Joint User Association and Interference Mitigation for D2D-Enabled Heterogeneous Cellular Networks

Tianqing Zhou · Yongming Huang · Luxi Yang

Published online: 9 July 2016
© Springer Science+Business Media New York 2016

Abstract The heterogeneous cellular networks (HCNs) with device-to-device (D2D) communications have been a promising solution to cost-efficient delivery of high data rates. A key challenge in such D2D-enabled HCNs is how to design an effective association scheme with D2D mode selection for load balancing. Moreover, the offloaded users and D2D receivers (RXs) would suffer strong interference from BSs, especially from high-power BSs. Evidently, a good association scheme should integrate with interference mitigation. Thus, we first propose an effective resource partitioning strategy that can mitigate the interference received by offloaded users from high-power BSs and the one received by D2D RXs from BSs. Based on this, we then design a user association scheme for load balancing, which jointly considers user association and D2D mode selection to maximize network-wide utility. Considering that the formulated problem is in a nonlinear and mixed-integer form and hard to tackle, we adopt a dual decomposition method to develop an efficient distributed algorithm. Simulation results show that the proposed scheme provides a load balancing gain and a resource partitioning gain.

Keywords Load balancing · User association · Heterogeneous cellular networks · Distributed algorithm · D2D communications

1 Introduction

Heterogeneous cellular networks (HCNs) have been widely regarded as a promising solution to improving area’s spectral efficiency, alleviating traffic congestions in hot-spots, and thus enhancing the end-user experience [1–4]. However, due to the limited backhaul connections between low-power base stations (BSs), some offloading techniques cannot fully balance the loads among different BSs. It is highly possible that some BSs are severely congested while adjacent BSs are very lightly loaded. To further alleviate congestions and increase system throughput, HCNs with device-to-device (D2D) communications have been a good option in recent years [5–7].

The D2D communication directly takes place between two closely located users. As a kind of proximity communication, it has attracted more and more attention due to its advantages such as offloading traffic, low power consumption and latency, and high data rate and spectral efficiency [8, 9]. Moreover, the D2D communications can also enhance the physical layer security [10]. Thus, the D2D communications have been advocated in various wireless networks. Although D2D communications can achieve many advantages, the D2D pairs may suffer severe interference from BSs, especially from high-power BSs. To fully exploit the potential of D2D communications, we need to consider some interference mitigation techniques such as power control and resource partitioning [11].

Since various BSs coexist in HCNs, the user association that assigns users to different BSs is a challenging topic.
[12, 13]. When the conventional signal strength-based association is applied to HCNs, the obtained load distribution is very imbalanced since most users are associated with high-power BSs. Thus, some association approaches that are well applied in traditional cellular networks may not be appropriate for HCNs. Moreover, when D2D communication techniques are incorporated into HCNs, the user association problem becomes more complicated. To make full use of novel network framework, we are required to design an association scheme with offloading capability. Next, we will focus on some offloading strategies for HCNs and D2D-enabled cellular networks.

1.1 Related work

To accommodate new characteristics of HCNs, many efforts in the literature toward load balancing for HCNs have been made. As a most frequently utilized method, the biasing method (cell range expansion) gives low-power BSs an offse/bias to attract more users for them. In work [14], authors give closed-form expressions of the downlink data rate and signal-interference-plus-noise ratio (SINR) distribution under a biasing method, but cannot achieve a closed-form expression of an optimal offset. Since this association approach only focuses on the load balancing for HCNs and doesn’t consider any interference mitigation measures, the offloaded users may receive the strong interference from high-power BSs. In order to avoid this interference, Singh et al. [15] propose a resource partitioning scheme so that the system throughput is greatly improved. Moreover, Jo et al. [16] also develop a tractable framework for the downlink SINR and rate analyses under a biasing method. Although the biasing method is simple for load balancing in HCNs, the closed-form expression of an optimal offset is not provided in existing works.

Besides the biasing method, there are other offloading schemes for HCNs. In work [17], authors perform the user association to maximize a sum-utility that relates to long-term rates, and develop an efficient distributed user association algorithm. Considering the cross-tier interference, Shen et al. [18] design a user association scheme with power control or beamforming based on work [17], and put forward an improved algorithm. Unlike work [17], authors in work [13] studies the interplay of user association and resource partitioning, and tries to design an association algorithm with guaranteeing the upper bounds of system performance. Recently, Cho et al. [19] adopt repulsive cell activation in the interfering daughter-cell network to balance cell loads, but this method appears to be high complicated.

Due to the limited backhaul connections between low-power BSs, most of aforementioned schemes may not achieve efficient load balancing. In other words, some offloading techniques cannot fully balance the loads among different BSs. To fully exploit the potential of D2D communications and thus alleviate network congestions, the user association for load balancing in D2D-enabled HCNs has been investigated in works [5, 6]. In work [5], authors propose an online offloading scheme that doesn’t consider interference mitigation. Moreover, the D2D pairs in work [5] just play the role of relay and cannot support direct data communication. Unlike work [5], the D2D receivers (RXs) in work [6] can directly communicate with D2D transmitters (TXs) or other BSs. Authors in work [6] jointly consider user association, D2D mode selection and power control for uplink D2D-enabled HCNs. Note that the D2D mode selection represents that a D2D RX selects some D2D TX or BS to connect. So far, few existing efforts jointly consider user association and D2D mode selection for load balancing in downlink D2D-enabled HCNs. It is necessary to design an efficient user association algorithm that can balance the loads among different BSs and fully utilize D2D communications to offload traffic.

1.2 Contributions and organization

In this paper, we propose a load balancing scheme for downlink D2D-enabled HCNs, which jointly considers the user association for cellular users and the D2D mode selection for potential D2D pairs. Moreover, in order to mitigate the interference received by offloaded users from high-power BSs and the one received by D2D RXs from BSs, we design a resource partitioning scheme. Specially, the whole frequency band is cut into three subbands including subbands 1, 2 and 3, where the high-power BSs monopolize subband 1, the D2D TXs can monopolize subband 3 and the low-power BSs can utilize subbands 1 and 2. In this way, some offloaded users can be associated with the subband 2 of low-power BSs to avoid the strong interference from high-power BSs, some D2D RXs can be associated with the subband 3 of D2D TXs to avoid the interference from all BSs. At last, the load balancing scheme is formulated as a network-wide utility maximization problem. Considering that the formulated problem is in a nonlinear and mixed-integer form and hard to tackle, we utilize a dual decomposition method to develop an efficient distributed algorithm.

The rest of this paper is organized as follows. In Section 2, we describe our system model including network model and resource partitioning model. In Section 3, we formulate the user association problem. In Section 4, we design a distributed algorithm using dual decomposition. In Section 5,
we give numerical results of proposed algorithms and other algorithms in terms of load balancing gains and resource partitioning gains. In Section 6, we present further discussions and conclusions.

