IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 30, 2016, accepted November 23, 2016, date of publication December 1, 2016, date of current version January 4, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2633488

Joint Caching Placement and User Association for Minimizing User Download Delay

YUE WANG[1], XIAOFENG TAO[1], (Senior Member, IEEE), XUEFEI ZHANG[1], AND GUOQIANG MAO[2], (Senior Member, IEEE)
[1]Wireless Technology Innovation Institute, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]School of Computing and Communications, University of Technology Sydney and National ICT Australia, Sydney, NSW 1466, Australia

Corresponding author: Y. Wang (wang_yue@bupt.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61461136002 and in part by the Joint Research Fund for Overseas Chinese, Hong Kong and Macao Young Scientists of the National Natural Science Foundation of China under Grant 61428102.

ABSTRACT To alleviate the backhaul burden and reduce user-perceived latency, content caching at base stations has been identified as a key technology. However, the caching strategy design at the wireless edge is challenging, especially when both wired backhaul condition and wireless channel quality are considered in the optimization. In this paper, taking into account the conditions of the backhaul in terms of delay and wireless channel quality, joint design and optimization of the caching and user association policy to minimize the average download delay is studied in a cache-enabled heterogeneous network. We first prove the joint caching and association optimization problem is NP-hard based on a reduction to the facility location problem. Furthermore, in order to reduce the complexity, a distributed algorithm is developed by decomposing the NP-hard problem into an assignment problem solvable by the Hungarian method and two simple linear integer subproblems, with the aid of McCormick envelopes and the Lagrange partial relaxation method. Simulation results reveal a near-optimal performance that performs up to 22% better in term of delay compared with those in the literatures at a low complexity of $O\left(nm^3/\varepsilon^2\right)$.

INDEX TERMS Caching placement, user association, backhaul condition, facility location problem, Lagrange partial relaxation method.

I. INTRODUCTION

According to the prediction of Cisco, global mobile data traffic will increase by a factor of 40 over the next five years, from the current level of 93 Petabytes to 3600 Petabytes per month [1]. The explosive growth of mobile data traffic, especially mobile video streaming, has imposed a heavy burden on backhaul links, which connect local base stations to the core network. Furthermore, in massive content delivery scenarios, e.g., in populated areas or during peak traffic hours, user may experience excessively long delay to content delivery due to the congestion in backhaul links, and thus the overall quality of experience (QoE) of users is degraded. To alleviate the backhaul burden and reduce user-perceived latency, one promising approach is to deploy caches at the small cell base stations (SBS) [2], [3].

The role of caching in the fifth generation (5G) has been recognized [2]–[4], and some decentralized caching architectures have been proposed [2]–[7]. The main idea of deploying caches at SBSs is to cache popular content items on the SBS closest to their respective users so that most of the requests can be served from local caches, instead of forwarding the user requests over the expensive and bandwidth-limited backhaul links. In the cache-enabled network, users (UE) can obtain the requested content from the candidate SBSs directly if the content is cached in the SBS, which is obviously beneficial to enhance the user experience. To get the better performance, whether the SBS caches the required content may be regarded as a novel important consideration of user association strategy. It follows that the operator may explicitly devise the user association strategy, together with the caching strategy, to improve the user perceived network performance (in terms of delay). In particular, the efficiency of the caching strategy depends largely on the user association rule such that there is a strong correlation between caching strategy and user association strategy.

So far, several literatures have investigated the design of caching policy to improve the efficiency of cache [6]–[8], where caching policies are developed taking into account the given user association rule. For example, In [6] and [7], Shanmugam et al propose firstly caching at small-cell base stations and design the optimal caching policy to maximize the cache-hit-ratio. In [8], a distributed caching placement algorithm is formulated to minimize the downloading latency with the aid of a factor graph. In [9], the UE-SBS association is formulated as a one-to-many matching game to maximize the average download rate based on the given caching policy. These literatures [6]–[9] don't optimize jointly the user association and cache-content management, leading the system to inefficient operating point.

There are also a few existing works, done on the joint design of cache policy and user association strategy in the cache-enabled heterogeneous network. Considering the bandwidth capacity constraints of SBS, [10] designs the joint user association and data caching strategy to minimize the requests served by the macro base stations (MBS). Reference [11] gives the joint design of video caching and user association scheme to minimize the user experienced delay, considering users with different quality requirements and video encoding policy. Reference [12] proposes an online algorithm to solve the optimum tradeoff between load balancing and content availability, in a way to design network costs. Reference [13] focuses on analyzing complexity of the joint user association and caching scheme. Reference [14] designs joint caching, routing, and channel assignment over coordinated small-cell cellular systems to maximize the throughput of the system by utilizing the column generation method.

However, most of these works ignore the heterogeneity of users, such as the difference of wireless channel quality of different users. Furthermore, They don't jointly take into account the wired backhaul condition and wireless channel quality when designing the caching and association strategy. Consequently, ignoring the effect of backhaul condition or wireless channel quality may result in inadequate performance gain.

In summary, to fully exploit the gain of cache, an efficient caching and association strategy needs to be designed jointly by properly considering backhaul condition and wireless channel quality. In this paper, joint design of the caching and user association policy is optimized to minimize the average delay of small cell users in the cache-enabled heterogeneous network. More specifically, **the main contributions of this paper are:**

1) The joint design of the optimal caching and association strategy is studied by formulating an integer non-linear optimization problem aiming at minimizing the average download delay. Specially, the optimized strategy takes wireless channel quality into consideration and is fully aware of the propagation delay over the backhaul. Further, we prove that the joint optimization problem is NP-Hard based on a reduction to the Unsplittable hard-Capacitated Metric Facility location problem.
2) To reduce the complexity and obtain a near-optimal solution, a distributed algorithm is proposed to decompose the NP-Hard problem into an assignment problem solved by Hungarian method and two simple linear integer subproblems, with the aid of McCormick envelopes and Lagrange partial relaxation method.
3) Simulations are conducted which show that the proposed algorithm has a low complexity and can achieve comparable performance to exhaustive search. Furthermore, the proposed algorithm can significantly reduce the average download delay, more specifically up to 22% less delay compared to that of the conventional scheme.

