Load Balancing for 5G Ultra-Dense Networks using Device-to-Device Communications

Hongliang Zhang, *Student Member, IEEE*, Lingyang Song, *Senior Member, IEEE*,
and Ying Jun (Angela) Zhang, *Senior Member, IEEE*

***Abstract*—Load balancing is an effective approach to address the spatial-temporal fluctuation problem of mobile data traffic for cellular networks. The existing schemes that focus on channel borrowing from neighboring cells cannot be directly applied to future 5G wireless networks, because the neighboring cells will reuse the same spectrum band in 5G systems. In this paper, we consider an orthogonal frequency division multiple access (OFDMA) ultra-dense small cell network, where Device-to-Device (D2D) communication is advocated to facilitate load balancing without extra spectrum. Specifically, the data traffic can be effectively offloaded from a congested small cell to other underutilized small cells by D2D communications. The problem is naturally formulated as a joint resource allocation and D2D routing problem that maximizes the system sum-rate. To efficiently solve the problem, we decouple the problem into a resource allocation subproblem and a D2D routing subproblem. The two subproblems are solved iteratively as a monotonic optimization problem and a complementary geometric programming problem, respectively. Simulation results show that the data sum-rate in the neighboring small cells increases 20% on average by offloading the data traffic in the congested small cell to the neighboring small cell base stations (SBSs).**

***Index Terms*—Device-to-device communications, ultra-dense small cell network, load balancing, sum-rate maximization, non-convex optimization.**

I. Introduction

The unprecedented growth in mobile devices and applications has triggered an explosion in the data traffic. According to Cisco's report, global mobile data traffic has grown 4000-fold over the last decade, and is expected to reach 30.6 exabytes per month by 2020 [2]. To meet the surge in the traffic volume, vendors and operators are looking into every tool at hand to enhance the spectrum efficiency and network capacity. Ultra-dense small cell deployment is one of the most promising solutions in this regard [3]–[5]. Noticeably, traffic fluctuates far more significantly in small cells than in large cells, since small cells do not benefit from the "law of large numbers". To address this issue, load balancing is recognized as an effective approach to alleviate traffic fluctuation in small cells [6], [7].

There has been a lot of research on load balancing in cellular networks. Most prior work focuses on borrowing channels from adjacent lightly loaded cells, such as selective borrowing [8], subcarrier borrowing without locking [9], and etc. Though effective, theses schemes do not work in the 5G ultra-dense small cell networks, where the same spectrum band is reused in neighboring small cells. Other techniques, including overlaying ad hoc relays on top of cellular networks [10], [11] and adapting cellular networks to the dedicated industrial scientific medical (ISM) band [12] or whitespaces spectrum [13], require the use of unlicensed spectrum to achieve load balancing. Moreover, due to uncoordinated channel access at the MAC layer, the interference from surrounding terminals is uncontrollable. Thus, operators cannot guarantee satisfactory Quality-of-Service (QoS) for mobile users.

In this paper, we advocate device-to-device (D2D) load balancing as a useful mechanism to alleviate traffic fluctuation among small cells. Specifically, D2D communication enables two devices in proximity to communicate directly with each other by reusing cellular spectrum under the control of the small cell base station (SBS) [14]. Due to its underlay property [15]–[20] and proximity gain [21]–[23], D2D communication can relay traffic from congested small cells to adjacent under-loaded small cells without the need of extra spectrum. D2D load-balancing for cellular networks has been considered in the literature, e.g. [24]–[27]. In particular, [24] discussed the technical feasibility of D2D load balancing in cellular networks, and designed a base station (BS) selection algorithm to determine to which BS a user should route its traffic. The work in [25] provided theoretical modeling and analysis to characterize the benefit of D2D load balancing, and derived a solution for allocating radio resources, including transmission power and time slots when the D2D route was fixed. D2D routing was considered in [26], where the destination BS was pre-determined, and a D2D link was only allowed to use one subchannel. The authors in [27] designed an online auction framework for the D2D routing and time slots allocation, when the subchannel allocation and the destination BS were given.

In this paper, we consider OFDMA ultra-dense small cell networks, where a user can either send its data through the SBS in its own small cell, or route its data to a neighboring SBS with the help of multi-hop D2D communications. In the network, D2D links can share the spectrum with cellular links. Due to spectrum sharing, radio resource allocation is particularly important for interference management. Besides, the optimal SBS (destination) selection and D2D routing are indispensables, if the overall system performance is to be optimized. Unlike the existing work in [24]–[27] that only considers part of these issues, our work jointly optimizes

Manuscript received November 22, 2017; revised March 16, 2018; accepted March 21, 2018. Part of this work has been presented in Proc. IEEE VTC-Spring 2017 [1]. This work was supported by the National Nature Science Foundation of China under grant number 61625101. The associate editor coordinating the review of this paper and approving it for publication was Prof. J. Lee.

H. Zhang and L. Song are with School of Electronics Engineering and Computer Science, Peking University, Beijing, China (email: {hongliang.zhang,lingyang.song}@pku.edu.cn).

