Joint User Association and Content Placement for D2D-Enabled Heterogeneous Cellular Networks

Yingying Li(✉), Rong Chai, Qianbin Chen, and Chun Jin

Key Lab of Mobile Communication Technology,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China
17783195324@163.com, {chairong,chenqb,jinchun}@cqupt.edu.cn

Abstract. The explosive increase of the multimedia traffic poses challenges on mobile communication systems. To stress this problem, caching technology can be exploited to reduce backhaul transmissions latency and improve content fetching efficiency. In this paper, we study the user association and content placement problem of device-to-device-enabled (D2D-enabled) heterogeneous cellular networks (HCNs). To stress the importance of the service delay of all the users, we formulate the joint user association and content placement problem as an integer-nonlinear programming problem. As the formulated NP-hardness of the problem, we apply the McCormick envelopes and the Lagrangian partial relaxation method to decompose the optimization problem into three subproblems and solved it by using Hungarian method and unidimensional knapsack algorithm. Simulation results validate the effectiveness of the proposed algorithm.

Keywords: Heterogeneous cellular networks · User association · Content placement · D2D communication

1 Introduction

The explosive increase of diversified high-speed traffics poses challenges on the transmission performance of the radio access networks (RANs) and backhaul links of the mobile networks [1,2]. To stress this problem, heterogeneous cellular networks (HCNs) with caching functionality are expected to offer users more high-quality communication links and locally fetch request content by interacting with the small cells and thus significantly reducing redundant content downloads through the backhaul links.

To further improve user QoS, device-to-device (D2D) communication technology which allows UEs communicate directly without the data forwarding by the BSs is proposed [3]. if D2D communication technology is considered in HCNs, transmission performance will be further improved. In D2D-enabled

X. Liu et al. (Eds.): ChinaCom 2018, LNICST 262, pp. 271–282, 2019.
https://doi.org/10.1007/978-3-030-06161-6_27

HCNs, UEs may associate with the macro BS (MBS), small cell BS (SBS) or D2D peer for information interaction. It is apparent that different user association or mode selection strategies may result in various network transmission performance based on different channel characteristics and network resources.

Some recent research works study the content placement strategy for cellular networks [4–6]. The authors in [4] consider the content placement in a femtocell network and design caching strategy which minimizes the average downloading delay of all the UEs. In [5], the optimal content placement problem in a femtocell network is formulated as an average bit error rate (BER) minimization problem and use greedy algorithm to solve it. In [6], the authors design a collaborative multi-tier caching framework and propose a joint content placement and routing scheme to maximize traffic offloading.

Content placement in D2D-enabled networks is studied as well. The authors in [7] examine the average caching failure rate in D2D communication network and propose a dual-solution search algorithm to solve content placement problem. In [8], the authors formulate the access selection and spectrum allocation problem as a utility function maximization problem and propose an efficient algorithm to obtain the optimal strategy.User association and content placement problem are jointly considered for HCNs [9–11]. In [9], the author formulate the joint content caching, routing and channel assignment problem as a throughput maximization problem and propose the column generation method. In [10], the authors examine the tradeoff between load balancing and backhaul traffic reduction and solving the problem iteratively. In [11], the authors consider the various condition of the backhaul links and propose a near-optimal distributed algorithm solving it.

However, few previous works consider the D2D-enabled HCNs, Meanwhile, most of these works focus on the performance optimization of throughput or network utility but fail to stress the importance of service delay, which should become a major concern especially for delay-sensitive services.

In this paper, we study joint user association and content placement problem of D2D-enabled HCNs. Jointly considering the constraints on wireless resources, storage capacity as well as user QoS requirements, we formulate the joint user association and cache content placement problem as a service delay minimization problem and propose an efficient algorithm by applying the McCormick envelopes and the Lagrangian partial relaxation method to obtain the solution.

This paper is organized as follows. The system model is described in Sect. 2. The proposed optimization problem is formulated in Sect. 3. In Sect. 4, the solution to the formulated optimization problem is presented. Simulation results are discussed in Sect. 5. Finally, the conclusions are drawn in Sect. 6.

2 System Model

In this paper, we consider the downlink transmission in a D2D-enabled HCN consisting of a MBS, N SBSs and a number of UEs. We assume that the MBS connects to IP network through a wired backhaul link and the SBSs access core network by associating to the MBS as shown in Fig. 1.