The rest of the paper is organized as follows. In Section II, the system model is presented and the joint caching and association optimization framework is formulated. In Section III, we present the reduction to the Unsplittable hard-Capacitated Metric Facility location problem. In Section IV, the decentralized algorithm is proposed. In Section V, the simulation results and the corresponding discussions are presented, and we conclude the paper in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

Consider a heterogeneous cellular network (HCN) consisting of a single MBS, N SBSs and U UEs randomly located in the network. The MBS is indexed by M. The set of the SBSs is denoted by $\mathcal{B}=\{B_1, B_2, \cdots, B_N\}$, where B_n, $n \in \mathcal{N} = \{1, 2, \cdots, N\}$ represents the n-th SBS. It is possible to have overlapping area between SBSs in ultra-dense deployment. Furthermore, we denote the set of the UEs by $\mathcal{J} = \{J_1, J_2, \cdots, J_U\}$, where J_u, $u \in \mathcal{U} = \{1, 2, \cdots, U\}$ represents the u-th UE. The MBS is connected to the core network through high-capacity backhaul such as optical fiber. Each SBS is connected to the core network through a wired backhaul link of limited capacity. Additionally, each SBS is equipped with a storage capacity of bytes $G_n \geq 0$. The two-layer architecture is described in Fig. 1.

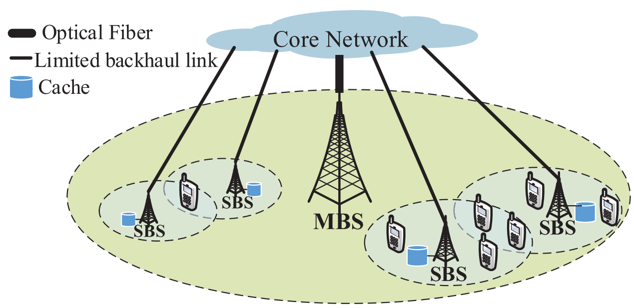

FIGURE 1. The two-layer HCN architecture.

The SBSs reuse the downlink resources of the MBS to serve the transmission to UE. As a result, there exists the interference between the SBSs and the MBS. Further, we assume that neighboring SBSs can be also allocated orthogonal frequency band or employ enhanced inter-cell interference coordination techniques (eICIC) proposed in LTE Rel.10 [15]. Each SBS B_n has a downlink bandwidth W_n, which is divided into A_n subchannel of bandwidth w. Each user access only one subchannel at a slot. Thus, the maximum number of active users of SBS B_n is A_n, where $A_n = W_n/w$.

To achieve load balancing, the "SBS-First" constraint is considered, such that each UE will try to download files from its adjacent SBSs unless the capacity of these adjacent SBSs is not sufficient. In this case, UE will turn to the MBS to deliver these files.

Denote the transmission power of the SBS B_n, the transmission power of the MBS M and the noise power at each UE as P_n, P_M and σ^2 respectively. Let $h_{n,u}$ be the channel gain between UE J_u and SBS B_n. Therefore, the signal-to-interference-plus-noise ratio (SINR) between UE J_u and SBS B_n is $\gamma_{u,n} = \frac{P_n h_{n,u}}{\sigma^2 + P_M h_{M,u}}$. Denote by $H(u)$ the set of available SBSs for UE J_u, which are capable of providing higher SINR for UE J_u.

UEs request files from a set $\mathcal{I} = \{1, 2 \cdots, F\}$ of $|\mathcal{I}| = F$ content items. Let $q_{u,i} \in \{0, 1\}$ denote whether user u requests file i. We have $q_{u,i} = 1$ if user u requests file i, and $q_{u,i} = 0$ otherwise. Assume that each request is entirely served by one base station. Without any loss of generality, we assume all these files have the same size L. This is because files can be divided into blocks of the same length or by leveraging advanced coding techniques [7]. Thus, each SBS B_n is equipped with a limited storage capacity of S_n files, where $S_n = G_n/L$.

B. PROBLEM FORMULATION

Let $x_{ni} \in \{0, 1\}$ be a binary decision variable, which represents whether the SBS B_n caches i-th file or not. We have $x_{ni} = 1$ if SBS B_n caches i-th file, and $x_{ni} = 0$ otherwise. The caching policy matrix is defined as follows:

$$\boldsymbol{x} = \{x_{ni} : n \in \mathcal{N}, i \in \mathcal{I}\}. \quad (1)$$

To indicate the association relationship between UE and SBS, we introduce binary decision variable $p_{u,n} \in \{0, 1\}$. The variable $p_{u,n}$ denotes whether UE J_u is associated with the SBS B_n. The UE-SBS association can be described through the following matrix:

$$\boldsymbol{p} = \left\{p_{u,n} : u \in \mathcal{U}, n \in \mathcal{N}\right\}. \quad (2)$$

Next, we need to calculate the delay for UE J_u to download file i when associating with SBS B_n. The main components of the delay are the wireless transmission delay and the backhaul delay. The wireless transmission delay between UE J_u and SBS B_n is calculated as:

$$D^1_{u,n} = \frac{L}{w_{u,n} \log_2(1 + \gamma_{u,n})}, \quad (3)$$

where L represents the file size, and $w_{u,n}$ indicates the bandwidth of UE J_u allocated by SBS B_n. The wireless transmission delay from SBS to UE depends on the bandwidth and SINR.

Another main component of delay is the backhaul delay. We denote the backhaul delay of UE J_u connected to SBS B_n as $D^B_{u,n}$. For wired backhaul, the backhaul delay of SBSs is related to the average link distance, the average traffic load and the average number of SBSs connecting to a single small cell gateway. It can be modeled to be an exponentially distributed random variable with a mean value of D_B [16].