Y. J. Zhang is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (email:yjzhang@ie.cuhk.edu.hk).

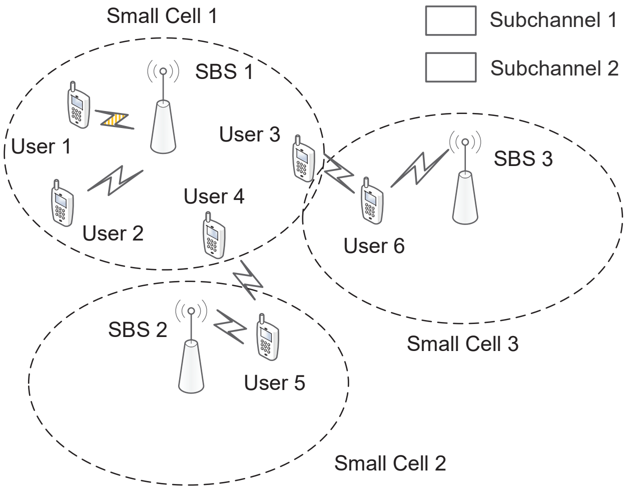


Fig. 1. System model for D2D load balancing in an ultra-dense small cell network sharing uplink resource.

resource allocation, D2D routing, and SBS selection.

Therefore, the main contributions of this paper are summarized as follows:

1) To solve this optimization problem efficiently, we decouple the D2D load balancing problem into resource allocation and D2D routing subproblems, and propose an I-RA-DR algorithm to solve these subproblems iteratively.
2) We transform the resource allocation subproblem into a Monotonic Optimization (MO) problem [28]. By exploiting monotonicity, the MO problem can be solved by an accelerated polyblock outer approximation efficiently.
3) The D2D routing subproblem consists of two steps: SBS selection and routing. For SBS selection, we search all the potential SBSs of each flow and select the SBS with the maximum total sum-rate obtained by the routing decision. Meanwhile, the routing problem is transformed into a Complementary Geometric Programming (CGP) problem [29], which can be efficiently solved via an iterative convex relaxation approach.
4) Simulation results show that our proposed I-RA-DR algorithm can achieve a better performance than a greedy routing scheme and the scheme without load balancing.

The rest of this paper is organized as follows. In Section II, we describe the system model. In Section III, problem formulation is introduced. Then we decouple the problem into resource allocation and D2D routing subproblems. In Section IV, the resource allocation subproblem is transformed into a MO problem. Section V presents SBS searching and routing decision which is transformed into a CGP problem for the D2D routing subproblem. In Section VI, we analyze the overall convergence and computational complexity of the proposed I-RA-DR algorithm. Section VII evaluates the proposed I-RA-DR algorithm through numerical simulations. Finally, Section VIII concludes this paper.

II. System Model

As illustrated in Fig. 1, we consider the uplink of an orthogonal frequency division multiple access (OFDMA) ultra-dense small cell network with B SBSs, denoted by $b \in \mathcal{B} = \{1, 2, \dots, B\}$. There exist M small cell users in the system, denoted by $i \in \mathcal{M} = \{1, 2, \dots, M\}$. We also assume that only F users, denoted by $\mathcal{F} = \{1, 2, \dots, F\}$, have their own data to transmit, and each user corresponds to a data flow. For simplicity, data flow generated by user f is called data flow f. To balance the load in each small cell, a user can either send its data through the SBS in its own small cell, or route its data to a neighboring SBS via multi-hop D2D communications[1]. In Fig. 1, for example, user 2's data is sent through SBS 1 by cellular communication, while the data generated by user 4 is sent to SBS 2 by a D2D route through user 5.

Let $\mathcal{K} = \{1, 2, \dots, K\}$ be the set of K subchannels to be allocated. In this system, we assume that the spectrum reuse factor is 1, which means that the same frequency band is reused in all small cells. Besides, the power allocated to user $i \in \mathcal{M}$ over subchannel k is denoted by $p_{k,i}$, satisfying $\sum\limits_{k \in \mathcal{K}} p_{k,i} \leq P$, where P is the total transmitted power of user i. In particular, $p_{k,i} > 0$ if subchannel k is allocated to user i, and $p_{k,i} = 0$ otherwise.

Suppose that the wireless channel is Rayleigh faded, and that the channel gain of link from node i to node j over subchannel k is expressed as

$$g_{i,j}^{k} = L_{i,j}^{k} |h_{i,j}^{k}|^2, \tag{1}$$

where $L_{i,j}^{k}$ denotes the corresponding distance-dependent path loss and $h_{i,j}^{k}$ denotes the corresponding small-scale fading. The nodes can be users or SBSs, i.e., $i, j \in \mathcal{M} \cup \mathcal{B}$. The path loss L is inversely proportional to the propagational distance d, i.e., $L = \kappa d^{-\alpha}$, where α is the path loss exponent, and κ is the constant power gains factor introduced by amplifier and antenna. The small-scale fading $h_{i,j}^{k} \sim \mathcal{CN}(0, 1)$ follows a complex Gaussian distribution. In addition, the thermal noise satisfies independent Gaussian distribution with zero mean and the variance σ^2.