By applying local caching scheme, we assume SBSs and some users has certain caching functionality. For simplicity, we refer cache-enabled users as serving users (SUs). Therefore, the RUs can access contents via three user association modes, i.e., MBS association, SBS association and D2D transmission mode.

In this paper, we denote the set of the SBSs as SBS= $\{\mathrm{SBS}_1, \ldots, \mathrm{SBS}_N\}$, where SBS_n represents the nth SBS, $1 \leq n \leq N$. Denote the storage capacity of SBS_n as $\mathrm{S}_n^{\mathrm{b}}$. Let RU = $\{\mathrm{RU}_1, \ldots, \mathrm{RU}_M\}$ denote the set of the RUs, where RU_m denotes the mth RU, $1 \leq m \leq M$, M is the number of RUs. Let SU= $\{\mathrm{SU}_1, \ldots, \mathrm{SU}_K\}$ denote the set of SUs where SU_k denotes the kth SU, $1 \leq k \leq K$, K is the number of the SUs. We assume RU_m can make random requests from the content library and let $F = \{F_1, \ldots F_L\}$ denotes the set of content files, where F_l represents the lth content file, $1 \leq l \leq L$, L is the number of content files.

Let binary variable $a_{m,l} \in \{0, 1\}$ describe the content request variable of RU_m, i.e., $a_{m,l} = 1$, if RU_m requests content F_l; otherwise, $a_{m,l} = 0$. We assume each user can request at most one content and SU_k can serve at most one RU during a time period. Denote $y_{k,l}^{\mathrm{d}}$ as the caching variable of SU_k, i.e., $y_{k,l}^{\mathrm{d}}$=1, if content F_l is stored in SU_k; otherwise, $y_{k,l}^{\mathrm{d}} = 0$.

To avoid transmission interference, we assume the bandwidth of the MBS and SBSs is divided into a number of subchannels with equal bandwidth and each RU can only be allocated one subchannel. Let W^{d} denotes the available bandwidth of each D2D communication link, and $W_0^{\max}$ and $W_n^{\max}$ denote the maximum available bandwidth of the MBS and SBS_n and W_0 and W_n denote the subchannel bandwidth of them. The maximum number of users associated to MBS and SBS_n can be calculated respectively as $A_0 = \lfloor W_0^{\max}/W_0 \rfloor$ and $A_n = \lfloor W_n^{\max}/W_n \rfloor$.

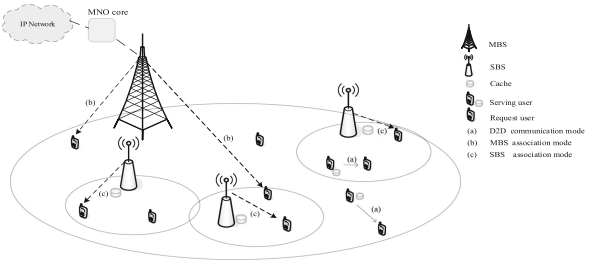

Fig. 1. System model

3 Optimization Problem Formulation

In this section, we formulate joint user association and content placement problem as a service delay minimization problem.

3.1 Objective Function

We express the service delay of the RUs as

$$D = D^{\mathrm{d}} + D^{\mathrm{m}} + D^{\mathrm{s}} \tag{1}$$

where D^{d}, D^{m} and D^{s} denote the service delay of the RUs when acquiring required contents through D2D transmission mode, MBS association mode and SBS association mode. The expressions of D^{d}, D^{m} and D^{s} will be described in following subsections.

Service Delay of RUs in D2D Communication Mode The service delay of the RUs in D2D communication mode, denoted by D^{d}, can be calculated as

$$D^{\mathrm{d}} = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} a_{m,l} x_{m,k}^{\mathrm{d}} D_{m,k,l}^{\mathrm{d}} \tag{2}$$

where $x_{m,k}^{\mathrm{d}}$ denotes the binary association variable for D2D transmission mode, i.e., $x_{m,k}^{\mathrm{d}} = 1$, if RU_m is associated with SU_k; otherwise, $x_{m,k}^{\mathrm{d}} = 0$, $D_{m,k,l}^{\mathrm{d}}$ denotes the service delay of RU_m acquiring F_l through associating with SU_k. We express $D_{m,k,l}^{\mathrm{d}}$ as

