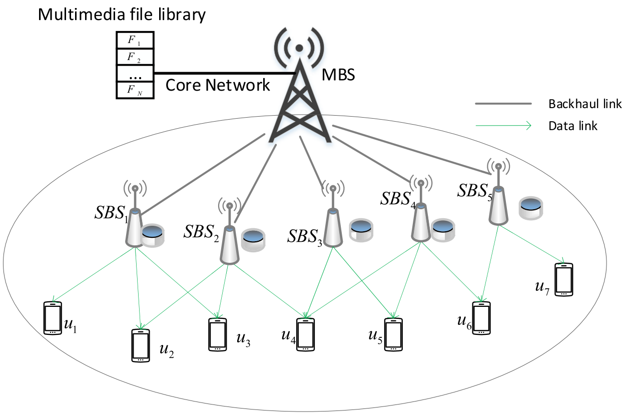


FIGURE 1. System model of the cache-enabled HetNets.

1) TRANSMISSION MODEL

As shown in Fig. 1, a two-tier macrocell-small cell HetNets where one MBS is overlaid by SBSs. Let $\mathcal{K}$ and $\mathcal{U}$ denote the set of K SBSs and the set of U users, respectively. The MBS has access to a library of N files, and each file $\mathcal{F} = \{F_1, \ldots, F_N\}$ has the same size of B (bits). Each SBS is equipped with a cache device which has a capacity of C bits. A SBS can be viewed as a relay and connects to the MBS via backhaul links. We assume that each user $u \in \mathcal{U}$ can be served simultaneously by $q_u(1 \leq q_u \leq K)$ SBSs. If the requested file is not completely retrieved from the associated SBSs, the missing proportion of the file has to be transported from the MBS via the backhaul links.

The SINR of the signal received by user $u \in U$ from the serving SBS $k \in K$ is given by

$$SINR_{ku} = \frac{P_{ku}g_{ku}}{\sum\limits_{v \in K, v \neq k} P_{vu}g_{vu} + \sigma^2} \tag{1}$$

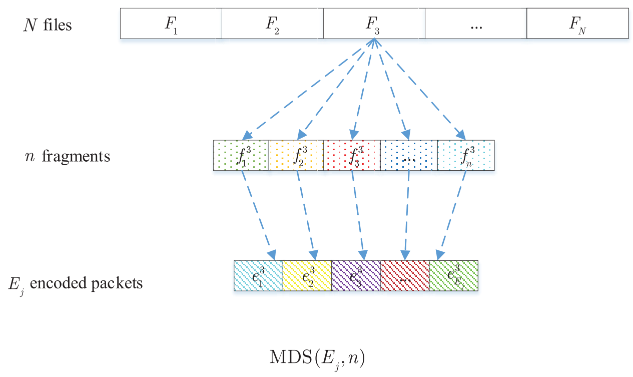


FIGURE 2. Illustration of MDS encoded caching in SBSs.

where P_{ku} is the transmit power from SBS k to user u, and g_{ku} denotes the channel gain from SBS k to user u. σ^2 is the variance of additive white Gaussian noise. $\sum\limits_{v \in K, v \neq k} P_{vu} g_{vu}$ denotes the inter-SBSs interference I_{vu} of user u received from all other nonserving SBSs. The achievable transmit rate from SBS k to user u is given by:

$$R_{ku} = W_{ku} \log_2(1 + SINR_{ku}) \tag{2}$$

where W_{ku} denotes the bandwidth that SBS k allocates to user u.

2) CACHE MODEL

We apply the MDS-coded caching scheme at SBSs. The cached content in different SBSs needs to be coordinated. In contrast to the case of caching uncoded fragments, the benefit from MDS codes is that the encoded packets are all independent from each other so that a certain number of randomly encoded packets will be sufficient to recover the file.

An illustration of the MDS coding scheme is shown in Fig. 2. Each file $F_j (1 \leq j \leq N)$ is split into n fragments, i.e., $F_j = \left\{ f_1^j, \ldots, f_n^j \right\}$. The n fragments are encoded into E_j packets $\left\{ e_1^j, \ldots, e_{E_j}^j \right\}$ [13]. In the *placement phase*, an encoded packet $e_{E_j}^j$ is sent to each SBS independently. Each SBS stores equal fraction of file F_j. Let x_{ku} be the association variable between user u and SBS k. If user u is associated with SBS k, $x_{ku} = 1$; otherwise $x_{ku} = 0$. In the *delivery phase*, a user requesting the file F_j contacts q_u SBSs, and $q_u = \sum\limits_{k \in K} x_{ku}$. we define q_{jk} as the cached fraction of the file F_j at each SBS k, i.e., the user u can receive a total of $\sum\limits_{k \in K} x_{ku} q_{jk}$ fraction of the requested file F_j from its associated $\sum\limits_{k \in K} x_{ku}$ SBSs. If $\sum\limits_{k \in K} x_{ku} q_{jk} \geq 1$, the user u can recover the requested file F_j according to the property of MDS coding. Otherwise, the MBS has to send the remaining fraction $(1 - \sum\limits_{k \in K} x_{ku} q_{jk})$ of the requested file to the SBSs.

For example, the user u_4 in Fig. 1 requests the file F_j from $\sum\limits_{k \in K} x_{ku} = 3$ SBSs, i.e., SBS_2, SBS_3 and SBS_4. The user receives at most a fraction $3q_{jk}$ of the encoded file F_j. If $3q_{jk} < 1$, the MBS has to send the missing $(1 - 3q_{jk})$ fraction of the requested file to the associated SBSs via backhaul links. After receiving the residual fraction of the file from the MBS, the associated SBSs act as relays, and send the missing encoded packets of the requested file to the user.

3) POWER CONSUMPTION MODEL

The system power consumption can be divided into four parts: transmission power, caching power, static power and backhaul power, which will be detailed in the following.

The transmission power of SBSs is determined by UA indicators x_{ku} and PA variables P_{ku}, which is given by

$$P_{tr} = \sum_{k \in K} \sum_{u \in U} P_{ku} x_{ku} \tag{3}$$

Moreover, the maximum transmission power of each SBS should not exceed its power constraint P_k^{max}.

The caching power of each SBS is proportional to the cache power efficiency $w_{ca}^k (watt/bit)$, the caching power[1] of all SBSs is given by

$$P_{ca} = C \cdot \sum_{k \in K} w_{ca}^k \tag{4}$$

The static power P_{st}^k is the power consumed by SBS k for baseband processing and cooling. The static power of all SBSs is given by

$$P_{st} = \sum_{k \in K} P_{st}^k \tag{5}$$

The power consumption of backhaul links is given by

$$P_{bh} = e_{MBS} \cdot R_0 \cdot \sum_{u \in U} \max(1 - \sum_{k \in K} x_{ku} q_{jk}, 0) \tag{6}$$

where R_0 denotes the minimum rate requirement of the user, e_{MBS} (J/bit) is the backhaul energy consumption efficiency, and q_{jk} indicates the fraction of the requested file F_j. Only if caching at SBSs reduces the amount of encoded packets to be transmitted from MBS to SBSs, the backhaul power P_{bh} can be reduced. Under the assumption that the cache fraction q_{jk} is fixed, P_{bh} mainly depends on $\sum\limits_{k \in K} x_{ku}$, i.e., the number of SBSs each user u is connected. That is to say, the UA indicators x_{ku} determine the backhaul power consumption as well.

Based on the analysis above, the power consumption of the network is given by

$$P_{total} = P_{tr} + P_{ca} + P_{st} + P_{bh} \tag{7}$$

[1] Since we focus on the joint UA and RA optimization problem under given MDS coded caching strategy, where cache updating is not involved, the energy consumption for content updating in the caches can be ignored.