2 System model

In this paper, we consider a D2D-enabled HCN that is the conventional macrocellular network overlaid with pico BSs (PBSs) and potential D2D pairs. This deployment is illustrated in Fig. 1, where MBS represents the macro BS, and each D2D pair contains D2D RX and D2D TX. In general, cellular users, D2D TXs and D2D RXs are called users, PBSs and MBSs are called BSs. Moreover, cellular users, D2D TXs and D2D RXs are also called receivers, PBSs, MBSs and D2D TXs are also called transmitters. Note that the D2D TXs are the receivers of BSs and they are also the transmitters of the corresponding D2D RXs.

To proceed, we need to make the following assumption.

Assumption 1 Each BS equally allocates power to all subbands being employed.

Remark This assumption has been widely used for downlink resource allocation due to its implementation simplicity and analytical tractability. Moreover, equal power allocation can achieve near-optimal solutions in many cases, especially at high SINR regime [20–22].

To reduce the interference received by offloaded users from MBSs and the one received by D2D RXs from BSs, we introduce a resource partitioning scheme. Specially, the whole frequency band is cut into three subbands including subband 1 (s = 1), subband 2 (s = 2) and subband 3 (s = 3). As illustrated in Fig. 1, MBSs just utilize subband 1, D2D TXs just use subband 3, and PBSs can utilize subbands 1 and 2. Note that the bandwidths of subbands 1, 2 and 3 are \((1 - \eta)W_1\), \(\eta W_1\) and \(W_2\) respectively, where \(W_1 = W - W_2\), \(W_2 = W_{prb}\), \(W\) is the system bandwidth and \(W_{prb}\) represents the bandwidth of one PRB (physical resource block). According to the descriptions of Long Term Evolution (LTE) [23], adjacent twelve subcarriers are grouped into one PRB with 180KHz, which is the smallest unit that can be allocated to each user. Considering that one D2D TX just serves one user (the corresponding D2D RX), one PRB may be sufficient for D2D TX. In this way, some offloaded users can be associated with the subband 2 of PBSs to avoid the strong interference from MBSs, some D2D RXs can be associated with the subband 3 of D2D TXs to avoid the interference from all BSs. Under the resource partitioning scheme, all users may be associated with the subband 1 of MBSs or some subband of PBSs. Moreover, any D2D RX may also be associated with the subband 3 of its corresponding D2D TX.

Now, we let the set of MBSs be \(N_m\), let the set of PBSs be \(N_p\), let the set of cellular users be \(K_c\), let the set of D2D TXs be \(K_d\) and denote the set of D2D RXs as \(K_r\), where \(N = N_m \cup N_p \cup N_d\), \(N_{mp} = N_m \cup N_p\) and \(K = K_c \cup K_d \cup K_r\). Meanwhile, we write the set of subbands as \(S = \{1, 2, 3\}\), and write the cardinalities of sets \(N, N_{mp}, N_m, N_p, S\) and \(K\) as \(N = |N|, N_{mp} = |N_{mp}|, N_m = |N_m|, N_p = |N_p|, S = |S|\) and \(K = |K|\) respectively. Then, the SINR received
by user \( k \in K \) from BS \( n \in N_m \) on subband 1 can be written as

\[
\text{SINR}_{nk1} = \frac{p_n g_{nk}}{\sum_{j \in N_m \setminus \{n\}} p_j g_{jk} + (1 - \eta) N_0},
\]

(1)

the SINR received by user \( k \in K \) from BS \( n \in N_p \) on subband 2 can be written as

\[
\text{SINR}_{nk2} = \frac{p_{nk} g_{nk}}{\sum_{j \in N_p \setminus \{n\}} p_{jk} g_{jk} + \eta N_0},
\]

(2)

and the SINR received by D2D RX \( k \in K_r \) from the corresponding D2D TX \( n_k \) on subband 3 can be written as

\[
\text{SINR}_{nk3} = \frac{p_{nk3} g_{nk3}}{\sum_{j \in N_r \setminus \{n_k\}} p_{jk3} g_{jk3} + W_2 N_0},
\]

(3)

where \( p_{ns} \) represents the transmit power of transmitter \( n \) on subband \( s \); \( g_{nsk} \) denotes the channel gain between receiver \( k \) and transmitter \( n \) on subband \( s \); \( N_0 \) is the noise power spectral density; the D2D TX \( n_k \) is the corresponding transmitter of D2D RX \( k \) in a potential D2D pair. Since the D2D TXs also need to be associated with BSs, the D2D TX \( k \in K_r \) can be regarded as the receiver of BSs, but it can also be seen as the transmitter of the corresponding D2D RX. By employing the assumption of equal power allocation, we can have \( p_{n1} = P_m \), \( p_{n2} = 0 \) mW and \( p_{n3} = 0 \) mW for any MBS \( n \in N_m \), \( p_{n1} = (1 - \eta) P_p \), \( p_{n2} = \eta P_p \) and \( p_{n3} = 0 \) mW for any PBS \( n \in N_p \), and \( p_{n1} = 0 \) mW, \( p_{n2} = 0 \) mW and \( p_{n3} = P_t \) for any D2D TX \( n \in N_r \), where \( P_m \), \( P_p \) and \( P_t \) are the transmit powers of MBS, PBS and D2D TX respectively.

In practical implement, the SINR received by any D2D RX on subband 3 should depend on the association results of other D2D RXs. In other words, the D2D RXs just receive the interference from the D2D TXs with associated users (D2D RXs) on subband 3. However, it will ensure that the design of effective association algorithms is difficult and the development of distributed algorithm is highly almost impossible. For ease of algorithm design, we just consider a special case, i.e., the D2D TXs always transmit signal in the association phase even if no users are served. That is to say, the interference produced by these D2D TXs always exists. After association, we can let D2D TXs without associated users no longer transmit signal to the corresponding D2D RXs, i.e., the interference produced by these D2D TXs don’t exist. In this way, our association algorithm may just achieve a sub-optimal solution, which means some D2D pairs are not utilized, i.e., the capacity of offloading traffic of D2D communication decreases. All the same, it can also relatively offload traffic that is the goal of introducing D2D communication into HCNs.