$$D_{m,k,l}^{\mathrm{d}} = y_{k,l}^{\mathrm{d}} \frac{S_l}{R_{m,k}^{\mathrm{d}}} \tag{3}$$

where $y_{k,l}^{\mathrm{d}}$ denotes the binary content caching index of the SUs, i.e., $y_{k,l}^{\mathrm{d}} = 1$, if F_l is cached in SU_k; otherwise, $y_{k,l}^{\mathrm{d}} = 0$, S_l denotes the size of F_l and $R_{m,k}^{\mathrm{d}}$ denotes the data rate of the transmission link between RU_m and SU_k, which is given by

$$R_{m,k}^{\mathrm{d}} = W^{\mathrm{d}} \log_2 \left(1 + \frac{P_k^{\mathrm{d}} g_{m,k}^{\mathrm{d}}}{\sigma^2} \right) \tag{4}$$

where P_k^{d} is the transmission power of SU_k, $g_{m,k}^{\mathrm{d}}$ is the channel gain between SU_k and RU_m, and σ^2 is the power of Guassian white noise.

Service Delay of RUs in MBS Association Mode The service delay of the RUs in MBS association mode, denoted by D^{m}, can be calculated as

$$D^{\mathrm{m}} = \sum_{m=1}^{M} \sum_{l=1}^{L} a_{m,l} x_m D_{m,l} \tag{5}$$

where x_m denotes the binary association variable for MBS association mode, i.e., $x_m = 1$, if RU_m is associated with the MBS; otherwise, $x_m = 0$, $D_{m,l}$ denotes the service delay of RU_m acquiring F_l by associating with the MBS. We express $D_{m,l}$ as

$$D_{m,l} = D_{m,l}^{\mathrm{t}} + D_{m,l}^{\mathrm{B}} + D_{m,l}^{\mathrm{w}} \tag{6}$$

where $D^{\mathrm{t}}_{m,l}$ denotes the transmission delay of RU_m acquiring F_l by associating with the MBS, $D^{\mathrm{w}}_{m,l}$ denotes the queuing delay at the MBS when RU_m acquiring F_l and $D^{\mathrm{B}}_{m,l}$ denotes the backhaul delay of the MBS. In this paper, we model the backhaul delay as an exponentially distributed random variable with a given mean value [11].

$D^{\mathrm{t}}_{m,l}$ in (6) can be expressed as

$$D^{\mathrm{t}}_{m,l} = \frac{S_l}{R_m} \tag{7}$$

where R_m denotes the data rate of the transmission link of between the MBS and RU_m, which can be expressed as

$$R_m = W_0 \log_2 \left(1 + \frac{P_m g_m}{\sigma^2}\right) \tag{8}$$

where P_m is the transmission power of the MBS when sending the content to RU_m and g_m is the channel gain between the MBS and RU_m.

$D^{\mathrm{w}}_{m,l}$ in (6) can be calculated as

$$D^{\mathrm{w}}_{m,l} = \frac{1}{\mu - \lambda} \tag{9}$$

where μ and λ are the service rate and arrival rate of the MBS, respectively.

Service Delay of RUs in SBS Association Mode where $x^{\mathrm{s}}_{m,n}$ denotes the binary association variable for SBS association mode, i.e., $x^{\mathrm{s}}_{m,n} = 1$ if RU_m is associated with SBS_n; otherwise, $x^{\mathrm{s}}_{m,n} = 0$, $D^{\mathrm{s}}_{m,n,l}$ denotes the service delay of RU_m when acquiring F_l through associating with SBS_n and can be computed as

$$D^{\mathrm{s}}_{m,n,l} = D^{\mathrm{s,t}}_{m,n,l} + \left(1 - y^{\mathrm{s}}_{n,l}\right)\left(D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}\right) \tag{10}$$

where $D^{\mathrm{s,t}}_{m,n,l}$ denotes the transmission delay of RU_m acquiring F_l by associating with SBS_n, $D^{\mathrm{s,t}}_{n,l}$ denotes the transmission delay between MBS_n and the SBS acquiring F_l, $y^{\mathrm{b}}_{n,l}$ denotes the binary content placement variable of the SBSs, i.e., $y^{\mathrm{b}}_{n,l} = 1$ if F_l is placed at SBS_n; otherwise, $y^{\mathrm{b}}_{n,l} = 0$.