B. PROBLEM FORMULATION

Given that the MDS-coded caching placement is known, our goal is to minimize the total power consumption by joint optimization of UA, PA and BA. We formulate a power consumption minimization (PCM) problem as follows:

$$\min_{x_{ku},P_{ku},W_{ku}} \quad P_{total} \tag{8}$$

$$\text{s.t.} \sum_{k\in K} R_{ku} \geq R_0, \quad \forall u \tag{8a}$$

$$\sum_{u\in U} P_{ku}x_{ku} \leq P_k^{\max}, \quad \forall k \tag{8b}$$

$$\sum_{u\in U} W_{ku}x_{ku} \leq W_k^{\max}, \quad \forall k \tag{8c}$$

$$\sum_{k\in K} x_{ku} \leq q_u, \quad \forall u \tag{8d}$$

$$P_{ku} \geq 0, \quad \forall (k,u) \in \mathcal{K}\times\mathcal{U} \tag{8e}$$

$$W_{ku} \geq 0, \quad \forall (k,u) \in \mathcal{K}\times\mathcal{U} \tag{8f}$$

$$x_{ku} \in \{0,1\}, \quad \forall (k,u) \in \mathcal{K}\times\mathcal{U} \tag{8g}$$

$$0 \leq q_{jk} \leq 1, \quad \forall j \in \mathcal{F},\ k \in \mathcal{K} \tag{8h}$$

where (8) guarantees the QoS requirement R_0 for each user. Constraint (8) is the maximum power limit for each SBS. (8) denotes that the total bandwidth that each SBS allocates to its associated users cannot exceed its available bandwidth. (8) states that each user is allowed to be served by $q_u(1 \leq q_u \leq K)$ SBSs. (8) and (8) impose non-negativity constraint on power and bandwidth variables. The constraint (8) keeps the association indicators x_{ku} binary. (8) limits the cache fraction of the requested file at each SBS.

The discrete nature of user association (i.e., indicators x_{ku}) and the continuous nature of resource assignment (i.e., variables P_{ku} and W_{ku}), lead the problem (8) to a MINLP problem. This type of problem even in conventional single-association HetNets is intractable. When it comes to the case of multi-association systems, due to the superimposition of multiple SBSs on the same user, the solution of this problem is more complicated. Therefore, to solve the problem (8), we propose two heuristic algorithms. In the first method, based on the assumption that the available bandwidth of a SBS is subdivided equally among its associated users, the original PCM problem is simplified to the UA and PA subproblems. It's found that the two sided matching model is appropriate to capture the structure of the UA subproblem. Since one user can be assigned with multiple SBSs and one SBS can serve multiple users, a many-to-many matching scheme is adopted. For the second method, under the unequal BA, we propose a three-phase iterative algorithm that optimizes UA, BA and PA successively. Based on the characteristic of the BA and PA subproblems, we adopt bisection and LP to solve the problem.

IV. JOINT OPTIMIZATION OF USER ASSOCIATION AND POWER ALLOCATION

To reduce the computational complexity, first, we adopt an equal share strategy to allocate the bandwidth of each SBS. The problem (8) is transformed into

$$\min_{x_{ku},P_{ku}} \quad P_{total} \tag{9}$$

$$\text{s.t.} \frac{W_k}{\sum\limits_{u\in U} x_{ku}} \log_2(1+SINR_{ku}) \geq \frac{R_0}{\sum\limits_{k\in K} x_{ku}}, \quad \forall u \tag{9a}$$

$$W_{ku} = \frac{W_k}{\sum\limits_{u\in U} x_{ku}}, \quad \forall (k,u) \in \mathcal{K}\times\mathcal{U} \tag{9b}$$

$$(8b), (8c), (8d), (8e), (8f), (8g), (8h) \tag{9c}$$

The constraint (8) sets the rate requirement of each user, i.e., the user u achieves equal rate $\frac{R_0}{\sum_{k\in K} x_{ku}}$ from its associated SBSs. $W_k^{\max}$ is the available bandwidth of SBS k, and $\sum_{u\in U} x_{ku}$ denotes the number of users served by the SBS k. (8) denotes that each user u associated with the same SBS k is allocated with equal bandwidth. The discrete UA decision x_{ku} and the PA variables P_{ku} should be determined by solving problem (8). Without loss of generality, the output of the UA decision x_{ku} is the prerequisite of the PA variables. The problem (8) is challenging because the objective is nonconvex and x_{ku} and P_{ku} are mix integer variables. In the following, we divide (8) into two subproblems, i.e, UA and PA. We resort to LP and VSU-based matching to cope with the two subproblems, respectively.

A. LINEAR PROGRAMMING FOR POWER ALLOCATION

If user association x_{ku} is fixed, the first term in the objective function (8) along with the following four constraints (8), (8), (8) and (8) forms the SBS transmission power problem which is independent of backhaul power consumption. Actually, the backhaul power consumption turns to a constant due to the given UA indicators x_{ku} and cache fraction q_{jk}. In addition, according to (7), the power consumption of each SBS k is determined by the dynamic transmission power consumption P_{tr}, with P_{ca} and P_{st} fixed for SBSs. Then, the reformulation of the PA subproblem for all SBSs can be rewritten as

$$\min_{P_{ku}} \sum_{k\in K}\sum_{u\in U} P_{ku}x_{ku} \tag{10}$$

$$\text{s.t.}\ c \sum_{j\in K, j\neq k} P_{ju}g_{ju} - P_{ku}g_{ku} + c\sigma^2 \leq 0, \quad \forall u \tag{10a}$$

$$\sum_{u\in U} P_{ku}x_{ku} \leq P_k^{\max}, \quad \forall k \tag{10b}$$

Here, constants a, b and c are defined as $a = \frac{W_k^{\max}}{\sum_{u\in U} x_{ku}}$, $b = \frac{R_0}{\sum_{k\in K} x_{ku}}$, $c = (2^{b/a} - 1)$, respectively. Obviously, (8) is a linear combination of power variables P_{ku}. (8) and (8) are affine functions of power variables P_{ku}. Therefore, the problem (8) is transferred to solving a tractable LP.

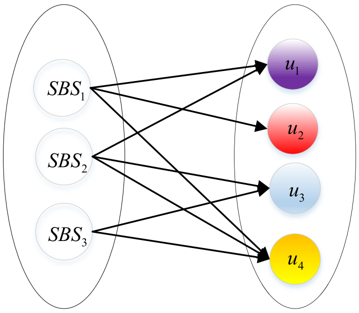


FIGURE 3. **Illustration of many to many matching.**

B. MANY TO MANY MATCHING GAME FORMULATION FOR USER ASSOCIATION

Given the resource constraint of each SBS, the UA subproblem is NP-hard as well. The characteristic of the multi-user association subproblem implies that many-to-many matching model is appropriate to solve this problem. We define users in set $\mathcal{U}$ and SBSs in set $\mathcal{K}$ as two sets of players in this many-to-many matching relation. In the following, we describe the method for preference calculation for different players and then demonstrate the VSU-based swap matching. Finally, the stability of the proposed VSU-based swap matching algorithm is analyzed.

We assume $\mathcal{U}$ and $\mathcal{K}$ are two disjoint sets of selfish and rational players, which aim to maximize their own utilities. If user u is assigned to SBS k, then we say u and k are matched with each other and a matching pair μ is formed [26]–[29]. The UA subproblem is reformulated as a many-to-many matching model (Fig. 3) defined as follows [26]–[28]:

Definition 1 (Many-to-Many Matching): A matching μ is the outcome of the considered user association subproblem and can be defined as a function from the sets $\mathcal{U} \bigcup \mathcal{K}$ into the set of all subsets of $\mathcal{U} \bigcup \mathcal{K}$ such that every $u \in \mathcal{U}$ and $k \in \mathcal{K}$ satisfy the following constraints:

a) $\mu\ (u) \subseteq \mathcal{K}$;

b) $\mu\ (k) \subseteq \mathcal{U}$;

c) $|\mu(u)| \leq q_u, \forall u \in \mathcal{U}$

d) $|\mu(k)| \leq \infty, \forall k \in \mathcal{K}$

e) $u \in \mu(k)$ if and only if $k \in \mu(u)$

where $\mu\ (u)$ is the set of partners for user u and $\mu\ (k)$ is the set of partners for SBS k under the matching model μ. The condition a) indicates that each user is matched with a subset of SBSs; b) implies that each SBS is matched with a subset of users; conditions c) and d) set the quoto of $\mu\ (u)$ and $\mu\ (k)$, respectively. Each user can be associated with up to q_u SBSs, and each SBS can serve multiple users at the same time.

To model the matching μ, the preferences of the players (i.e., users and SBSs) need to be well defined, which are described as follows.

1) USER's PREFERENCES

From the user's perspective, each user u seeks to maximize its own utility $V_u(k)$ which is determined by its achieved rate R_{ku}. In order to maximize the total utilities, each user tends to calculate the preferences over the SBS by ranking the SBS set depending on the value of their utilities.

2) SBS's PREFERENCES

From the SBS's perspective, the preference lists of SBSs can be computed based on the transmission power that can provide to it's associated users. Note that, the preference of SBS k is based on the benefit $V_k(u)$ (i.e., the reverse of power cost), i.e., if the SBS k chooses the user u, this user accepts this SBS if and only if the power consumption is reduced by this assignment.