When the SINRs are given, the achievable rate [in bps] received by user \( k \in K \) from BS \( n \in N_m \) on subband 1 can be written as

\[
r_{nk1} = (1 - \eta) W_1 \log_2 (1 + \text{SINR}_{nk1}),
\]

(4)

the achievable rate [in bps] received by user \( k \in K \) from BS \( n \in N_p \) on subband 2 can be written as

\[
r_{nk2} = \eta W_1 \log_2 (1 + \text{SINR}_{nk2}),
\]

(5)

and the achievable rate [in bps] received by D2D RX \( k \in K_r \) from the corresponding D2D TX \( n_k \) on subband 3 can be written as

\[
r_{nk3} = W_2 \log_2 (1 + \text{SINR}_{nk3}).
\]

(6)

Moreover, since some subbands of some transmitters cannot be utilized by receivers, the achievable rates can be set to 0 in these cases. In order to meet the demand of algorithm design, i.e., avoid the case log(0), we add a very small constant \( \vartheta \) to the achievable rate. Thus, we have

\[
r_{nk} = r_{nk} + \vartheta, \text{ e.g., } \vartheta = 10^{-20}.
\]

To proceed, we need to give the following definitions.

**Definition 1** [17]. The effective load of transmitter \( n \) on subband \( s \) is \( y_{ns} = \sum_{k \in K} x_{nsk} \), where \( x_{nsk} \) represents the association indicator, i.e., \( x_{nsk} = 1 \) when receiver \( k \) is associated with subband \( s \) of transmitter \( n \), 0 otherwise.

**Definition 2** [17, 24]. If the load of transmitter \( n \) on subband \( s \) is \( y_{ns} \), the effective (long-term) rate of receiver \( k \) who is associated with subband \( s \) of transmitter \( n \) is given by \( R_{nsk} = r_{nk}/y_{ns} \).

### 3 Problem formulation

In this section, we formulate the association scheme with resource partitioning as a network-wide utility maximization problem. Mathematically, it is given by

\[
\max_{\mathbf{x}} \quad F(\mathbf{x}) = \sum_{n \in N} \sum_{s \in S} \sum_{k \in K} x_{nsk} U_{nsk} (R_{nsk})
\]

s.t. \( \sum_{n \in N} \sum_{s \in S} x_{nsk} = 1, \forall k \in K \),

(7)

\[
\sum_{n \in N} \sum_{s \in S} x_{nsk} \in \{0, 1\}, \forall n \in N, \forall s \in S, \forall k \in K.
\]

where \( \mathbf{x} = \{x_{nsk}, n \in N, s \in S, k \in K\} \); \( U_{nsk} \) denotes the utility received by receiver \( k \) from transmitter \( n \) on subband \( s \); the first constraint means that a receiver can be only connected to one transmitter on some subband.

**Proposition 1** Under resource partitioning, if a logarithmic function is adopted as the utility function of problem (7), then the equivalent form of the problem (7) is given by

\[
\max_{\mathbf{x}} \quad G(\mathbf{x})
\]

s.t. \( \sum_{n \in N} \sum_{s \in S} x_{nsk} = 1, \forall k \in K \),

\[
\sum_{n \in N} \sum_{s \in S} x_{nsk} \in \{0, 1\}, \forall n \in N, \forall s \in S, \forall k \in K,
\]

(8)
where \( c_{nk} = \log r_{nk} \) and

\[
G (x) = \sum_{n \in N_{mp}} \sum_{k \in K} x_{nk1} \left( c_{n1k} - \log \sum_{j \in K} x_{nj1} \right) \\
+ \sum_{n \in N_p} \sum_{k \in K} x_{nk2} \left( c_{n2k} - \log \sum_{j \in K} x_{nj2} \right) \\
+ \sum_{k \in K_r} x_{n3k} \left( c_{n3k} - \log x_{n3k} \right),
\]

(9)

Proof According to the definition of achievable rate, we know that the achievable rates of MBSs on subband 1, the ones of PBSs on subbands 1 and 2, and the ones received by any D2D RX \( k \in K_r \) from the corresponding D2D TX \( n_k \) on subband 3 are larger than \( \vartheta \). In other cases, the achievable rates equal to \( \vartheta \). When the achievable rates are \( \vartheta \), these terms in the objective function \( F (x) \) of the problem (7) don’t need to be considered. In fact, the achievable rates with \( \vartheta \) represent that some subbands of transmitters are unavailable. Thus, we have

\[
F (x) \equiv G (x) = \sum_{n \in N_{mp}} \sum_{k \in K} x_{nk1} U_{nk1} (R_{nk1}) \\
+ \sum_{k \in K_r} x_{nk3} U_{nk3} (R_{nk3}) \\
+ \sum_{n \in N_p} \sum_{k \in K} x_{nk2} U_{nk2} (R_{nk2}).
\]

(10)

where \( a \equiv b \) represents that \( a \) is equivalent to \( b \).

To guarantee the user fairness, we introduce a logarithmic function as the mentioned utility function of problem (7). Then, we have the expression (9).

Seen from the objective function (9), the receiver needs to trade off the load and achievable rate when it selects the subband 1 of transmitter (BS) \( n \in N_{mp} \) or the subband 2 of BS \( n \in N_p \). In other words, the proposed scheme is not the maximal achievable rate association but the association that owns an offloading capability.

According to the definition of effective load, we can further convert problem (8) into

\[
\max_{x, y} \quad H (x, y) \\
\text{s.t.} \quad \sum_{n \in S} x_{nk} = 1, \forall k \in K, \\
\quad \sum_{k \in K_r} x_{nk1} = y_{n1}, \forall n \in N_{mp}, \\
\quad \sum_{k \in K} x_{nk2} = y_{n2}, \forall n \in N_p, \\
\quad x_{nk} \in [0, 1], \forall n \in N, \forall s \in S, \forall k \in K,
\]

(11)

where \( y = \{ y_{ns}, \forall n \in N, \forall s \in S \} \) and

\[
H (x, y) = \sum_{n \in N_{mp}} \sum_{k \in K} x_{nk1} c_{n1k} - \sum_{n \in N_{mp}} y_{n1} \log y_{n1} \\
+ \sum_{n \in N_p} \sum_{k \in K} x_{nk2} c_{n2k} - \sum_{n \in N_p} y_{n2} \log y_{n2} \\
+ \sum_{k \in K_r} x_{n3k} (c_{n3k} - \log x_{n3k}).
\]

(12)

\( \Box \)

4 Association algorithm

To find the global optimal solutions of the problem (11), global network information should be collected. That means a centralized controller is required to perform user association and coordination. Considering this case, we design a distributed algorithm for the proposed problem to achieve suboptimal solutions, which doesn’t require any coordination among BSs.

As a general method to solve an optimization problem, the dual decomposition has attracted increasing attention in the literature since it can simplify a highly complex and large-scale problem. Specially, this method breaks the original optimization problem up into smaller subproblems and separately solves these smaller ones in a distributed manner. So far, the dual decomposition method has been widely used in user association [16, 17], simultaneous routing and resource allocation [25], sum capacity maximization [26] and distributed control [27]. In this section, we design a highly efficient distributed algorithm by dual decomposition. Then, users and BSs can separately solve their two subproblems into which the dual problem is cut.