$D^{\mathrm{s,t}}_{m,n,l}$ in (10) can be computed as

$$D^{\mathrm{s,t}}_{m,n,l} = \frac{S_l}{R^{\mathrm{s}}_{m,n}} \tag{11}$$

where $R^{\mathrm{s}}_{m,n}$ denotes the data rate of the transmission link between RU_m and SBS_n, which can be expressed as

$$R^{\mathrm{s}}_{m,n} = W_n \log_2 \left(1 + \frac{P^{\mathrm{s}}_n g^{\mathrm{s}}_{m,n}}{\sigma^2}\right) \tag{12}$$

where P_n^{s} denotes the transmission power of SBS_n and $g_{m,n}^{\mathrm{s}}$ is the channel gain between RU_m and SBS_n.

$D_{n,l}^{\mathrm{s,t}}$ in (10) can be expressed as

$$D_{n,l}^{\mathrm{s,t}} = \frac{S_l}{R_n^{\mathrm{s}}} \tag{13}$$

where R_n^{s} denotes the data rate of the transmission link between SBS_n and the MBS, which can be computed as

$$R_n^{\mathrm{s}} = W_0 \log_2 \left(1 + \frac{P_n^{\mathrm{s}} g_n^{\mathrm{s}}}{\sigma^2}\right) \tag{14}$$

where P_n^{s} is the transmission power of the MBS when transmitting to SBS_n and g_n^{s} is the channel gain between the MBS and SBS_n.

3.2 Optimization Constraints

To design the optimal joint user association and content placement policy which minimizes the service delay of all the RUs, a number of optimization constraints have to be considered.

User Association Constraints In this paper, we assume that each RU can acquire the required content by means of at most one association mode, i.e.,

$$\mathrm{C1}: \sum_{k=1}^{K} x_{m,k}^{\mathrm{d}} + \sum_{n=1}^{N} x_{m,n}^{\mathrm{s}} + x_m \leq 1. \tag{15}$$

As for D2D communication, we assume each SU can only offer service for at most one RU provided that the SU has cached the required content of the RU. Hence, we can express the constraints as

$$\mathrm{C2}: \sum_{m=1}^{M} x_{m,k}^{\mathrm{d}} \leq 1, \tag{16}$$

$$\mathrm{C3}: \ a_{m,l}\, x_{m,k}^{\mathrm{d}} \leq y_{k,l}^{\mathrm{d}}. \tag{17}$$

Accounting for the bandwidth capacity constraints of the MBS and the SBSs, the number of RUs associating with each SBS or the MBS should not exceed the maximum number of the subchannels of the corresponding BS, which can be formulated as

$$\mathrm{C4}: \sum_{m=1}^{M} x_{m,n}^{\mathrm{s}} \leq A_n, \tag{18}$$

$$\mathrm{C5}: \sum_{m=1}^{M} x_m \leq A_0. \tag{19}$$

Data Rate Constraints We assume that the RUs with certain content demand may have different minimum data rate requirements, thus the data rate constraint of RU_m can be expressed as

$$\mathrm{C6}: \sum_{k=1}^{K} x_{m,k}^{\mathrm{d}} R_{m,k}^{\mathrm{d}} + \sum_{n=1}^{N} x_{m,n}^{\mathrm{s}} R_{m,n}^{\mathrm{s}} + x_m R_m \geq R_m^{\min} \tag{20}$$

where $R_m^{\min}$ denotes the minimum data rate requirement of RU_m.

Caching Storage Constraints of the SBSs Considering the limited and various cache capacity of the SBSs, the number of contents placed in the cache of the SBSs should be limited to the maximum cache storage constraint, which can be expressed as

$$\mathrm{C7}: \sum_{l=1}^{F} y_{n,l}^{\mathrm{s}} S_l \leq \mathrm{S}_n^{\mathrm{s}} \tag{21}$$

where $\mathrm{S}_n^{\mathrm{s}}$ denotes the maximum cache capacity of SBS_n.