When the requested content is cached in the nearby SBS, the user can fetch directly the content from the local caches of SBS, without the need for going through the backhaul. Thus, it doesn't incur extra delay over the backhaul. In other words, whether the delay of UEs contains the backhaul delay depends on whether the requested content is cached. Thus, when user requests file i, the backhaul delay between UE J_u and SBS B_n is calculated as

$$D^2_{u,n} = (1 - x_{ni}) D^B_{u,n}. \quad (4)$$

Consequently, the delay for UE J_u to download file i when associating with SBS B_n is written as

$$\begin{aligned} D^u_{i,n} &= D^1_{u,n} + D^2_{u,n} \\ &= \frac{L}{w_{u,n} \log_2(1 + \gamma_{u,n})} + (1 - x_{ni}) D^B_{u,n}. \quad (5) \end{aligned}$$

The average delay of small cell users can be calculated as

$$\overline{D} = \frac{1}{|U|} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{I}} q_{u,i} p_{u,n} D^u_{i,n}. \quad (6)$$

With the consideration of transmission bandwidth capacity constraint and storage capacity constraint, the joint caching and user association problem to minimize the average delay of small cell users is formulated as

$$\begin{aligned} &\min_{\boldsymbol{p},\boldsymbol{x}} \ \bar{D} && (7) \\ &\text{Subject to: } \sum_{i \in \mathcal{I}} x_{ni} \leq S_n, \ \forall n \in \mathcal{N}, && (8) \\ &\qquad \sum_{n \in H(u) \cup \{M\}} p_{u,n} = 1, \ \forall u \in \mathcal{U}, && (9) \\ &\qquad \sum_{u \in \mathcal{U}} p_{u,n} \leq A_n, \ \forall n \in \mathcal{N}, && (10) \\ &\qquad x_{ni} \in \{0, 1\}, \ \forall n \in \mathcal{N}, i \in \mathcal{I}, && (11) \\ &\qquad p_{u,n} \in \{0, 1\}, \ \forall u \in \mathcal{U}, n \in \mathcal{N}. && (12) \end{aligned}$$

The objective of the optimization problem is to minimize the average download delay. The constraints of the optimization are specified in (8)-(12). The inequality (8) denotes the storage capacity constraint of each SBS. The equality (9) indicates that each UE can only associate with one SBS in $H(u)$ or MBS M and avoid partial association. The inequality (10) reveals the transmission bandwidth constraint of each SBS. Finally, (11) and (12) dictate discrete and binary nature of optimization variables.

Note that the optimization problem defined in (7)-(12) is non-linear combination optimization problem since both of the caching variable and user association variable are integer values. Furthermore, the objective function is a non-linear function since there is mutual dependency between the caching variable and user association variable. In the next section, by resorting to a reduction to facility location problem, we prove that the optimization problem is NP-Hard.

III. THE REDUCTION TO FACILITY LOCATION PROBLEM

The connection between the unsplittable hard-capacity facility location problem and the joint caching and user association problem is non-trivial. In fact, previous work in the literature that established reductions of caching problem to facility location problem focused on the simple case that users only are connected to the base station with the requested file already cached, and the cost of communication between any base station and user pair is same [10]. Our model considers the case that users with different wireless channel quality may be associated with any base station within its communication range. Thus, the connection relationship and cost value of the facility location problem need to be redesigned.

Lemma 1: The optimization problem is polynomial-time reducible to the unsplittable hard-capacity facility location problem.

Proof: The unsplittable hard-capacity facility location problem is described as follows. Given a set of locations $\mathcal{L}$, there is a subset $\mathcal{A} \subseteq \mathcal{L}$ of facilities and a subset $\mathcal{B} \subseteq \mathcal{L}$ of clients that must be assigned to some open facilities. For each client $j \in \mathcal{B}$, there is a positive integer demand d_j, which can only be served by a single facility (unsplittable). For each facility $i \subseteq \mathcal{A}$, it can serve a total demand at most $C_i \geq 0$ (hard-capacity). The cost of serving one unit of demand of client j by facility i is $c_{i,j} \geq 0$. The cost of opening facility $i \subseteq \mathcal{A}$ is $f_i \geq 0$. The facility location problem aims to decide the set of facilities and find the optimal assignment of each client to facilities so as to minimize the total cost incurred.

The reduction of the optimization problem to the unsplittable hard-capacity facility location problem is as follows:

The set of facility $\mathcal{A}$ contains two parts: the first part is named a_M for the MBS, and the second part is a_{ni}, which is for every SBS $n \in \mathcal{N}$ and every file $i \in \mathcal{I}$. The set of client $\mathcal{B}$ consists of the following subsets: (i) $\mathcal{B}_1$ contains $|U|$ clients, denoted as b_u, $b_u \in \mathcal{U}$. Those clients in $\mathcal{B}_1$ indicates the cellular users. (ii) $\mathcal{B}_2$ contains $|F - S_n|$ virtual clients, denoted as $b'_{n,1}$ $b'_{n,2}$ etc, $\forall n \in \mathcal{N}$. (iii) the subset of $\mathcal{B}_3$ contains $|(S_n - 1) * A_n|$ virtual clients, denoted as $b''_{n,1}$, $b''_{n,2}$ etc, $\forall n \in \mathcal{N}$. For each facility, the capacity of the facility a_M is equal to $+\infty$ and the capacity of the facility a_{ni} for each SBS $n \in \mathcal{N}$ and each file $i \in \mathcal{I}$ is set to A_n. For each client, the demand of the client $b_u \in \mathcal{B}_1$ and $b''_q \in \mathcal{B}_3$ is equal to 1. In addition, the demand of the client $b'_c \in \mathcal{B}_2$ is set to A_n, which is unsplittable. UE J_u only can have a relationship of connection with these SBSs in $H(u)$. The cost of opening facility is equal to 0. The cost for each pair of facility and client is specified as follows:

1) The cost of each pair of the form (a_{ni}, b_u) is calculated as the delay of UE J_u connected to the SBS B_n with file i cached. Therefore, the cost is calculated as $cost(a_{ni}, b_u) = \frac{L}{w_{u,n}\log(1+\gamma_{u,n})} + (1 - q_{u,i})D^B_{u,n}$.