Considering the coupling constraints in the problem (11), i.e., the second and third constraints, we introduce the Lagrange multipliers to relax them. For any BS \( n \in N_{mp} \), we introduce \( \mu_{n1} \) for the corresponding (second) constraint; For any BS \( n \in N_p \), we introduce \( \mu_{n2} \) for the corresponding (third) constraint. Thus, the Lagrange function with respect to these constraints is

\[
\mathcal{L} (x, y, \mu) = \sum_{n \in N_{mp}} \sum_{k \in K} x_{nk1} c_{n1k} - \sum_{n \in N_{mp}} y_{n1} \log y_{n1} \\
+ \sum_{n \in N_p} \sum_{k \in K} x_{nk2} c_{n2k} - \sum_{n \in N_p} y_{n2} \log y_{n2} \\
+ \sum_{k \in K_r} x_{n3k} (c_{n3k} - \log x_{n3k}) \\
+ \sum_{n \in N_{mp}} \mu_{n1} \left( y_{n1} - \sum_{k \in K} x_{nk1} \right) \\
+ \sum_{n \in N_p} \mu_{n2} \left( y_{n2} - \sum_{k \in K} x_{nk2} \right).
\]

(13)
Then, the dual function can be written as

\[
I (\mu) = \begin{cases}
\max_{x, y} & L (x, y, \mu) \\
\text{s.t.} & \sum_{n \in S} \sum_{s} x_{nsk} = 1, \; \forall k \in K, \\
& x_{nsk} \in [0, 1], \; \forall n \in N, \; \forall s \in S, \; \forall k \in K,
\end{cases}
\]

(14)

and the dual problem of (11) is given by

\[
\min_{\mu} I (\mu).
\]

(15)

Considering the problem (15) is not coupling with respect to \(x\) and \(y\), we can separately obtain the primal optimal solutions. Thus, the problem (15) can be decomposed into

\[
I_1 (\mu) = \begin{cases}
\max_{x} & L_1 (x, \mu) \\
\text{s.t.} & \sum_{n \in S} \sum_{s} x_{nsk} = 1, \; \forall k \in K, \\
& x_{nsk} \in [0, 1], \; \forall n \in N, \; \forall s \in S, \; \forall k \in K,
\end{cases}
\]

(16)

and

\[
I_2 (\mu) = \max_{y} L_2 (y, \mu),
\]

(17)

where

\[
L_1 (x, \mu) = \sum_{n \in N_{mp}} \sum_{k \in K_c} x_{nk1} (c_{nk1} - \mu_{nk1}) + \sum_{n \in N_p} \sum_{k \in K_c} x_{nk2} (c_{nk2} - \mu_{nk2}) + \sum_{k \in K_r} x_{nk3k} (c_{nk3k} - \log x_{nk3k}),
\]

(18)

\[
L_2 (y, \mu) = \sum_{n \in N_{mp}} \gamma_{n1} (\mu_{n1} - \log \gamma_{n1}) + \sum_{n \in N_p} \gamma_{n2} (\mu_{n2} - \log \gamma_{n2}).
\]

(19)

When the dual optimal \(\mu^*\) is given, the optimal solutions of Eq. 11 can be obtained by separately solving its two subproblems.

Now, we solve the outer problem (15) by a gradient projection method [28]. For any BS \(n \in N_{mp}\), we search the optimal \(\mu_{n1}\) in the direction of negative gradient, i.e., \(-\nabla L (\mu_{n1})\). Similarly, for any BS \(n \in N_p\), we search the optimal \(\mu_{n2}\) in the direction of negative gradient, 

\[
\nabla I (\mu_{n2}).
\]

To obtain these gradients, we need to solve subproblem (16) and Eq. 17. According the condition of resource utilization, we can give the expanded form of the expression (18). Then, we have

\[
L_1 (x, \mu) = \sum_{n \in N_{mp}} \sum_{k \in K_c} x_{nk1} (c_{nk1} - \mu_{nk1}) + \sum_{n \in N_p} \sum_{k \in K_c} x_{nk2} (c_{nk2} - \mu_{nk2}) + \sum_{k \in K_r} x_{nk3k} (c_{nk3k} - \log x_{nk3k}).
\]

(20)

Moreover, the Eq. 20 is equivalent to

\[
L_1 (x, \mu) = \sum_{k \in K_c} \left\{ \sum_{n \in N_{mp}} x_{nk1} (c_{nk1} - \mu_{nk1}) + \sum_{n \in N_p} x_{nk2} (c_{nk2} - \mu_{nk2}) \right\} + \sum_{k \in K_r} x_{nk3k} (c_{nk3k} - \log x_{nk3k}).
\]

(21)

According the form of the subproblem (16), we can easily develop its algorithm whose detailed procedure can be found in Algorithm 1. Since this algorithm is performed by users including cellular users, D2D RXs and D2D TXs, it can be regarded as the algorithm on user’s side. In the steps 6-10 of Algorithm 1, any cellular user \(k \in K_c\)
selects the subband 1 of some BS or the subband 2 of some PBS to maximize its utility, i.e., achieve the maximal utility \( c_{n^*s^k} - \mu_{n^*s^k} \). In other words, any user \( k \) selects the subband \( s^* \) of BS \( n^* \) if \( c_{n^*s^k} - \mu_{n^*s^k} \) is the maximal value among possible associations. As mentioned in previous section, any D2D RX \( k \in K_r \) can be associated with the subband 1 of some BS or the subband 2 of some PBS or the corresponding D2D TX \( n_k \). When any D2D TX \( n_k \) is not utilized by the corresponding D2D RX \( k \in K_r \), the term \( x_{n_k3k} (c_{n_k3k} - \log x_{n_k3k}) \) in the association object can be neglected. However, when any D2D TX \( n_k \) is selected by the corresponding D2D RX \( k \in K_r \), the term \( x_{n_k3k} (c_{n_k3k} - \log x_{n_k3k}) \) can be simplified into \( c_{n_k3k} \). Thus, in the steps 11-18 of Algorithm 1, any D2D RX \( k \in K_r \) first selects the subband 1 of some BS or the subband 2 of some PBS to achieve the maximal utility \( c_{n^*s^k} - \mu_{n^*s^k} \), then selects the subband 3 of the corresponding D2D TX \( n_k \) if \( c_{n^*s^k} - \mu_{n^*s^k} < c_{n_k3k} \). Similar to the cellular users, any D2D TX \( k \in K_t \) performs user association in the steps 19-23 of Algorithm 1.