User Content Request Constraints In this paper, we assume that each RU can only access one content, i.e., $\sum_{l=1}^{L} a_{m,l} \leq 1$. It is apparent that user association and content placement should subject to user requirement on certain content. More specifically, in the case that one RU does not pose requirement on one particular content, no corresponding user association and content placement strategy should be designed, i.e., if there have no request for F_l, we set $x_{m,k}^{\mathrm{d}}$, $x_{m,n}^{\mathrm{s}}$, x_m and $y_{n,l}^{\mathrm{s}} = 0$, otherwise, $x_{m,k}^{\mathrm{d}}$, $x_{m,n}^{\mathrm{s}}$, x_m and $y_{n,l}^{\mathrm{s}} = 0$ or 1. We can rewrite the constraints of user content request as

$$\mathrm{C8}: (1 - \sum_{l=1}^{L} a_{m,l}) x_{m,k}^{\mathrm{d}} \leq 0, \tag{22}$$

$$\mathrm{C9}: (1 - \sum_{l=1}^{L} a_{m,l}) x_{m,n}^{\mathrm{s}} \leq 0, \tag{23}$$

$$\mathrm{C10}: (1 - \sum_{l=1}^{L} a_{m,l}) x_m \leq 0, \tag{24}$$

$$\mathrm{C11}: (1 - \sum_{l=1}^{L} a_{m,l}) y_{n,l}^{\mathrm{s}} \leq 0. \tag{25}$$

3.3 Optimization Problem

Jointly considering the optimization objective and the constraints, we can formulate the optimization problem as follows.

$$\min_{x^{\mathrm{d}}_{m,k}, x^{\mathrm{s}}_{m,n}, x_m, y^{\mathrm{s}}_{n,l}} \quad D \qquad (26)$$
$$\text{s.t.} \qquad \mathrm{C1} - \mathrm{C11}.$$

4 Solution of the Optimization Problem

The problem in (26) is an integer-nonlinear programming problem, it is difficult to solve it directly. To solve the problem, in this section, we apply McCormick envelopes to remove the coupling among optimization variables in (26) and equivalently transform the optimization problem into three subproblems by applying Lagrangian partial relaxation, then we solve the subproblems by using Hungarian method and unidimensional knapsack algorithm respectively.

4.1 Reformulation of the Optimization Problem

To decouple the user association variables $x^{\mathrm{s}}_{m,n}$ and the content placement variables $y^{\mathrm{s}}_{n,l}$ in the objective function in (26), we introduce a new variable, $z^{\mathrm{s}}_{m,n,l} = x^{\mathrm{s}}_{m,n} y^{\mathrm{s}}_{n,l}$ and rewrite the optimization problem by using McCormick envelopes [11]. For convenience, we set

$$\mathbf{X} = \{x^{\mathrm{d}}_{m,k}, x^{\mathrm{s}}_{m,n}, x_m | \mathrm{RU}_m \in \mathrm{RU}, \mathrm{SBS}_n \in \mathrm{SBS}, \mathrm{SU}_k \in \mathrm{S}_U\}, \qquad (27)$$
$$\mathbf{Y} = \{y^{\mathrm{s}}_{n,l} | \mathrm{SBS}_n \in \mathrm{SBS}, F_l \in F\}, \qquad (28)$$
$$\mathbf{Z} = \{z^{\mathrm{s}}_{m,n,l} | \mathrm{RU}_m \in \mathrm{RU}, \mathrm{SBS}_n \in \mathrm{SBS}, F_l \in F\}. \qquad (29)$$

The original optimization problem can be rewritten as

$$\min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} a_{m,l} x^{\mathrm{d}}_{m,k} y^{\mathrm{d}}_{k,l} D^{\mathrm{d}}_{m,k,l} + \sum_{m=1}^{M}\sum_{l=1}^{L} a_{m,l} x_m \{D^{\mathrm{t}}_{m,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}\}$$
$$\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{L} \{a_{m,l} x^{\mathrm{s}}_{m,n} (D^{\mathrm{s,t}}_{m,n,l} + D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}) -$$
$$a_{m,l} z^{\mathrm{s}}_{m,n,l} (D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l})\} \qquad (30)$$
$$\text{s.t.} \quad \mathrm{C1} - \mathrm{C11 \text{ in } (26)},$$
$$\mathrm{C12}: \ z^{\mathrm{s}}_{m,n,l} \geq 0,$$
$$\mathrm{C13}: \ z^{\mathrm{s}}_{m,n,l} \geq x^{\mathrm{s}}_{m,n} + y^{\mathrm{s}}_{n,l} - 1,$$
$$\mathrm{C14}: z^{\mathrm{s}}_{m,n,l} \leq y^{\mathrm{s}}_{n,l},$$
$$\mathrm{C15}: z^{\mathrm{s}}_{m,n,l} \leq x^{\mathrm{s}}_{m,n}.$$