Remark 1: The UA subproblem reformulated as many to many matching game has *externalities*, also known as *peer effects*.

Since each user can be associated with q_u SBSs, and each SBS can be matched with a subset of users, the users rank the SBSs using their predefined utilities and then form their preference lists.

According to the equation (2), the preference value $V_u(k)$ for the user u is a function of the interference from other users. Due to the interference, the preference value $V_u(k)$ not only depends on the SBSs they matched with, but relates to other users. We define this type of matching as the matching game with externalities [26], where each player has a dynamic preference list over the opposite set of players. This is different from the conventional matching games in which players have fixed preference lists [30]. However, the matching problem of dynamic quotas motivates us to develop a new scheme that significantly differs from existing deferred acceptance (DA) approaches in wireless networks. At first, we introduce the concept of *swap matching*, which is defined as follows:

Definition 2 (Swap Matching): For a given pair of association $u_p \in \mu(k_i)$, $u_q \in \mu(k_j)$ in a matching μ, where $u_p \notin \mu(k_j)$, $u_q \notin \mu(k_i)$, a swap matching $\mu_{jq}^{ip} = \mu \backslash \left\{(k_i, u_p), (k_j, u_q)\right\} \cup \left\{(k_i, u_q), (k_j, u_p)\right\}$ can be defined as $u_p \in \mu_{jq}^{ip}(k_j)$, $u_q \in \mu_{jq}^{ip}(k_i)$ and $u_p \notin \mu_{jq}^{ip}(k_i)$, $u_q \notin \mu_{jq}^{ip}(k_j)$.

In the definition 2, the word "*swap matching*" has two transformation forms, i.e., "swapping" two users accessed to different SBSs or "switching" one user to other available SBS [28]. For brevity, we use "*swap matching*" to characterize all types of "transfer matching" in this paper. In other words, a *swap matching* enables users u_p and u_q to swap one of their matched SBSs, while keeping other users' and SBSs' matchings unchanged. In particular, since we introduce "virtual" users (SBSs), one of the users (SBSs) involved in the swap can be a "virtual" user (SBS). However, due to the players' selfish interests, players involved in the swap operation may not be approved. Next, by introducing "*swap blocking pair*", we are interested to obtain a stable state, in which there are no players that are not matched to one another but they all prefer to be partners.

Definition 3 (Swap Blocking Pair): Given a matching μ and a SBS pair $(k_i, k_j) \in \mathcal{K}$ with k_i and k_j matched in $\mu(k_i)$ and $\mu(k_j)$, respectively. If there exist the user $u_p \in \mu(k_i)$ and the user $u_q \in \mu(k_j)$ satisfying:

a)$\forall x \in \left\{k_i, k_j, u_p, u_q\right\}, V_x(\mu_{jq}^{ip}) \geq V_x(\mu)$

b)$\exists x \in \left\{k_i, k_j, u_p, u_q\right\}, V_x(\mu_{jq}^{ip}) > V_x(\mu)$

the swap matching μ_{jq}^{ip} is approved. (k_i, k_j) or (u_p, u_q) is defined as a *swap blocking pair* in μ.

The definition 3 implies that the swap operations are executed between *swap blocking pairs*, and the utilities of all the involved players should not be decreased after swap operations or at least one player's utility will increase, i.e., swap operations allow a given SBS k to swap its matching if and only if it is beneficial for achieving high utility.

The swap operations are performed within a finite number of iterations. The user preferences do not change in one iteration, and each user should rebuild its preference list once the matching game is over for all users. Thus, the iterative swap operations continue until no new *swap blocking pairs* are required. Since *swap matching* allows swap operations after a matching decision has been made, users can update their preference lists based on the new interference conditions resulting from other users' matching. This helps the users to try for a different SBS k_j that may provide lower transmission power than the current association SBS k_i.

Remark 2: The objective is to minimize the total power consumption for downlink transmission in the network. Since the centralized scheme assumes the availability of all information and calculates the associations centrally, the incentives of the users or SBSs do not need to be considered separately. During swap operations, the MBS checks whether the two users (SBSs) can benefit each other by swap their current matchings. In a nutshell, two arbitrary users (SBSs) can be arranged by the MBS to form a *swap blocking pair* and the *externalities* are well handled [26].

Consequently, the MBS ensures that the players keep executing approved swap operations in order to achieve a stable state, called as a *two-sided swap stable*, which is defined as follows [29]:

Definition 4 (Two-Sided Swap Stable): A matching μ is *two-sided swap stable* if there does not exist a *swap-blocking pair* (k_i, k_j).

Proposition 1: The swap matching algorithm is guaranteed to converge to a pair wise stable matching.

Proof: We prove this proposition by contradiction. Suppose that there exist a user u and a SBS k with $u \notin \mu(k)$ and $k \notin \mu(u)$ that block our matching μ, such that both $u \succ_k \mu(k)$ and $k \succ_u \mu(u)$ are satisfied simultaneously. If $u \succ_k \mu(k)$ is true, it means that the SBS k must propose to be paired with the user u in some earlier swap operations. However, at the same time, both $u \notin \mu(k)$ and $k \notin \mu(u)$ are true. Then, in the latter swap operations, at the proposal time of SBS k, only user u that has a higher utility value than $V_k(u)$ can be associated to the SBS k which means $\mu(k) \succ_k u$. This contradicts the initial supposition $u \succ_k \mu(k)$. Therefore, there is no blocking pair in the final matching μ, i.e., μ is stable. ■

C. THE PROPOSED VSU-BASED MATCHING ALGORITHM

Based on the above definitions, we propose a novel VSU-based swap matching algorithm to solve the UA subproblem. Note that, the virtual SBSs and users are named as $\mathcal{V}_s$ and $\mathcal{V}_u$, respectively. The proposed VSU-based matching algorithm is detailed in Step 1 of Algorithm 1, which consisting of three steps: initialization phase, VSU-based swap-matching phase, and final matching phase output.

Algorithm 1 Joint User Association and Power Allocation With VSU-Based Swap Matching Algorithm (JUPVA)

```
Step 1: User Association
    Step 1.1: Initialization
        1) Each user $u$ discovers the SBSs in the vicinity;
        2) SBSs $\mathcal{K} \cup \mathcal{V}_s$ and users $\mathcal{U} \cup \mathcal{V}_u$ are randomly matched with each other subject to constraints.
    Step 1.2: VSU-based swap-matching phase
        In each round, for each matched user $u_p \in \mathcal{U} \cup \mathcal{V}_u$
        1) The SBS $k_i \in \mathcal{K} \cup \mathcal{V}_s$ searches for another user $u_q \in \mathcal{U} \cup \mathcal{V}_u \backslash \{u_p\}$ to form a swap-blocking pair $(u_p, u_q)$ along with $u_p \in \mu(k_i)$, $u_q \in \mu(k_j)$.
        2) If $\mu_{jq}^{ip}$ is approved, user $u_p$ exchange its match $k_i$ with $u_q$ for $k_j$, update the current matching state to $\mu = \mu_{jq}^{ip}$
        3) Else if there does not exist such a blocking pair, $u_p$ keeps it matching state.
        4) Repeat step 2 until there is no blocking pair in the current round, iterations will not stop until no user can form a swap blocking pair with any other users in a new round.
Step 2: Power Allocation.
        Repeat
        1) Update the user association matrix $\mathbf{X}$ by solving the VSU-based swap matching problem above.
        2) Update power level $\mathbf{P}$ by solving LP using the matlab software.
        Until convergence
Step 3 :End of algorithm.
```

Initially, the SBSs and users randomly match with each other subject to $|\mu(u)| \leq q_u$ and $|\mu(k)| \leq \infty$ [31]. Then, each user u_p keeps searching for other users $\mathcal{U} \cup \mathcal{V}_u \backslash u_p$ to check whether there is a *swap-blocking pair*, i.e., the swap process is to keep searching approved swap matchings μ_{jq}^{ip} among the players so as to reach a *two-sided swap stable* matching. Unlike traditional matching models reformulated to solve UA subproblem, users (SBSs) involved in swap matchings may be a virtual user (SBS) in this paper. However, we do not consider the utilities of $\mathcal{V}_s$ and $\mathcal{V}_u$. Therefore, different from [26], [27], since we introduce $\mathcal{V}_s$ and $\mathcal{V}_u$, the degree of freedom for swap matchings is more extensive, and the swap space contains: both SBSs and users that belong to potential *swap-blocking pairs* are virtual; one of those SBSs (users) that belongs to potential *swap-blocking pairs* is virtual; neither SBSs nor users that belongs to potential *swap-blocking pairs* is virtual, i.e., the diversity of potential *swap-blocking pairs* keeps searching the lower total power consumption after each swap operation. Finally, the swap matching operations terminate when there is no *swap blocking pair*, and the final *two-sided swap stable* state is reached.