Algorithm 1 At user terminal \( k \)

1: If \( t = 0 \)
   2:   Estimate \( c \) using pilots signals from all transmitters.
3: Else
4:   Receive the information \( \mu_{n1} \) broadcasted by any BS \( n \).
5:   Receive the information \( \mu_{n2} \) broadcasted by any PBS \( n \).
6:   If \( k \in K_r \)
7:     Select the subband 1 of some BS or the subband 2 of some
8:     PBS to achieve the maximal utility: \( c_{n^*s^k} - \mu_{n^*s^k} \).
9:     \( x_{n^*s^k} = 1 \).
10: EndIf
11: If \( k \in K_r \)
12:     Select the subband 1 of some BS or the subband 2 of some
13:     PBS to achieve the maximal utility: \( c_{n^*s^k} - \mu_{n^*s^k} \).
14:     If \( c_{n^*s^k} - \mu_{n^*s^k} < c_{n_k3k} \)
15:       \( n^* = n_k; s^* = 3 \).
16: EndIf
17: \( x_{n^*s^k} = 1 \).
18: EndIf
19: If \( k \in K_t \)
20:     Select the subband 1 of some BS or the subband 2 of some
21:     PBS to achieve the maximal utility: \( c_{n^*s^k} - \mu_{n^*s^k} \).
22:     \( x_{n^*s^k} = 1 \).
23: EndIf
24: Feedback association information \( x_{n^*s^k} = 1 \) to the BS \( n^* \).
25: EndIf

In the subproblem (17), the optimal load \( y_{n1} \) of BS \( n \in N_{mp} \) on subband 1 can be calculated according to Karush-Kuhn-Tucker (KKT) [28] conditions and given by

\[
y_{n1}^{t+1} = \exp \left( \mu_{n1}^{t} - 1 \right).
\]

(22)

Similarly, the optimal load \( y_{n2} \) of BS \( n \in N_p \) on subband 2 can be given by

\[
y_{n2}^{t+1} = \exp \left( \mu_{n2}^{t} - 1 \right).
\]

(23)

After getting the optimal load at time slot \( t \), we can update the multiplier \( \mu_{n1} \) of BS \( n \in N_{mp} \) on subband 1 using

\[
\mu_{n1}^{t+1} = \mu_{n1}^{t} - \xi^{t} \left( y_{n1}^{t} - \sum_{k \in K} x_{n1k}^{t} \right).
\]

(24)

Similarly, the multiplier \( \mu_{n2} \) of BS \( n \in N_p \) on subband 2 can be updated by

\[
\mu_{n2}^{t+1} = \mu_{n2}^{t} - \xi^{t} \left( y_{n2}^{t} - \sum_{k \in K} x_{n2k}^{t} \right).
\]

(25)

where \( \xi^{t} \) is a small enough stepsize for updating \( \mu_{ns} \) at time slot \( t \). To this end, we can adopt Bertsekas's stepsize rule, i.e., equation (6.60) in [29]. Moreover, Shen et al. [18] propose a dual coordinate method to find optimal \( \mu \). These approaches ensure a faster convergence rate for the whole algorithm, but it may occupy a higher calculation complexity. For simplicity, we just consider a constant stepsize (special case for Bertsekas’s stepsize rule). As for other rules, we can easily apply them into the proposed algorithm, and thus we will no longer take them into account.

Evidently, the updating procedure of \( y_{ns} \) and \( \mu_{ns} \) takes place on BS’s side, which is listed in Algorithm 2. When the BS \( n \) represents a MBS, the load \( y_{n1} \) and the multiplier \( \mu_{n1} \) are updated using Eqs. 22 and 24 respectively, which can be found in steps 1-9 of Algorithm 2; when the BS \( n \) represents a PBS and adopts subband 1, the load \( y_{n1} \) and the multiplier \( \mu_{n1} \) are updated using Eqs. 22 and 24 respectively, which can be found in steps 16-18 of Algorithm 2; when the BS \( n \) represents a PBS and adopts subband 2, the load \( y_{n2} \) and the multiplier \( \mu_{n2} \) are updated using Eqs. 23 and 25 respectively, which can be found in steps 20-22 of Algorithm 2.
Algorithm 2 At base station \( n \)

1. If \( n \in \mathcal{N}_m \)
2. \hspace{0.5cm} If \( t = 0 \)
3. \hspace{1.5cm} Initialize stepsize \( \xi^t \) and \( \mu_{n1}^t \).
4. \hspace{1.5cm} Else
5. \hspace{2.5cm} Receive the information \( x_{n1k}^t = 1 \) from any user \( k \in \mathcal{K} \).
6. \hspace{2.5cm} Calculate \( y_{n1}^{t+1} \) using (22) and update \( \mu_{n1}^{t+1} \) using (24).
7. \hspace{2.5cm} Broadcast information \( \mu_{n1}^{t+1} \) to all users.
8. \hspace{1.5cm} EndIf
9. EndIf
10. If \( n \in \mathcal{N}_p \)
11. \hspace{1.5cm} For \( s \in \{1, 2\} \)
12. \hspace{2.5cm} If \( t = 0 \)
13. \hspace{3.5cm} Initialize stepsize \( \xi^t \) and \( \mu_{ns}^t \).
14. \hspace{2.5cm} Else
15. \hspace{3.5cm} If \( s = 1 \)
16. \hspace{4.5cm} Receive the information \( x_{n1k}^t = 1 \) from any user \( k \in \mathcal{K} \).
17. \hspace{4.5cm} Calculate \( y_{n1}^{t+1} \) using (22) and update \( \mu_{n1}^{t+1} \) using (24).
18. \hspace{4.5cm} Broadcast information \( \mu_{n1}^{t+1} \) to all users.
19. \hspace{3.5cm} Else
20. \hspace{4.5cm} Receive the information \( x_{n2k}^t = 1 \) from any user \( k \in \mathcal{K} \).
21. \hspace{4.5cm} Calculate \( y_{n2}^{t+1} \) using (23) and update \( \mu_{n2}^{t+1} \) using (25).
22. \hspace{4.5cm} Broadcast information \( \mu_{n2}^{t+1} \) to all users.
23. \hspace{2.5cm} EndIf
24. \hspace{1.5cm} EndFor
25. EndIf
26. EndIf

In the Eqs. 24 and 25, there are some interesting meanings. Specially, the multiplier \( \mu_{n1} \) can be regarded as a message between user \( k \in \mathcal{K} \) and BSs \( n \in \mathcal{N}_{mp} \) on subband 1. Furthermore, it can also be seen as the service cost of BSs \( n \in \mathcal{N}_{mp} \) on subband 1, which should be dependent on the load distribution. When \( \sum_{k \in \mathcal{K}} x_{n1k} \) and \( y_{n1} \) are deemed to be the serving demand and available service of BSs \( n \in \mathcal{N}_{mp} \) on subband 1 respectively, the cost \( \mu_{n1} \) can tradeoff supply and demand. Consequently, the cost \( \mu_{n1} \) will go up if the demand \( \sum_{k \in \mathcal{K}} x_{n1k} \) exceeds the supply \( y_{n1} \) and vice versa. Similarly, the multiplier \( \mu_{n2} \) represents a message between user \( k \in \mathcal{K} \) and BSs \( n \in \mathcal{N}_p \) on subband 2, and meanwhile it can be also interpreted as the service cost of BSs \( n \in \mathcal{N}_p \) on subband 2. When \( \sum_{k \in \mathcal{K}} x_{n2k} \) and \( y_{n2} \) are denoted as the serving demand and available service of BSs \( n \in \mathcal{N}_p \) on subband 2 respectively, the cost \( \mu_{n2} \) can tradeoff supply and demand. Thus, the cost \( \mu_{n2} \) will go up if the demand \( \sum_{k \in \mathcal{K}} x_{n2k} \) exceeds the supply \( y_{n2} \) and vice versa. In the whole association procedure, some users may not be associated with some overloaded BS when the latter increases its service price, but may be connected to some underloaded BS with decreasing price.