4.2 Lagrangian Partial Relaxation and Dual Problem Formulation

Obviously the optimization problem in (30) is convex which can be solved using traditional optimization tools. In this subsection, we apply the method of Lagrange partial relaxation [12] to incorporate C13–C15 into the function (30), which can be calculated as

$$\begin{aligned}
&\max_{\boldsymbol{\varphi},\boldsymbol{\nu},\boldsymbol{\eta}} \min_{\mathbf{X},\mathbf{Y},\mathbf{Z}} L\left(\varphi_{m,n,l}, \nu_{m,n,l}, \eta_{m,n,l}, x^{\mathrm{d}}_{m,k}, x^{\mathrm{s}}_{m,n}, x_m, y^{\mathrm{s}}_{n,l}, z^{\mathrm{s}}_{m,n,l}\right). \\
&= \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{L}\{a_{m,l}x^{\mathrm{s}}_{m,n}(D^{\mathrm{s,t}}_{m,n,l} + D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}) - \\
&\quad a_{m,l}z^{\mathrm{s}}_{m,n,l}(D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}) + \\
&\quad \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} a_{m,l}x^{\mathrm{d}}_{m,k}y^{\mathrm{d}}_{k,l}D^{\mathrm{d}}_{m,k,l} + \sum_{m=1}^{M}\sum_{l=1}^{L} a_{m,l}x_m\{D^{\mathrm{t}}_{m,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}\} + \\
&\quad \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{F}\{a_{m,l}\varphi_{m,n,l}(x^{\mathrm{s}}_{m,n} + y^{\mathrm{s}}_{n,l} - 1 - z^{\mathrm{s}}_{m,n,l}) + a_{m,l}\nu_{m,n,l}(z^{\mathrm{s}}_{m,n,l} - y^{\mathrm{s}}_{n,l}) + \\
&\quad a_{m,l}\eta_{m,n,l}(z^{\mathrm{s}}_{m,n,l} - x^{\mathrm{s}}_{m,n})\}
\end{aligned} \tag{31}$$

where $\varphi \geq 0$, $\nu \geq 0$ and $\eta \geq 0$ are the corresponding lagrange multipliers for C13, C14 and C15.

4.3 Dual Decomposition and Solution

By examining the optimization problem formulated in (31), it can be validated that both the objective problem and the constraints are separable in terms of $x^{\mathrm{d}}_{m,k}$, $x^{\mathrm{s}}_{m,n}$, x_m, $y^{\mathrm{s}}_{n,l}$ and $z^{\mathrm{s}}_{m,n,l}$. We can further decompose the problem into three subproblems, that is

$$\begin{aligned}
P_1 : \min_{\mathbf{X}} &\sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} a_{m,l}x^{\mathrm{d}}_{m,k}y^{\mathrm{d}}_{k,l}D^{\mathrm{d}}_{m,n,l} + \sum_{m=1}^{M}\sum_{l=1}^{L} a_{m,l}x_m\{D^{\mathrm{t}}_{m,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l}\} \\
&\sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{L} a_{m,l}x^{\mathrm{s}}_{m,n}\left(D^{\mathrm{s,t}}_{m,n,l}\ D^{\mathrm{s,t}}_{n,l} + D^{\mathrm{B}}_{m,l} + D^{\mathrm{w}}_{m,l} + \varphi_{m,n,l} - \eta_{m,n,l}\right)
\end{aligned} \tag{32}$$

$$\text{s.t.}\ \ \mathrm{C1} - \mathrm{C6},\ \mathrm{C8} - \mathrm{C10}.$$

$$P_2 : \max_{\mathbf{Y}} \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{l=1}^{L} a_{m,l}y^{\mathrm{s}}_{n,l}\left(\nu_{m,n,l} - \varphi_{m,n,l}\right) \tag{33}$$

$$\text{s.t.}\ \ \mathrm{C7}, \mathrm{C11}.$$

$$P_3 : \min_{\mathbf{z}} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{l=1}^{L} a_{m,l} z^{\mathrm{s}}_{m,n,l} (v_{m,n,l} + \eta_{m,n,l} - \varphi_{m,n,l} - D^{\mathrm{s,t}}_{n,l} - D^{\mathrm{B}}_{m,l} - D^{\mathrm{w}}_{m,l}) \tag{34}$$

s.t. C12.