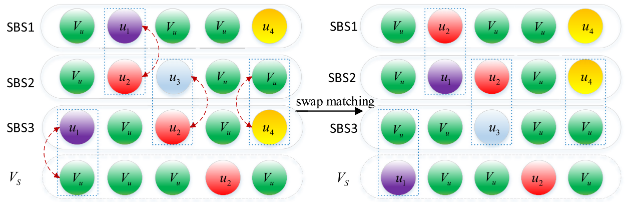


FIGURE 4. **Illustration of swap matching operations.**

For a better understanding of the VSU-based matching, we show the illustration of swap matching operations in Fig. 4. Through greedily swap operations among all users $(\mathcal{U} \cup \mathcal{V}_u)$ within infinite domain, the VSU-based matching algorithm proceeds with reducing the total power consumption from initial random matching. The proposed algorithm must obey the quoto criterion and guarantee each user's rate requirement R_0 after each swap operation.

Proposition 2: Starting from the initial randomly association state, the proposed VSU-based matching algorithm is guaranteed to converge to a final matching.

Proof: The proposed VSU-based swap matching is developed to achieve the two-sided swap stability. Since the number of SBSs and users in the HetNets are finite, a SBS cannot have infinite gains, which implies a SBS can carry on a fixed number of swaps to reach the maximum gain, i.e., the minimum power consumption. Therefore, the algorithm is guaranteed to terminate after a finite number of iterations and the termination occurs when there is no *swap blocking pair* to improve his current association, i.e., no user will have desire to deviate from the current associations. That is, the swap operation continues if and only if at least one association (SBS-user matching) is executed. Therefore, in every swap, if the system power consumption is reduced, eventually, the process approaches the convergence state as the solution of the problem exists in the finite domain. Consequently, after finite number of iterations, the algorithm converges and terminates. ■

V. JOINT OPTIMIZATION OF USER ASSOCIATION, BANDWIDTH ALLOCATION, AND POWER ALLOCATION

The problem (8) is non-convex with respect to variables (i.e., P_{ku} and W_{ku}) and is also tightly coupled with UA indicators X_{ku}. Compared with the problem (8), the problem (8) has a higher computation complexity. To achieve its global optimal solution, we need to fully search the feasible resource space along with all possible combinations of UAs. Thus, even for a centralized system, it may be infeasible to solve the problem (8) at each association slot. In this section, we propose a three-phase iterative algorithm (JURVA) to solve the problem (8). First, by adopting the Lagrange decomposition dual method, the BA is transformed to a convex optimization problem, and then solved by bisection algorithm. Second, we design the iterative algorithm for the joint UA, BA and PA optimization problem.

A. JOINT OPTIMIZATION OF USER ASSOCIATION AND BANDWIDTH ALLOCATION

Similar as the assumption in [32], [33], we assume that the UA is carried out in a large time scale compared to the change of channel. That is, the channel can be regarded as almost stationary during the UA and RA period. Thus, we use the maximum tolerable interference to simplify the interference value $\sum_{v \in K, v \neq k} P_{vu} g_{vu}$ and assume the interference experienced by the user u when associated with SBS k as a constant [32], [34]. With the help of this approximation method and Lagrange decomposition dual method, the original transmission power consumption of all SBSs can be transformed into subproblems that can be independently solved by each SBS.

Thus, under fixed user association, the original PCM problem can be transformed to optimize the transmission power of each SBS individually. For the SBS $k \in \mathcal{K}$, we denote $\mathcal{K}_k$ as the set of users associated with it. The optimal BA subproblem to minimize the power consumption for each SBS k can be formulated as

$$\min_{W_{ku}} \sum_{u \in U} x_{ku} P_{ku} = \sum_{u \in \mathcal{K}_k} \left(2^{\bar{R}_u / W_{ku}} - 1\right) \frac{\sigma^2 + I_{vu}}{g_{ku}} \tag{11}$$

$$\text{s.t. } W_{ku} \log_2(1 + SINR_{ku}) = \bar{R}_u, \quad \forall u \tag{11a}$$

$$\sum_{u \in \mathcal{K}_k} W_{ku} = W_k^{\max} \tag{11b}$$

$$W_{ku} > 0, \quad \forall u \in \mathcal{K}_k \tag{11c}$$

$$P_{ku} > 0, \quad \forall u \in \mathcal{K}_k \tag{11d}$$

where, $\bar{R}_u = \frac{R_0}{\sum_{k \in K} x_{ku}}$ denotes the average rate requirement of user u. The equality constraint (8) replaces the inequality constraint (8) in (8) to fully exploit the available bandwidth of SBS k. (8) and (8) denote all associated users in SBS k should be assigned bandwidth resource and power resource, respectively.

Theorem: (8) is a convex problem.

Proof: When $W_{ku} > 0$,

$$\frac{\partial^2 [(2^{\bar{R}_u / W_{ku}} - 1) \frac{\sigma^2 + I_{vu}}{g_{ku}}]}{\partial^2 W_{ku}} > 0 \tag{12}$$

Thus, $P_{ku} = (2^{\bar{R}_u / W_{ku}} - 1) \frac{\sigma^2 + I_{vu}}{g_{ku}}$ is a convex function with respect to W_{ku}. According to [35], the constraints (8) and (8) of the problem (8) satisfy Slaters conditions. It is clear that (8) defines a convex problem. ■

We adopt Karush-Kuhn-Tucker (KKT) conditions of (8) to obtain the optimal solution. Let λ and μ_{ku} denote the dual variables for the constraint (8) and (8), respectively. The Lagrangian function is given as follows

$$\mathcal{L}(W_{ku}, \lambda, \mu_{ku}) = \sum_{u \in \mathcal{K}_k} \left(2^{\bar{R}_u / W_{ku}} - 1\right) \frac{\sigma^2 + I_{vu}}{g_{ku}} + \lambda \left(\sum_{u \in \mathcal{K}_k} W_{ku} - W_k^{\max} \right) - \sum_{u \in \mathcal{K}_k} \mu_u W_{ku} \tag{13}$$

Here, we define W_{ku}^*, λ^* and μ_{ku}^* as the primal and the dual optimal points with zero duality gap for user $u \in \mathcal{K}_k$. Based on the KKT conditions, the following equations should be satisfied:

$$2^{\bar{R}_u/W_{ku}^*}\frac{(\sigma^2+I_{vu})\bar{R}_u \ln 2}{g_{ku}{W_{ku}^*}^2} = \lambda^* \tag{14}$$

$$\sum_{u\in\mathcal{K}_k} W_{ku}^* = W_k^{\max} \tag{15}$$

$$W_{ku}^* > 0 \tag{16}$$

We define the equation

$$\lambda = f(W_{ku}) = 2^{\bar{R}_u/W_{ku}}((\sigma^2+I_{vu})\bar{R}_u \ln 2/g_{ku}W_{ku}^2 \tag{17}$$

Obviously, $\partial f(W_{ku})/\partial W_{ku} < 0$, when $W_{ku} > 0$, i.e., (17) is monotonically decreasing with respect to W_{ku}. Thus, Eqs. (14)-(16) can be solved via the bisection method: Starting from a random λ, W_{ku} can be calculated for users $u \in \mathcal{K}_k$. If $\sum\limits_{u\in\mathcal{K}_k} W_{ku} > W_k^{\max}$, λ is increased; otherwise, λ is decreased. This algorithm terminates when the gap between the sum of W_{ku} and $W_k^{\max}$ is lower than a given threshold. Moreover, this procedure is repeated until the proposed algorithm converges. Given that the binary UA and BA decision variables, i.e., x_{ku} and W'_{ku} are now determined, the next step is to solve the PA subproblem demonstrated in section V-B.