The aforementioned two algorithms are listed in the pattern [30], which can give us a clear insight on exchanged information and overhead. Evidently, the whole association procedure should interactively execute user’s algorithm (Algorithm 1) and BS’s algorithm (Algorithm 2). As for the proposed algorithm (whole association procedure), we can give the convergence proof for it by employing the following theorem.

Theorem 1 The proposed algorithm (whole association procedure) will converge to the optimum of dual problem (15).

Proof The derivatives of function \( I (\mu) \) are calculated by

\[
\frac{\partial I}{\partial \mu_{n1}} (\mu) = y_{n1} (\mu_{n1}) - \sum_{k \in \mathcal{K}} x_{n1k} (\mu_{n1}), \forall n \in \mathcal{N}_{mp},
\]

\[
\frac{\partial I}{\partial \mu_{n2}} (\mu) = y_{n2} (\mu_{n2}) - \sum_{k \in \mathcal{K}} x_{n2k} (\mu_{n2}), \forall n \in \mathcal{N}_p.
\]

Since the number of users scattered in network is limited in real life, \( \sum_{k \in \mathcal{K}} x_{n1k} (\mu_{n1}), \sum_{k \in \mathcal{K}} x_{n2k} (\mu_{n2}), y_{n1} (\mu_{n1}) \) and \( y_{n2} (\mu_{n2}) \) are bounded. Thus, we can easily conclude that the subgradients of dual objective function (14) are also bounded.

\[
\sup_{t} \| \frac{\partial I (\mu)}{\partial \mu_{n1}} \| \leq u, \forall n \in \mathcal{N}_{mp},
\]

\[
\sup_{t} \| \frac{\partial I (\mu)}{\partial \mu_{n2}} \| \leq u, \forall n \in \mathcal{N}_p,
\]

where \( u \) is a constant number. Evidently, our problem meets the necessary conditions of the convergence proof of subgradient decent method in [31], whereas we can prove Theorem 1.

Next, we will give some analyses for the algorithm complexity. As shown in Algorithm 1, any user has a computation complexity of \( O (K_{mp}) \), and thus this algorithm has a computation complexity of \( O (N_{mp}) \). In the Algorithm 2, any MBS has a computation complexity of \( O (K) \), and any PBS has the one of \( O (2K) \), and thus this algorithm has a computation complexity of \( O (2K) \). However, when a centralized algorithm is adopted, the computation complexity may be \( O (N_{mp} K) \) at each iteration. Evidently, a centralized algorithm for solving the formulated problem should be more complicated than the advocated algorithm.

Moreover, the Eqs. 24 and 25 show that BSs only require very little local information to adjust the multiplier \( \mu \) in a completely distributed manner. Specially, each BS broadcasts its service cost that only contains very little information to all users, and each user should send its service
demand to the BS to which it expects to connect. Evidently, the amount of exchanged information of the proposed algorithm should be $N_m + 2N_p + 3K$ at each iteration. Unlike the proposed algorithm, a centralized algorithm may have the amount of exchanged information that is proportional to $N_mK + 2N_pK$ at each iteration. Thus, the proposed algorithm should be more practical and favored for some cases, especially for large-scale problems.

5 Numerical results

In D2D-enabled HCNs, the transmit powers of MBS, PBS and D2D TX are 46 dBm, 30 dBm and 20 dBm respectively. In addition, MBSs are deployed into a conventional cellular network, potential D2D pairs, PBSs and cellular users are uniformly and independently scattered into each macrocell. We assume that the distance between any two MBSs is 1000 m, the distance between D2D RX and D2D TX is greater than 10 m and less than 50 m, and the noise power spectral density is -174dBm/Hz. In the propagation environment, we adopt 128 and 140 dB as the pathloss model of MBSs and use 128.1 + 37.6 log 10 (d) dB as the pathloss model of PBSs and D2D pairs [32], where $d$ represents the distance between receiver and transmitter in kilometers. Moreover, MBSs and PBSs own log-normal shadowing with standard deviation 10 dB, and D2D pairs have the one with standard deviation 12 dB [32].

Considering that the proposed association scheme jointly performs user association and resource partitioning, we can simply call it association with resource partitioning (ARP). To highlight the effectiveness of scheme ARP, we introduce other association schemes for comparison, which mainly include two types of association schemes, i.e., conventional association and load balancing association. Specially, the former refers to maximal achievable rate association (MARA) and maximal SINR association (MSA), and the latter includes biasing rate association (BRA) and biasing SINR association (BSA). Evidently, the schemes MARA and BRA are closely related to available bandwidth, but the schemes MSA and BSA are not. The detailed descriptions for them are listed as follows. Note that these introduced algorithms can be referred to the work [17], but they slightly differ from the algorithms in work [17].

Maximal Achievable Rate Association (MARA) In the scheme MARA, we replace the utility $c_{n*s^k} - \mu_{n*s^*}$ in Algorithm 1 by $r_n*s^k$ and meanwhile replace $c_{n_3k}$ by $r_{n_3k}$, then perform the steps 6-23 of Algorithm 1.

Maximal SINR Association (MSA) In the scheme MSA, we replace the utility $c_{n*s^k} - \mu_{n*s^*}$ in Algorithm 1 by SINR$_{n*s^k}$ and meanwhile replace $c_{n_3k}$ by SINR$_{n_3k}$, then perform the steps 6-23 of Algorithm 1.

Biasing Rate Association (BRA) In the scheme BRA, we replace the utility $c_{n*s^k} - \mu_{n*s^*}$ in Algorithm 1 by $r_n*s^k e^{-\mu_{n}s^*}$ and meanwhile replace $c_{n_3k}$ by $r_{n_3k}$, then perform the steps 6-23 of Algorithm 1. Note that $\mu_{n}s^*$ is the optimal solution obtained by the proposed algorithm. Evidently, we can easily find that this scheme should be equivalent to scheme ARP according to the association rule of Algorithm 1.