2) The cost of each pair of the form (a_{ni}, b'_c) and the form (a_{ni}, b''_q) is set to very small positive constant d, $d \ll \min(cost(a_{ni}, b_u))$. The setting of parameter d is to ensure that all clients in the subset of $\mathcal{B}_2$ and $\mathcal{B}_3$ are associated with the facility a_{ni}. Consequently, exactly S_n of the facilities are uncovered by the virtual users of $\mathcal{B}_2$, corresponding to the cached files. Meanwhile, a total of A_n cellular users can be accessed to SBS B_n, corresponding to the capacity constraint of SBSs.

3) The cost of each pair of the form (a_M, b_u) is set to very large positive constant h, $h \gg \max(cost(a_{ni}, b_u))$. The setting of parameter h is to ensure that all clients in the subset of $\mathcal{B}_1$ will choose firstly to access the facility a_{ni}. Only when a_{ni} can't serve more clients, the client choose to access the facility a_M, which is consistent with the hypothesis of "SBS-First".

Based on the above description, we formulate the unsplittable facility location problem. In addition, based on the proof in [10], we can obtain the following two conclusions:

1) When the cost of the feasible solution for the facility location problem is D, there exists a corresponding feasible solution for the optimization problem with cost C, satisfying

$$D = C + \left(U - \sum_{n=1}^{N} A_n\right) h + \sum_{n=1}^{N} ((F - S_n) A_n + |(S_n - 1) A_n|) d. \tag{13}$$

2) When the cost of the feasible solution for the optimization problem is C, there exists a corresponding feasible solution for the facility location problem at cost D, satisfying

$$C = D - \left(U - \sum_{n=1}^{N} A_n\right) h - \sum_{n=1}^{N} ((F - S_n) A_n + |(S_n - 1) A_n|) d. \tag{14}$$

Thus, the reduction from the optimization problem defined in (7)-(12) to the above proposed unsplittable facility location problem holds. There exists a reduction from the optimization problem to the unsplittable hard-capacity facility location problem, which is known to be NP-Hard [17]. ■

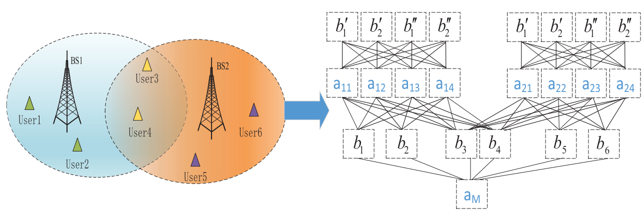

FIGURE 2. A example of the reduction to the facility location problem.

Fig. 2 presents an example of the reduction based on the above description. Here, the parameters of the system are set

TABLE 1. **The proof of the transformation from equality to inequality.**

$p_{u,n}$	x_{ni}	$z^u_{i,n}$	$\max\left(p_{u,n}-x_{ni},0\right)$	$\min\left(p_{u,n},1-x_{ni}\right)$
0	0	0	0	0
0	1	0	0	0
1	0	1	1	1
1	1	0	0	0

as follows: $|\mathcal{N}| = 2$, $|\mathcal{U}| = 6$, $|\mathcal{I}| = 4$, $|S_1| = |S_2| = 2$, $|A_1| = |A_2| = 2$. Therefore, each SBS B_n contains four facilities. In addition, user 3 and user 4 are in overlapping coverage area of SBS 1 and SBS 2, so these users have relationship of connection with the facility a_{1i} and a_{2i}.

IV. DECENTRALIZED ALGORITHM

The problem defined in (7)-(12) is NP-Hard and the complexity is extremely high. To reduce the complexity of the problem, a distributed algorithm is proposed in this section. Firstly, the optimization problem is transformed equivalently with the aid of McCormick envelopes. Secondly, we use the method of Lagrange partial relaxation to solve the transformed problem and decompose the problem into several subproblems.

It can be shown that the caching variable and user association variable are tightly coupled in the objective function of the optimization problem, which causes the problem hard to solve. To conquer the challenge, we introduce a new variable $z^u_{i,n}$, $z^u_{i,n} = (1-x_{ni})p_{u,n}$ that allows us to rewrite the optimization problem defined in (7)-(12) as follows:

$$\min_{p,x,z} \frac{1}{|U|} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} q_{u,i} \left[\frac{Lp_{u,n}}{w_{u,n}\log_2(1+\gamma_{u,n})} + z^u_{i,n} D^B_{u,n} \right] \tag{15}$$

Subject to: (8)-(12),

$$z^u_{i,n} = (1-x_{ni})p_{u,n}, \forall u \in \mathcal{U}, i \in \mathcal{I}, n \in \mathcal{N}. \tag{16}$$

To obtain the convex relaxation, we replace the non-convex constraint $z^u_{i,n} = (1-x_{ni})p_{u,n}$ with its McCormick convex relaxation by using McCormick envelopes [18], which is given by

$$z^u_{i,n} \geq p_{u,n} - x_{ni}, \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}, \tag{17}$$
$$z^u_{i,n} \geq 0, \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}, \tag{18}$$
$$z^u_{i,n} \leq p_{u,n}, \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}, \tag{19}$$
$$z^u_{i,n} \leq 1 - x_{ni}, \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}. \tag{20}$$

Specially, due to the discrete and binary nature of optimization variables x_{ni} and $p_{u,n}$, it can be readily established that the equality $z^u_{i,n} = (1-x_{ni})p_{u,n}$ is equivalent strictly to the constraints (17)-(20), which is shown in Table I.

Thus, the optimization problem can be further expressed as

$$\min_{p,x,z} \frac{1}{|U|} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} q_{u,i} \left[\frac{Lp_{u,n}}{w_{u,n}\log_2(1+\gamma_{u,n})} + z^u_{i,n} D^B_{u,n} \right] \tag{21}$$

Subject to: (8)-(12), (17)-(20).