The subproblem P_1 involves only the association variables $x^{\mathrm{d}}_{m,k}$, $x^{\mathrm{s}}_{m,n}$, x_m, which can be transformed into a balanced assignment problem and solved by using Hungarian method [13]. The subproblem P_2 involves the content placement variables $y^{\mathrm{s}}_{n,l}$, which can be decomposed into $|N|$ unidimensional knapsack problems, one for each SBS_n, which can be solved independently. Similar to subproblem P1, P3 can also be solved by using Hungarian method.

4.4 Updating Lagrange Multipliers

Lagrange multipliers can be updating based on a subgradient method for finding the locally optimal solution of above three subproblems via the form of an iterative operation, i.e.,

$$\varphi_{m,n,l}(t+1) = [\varphi_{m,n,l}(t) + \alpha(t) d(\varphi_{m,n,l}(t))]^{+} \tag{35}$$

$$v_{m,n,l}(t+1) = [v_{m,n,l}(t) + \alpha(t) d(v_{m,n,l}(t))]^{+} \tag{36}$$

$$\eta_{m,n,l}(t+1) = [\eta_{m,n,l}(t) + \alpha(t) d(\eta_{m,n,l}(t))]^{+} \tag{37}$$

where $[\mathrm{z}]^{+} = \max\{0, \mathrm{z}\}$ and $\alpha(t) = \varepsilon \frac{u_b - l_b}{||g(t)||^2}$ [11] is the step-size in tth iteration. ε is the positive control parameter. u_b and l_b are the upper bound and lower bound respectively. Conducting the above process iteratively, the algorithm will achieve convergence and then obtain the globally near-optimal user association and content placement strategy.

5 Simulation Results

In this section, we examine the performance of the proposed algorithm and compare the algorithm with other two algorithms via simulation. In the simulation, we consider a D2D-enabled HCN consisting of one MBS, two SBSs and a number of UEs. We consider 10 SUs and other parameters are summarized in Table 1. We initially set ε to 2.0 and update it to $\varepsilon = \varepsilon/2$ if there is no variation in the upper bound for about 50 successive iterations.

Figure 2 shows the upper bound and lower bound versus the number of iterations obtained from our proposed algorithm. The number of RUs is chosen as 40 and we set subchannel bandwidth is 1 MHz. We assume each RU makes random requests from the content library and the SUs randomly pre-caching some contents. It can be observed that the upper bound and lower bound nearly simultaneously converges within less than 140 iterations.

Table 1. Simulation parameters

Parameters	Value
Transmission power of SU_k (P_k^d), SBS_n(P_n^s), MBS (P_n)	0.25 W, 2 W, 40 W
Storage capacity (SBS_n^s)	10
The number of the contents (L)	50
Content size (S_l)	[1, 10] Mbits
Noise power (σ^2)	–174 dBm/HZ
Maximum numbers of associated users of the MBS (A_0), SBS_n (A_n)	40, 10
Minimum data rate requirements of users (R_m^{min})	100 Kbps
Backhaul delay ($D_{m,l}^B$)	$\{1, 2\}$
Channel path loss model	128.1 + 27log(d) dB, d denotes the distance

Figure 3 shows the service delay versus different number of RUs. we compare the performance of proposed scheme with two algorithms: Scheme 1 proposed in [8] and Scheme proposed in [10]. From the figure, we can see that the service delay increases with the increase of backhaul delay and our proposed scheme outperforms them and the performance gap is lager with the increasing number of the RUs. This is because Scheme 1 only considers cache-enabled BSs, no D2D transmission is allowed. In Scheme 2, the popular contents are pre-caching into BSs and UEs and request UEs associate to BSs or UEs providing its maximum received power. However, separately designing content placement without jointly consider user association will not efficient improve system performance.