B. POWER ALLOCATION FOR FIXED USER ASSOCIATION AND BANDWIDTH ALLOCATION

This stage allocates the minimum power to each user by taking both network resources and QoS requirements into account. For the given UA and BA, the PA subproblem can be formulated as

$$\min_{P_{ku}} P_{total} \tag{18}$$

$$\text{s.t. } W'_{ku}\log_2(1+SINR_{ku}) \geq \bar{R}_u, \quad \forall u \tag{18a}$$

$$(8b), (8c), (8d), (8e), (8f), (8g), (8h) \tag{18b}$$

Here, we substitute the equally BA variables $W_{ku} = \frac{W_k}{\sum\limits_{u\in U} x_{ku}}$ in (8) with the value W'_{ku}, which has already been worked out via bisection algorithm in V-A. Then, by the aid of the approach proposed in Section III. A, we convert (18) into a LP problem, which can be solved by available software packages [35].

C. JOINT USER ASSOCIATION AND RESOURCE ALLOCATION WITH VSU-BASED MATCHING

In the previous subsections, we have obtained the UA variables through VSU-based swap matching algorithm, and a KKT optimal bandwidth allocation for fixed user association in V-A, as well as the PA with fixed UA and BA in V-B. The proposed three-phase iterative algorithm for joint UA, BA and PA is summarized in Alg. 2, which is referred as JURVA. Note that, during the execution of JURVA, the MBS determines UA, BA and PA, and the decision made by the MBS is broadcasted to all SBSs and all users.

Algorithm 2 Joint User Association, Bandwidth Allocation and Power Allocation With VSU-Based Swap Matching Algorithm (JURVA)

```
Initialization: $t = t + 1$
    1) UA Phase:
        Run the step 1 in Algorithm 1 to obtain association vectors $X_{ku}$.
    2) BA Phase:
        For fixed user association vectors $X_{ku}$, according to (11), obtain the optimal bandwidth allocation indicators $W'_{ku}$ via the bisection method.
    3) PA Phase:
        For the fixed bandwidth allocation indicators $W'_{ku}$, obtain the optimal power allocation $P'_{ku}$ by using Linear programming similar to step 2 in Algorithm 1.
Until convergence
```

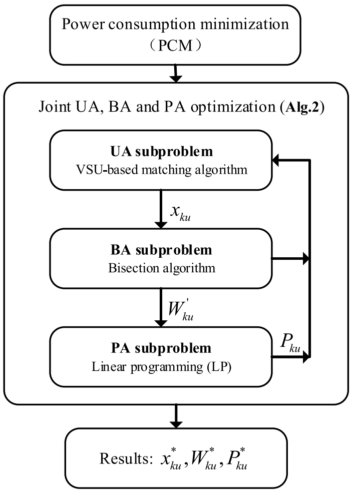

Power consumption minimization （PCM）
Joint UA, BA and PA optimization (Alg.2)
UA subproblem
VSU-based matching algorithm
x_{ku}
BA subproblem
Bisection algorithm
W'_{ku}
PA subproblem
Linear programming (LP)
P_{ku}
Results: $x_{ku}^*, W_{ku}^*, P_{ku}^*$

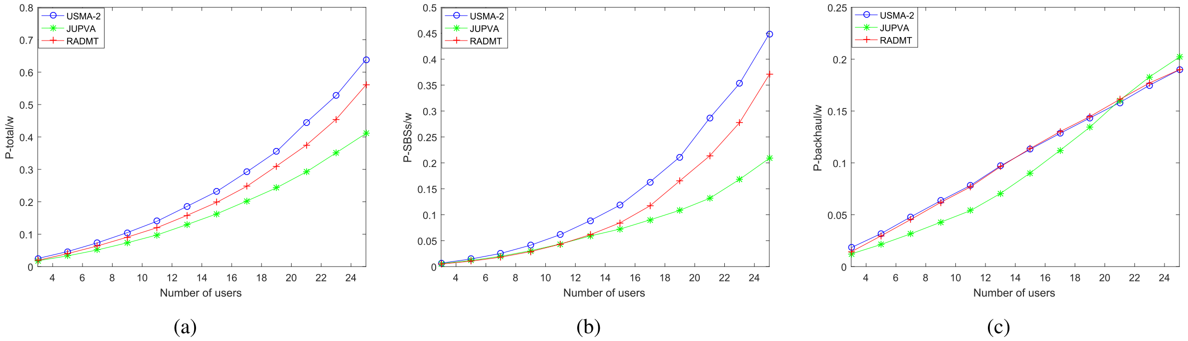
FIGURE 5. Proposed framework for solving the PCM problem.

The proposed centralized framework corresponding to JURVA for joint UA, BA, and PA phases is shown in Fig. 5. The proposed JURVA comprises three main phases: the UA phase, BA phase and PA phase. The UA phase matches the users to the SBSs. Then, the BA phase focuses on the BA of users in the associated SBSs. Finally, the PA phase performs admission controls, bandwidth updating and transmit power allocation. In other words, the MBS allocates the SBSs to each user randomly in the initialization phase. During the resource allocation phase, the MBS performs UA, BA and PA iteratively so as to obtain a joint solution, i.e., x_{ku}^*, W_{ku}^* and P_{ku}^*, which minimize the total power consumption.

VI. SIMULATION RESULTS

A. SIMULATION SETTING

This section presents simulation results to evaluate the proposed algorithms along with other benchmarks for comparison on top of MATLAB and CVX. In the following simulations, six BSs are deployed in a 1km × 1km area. Due to the property of MDS codes, each user is allowed to be served



(a) (b) (c)

FIGURE 6. Comparison on power consumption: (a) Total power consumption vs. number of users (b) Transmission power consumption of SBSs vs. number of users (c) Backhaul power consumption vs. number of users.

TABLE 2. Simulation parameters.

Parameter	Value
Number of MBS	1
Number of SBSs	5
q_u	2,3,4
$W_k^{\max}$	20 MHz
$P_k^{\max}$	25dBm
Shadow fading	8 dB
noise	-111.45 dBm
R_0	5 Mbps
e_{MBS}	$1 \times 10^{-8} J/bit$
Pathloss between SBS and user	$34 + 40\log_{10}d\ (m)$

by multiple SBSs, and the overlapped SBSs are deployed to meet the multi-association condition, which can provide higher probability of re-constructing the content. To realize the overlapping coverage of SBSs, one MBS is located at the center and five SBSs are evenly distributed on the circle with radius $r_1 = 100$ m. Each SBS has a coverage area $\mathcal{S}$ of radius $r_2 = 100$ m. Simulation parameters including channel model and system assumptions are summarized in Table 2.

We introduce two many-to-many matching algorithms [26], [27] to solve the proposed UA subproblem, and compare them to the proposed VSU-based matching algorithm in terms of power consumption. It's worth mentioning that, the matching theory is used to allocate users to sub-channels for NOMA networks [26] and assign D2D pairs to resource blocks in D2D communications [27], respectively. Both of them apply the matching-based algorithms to improve the rate rather than the power cost. However, we adopt them as comparison baselines, because: i) To the best of our knowledge, this is the first work that discuss the joint multi-UA and RA in MDS coded cache-enabled HetNets so as to minimize the total power consumption in terms of transmission and backhaul power consumption. Since the optimized model we considered is different from previous works, we can only choose some conventional technology algorithms [26], [27] as the benchmarks, though it is unfair to do so. ii) The swap matching [26], [27] is an efficient algorithm to tackle a kind of optimization problems. Since the multi-UA problem which is formulated as a many to many two-sided matching with externalities, is similar to the matching models in [26], [27]. This type of matching is more complex than traditional matching problems without externalities, traditional DA matching algorithm could not be applied to solve it. Therefore, [26] and [27] introduce the concept of swap matching. Actually, we improve the swap matching algorithms [26], [27] comprehensively, and propose a novel VSU-based swap matching algorithm, which is proven to be a pairwise stable. Based on the above reasons, we choose [26], [27] as the benchmarks in our study. For convenience, we abbreviate the two matching algorithms [26], [27] for USMA-2 and RADMT, respectively. Note that, the simulation results do not consider the fixed cache and static power consumption.[2]

B. PERFORMANCE EVALUATION

To evaluate the performance of the proposed VSU-based matching scheme, we first set the user's quoto constraint $q_u = 3$. Fig. 6 plots the total power consumption of JUPVA, USMA-2 and RADMT under different number of users. As shown in Fig. 6a, with the number of users increases, the sum power consumption of the three schemes increase, and the proposed JUPVA yields a better performance than the USMA-2 and RADMT schemes. Note that, we adopt the same LP scheme for PA (i.e., step 2 of Alg.1) to the two benchmarks as well. Specifically, when the number of network users is 25, the gains are about 35.25% and 23.64% compared to the USMA-2 and RADMT scheme. The reason is that the VSU-based swap matchings not only contain all types of swap blocking pairs such as "hole" and "dummy" in [26]–[28], but superinduce other "virtual" swap probabilities. In another word, under the circumstances of $\mathcal{V}_s$ and $\mathcal{V}_u$, the proposed JUPVA extends the freedom of swap matching operations and thus results in lower power consumption.