In order to solve the new optimization problem, we use the method of Lagrange partial relaxation [19]. Specially, we relax the constraints (17), (19), (20) and introduce the respective set of dual Lagrange multipliers:

$$\mu^u_{i,n} \geq 0 \quad \forall u \in \mathcal{U},\ \forall i \in \mathcal{I},\ \forall n \in \mathcal{N}, \tag{22}$$
$$\lambda^u_{i,n} \geq 0 \quad \forall u \in \mathcal{U},\ \forall i \in \mathcal{I},\ \forall n \in \mathcal{N}, \tag{23}$$
$$\psi^u_{i,n} \geq 0 \quad \forall u \in \mathcal{U},\ \forall i \in \mathcal{I},\ \forall n \in \mathcal{N}. \tag{24}$$

Hence, the Lagrange function is expressed as

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{p}, \boldsymbol{x}, z) &= \frac{1}{|U|} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} [\frac{q_{u,i}Lp_{u,n}}{w_{u,n}\log_2(1+\gamma_{u,n})} \\ &\quad + q_{u,i}z^u_{i,n}D^B_{u,n} + \mu^u_{i,n}(p_{u,n}-x_{ni}-z^u_{in}) \\ &\quad + \lambda^u_{i,n}(z^u_{i,n}-p_{u,n}) + \psi^u_{i,n}(z^u_{i,n}+x_{ni}-1)]. \end{aligned} \tag{25}$$

Thus, the dual problem can be given by

$$\max_{\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\psi}} \min_{\boldsymbol{p},\boldsymbol{x},z} L\left(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{p}, \boldsymbol{x}, z\right),$$

Subject to: (8)-(12), (18), (22)-(24).

Interestingly, given the dual variables $\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\varphi}$, the Lagrange function can be written as

$$L\left(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{p}, \boldsymbol{x}, z\right) = f\left(\boldsymbol{p}\right) + g\left(\boldsymbol{x}\right) + h\left(z\right), \tag{26}$$

where $f(\boldsymbol{p})$, $g(\boldsymbol{x})$ and $h(z)$ are the objective functions of P1, P2, P3 respectively. Furthermore, the feasible region of dual problem can be decomposed into three independent regions (i.e. {(9), (10), (12)}, {(8), (11)} and {(18)}). Therefore, the dual problem can be decomposed into three subproblems, named as P1, P2, P3 respectively. The three subproblems are given as follows:

$$\begin{aligned} P1: \min_{\boldsymbol{p}} &\sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} q_{u,i} \left[\frac{L}{w_{u,n}\log_2(1+\gamma_{u,n})} \right] p_{u,n} \\ &+ \mu^u_{i,n}p_{u,n} - \lambda^u_{i,n}p_{u,n} \end{aligned}$$

Subject to: (9), (10), (12).

$$P2: \max_{\boldsymbol{x}} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} \mu^u_{i,n}x_{ni} - \psi^u_{i,n}x_{ni}$$

Subject to: (8), (11).

$$P3: \min_{z} \sum_{u\in\mathcal{U}} \sum_{n\in\mathcal{N}} \sum_{i\in\mathcal{I}} \left(q_{u,i}D^B_{u,n} - \mu^u_{i,n} + \lambda^u_{i,n} + \psi^u_{i,n} \right) z^u_{i,n}$$

Subject to: (18).

Particularly, after the decomposition, the joint optimization problem becomes essentially separate optimization problems and the coupling between the association variable and the caching variable disappears.

The first subproblem only involves the UE-SBS association variable $\boldsymbol{p}$. Here, we model the first subproblem as the assignment problem. We view each base station B_n as a machine of processing capacity A_n, and each UE J_u as a job that requires one units of processing. When UE J_u

is assigned to BS B_n, it incurs a cost of d_{un}, $d_{un} = \sum_{i \in I} \frac{q_{u,i} L}{w_{u,n} \log_2(1+\gamma_{u,n})} + \mu_{i,n}^u - \lambda_{i,n}^u$. Because the total processing capacity of all machines is not equal to the number of jobs, a dummy variable is introduced, either for a machine or a job, to make it balanced. In other words, if $\sum_{n \in \mathcal{N}} A_n > U$, we add $\sum_{n \in \mathcal{N}} A_n - U$ virtual jobs to the job sets. The cost of these virtual jobs is zero. On the other hand, if $\sum_{n \in \mathcal{N}} A_n < U$, we need to introduce a virtual machine of processing capacity $U - \sum_{n \in \mathcal{N}} A_n$. Due to the special structure of the assignment problem, the solution can be found using a more convenient method called Hungarian method [20]. The second subproblem only involves the caching variable $\boldsymbol{x}$ and the third subproblem only involves the added new variable $\boldsymbol{z}$. Both subproblems are the linear integer optimization problem, which can be solved by the generic linear integer programming method [17].

By solving the three subproblems and obtaining the values of $\boldsymbol{p}, \boldsymbol{x}, \boldsymbol{z}$, we use the subgradient method to update the dual variables. In the t-th iteration, for $\forall u \in \mathcal{U}, i \in \mathcal{I}, n \in \mathcal{N}$, the dual variables are updated as follow:

$$\mu_{i,n}^u (t+1) = \left[\mu_{i,n}^u (t) + \sigma (t) d \left(\mu_{i,n}^u (t)\right)\right]^+, \quad (27)$$

$$\lambda_{i,n}^u (t+1) = \left[\lambda_{i,n}^u (t) + \sigma (t) d \left(\lambda_{i,n}^u (t)\right)\right]^+, \quad (28)$$

$$\psi_{i,n}^u (t+1) = \left[\psi_{i,n}^u (t) + \sigma (t) d \left(\psi_{i,n}^u (t)\right)\right]^+, \quad (29)$$

where $[x]^+ = \max\{0, x\}$ and $\sigma(t)$ is the step size of the t-th iteration. And $d(\boldsymbol{\mu}(t))$, $d(\boldsymbol{\lambda}(t))$, $d(\boldsymbol{\psi}(t))$ are the subgradient of dual problem with respect of $\mu_{i,n}^u(t)$, $\lambda_{i,n}^u(t)$, $\psi_{i,n}^u(t)$, given by

$$d\left(\mu_{i,n}^u(t)\right) = p_{u,n}(t) - x_{ni}(t) - z_{i,n}^u(t), \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}, \quad (30)$$

$$d\left(\lambda_{i,n}^u(t)\right) = z_{i,n}^u(t) - p_{u,n}(t), \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}, \quad (31)$$

$$d\left(\psi_{i,n}^u(t)\right) = z_{i,n}^u(t) + x_{ni}(t) - 1, \quad \forall u \in \mathcal{U},\ i \in \mathcal{I},\ n \in \mathcal{N}. \quad (32)$$