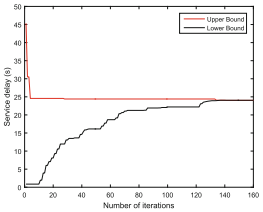

Fig. 2. Service delay vs number of iterations.

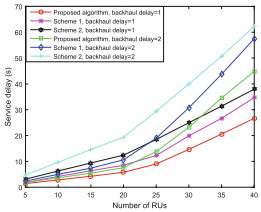

Fig. 3. Service delay vs the number of RUs.

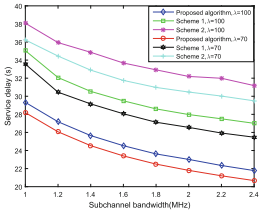

Fig. 4. Service delay vs the subchannel bandwidth

Figure 4 shows the service delay versus the subchannel bandwidth of the MBS. We set backhaul delay as 1, the service rate of the MBS, i.e., μ=110 and examine the service delay performance corresponding to different packet arrival rate. From the figure we can see that the service delay increases with the increase

of the arrival rate and our proposed algorithm achieves better performance than other two schemes. This is because our proposed algorithm can make a suitable adjustment to deal with different bandwidth status while the other two schemes fail to consider it, thus result in relatively lower performance.

6 Conclusions

In this paper, we investigated the joint user association and content placement strategy in D2D-enabled HCN and formulate joint user association and content caching problem as the service delay minimization problem of all the RUs. By applying McCormick envelopes and Lagrangian partial relaxation, we decoupling the original problem into three subprolbems and then solve it by using Hungarian method and unidimensional knapsack algorithm. Numerical results demonstrated the proposed algorithm outperforms previously proposed schemes.

References

1. Wang, X., Li, X., Leung, V.C.M., Nasiopoulos, P.: A framework of cooperative cell caching for the future mobile networks. In: Hawaii International Conference on System Sciences, Kauai, HI, pp. 5404–5413 (2015)
2. Li, X., Wang, X., Li, K., Leung, V.C.M.: Collaborative hierarchical caching for traffic offloading in heterogeneous networks. In: IEEE ICC, pp. 1–6(2017)
3. Tan, Z., Li, X., Yu, F.R., Chen, L., Ji, H., Leung, V.C.M.: Joint access selection and resource allocation in cache-enabled HCNs with D2D communications. In: IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6 (2017)
4. Shanmugam, K., Golrezaei, N., Dimakis, A., Molisch, A.F., Gaire, G.: FemtoCaching: wireless content delivery through distributed caching helpers. IEEE Trans. Inform. Theory **59**(12), 8402–8413 (2013)
5. Song, J., Song, H., Choi, W.: Optimal content placement for wireless Femto-caching network. IEEE Trans. Wirel. Commun. **16**(7), 4433–4444 (2017)
6. Li, X., Wang, X., Li, K., Han, Z., Leung, V.C.M.: Collaborative multi-tier caching in heterogeneous networks: modeling, analysis, and design. IEEE Trans. Wirel. Commun. **16**(10), 6926–6939 (2017)
7. Kang, H.J., Kang, C.G.: Mobile device-to-device (D2D) content delivery networking: a design and optimization framework. J. Commun. Netw. **16**(5), 568–577 (2014)
8. Dai, B., Yu, W.: Joint user association and content placement for cache-enabled wireless access networks. In: IEEE ICASSP, pp. 3521–3525 (2016)
9. Khreishah, A., Chakareski, J., Gharaibeh, A.: Joint caching, routing, and channel assignment for collaborative small-cell cellular networks. IEEE J. Sel. Areas Commun. **34**(8), 2275–2284 (2016)
10. Yang, C., Yao, Y., Chen, Z., Xia, B.: Analysis on cache-enabled wireless heterogeneous networks. In IEEE Trans. Wirel. Commun. **15**(1), 131–145 (2016)
11. Wang, Y., Tao, X., Zhang, X., Mao, G.: Joint caching placement and user association for minimizing user download delay. In: IEEE Access, vol. 4, pp. 8625–8633 (2016)
12. Bertsekas, D., Nedic, A., Ozdaglar, A.: Convex analysis and optimization. In: Athena. Scientific Press (2003)
13. Kuhn, H.W.: The Hungarian method for the assignment problem. Nav. Res. Logist. **52**(1), 7–21 (2005)