[2] We assume homogeneous SBSs have the same storage space and caching hardware in this paper, during the *cache placement* phase, the storage space in each SBS is fully occupied with MDS encoded packets. i.e. each SBS has equal cache power consumption. Thus, the assumption that the static power and cache power consumption are equal at each SBS is justifiable, and neither of them will affect the resource variables (e.g., x_{ku}, P_{ku} and W_{ku}) in this paper.

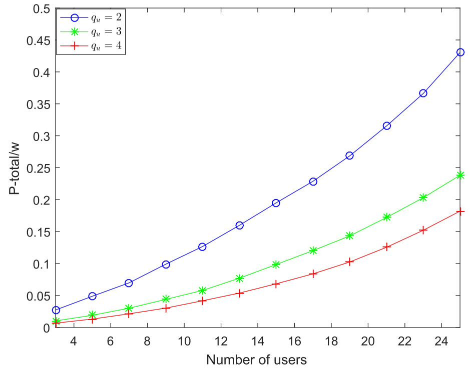


FIGURE 7. Total power consumption v.s. the quoto constraint q_u.

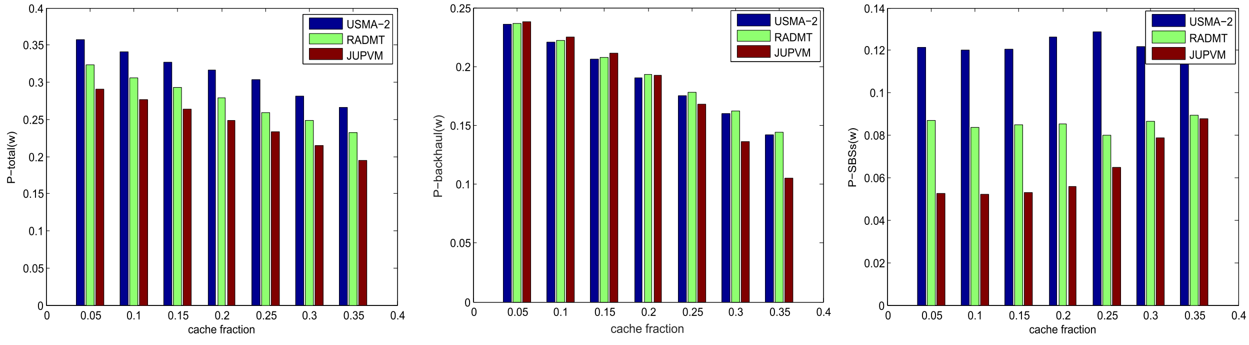
Fig. 6b shows the transmission power consumption versus the number of users. As the number of users increases, the transmission power of SBSs increases, but the power of the proposed JUPVA keeps increasing with the smoothest speed than benchmarks. A key observation from Fig. 6b is that, the transmission power of SBSs is reduced by up to 53.22% and 44.15% compared to RADMT and USMA-2. The reason is that, the introduction of $\mathcal{V}_s$ and $\mathcal{V}_u$ leads to more potential swap blocking pairs between SBSs and users, and the VSU-based matching optimizes the number of swap links effectively, which ensures that the proposed JUPVA keep tracking the optimal SBSs-users matching found so far, and finally obtain the minimum power value. In summary, the VSU-based matching combined with the PA at each SBS, offers a significant advantage in energy-saving at SBSs.

Fig. 6c shows the backhaul power consumption for different schemes. Since we focus on the UA and PA in the proposed JUPVA, from the equation (6), we observe that the backhaul power consumption between the MBS and SBSs only depend on the UA states x_{ku} except for the given q_{jk}, constant e_{MBS} and R_0. Thus, with the increase number of users, the backhaul power value of the three algorithms are almost equal. However, due to the fewer swap matching links in USMA-2 or RADMT, the curves of those are monotonous.

Fig. 7 presents the total power consumption of the proposed JUPVA with different quoto constraints of each user. The quoto constraints of each user are set as $q_u = 2, 3, 4$ and other parameters are the same as those in Table 2. In this paper, since the motivation of the proposed RA scheme is to exploit multi-association property for encoded cache HetNets, with the quoto q_u increasing from 2 to 4, it is possible to reduce the total power consumption, and so thus our algorithm does.

Fig. 8 illustrates the power consumption varying with the cache size. In order to observe that how q_{jk} affects the power consumption of different algorithms, we set the fraction q_{jk} according to a certain distribution to simplify the analysis. When the user's quoto constraint $q_u = 3$ and the number of users $u = 10$, we set the cache fraction of the requested content q_{jk} to obey uniform distribution, and the mean value of q_{jk} are set 0.05, 0.1,..., 0.35, respectively. From Fig. 8a, we observe that the increase of q_{jk} leads to the decrease of the total power consumption. The simulation results are in line with our intuition. This is because the cache fraction of the requested file in the associated SBSs increases, i.e., the increased content hit ratio can reduce the backhaul congestion. As shown in Fig. 8b, the MBS only needs to transmit less fraction content through the backhaul links, which reduces the backhaul power consumption. All of the three algorithms achieve the reduced backhaul power consumption with the increase of content hit ratio.

In Fig. 8c, the proposed JUPVA presents more stable performance than other two benchmarks. The reasons for this result can be summarized into the following aspects. Firstly, though the transmission power of SBSs which depends on the PA variables P_{ku} and UA indicators x_{ku} is independent of the cache fraction value q_{jk}, users can fetch more encoded file packets from the cache of SBSs with the increased cache size. Secondly, the virtual users and SBSs can introduce more potential swap links and then hold the swap matching operations smoothly. Thirdly, the LP applied to PA subproblem optimizes the transmission power of SBSs effectively. Besides, the potential swap blocking pairs in USMA-2 and RADMT are finite, thus, the USMA-2 and RADMT maintain fluctuate with the increasing cache fraction, but the proposed JUPVA maintains an upward trend.



(a) (b) (c)

FIGURE 8. Comparison on power consumption: (a) Total power consumption vs. cache fraction (b) Backhaul power consumption vs. cache fraction (c) Transmission power consumption of SBSs vs. cache fraction.

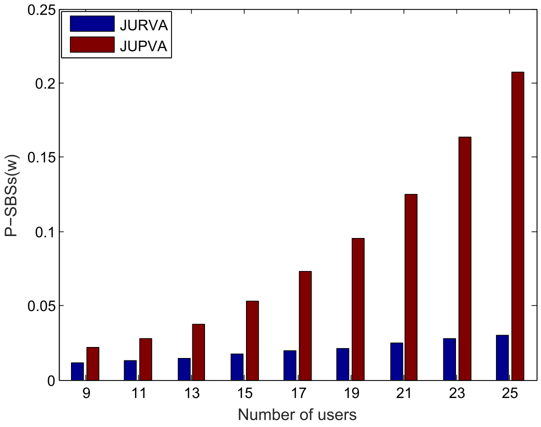


FIGURE 9. The transmission power of SBSs for proposed algorithms.

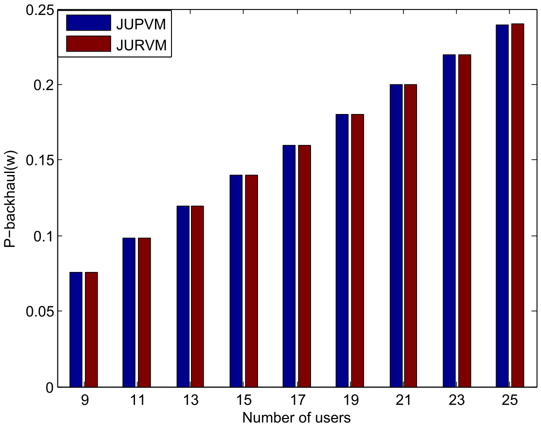


FIGURE 10. The backhaul power consumption for proposed algorithms.