Denote $g(t) = [d(\boldsymbol{\mu}(t)), d(\boldsymbol{\lambda}(t)), d(\boldsymbol{\psi}(t))]^T$ and set the step size as $\sigma(t) = v\frac{UB - q(t)}{\|g(t)\|^2}$ [21], where UB is the upper bound on each iteration and v is a positive constant and $q(t)$ is the value of Lagrange function in the t-th iteration. The UB can be found by simply finding a feasible solution of the primary problem. Note that the step size is nonsummable diminishing step length. Based on the proof in [22], the algorithm is guaranteed to converge to the optimal value. The method is summarized in Algorithm 1.

V. SIMULATION

In this section, numerical results of the proposed algorithm are presented. In Section V.A, we compare the performance of the proposed algorithm with that of the exhaustive search, establishing the performance of the proposed algorithm. In Section V.B, we present the convergence analysis and discuss the impact of various parameters on the proposed algorithm. In Section V.C, the proposed algorithm is compared with conventional scheme.

Algorithm 1 Decentralized Algorithm for the Primal Optimization Problem

Require:
$t = 1$, $\mu_{i,n}^u(1) = 0$, $\lambda_{i,n}^u(1) = 0$, $\psi_{i,n}^u(1) = 0$, $q(1) = 0$, $UB = +\infty$, $\varepsilon = 0.01$, and $t_{\max} = 2000$.
Ensure:
while $\left|\frac{UB - q(t)}{UB}\right| \geq \varepsilon$ and $t \leq t_{\max}$ **do**
  Solve P1 and find the solution of $p_{u,n}$.
  Solve P2 and find the solution of x_{ni}.
  Solve P3 and find the solution of $z_{i,n}^u$.
  Update UB.
  $q(t) = L(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{p}, \boldsymbol{x}, \boldsymbol{z})$ and $\sigma(t) = v\frac{UB - q(t)}{\|g(t)\|^2}$.
  Update the dual variable $\mu_{i,n}^u(t+1)$, $\lambda_{i,n}^u(t+1)$, $\psi_{i,n}^u(t+1)$ by using (27), (28), (29).
  Update $t = t + 1$.
end while

TABLE 2. Parameter values used in numerical results.

Macrocell radius	400 (m)
Transmit power of SBS	23 (dBm)
Pass-loss model	ITU-UMi model
Noise power spectrum density	-174(dBm/Hz)
Shape parameter η	0.6
SINR threshold δ	0.1
Bandwidth of base station W	20(MHz)
File size L	10(Mbits)
Maximum number of active users A_n	20
Backhaul delay D_B	[0, 3]
Number of files F	6(small-scale system) 50(large-scale system)
Number of users U	50(small-scale system) 200(large-scale system)
Number of base stations N	2(small-scale system) 8(large-scale system)
Storage capacity S_n	1(small-scale system) 3(large-scale system)

We numerically evaluate the algorithm by fixing the location of MBS at the center of a macrocell with a radius 400m and distribute SBSs randomly throughout the MBS coverage area. The physical layer parameters such as the transmit power of SBSs, the path-loss model, noise power are chosen according to 3GPP standards. Each user requests one file based on the Zipf distribution with shape parameter $\eta = 0.6$, where the request probability of the i-th file is $\rho_i = \frac{1/i^\eta}{\sum_{i=1}^F 1/i^\eta}$ [23]. The range for the mean of the backhaul delivery delay D_B is selected based on measurements obtained from a practical network [24]. To investigate the impact of backhaul delay, we choose $D_B \in [0, 3]$. The parameter v of Algorithm 1 is set to 0.5. The system parameters are summarized in Table II.

A. OPTIMALITY TEST OF THE PROPOSED ALGORITHM

The performance of the proposed algorithm is evaluated firstly. We compare the performance of the proposed algorithm with the exhaustive search in a small-scale system.

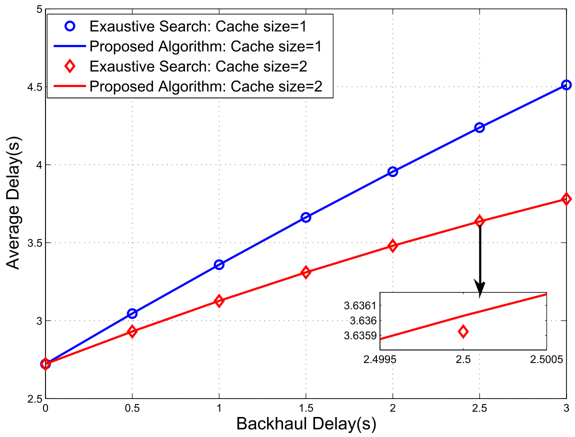

FIGURE 3. Performance comparison of the proposed algorithm and exhaustive search.

The result obtained from the exhaustive search is adopted as a benchmark, which is the lower bound of the average delay. In the small-scale system, the file library has six files. There are two SBSs and each has a capacity of one file. A total of 50 users are placed randomly, independently and uniformly in the cell. We consider the performance averaged over five thousand network instances. Fig. 3 shows that the performance of the proposed algorithm is very close to that obtained using the exhaustive search. In addition, it also can be observed that as cache size increases slightly, the average download delay reduce significantly, which shows that caching is beneficial to enhance wireless network performance.

B. CONVERGENCE AND COMPLEXITY

1) Convergence

The convergence of the proposed algorithm in a large-scale system is depicted in Fig. 4. In the large-scale system, the file library has 50 files. There are 8 SBSs and each has a capacity of 3 files. A total of 200 users are placed randomly, independently and uniformly in the cell. As it can be seen, the proposed algorithm gradually improves the obtained result and converges rapidly in less than a few hundred steps.