Biasing SINR Association (BSA) In the scheme BSA, we replace the utility $c_{n*s^k} - \mu_{n*s^*}$ in Algorithm 1 by SINR$_{n*s^k}/p_{n*s^*}$ and meanwhile replace $c_{n_3k}$ by SINR$_{n_3k}/p_{n_3k}$, then perform the steps 6-23 of Algorithm 1. Note that the subband $s$ will be not considered in the association when $p_{ns} = 0$ mW.

As for association performance, we mainly focus on the load balancing level, the cumulative distribution function (CDF) of effective rates, the coverage probability of effective rates and convergence of proposed algorithm.

Figure 3 shows the load distributions for different association schemes. The schemes MSA and MARA result in very unbalanced load distributions: most users are associated with the macrotier consisting of MBSs, and very few users can be served by the picotier consisting of PBSs. That’s because the MBS has higher transmit power than the PBS, and thus users associated with MBSs often have higher SINRs/rates than them associated with PBSs. In Fig. 3, the scheme BSA achieves a relatively high balancing level: more users favour picotier since they may achieve higher SINRs on subband 2 of PBSs than the ones on subband 1 of MBSs. As illustrated in Fig. 3, the schemes ARP and BRA can balance the loads among different network tiers and achieve almost the same effect. These schemes can offload the low-rate users associated with overloaded MBSs in the schemes MSA and MARA to the adjacent underloaded BSs.

To measure the status of the system load balancing level in a more refined metric, we introduce the Jain’s fairness index [33], which is given by

$$J = \left( \frac{\sum_{n \in N_{mp}} y_n}{N_{mp} \sum_{n \in N_{mp}} y_n^2} \right)^{2},$$

where $\sum_{k \in K} \sum_{s \in S} x_{nsk} = y_n$ represents the load of BS $n$. The larger $J$ that belongs to the interval $[1/N_{mp}, 1]$ means more balanced load distribution among the given cells. Thus, the Jain’s fairness index in this paper can be also named as the load balancing index (LBI). Significantly, we just consider the load balancing level of all BSs, which doesn’t refer to D2D TXs. Since any D2D TX just has one
or zero served user (its corresponding D2D RX), the load balancing level of D2D TXs doesn’t need to be considered.

Figure 4 shows the load balancing indices for different association schemes. Since the scheme BSA doesn’t depend on transmit power of BSs and performs user association in a very random manner, it achieves the highest LBI among all schemes. Unlike the scheme BSA, the schemes MSA and MARA should achieve the lowest LBI among all schemes since they attract most users for MBSs and let very few users select PBSs. Compared with the schemes MSA and MARA, the schemes ARP and BRA achieve a higher LBI because of their offloading capabilities.

The main goal of introducing D2D communications into HCNs is to offload traffic. As for this, we will show the offloading capabilities (numbers of supporting D2D pairs) of different association schemes.

Figure 5 shows the numbers of D2D RXs served by BSs or D2D TXs for different association schemes. The scheme MSA supports the most D2D pairs among all schemes, and the schemes ARP and BRA support more D2D pairs than schemes MSA and MARA. When the number of D2D pairs is relatively small, the D2D RXs can receive the weaker interference from D2D TXs than from BSs. Thus, more D2D RXs can be served by the corresponding D2D TXs. However, due to the limited bandwidth, the achievable rates of D2D TXs may have not enough superiority, which leads to more fewer D2D pairs are supported. It is noteworthy that the numbers of D2D pairs supported by schemes MSA and MARA will decrease with the number of D2D pairs due to the stronger and stronger interference. In the scheme BSA, more D2D RXs select the corresponding D2D TXs because of large shadowing fading. Unlike other
schemes, the schemes ARP and BRA trade off load and achievable rate, i.e., reduce the achievable rate by offloading. Thus, this operation will be beneficial to the utilization of D2D pairs. Significantly, the capabilities of supporting D2D pairs represents their offloading capabilities. As we know, more balanced load distribution is often beneficial to the full utilization of network resources and improving user experience.

The main goal of balancing network loads among different BSs is to achieve the load balancing gain that the association scheme improves (edge) user experience by balancing network loads. To show the load balancing gain, we investigate the cumulative distribution function (CDF) of effective rates of associated users.

Figure 6 plots the CDFs of effective rates of associated users for different association schemes. In the scheme MARA, most users select MBSs, which results in very few resources that can be utilized by users associated with MBSs. Thus, compared with other schemes, the scheme MARA has more low-rate users. Unlike scheme MARA, the scheme MSA doesn’t consider the available bandwidth of BSs, which makes many D2D RXs select their corresponding D2D TXs. According to the system assumption, we know that these D2D TXs often have high SINRs. Thus, the scheme MSA may have fewer low-rate users than scheme MARA. Although the scheme BSA has an offloading capacity, it cannot guarantee that the offloading operation can improve user experience because of highly
random association. As illustrated in Fig. 6, the scheme BSA hardly has any superiority over the scheme MARA. Unlike the scheme BSA, the schemes ARP and BRA can offload the low-rate users associated with overloaded MBSs in the schemes MSA and MARA to the adjacent underloaded BSs, which can improve user experience. Therefore, the schemes ARP and BRA own the fewest low-rate users among all schemes.

Figure 7 plots the CDFs of effective rates of users associated with macrotier for different association schemes. As mentioned in previous section, the schemes ARP and BRA can offload the low-rate users associated with overloaded MBSs in the schemes MSA and MARA to the adjacent underloaded BS. Evidently, this operation can improve user experience, and thus the schemes ARP and BRA have fewer low-rate users than schemes MSA and MARA. Similar to Fig. 6, the scheme MSA has fewer low-rate users than scheme MARA.

Evidently, the schemes ARP and BRA achieve more higher load balancing gains than other schemes.

By partitioning resource, we mitigate the interference received by offloaded users from MBSs and the one received by D2D RXs from all BSs. Thus, the resource partitioning gain represents that the association scheme improves user experience by partitioning resource. To reveal the resource partitioning gain, we investigate the impact of resource partitioning factor ($\eta$) on the rate coverage (coverage probability of effective rates) of associated users. Note that the rate coverage represents the proportion of the users whose effective rates are greater than target rate $\rho$ in all users.

Fig. 7 The CDFs of effective rates of users associated with macrotier for different association schemes

Fig. 8 The coverage probabilities of effective rates for four different target rates
Figure 8 shows the coverage probabilities of effective rates of associated users for different target rates. Seen from Fig. 8, we can find that the coverage probability initially increases with \( \eta \) and then decreases with it. That’s because the offloaded users may receive weaker and weaker interference from MBSs as this fraction increases, but then their effective rates should decrease with this fraction due to fewer available bandwidth. Through a direct observation, we can easily find the resource partitioning gain, i.e., the association scheme with resource partitioning provides a higher coverage probability than the one without resource partitioning. Moreover, the coverage probability should be higher and higher as the target rate becomes lower and lower.