We also set the quoto of user $q_u = 3$, and other parameters are the same as those in Table 2. As shown in Fig. 9, the transmission power consumption of the JURVA scheme can reach up to 46.12%, 66.04%, 85.51% gains over the JUPVA for 9 users, 15 users, 25 users, respectively. As the users gets more densified, the power consumption of SBSs increases. Since our joint algorithms aims to minimize the system power consumption, the associated SBSs that serve users consume the lowest transmission power. Thus, the results verify that the JURVA is more advantageous than JUPVA when BA is combined with UA and PA.

Fig. 10 shows the total backhaul power consumption versus the number of users for the proposed JURVA and JUPVA. In Fig. 10, with the number of users increases, the backhaul power consumption also increases. Actually, for given cache fraction of the requested files, the backhaul power consumption is only related to the actual UA indicators x_{ku}, and RA (i.e., BA and PA) only determines the transmission power. Thus, the backhaul power consumption of JURVA is almost the same as JUPVA. However, since UA indicators x_{ku} are statistically for all users over the backhaul links, the optimized UA links accessed to the SBSs also lead to the lowest backhaul power consumption.

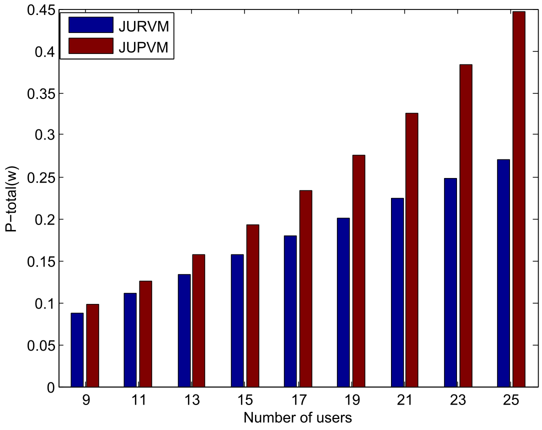


FIGURE 11. The total power consumption for proposed algorithms.

Fig. 11 demonstrates the overall power consumption in terms of transmission power consumed by all SBSs and backhaul power consumption between the MBS and SBSs. Quite evidently, the proposed JURVA consumes less power than JUPVA. As shown in Fig. 10 and Fig. 11, the difference between the overall power consumption and the backhaul power is trivial, which means that the backhaul power accounts for a large proportion of the total power consumption in JURVA. In other words, both the proposed JURVA and JUPVA effectively reduce the transmission power rather than backhaul power consumption. In view of these, we will put forward to the backhaul power minimization problem in future work. Moreover, Fig. 11 further justifies that JURVA is superior to JUPVA in terms of transmission power consumption as well, i.e., considering both BA and PA results in a more power-saving solution than taking the PA into account solely.

C. CONVERGENCE ANALYSIS OF JUPVA AND JURVA

In order to provide practical evidence of the convergence event, Fig. 12 shows the convergence behaviors of the two proposed algorithms within a single snapshot. Fig. 12a shows the convergence of JUPVA and Fig. 12b shows the convergence of JURVA algorithm. In each subfigure of the Fig. 12, we show the system power consumption with the increasing outer-most loop iterations for different number of users U. As observed, in each iteration, the UA and the PA/RA algorithm are implemented once, respectively, and the system power consumption is reduced little by little before reaching the convergence state. As expected, as the number of iterations increases, the power consumption gradually decreases after each iteration, both of those two algorithms converge within 200 iterations even if the number of users $U = 20$. Since the PCM problem discussed herein is non-convex and NP-hard, the optimal solution needs to search all possible feasible solution space. Moreover, in the proposed VSU-based matching model, since the number of players, i.e., (virtual) users and (virtual) SBSs, are large in terms of their associations, it is not feasible to implement in a wireless system. Thus, we mostly focused on *two-sided swap stable*,

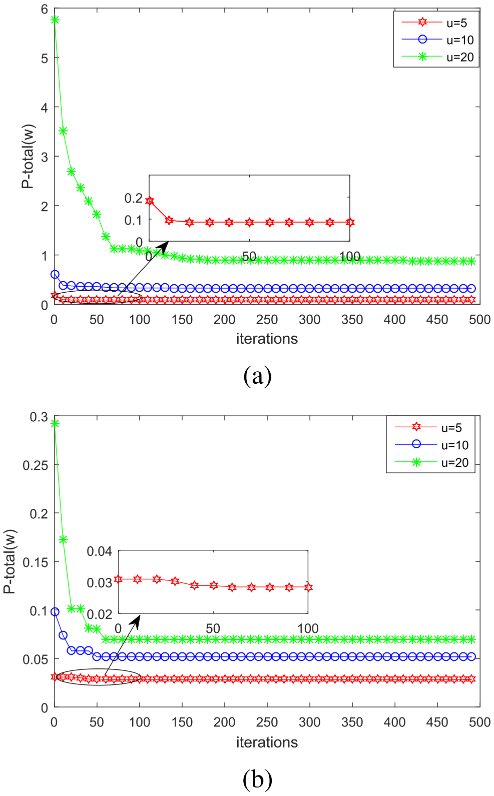


FIGURE 12. **Convergence of proposed algorithm. (a) The convergence of JUPVA. (b) The convergence of JURVA.**

and the VSU-based swap matching algorithm that applied to UA subproblem is developed based on this concept.

In the proposed JUPVA and JURVA, each swap matching reduces the system power consumption compared to the power before that particular matching. And the **step 1.2** of JUPVA is the evidence of this statement. Through swap operations, for a user, we define possible preferred SBSs for the tagged user. This implies that the tagged user will associate to its most preferred SBSs or replace an existing user of that SBSs only if the system power is minimized by this swap operation. Therefore, after each swap process, the system power consumption is minimized. Eventually, the process approaches the convergence state, i.e, the solution of the PCM problem exists in a finite domain. Consequently, after finite number of iterations, the JUPVA/JURVA algorithm converges and terminates.

D. COMPLEXITY ANALYSIS OF JUPVA AND JURVA

Given the convergence of the proposed two algorithms, we then discuss the computational complexity of the proposed VSU-based many to many matching algorithm. For the initialization phase, the complexity mainly lies in the process of sorting the usersąŕ SINR, which is $\mathcal{O}(U^2)$ in average. Note that during the swap-matching phase, a number of iterations are operated to reach the final matching. In every iteration, the MBS searches for swap-blocking pairs and the users execute all the approved swap operations over corresponding SBSs. So the complexity of the swap-matching phase lies in the number of both iterations and attempts of swap matchings in each iteration.

Firstly, we define V_s and V_u as the number of virtual SBSs and users, respectively. And we assume that each user associates to a maximum of q_u SBSs and a SBS can serve q_s users at the same time. Thus, in each iteration of swap-matching phase, at most $\frac{1}{2}\left(U+V_u\right)q_sq_u\left(K+V_s-q_u\right)$ swap matchings need to be considered when $Uq_u = Kq_s$.

Proposition 3: Given the number of total iterations L, the computational complexity of UA process can be approximated as $\mathcal{O}(L\left(U+V_u\right)q_sq_u\left(K+V_s-q_u\right))$.

Proof: When $Uq_u = Kq_s$, each player remains matched before and after every matching, and thus, any swap matching μ_{jq}^{ip} consists of two users and two SBSs. For the user u_p, there exist $q_u\left(K+V_s-q_u\right)$ possible combinations of k_i and k_j in μ_{jq}^{ip} since there are $\left(K+V_s\right)$ SBSs and each user can associate q_u SBSs. On the other hand, for the SBS k_j, at most q_s possible users need to be considered. That is to say, a swap matching μ_{jq}^{ip} with u_p fixed has $q_sq_u\left(K+V_s-q_u\right)$ possible combinations. Since there are $\left(U+V_u\right)$ users, at most $\frac{1}{2}\left(U+V_u\right)q_sq_u\left(K+V_s-q_u\right)$ swap matchings need to be considered in each iteration of UA process. Therefore, given the number of total iterations L, the computational complexity of UA process can be presented by $\mathcal{O}(L\left(U+V_u\right)q_sq_u\left(K+V_s-q_u\right))$. ■

In addition to the UA process, all RA operations occur in constant time, so we can ignore the complexity of the bandwidth and power allocation operations. Mainly, the running time of the UA process dominates the computation time of the entire JUPVA and JURVA algorithms. As shown in Fig. 12a and Fig. 12b, the iterations of this outermost loop is proportional to the number of users U. When the number of users in the system is less, the number of swap operations is relatively less compared to the case when the number of users in the system is higher. In this case, the outermost loop terminates in less number of iterations, which is obvious in Fig. 12a and Fig. 12b.