2) Complexity

To guarantee the accuracy ε of subgradient method, the proposed algorithm need $O\left(1/\varepsilon^2\right)$ iterations [19]. Furthermore, the time complexity of the proposed algorithm in each iteration is the same, namely $O\left(nm^3\right)$ [20], where n denotes the maximum number of neighboring BSs a user can be connected to and m denotes the number of users. As a result, the complexity of the proposed algorithm is $O\left(nm^3/\varepsilon^2\right)$. In Table III, the number of iterations and time complexity per iteration of the proposed algorithm and exhaustive search are summarized.

TABLE 3. Number of iterations and time complexity of algorithms.

	Number of iterations	Time complexity per iteration
Exhaustive search	$\left(C_F^{S_n}\right)^N$	$O\left(nm^3\right)$
The proposed algorithm	$O\left(1/\varepsilon^2\right)$	$O\left(nm^3\right)$

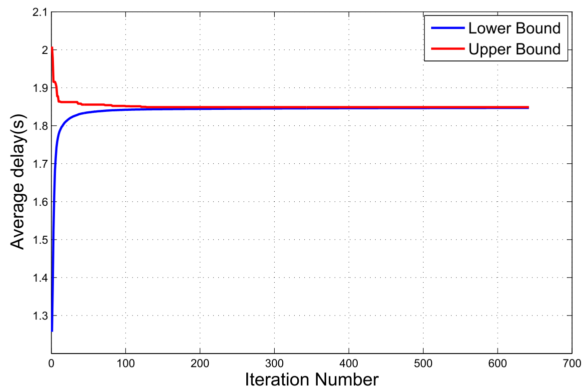

FIGURE 4. The convergence of the proposed algorithm.

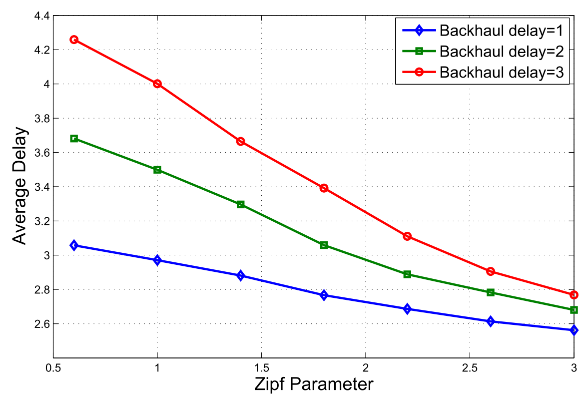

FIGURE 5. The effect of Zipf Parameter.

C. PARAMETER IMPACT ANALYSIS OF THE PROPOSED ALGORITHM

We explore the effect of the steepness of the file request pattern on the performance of the proposed algorithm in a small-scale system. The shape parameter of the file popularity is varied from the value 0.6 to 3. Fig. 5 shows the effect of Zipf parameter on the average delay. It can be observed that as the Zipf parameter increases, the average delay decreases. In addition, it can be seen that as the Zipf parameter increases, the effect of the backhaul delay on the average delay decreases. This is because as popularity distribution gets steeper, a small number of contents are more popular when Zipf parameter is high, which improves the caching effectiveness. Thus, more contents can be served directly from the local caches of BSs and don't have to travel through the backhaul, which decreases the effect of backhaul delay.

D. COMPARISON WITH OTHER SCHEMES

We compare the proposed algorithm with the Most Popular Content-Maximum SINR (MPC-MS) scheme in a large-scale system. The MPC-MS scheme is to cache the most popular contents, which is a standard caching placement strategy [25], [26], and users are associated with the SBS with the maximum-SINR without considering the backhaul conditions [27].

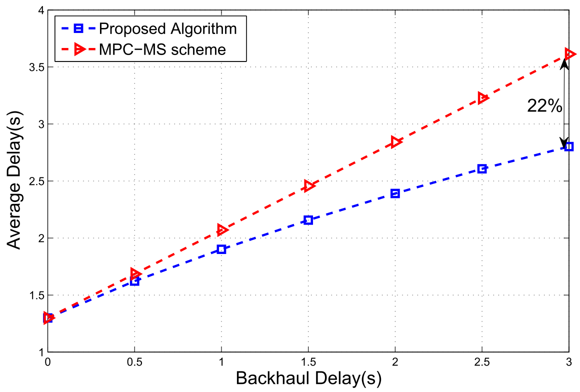

FIGURE 6. Performance comparison of different schemes.

Fig. 6 demonstrates that the proposed scheme outperforms the MPC-MS scheme and some important insights are also revealed. Firstly, the backhaul delay affects significantly the caching policy and user association scheme. When the backhaul delay is very small, the proposed algorithm has a similar performance as that achieved by the MPC-MS scheme. On the other hand, when the backhaul delay is large, the performance gap of the proposed algorithm and the MPC-MS scheme increases. This is because backhaul delay becomes a major component of delivery delay but the MPC-MS scheme ignores the backhaul conditions, thereby achieving a higher average download delay. The simulation result shows that the proposed algorithm can reduce delay by up to 22% than the conventional scheme.

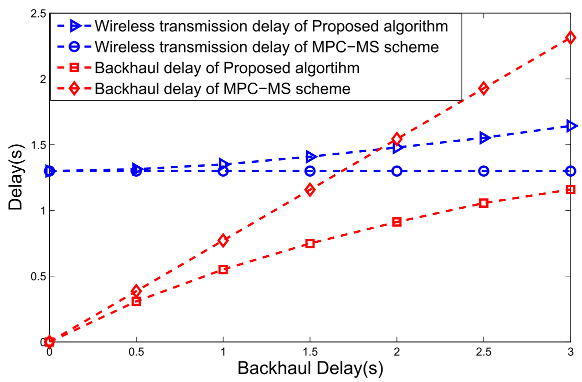

FIGURE 7. Delay allocation of different schemes.