Figure 9 illustrates the convergence of proposed algorithm, where parameter \( t \) is \( t \)-th iteration. To achieve optimal solutions and implementation simplicity, we only consider a constant stepsize for updating multiplier \( \mu \) in the proposed algorithm. As illustrated in Fig. 9, the proposed algorithm can converge in just a few iterations, which means it can be well applied in reality, especially in large-scale case.

6 Conclusion

For the D2D-enabled HCNs, we propose an offloading scheme with maximizing network-wide utility, and then design a highly effective distributed algorithm by dual decomposition. Numerical results show that the proposed association scheme can provide a load balancing gain, and meanwhile reveals the offloading capacity of D2D pairs. Moreover, the proposed resource partitioning scheme can also provide its gain. Future work can include designing a dynamic association algorithm, introducing power control and considering uplink scenario.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grants 61372101, 61422105 and 61271018, the National Science and Technology Major Project of China under Grants 2013ZX03003006-002 and 2012ZX03004-005-003, the Natural Science Foundation of Jiangsu Province under Grant BK20130019, the Research Project of Jiangsu Province under Grant BE2012167, and the Program for New Century Excellent Talents in University under Grant NCET-11-0088.

References

1. Wang B, Kong Q, Yang L (2015) Context-aware user association for energy cost saving in a green heterogeneous network with hybrid energy supplies. Mobile Network Appl, pp 1–15
2. Ghosh A, Mangalvedhe N et al. (2012) Heterogeneous cellular networks: From theory to practice. IEEE Commun Mag 50(6):54–64
3. Andrews JG (2013) Seven ways that HetNets are a cellular paradigm shift. IEEE Commun Mag 51(3):136–144
4. Damnjanovic A, Montojo J et al. (2011) A survey on 3GPP heterogeneous networks. IEEE Wirel Commun 18(3):10–21
5. Liu J, Kawamoto Y et al. (2014) Device-to-device communications achieve efficient load balancing in LTE-advanced networks. IEEE Wirel Commun 21(2):57–65
6. Cheng Y, Gu Y, Lin X (2013) Combined power control and link selection in deviceto-device enabled cellular systems. IET Commun 7(12):1221–1230
7. Doppler K, Rinne M et al. (2009) Device-to-device communication as an underlay to LTE-advanced networks. IEEE Commun Mag 47(12):42–49
8. Andreev S, Pyattaev A et al. (2014) Cellular traffic offloading onto network-assisted device-to-device connections. IEEE Commun Mag 52(4):20–31
9. Phunchongharn P, Hossain E, Kim D (2013) Resource allocation for device-to-device communications underlaying LTE-advanced networks. IEEE Wirel Commun 20(4):91–100
10. Alam M, Yang D et al. (2014) Secure device-to-device communication in LTE-a. IEEE Commun Mag 52(4):66–73
11. Zhu D, Wang J et al. (2014) Downlink resource reuse for device-to-device communications underlaying cellular networks. IEEE Signal Process Lett 21(5):531–534
12. Andrews JG, Singh S et al. (2014) An overview of load balancing in HetNets: Old myths and open problems. IEEE Wirel Commun 21(2):18–25
13. Fooladivanda D, Rosenberg C (2013) Joint resource allocation and user association for heterogeneous wireless cellular networks. IEEE Trans Wirel Commun 12(1):248–257
14. Singh S, Dhillon H, Andrews J (2013) Offloading in heterogeneous networks: modeling, analysis, and design insights. IEEE Trans Wirel Commun 12(5):2484–2497
15. Singh S, Andrews J (2014) Joint resource partitioning and offloading in heterogeneous cellular networks. IEEE Trans Wirel Commun 13(2):888–901
16. Jo HS, Sang YI, Xia P, Andrews J (2012) Heterogeneous cellular networks with flexible cell association: a comprehensive downlink SINR analysis. IEEE Trans Wirel Commun 11(10):3484–3495
17. Ye Q, Rong B et al. (2013) User association for load balancing in heterogeneous cellular networks. IEEE Trans Wirel Commun 12(6):2706–2716
18. Shen K, Yu W (2014) Distributed pricing-based user association for downlink heterogeneous cellular networks. IEEE J Sel Areas Commun 32(6):1100–1113
19. Cho S, Choi W (2013) Energy-efficient repulsive cell activation for heterogeneous cellular networks. IEEE J Sel Areas Commun 31(5):870–882
20. Viswanath P, Tse D, Anantharam V (2001) Asymptotically optimal water-filling in vector multiple-access channels. IEEE Trans Inf Theory 47(1):241–267
21. Jang J, Lee KB (2003) Transmit power adaptation for multiuser OFDM systems. IEEE J Sel Areas Commun 21(2):171–178
22. Son K, Chong S, Veciana G (2009) Dynamic association for load balancing and interference avoidance in multi-cell networks. IEEE Trans Wirel Commun 8(7):3566–3576
23. 3GPP (2010) TS 36.201: LTE physical layer: general description (release 9), v9.1.0. Online: http://www.3gpp.org/specifications/releases, Accessed March 2010
24. Singh S (2012) Belief propagation for optimal user association in wireless networks. Online: http://sarabjotsingh.com/downloads/SIN_BP12.pdf, Accessed May 2012
25. Xiao L, Johansson M, Boyd S (2004) Simultaneous routing and resource allocation via dual decomposition. IEEE Trans Wirel Commun 52(7):1136–1144
26. Codreanu M, Juntti M, Latva-aho M (2007) On the dual-decomposition-based sum capacity maximization for vector broadcast channels. IEEE Trans Veh Technol 56(6):3577–3581
27. Rantzer A (2009) Dynamic dual decomposition for distributed control. American Control Conference, MO USA, pp 884–888
28. Boyd SP, Vandenberghe L (2004) Convex optimization. Cambridge University Press
29. Bertsekas D (2009) Convex optimization theory. Athena Scientific
30. Boyd S, Xiao L, Mutapcic A (2003) Subgradient methods. Online: https://web.stanford.edu/class/ee392o/subgrad_method.pdf, Accessed October 2003
31. Hakola S, Chen T et al (2010) Device-to-device (D2D) communication in cellular network - performance analysis of optimum and practical communication mode selection. In: Proceedings of the IEEE WCNC, NSW, Australia, pp 1–6
32. Sedq A, Gohary R et al (2013) Optimal tradeoff between sum-rate efficiency and Jain’s fairness index in resource allocation. IEEE Trans Wirel Commun 12(7):3496–3509