VII. CONCLUSION

In this paper, a joint optimization of user association, power and bandwidth allocation for MDS encoded cache-enabled HetNets is considered. The aim is to minimize the system power consumption including the transmission power at SBSs and the backhaul power between the MBS and SBSs. At first, by adopting the equal bandwidth allocation approach, we decompose the original problem into a lower level resource allocation problem, i.e., power allocation subproblem via linear programming after solving the user association subproblem by virtual SBSs and users-based many-to-many matching game. Secondly, considering the unequal BA, we propose a three-phase based algorithm to further reduce the power consumption in a centralized and iterative way. Simulation results show the fast convergence of the proposed algorithms, and the advantages of caching coded contents in SBSs as well as the benefits of utilizing virtual SBSs and users-based matching.

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021*, Cisco, San Jose, CA, USA, 2017.

[2] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[3] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[4] Y. Li, T. Jiang, K. Luo, and S. Mao, "Green heterogeneous cloud radio access networks: Potential techniques, performance trade-offs, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 33–39, Nov. 2018.

[5] T. Zhou, Z. Liu, J. Zhao, C. Li, and L. Yang, "Joint user association and power control for load balancing in downlink heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2582–2593, Mar. 2018.

[6] Q. Han, B. Yang, G. Miao, C. Chen, X. Wang, and X. Guan, "Backhaul-aware user association and resource allocation for energy-constrained HetNets," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 580–593, Jan. 2017.

[7] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[8] Z. Zhang, D. Liu, and Y. Yuan, "Layered hierarchical caching for SVC-based HTTP adaptive streaming over C-RAN," in *Proc. IEEE WCNC*, Mar. 2017, pp. 1–6.

[9] X. Song, Y. Geng, X. Meng, J. Liu, W. Lei, and Y. Wen, "Cache-enabled device to device networks with contention-based multimedia delivery," *IEEE Access*, vol. 5, pp. 3228–3239, 2017.

[10] S. Zhang, P. He, K. Suto, P. Yang, and X. S. Shen, "Cooperative edge caching in user-centric clustered mobile networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 8, pp. 1791–1805, Aug. 2018.

[11] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.

[12] J. Liao, K.-K. Wong, Y. Zhang, Z. Zheng, and K. Yang, "Coding, multicast, and cooperation for cache-enabled heterogeneous small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

[13] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3288–3298, Dec. 2016.

[14] Q. Jia, R. Xie, T. Huang, J. Liu, and Y. Liu, "Energy-efficient cooperative coded caching for heterogeneous small cell networks," in *Proc. IEEE INFOCOM WKSHPS*, May 2017, pp. 468–473.

[15] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1794–1808, Feb. 2018.

[16] F. Guo, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Joint optimization of caching and association in energy-harvesting-powered small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6469–6480, Jul. 2018.

[17] K. Guo, C. Yang, T. Liu, and Z. Xiong, "Jointly optimizing user association and BS muting for cache-enabled networks with network-coded multicast and reconstructed interference cancelation," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5539–5553, Nov. 2018.

[18] G. Ren, H. Qu, J. Zhao, S. Zhao, and Z. Luan, "A distributed user association and resource allocation method in cache-enabled small cell networks," *China Commun.*, vol. 14, no. 10, pp. 95–107, 2017.

[19] X. Huang and N. Ansari, "Content caching and distribution in smart grid enabled wireless networks," *IEEE Internet Things J.*, vol. 4, no. 2, pp. 513–520, Apr. 2017.

[20] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: Fundamentals and applications," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 52–59, May 2015.

[21] Q. Pham, T. LeAnh, N. H. Tran, B. J. Park, and C. S. Hong, "Decentralized computation offloading and resource allocation for mobile-edge computing: A matching game approach," *IEEE Access*, vol. 6, pp. 75868–75885, 2018.

[22] A. Abouaomar, A. Kobbane, and S. Cherkaoui, "Matching-game for user-fog assignment," in *Proc. IEEE GLOBECOM*, Dec. 2018, pp. 1–6.

[23] S. Sardellitti, M. Merluzzi, and S. Barbarossa, "Optimal association of mobile users to multi-access edge computing resources," in *Proc. IEEE ICC WKSHPS*, May 2018, pp. 1–6.

[24] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *Proc. 12th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw.*, Hammamet, Tunisia, May 2014, pp. 569–574.

[25] Z. Zhang, D. Liu, and X. Wang, "Joint carrier matching and power allocation for wireless video with general distortion measure," *IEEE Trans. Mobile Comput.*, vol. 17, no. 3, pp. 577–589, Mar. 2018.

[26] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[27] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Many-to-many matching with externalities for device-to-device communications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 1, pp. 138–141, Feb. 2017.

[28] H. Shao, H. Zhao, Y. Sun, J. Zhang, and Y. Xu, "QoE-aware downlink user-cell association in small cell networks: A transfer-matching game theoretic solution with peer effects," *IEEE Access*, vol. 4, pp. 10029–10041, 2016.

[29] E. Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. Int. Symp. Algorithmic Game Theory (SAGT)*, Amalfi, Italy, Oct. 2011, pp. 117–129.

[30] F. Pantisano, M. Bennis, W. Saad, S. Valentin, and M. Debbah, "Matching with externalities for context-aware user-cell association in small cell networks," in *Proc. Globecom Workshops (GC Wkshps)*, Atlanta, GA, USA, Dec. 2013, pp. 4483–4488.

[31] A. E. Roth, "A natural experiment in the organization of entry-level labor markets: Regional markets for new physicians and surgeons in the United Kingdom," *Amer. Econ. Rev.*, vol. 81, pp. 415–425, Jun. 1991.

[32] B. Wang, Q. Kong, W. Liu, and L. T. Yang, "On efficient utilization of green energy in heterogeneous cellular networks," *IEEE Syst. J.*, vol. 11, no. 2, pp. 846–857, Jun. 2017.

[33] G. Ye, H. Zhang, H. Liu, J. Cheng, and V. C. M. Leung, "Energy efficient joint user association and power allocation in a two-tier heterogeneous network," in *Proc. IEEEE Globecom*, Washington, DC, USA, Dec. 2016, pp. 1–5.

[34] Z. Tan, X. Li, F. R. Yu, L. Chen, H. Ji, and V. C. M. Leung, "Joint access selection and resource allocation in cache-enabled HCNs with D2D communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[35] CVX Research Inc. (2015). *CVX: MATLAB Software for Disciplined Convex Programming, Version 3.0 Beta*. [Online]. Available: http://cvxr.com/cvx

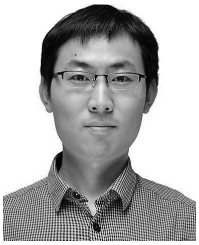
[36] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *Proc. IEEE ICC*, Jun. 2012, pp. 2889–2894.

[37] *Further Advancements for E-UTRA Physical Layer Aspects*, document 3GPP TR36.814, 2017.

FANGFANG YIN received the M.S. degree from the Kunming University of Science and Technology, in 2010. She is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her current research interests include resource allocation for HetNets, convex optimization, and matching theory.

ANYUE WANG received the bachelor's and master's degrees in communication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015 and 2018, respectively. He is currently pursuing the Ph.D. degree with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg. His research interests include resource management in space-air-terrestrial integrated systems, content delivery networks, and NOMA systems for 5G networks and beyond.

DANPU LIU received the Ph.D. degree in communication and electrical systems from the Beijing University of Posts and Telecommunications, Beijing, China, in 1998. She was a Visiting Scholar with the City University of Hong Kong in 2002, University of Manchester in 2005, and Georgia Institute of Technology, in 2014. She is currently with the Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications. Her research involved MIMO, OFDM, and broadband wireless access systems. She has published over 100 papers and three teaching books, and submitted 26 patent applications. Her current research interests include 60-GHz mmWave communication, wireless high definition video transmission, and wireless sensor networks.



ZHILONG ZHANG received the B.E. degree in communication engineering from the University of Science and Technology, Beijing, China, in 2007, and the M.S. and Ph.D. degrees in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, in 2010 and 2016, respectively, where he is currently a Lecturer. From 2010 to 2012, he was a Software Engineer with TD Tech Ltd., Beijing. From 2014 to 2015, he was a Visiting Scholar with Stony Brook University, Stony Brook, NY, USA. His research interests include optimization theory and its applications in wireless video transmission, cross-layer design, and wireless networks.