Further, Fig. 7 shows the advantage of the proposed algorithm from the perspective of delivery delay. It can be observed that as backhaul delay is relatively small, wireless transmission delay will dominate the average delay and becomes the limiting factor. In this case, the gap of the MPC-MS scheme with the proposed algorithm is relatively small. On the other hand, as backhaul delay increases gradually, the average delay is mainly contributed by the backhaul delay caused by constrained backhaul link. In this case, the proposed algorithm is fully aware of the backhaul conditions and reduce the larger backhaul delay. Therefore, it can be concluded that the proposed algorithm achieves the efficient tradeoff between the wireless transmission delay and backhaul delay.

VI. CONCLUSION

This paper designs the joint caching and association strategy to minimize the average download delay. The joint strategy takes into account wireless channel quality and is aware of the transmission delay over the backhaul. We analyze the joint optimization problem by formulating an integer nonlinear optimization problem. The problem is proved to be NP-Hard based on a reduction from the facility location problem. In order to reduce the complexity, a distributed algorithm is proposed by decomposing the NP-hard problem into an assignment problem solved by Hungarian method and two simple linear integer subproblems, with the aid of McCormick envelopes and Lagrange partial relaxation method. Simulation results show that the proposed algorithm can significantly reduce the average download delay, approaching the lower bound of the average download delay but with a low complexity. Moreover, the simulation results demonstrate the necessity to consider the cache condition, i.e, whether the BS caches the requested contents when deciding the best UE-SBS association, especially when the backhaul condition is poor. Therefore, it can be concluded that our work gives a promising method to determine the optimal caching policy and user association scheme, and provides some important insights for understanding the complicated interaction between the caching policy and user association strategy.

REFERENCES

[1] C. V. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update, 2009–2014," Cisco Public Inf., San Jose, CA, USA, Tech. Rep., Feb. 2010, vol. 9.

[2] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[3] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[5] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[6] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[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] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[9] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. 12th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2014, pp. 37–42.

[10] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.

[11] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2014, pp. 1078–1086.

[12] K. Naveen, L. Massoulie, E. Baccelli, A. C. Viana, and D. Towsley, "On the interaction between content caching and request assignment in cellular cache networks," in *Proc. 5th Workshop Things Cellular, Oper., Appl. Challenges (AllThingsCellular)*, New York, NY, USA, 2015, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/2785971.2785975

[13] M. Dehghan *et al.*, "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr./May 2015, pp. 936–944.

[14] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.

[15] D. Astely, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström, and S. Parkvall, "LTE: The evolution of mobile broadband," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 44–51, Apr. 2009.

[16] D. C. Chen, T. Q. S. Quek, and M. Kountouris, "Backhauling in heterogeneous cellular networks: Modeling and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3194–3206, Jun. 2015.

[17] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*. Berlin, Germany: Springer, 2008.

[18] L. Liberti and C. C. Pantelides, "An exact reformulation algorithm for large nonconvex NLPs involving bilinear terms," *J. Global Optim.*, vol. 36, no. 2, pp. 161–189, Oct. 2006.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[20] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.

[21] T. Bektaş, J.-F. Cordeau, E. Erkut, and G. Laporte, "Exact algorithms for the joint object placement and request routing problem in content distribution networks," *Comput. Oper. Res.*, vol. 35, no. 12, pp. 3860–3884, Dec. 2008.

[22] D. P. Bertsekas,*Nonlinear programming*. Belmont, MA, USA: Athena Scientific, 1999.

[23] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 1999, pp. 126–134.

[24] D. B. West, *Introduction to Graph Theory*, vol. 2. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[25] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[26] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[27] S. Mukherjee, "Distribution of downlink SINR in heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 575–585, Apr. 2012.

YUE WANG received the B.S. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2014. She is currently pursuing the Ph.D. degree in communications and information systems with the Beijing University of Posts and Telecommunications.

Her research interests are in the area of wireless communications and networks, with current emphasis on the analysis and optimization of wireless caching technique for 5G and big data.

XIAOFENG TAO (SM'13) received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1993, and the M.S.E.E. and Ph.D. degrees in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1999 and 2002, respectively.

He is currently a Professor with BUPT, a Director of the National Engineering Laboratory, and a Chair of the IEEE ComSoc Beijing Chapter. He was the Inventor or Co-Inventor of 80 patents, the Author or Co-Author of 200 papers and two books, in wireless communication areas. He was a Co-Author of Honored Mention Award at the ACM MobiCom 2009, the Best Paper Awards at the ISCIT 2012, and the WCNC 2014, also a winner of the Chinese National Invention Awards (2008 and 2013). He currently focuses on 5G research. He is a Fellow of the IET.

XUEFEI ZHANG received the B.S. and Ph.D. degrees in telecommunications engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2010 and 2015, respectively. She is currently with the National Engineering Lab, BUPT. Her research area includes heterogeneous network, machine technology communications, stochastic geometry, and optimization theory.

GUOQIANG MAO (S'98–M'02–SM'08) received the Ph.D. degree in telecommunications engineering from Edith Cowan University in 2002. He was with the School of Electrical and Information Engineering, The University of Sydney, from 2002 to 2014. He joined the University of Technology Sydney in 2014 as a Professor of Wireless Networking and the Director of Center for Real-time Information Networks. The Center is among the largest university research centers in Australia in the field of wireless communications and networking. He has authored over 200 papers in international conferences and journals, which have been cited over 4000 times.

His research interest includes intelligent transport systems, applied graph theory and its applications in telecommunications, Internet of Things, wireless sensor networks, wireless localization techniques, and network performance analysis. He is currently an Editor of the IEEE Transactions on Vehicular Technology (since 2010) and the IEEE Transactions on Wireless Communications (since 2014). He received the Top Editor Award for outstanding contributions to the IEEE Transactions on Vehicular Technology in 2011, 2014 and 2015. He is a Co-Chair of the IEEE Intelligent Transport Systems Society Technical Committee on Communication Networks. He has served as a Chair, a Co-Chair, and a TPC member in a large number of international conferences.