Define a flow matrix $\boldsymbol{X}_{(M+B) \times (M+B) \times F} = [x_{i,j}^{f}]$, where

$$x_{i,j}^{f} = \begin{cases} 1, & \text{when flow } f \text{ transmits on link } i - j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Thus, we can derive the link matrix $x_{i,j}$ by

$$x_{i,j} = 1 - \prod_{f \in \mathcal{F}} (1 - x_{i,j}^{f}) \tag{3}$$

to indicate whether link $i - j$ is active.

Besides, define a destination matrix $\boldsymbol{D}_{(M+B) \times F} = [d_{i,f}]$ to indicate whether a node is the destination of a flow. Specifically,

$$d_{i,f} = \begin{cases} 1, & \text{when node } i \text{ is the destination of flow } f, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

[1] The number of D2D links for one flow is not predefined, instead, it is determined by the solution of the formulated problem. If the solution suggests that multi-hop D2D communications can achieve a better performance for the flow, then it will set up the multi-hop D2D communications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2819648, IEEE Transactions on Wireless Communications

Likewise, define a source matrix $\boldsymbol{S}_{(M+B)\times F} = [s_{i,f}]$ to indicate whether a node is the source of a flow, where

$$s_{i,f} = \begin{cases} 1, & \text{when node } i \text{ is the source of flow } f, \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

respectively. It is worth mentioning that the entries in the destination matrix are decision variables, and those in the source matrix are constants. Besides, we define N_i as the neighbor set of node i, where

$$j \begin{cases} \in N_i, & \text{node } j \text{ is in the communication range of node } i, \\ \notin N_i, & \text{otherwise.} \end{cases} \tag{6}$$

In this paper, a link is allowed to transmit on multiple subchannels, which can be realized by carrier aggregation technique [30], [31]. The signal-to-interference-plus-noise ratio (SINR) at the link $i-j$ over subchannel k, denoted by $\gamma_{i,j}^k$, is calculated as

$$\gamma_{i,j}^k = \frac{x_{i,j}p_{i,j}^k g_{i,j}^k}{\sigma^2 + \sum\limits_{i'\neq i, j'\neq j} x_{i',j'}p_{i',j'}^k g_{i',j}^k + \sum\limits_{i,j'\neq j} x_{i,j'}p_{i,j'}^k g_{i,j'}^k + \sum\limits_{i'\neq i,j} x_{i',j}p_{i',j}^k g_{i',j}^k}. \tag{7}$$

According to the Shannon capacity formula, the achievable data rate $R_{i,j}$ is given by

$$R_{i,j} = \sum_{k\in\mathcal{K}} \log_2(1+\gamma_{i,j}^k). \tag{8}$$

We denote the rate of flow f by r_f, and the rate of flow f on link $i-j$ by $r_{i,j}^f$. As the total transmission rate of all the flows in link $i-j$ cannot exceed the capacity of this link, the rate $r_{i,j}^f$ needs to satisfy

$$\sum_{f\in\mathcal{F}} r_{i,j}^f \leq R_{i,j}, \forall i \in \mathcal{M}, \forall j \in \mathcal{M}\cup\mathcal{B}. \tag{9}$$

Moreover, since the rate of a flow on each link is equal to the input rate from the source, the rate of a flow needs to be the same on all links in the path. Therefore, the data rate of a flow can be defined as the input rate from the source, i.e.,

$$r_f = \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r_{i,j}^f s_{i,f}, \forall f \in \mathcal{F}. \tag{10}$$

In addition, the incoming and outgoing rates for a flow must satisfy flow conservation constraints:

$$\sum_{j\in N_i} r_{j,i}^f - \sum_{j\in N_i} r_{i,j}^f = r_f(d_{i,f}-s_{i,f}), \forall i\in\mathcal{M}, \forall f\in\mathcal{F}. \tag{11}$$

Note that the data rate for a flow is positive on a link only if this flow goes through this link. Thus, we also have the following constraint:

$$r_{i,j}^f = 0 \Leftrightarrow x_{i,j}^f = 0, r_{i,j}^f > 0 \Leftrightarrow x_{i,j}^f = 1. \tag{12}$$

Finally, the destination of a flow path must be a SBS. That is to say,

$$\sum_{i\in\mathcal{B}} d_{i,f} = 1, \forall f\in\mathcal{F}, \quad \sum_{i\notin B} d_{i,f} = 0, \forall f \in \mathcal{F}. \tag{13}$$

III. Problem Formulation

The proposed D2D load balancing scheme aims to maximize the total sum-rate by optimizing $x_{i,j}^f, d_{i,f}$, and $p_{i,j}^k$ for all flows and nodes. The optimization problem is formulated as:

$$\text{P1} : \underset{\boldsymbol{X};\boldsymbol{D};\boldsymbol{P}}{\text{maximize}} \sum_{f\in\mathcal{F}} r_f, \tag{14a}$$

$$s.t. \sum_{j\in N_i} r_{j,i}^f - \sum_{j\in N_i} r_{i,j}^f = r_f(d_{i,f}-s_{i,f}), \forall i\in\mathcal{M}, f\in\mathcal{F}, \tag{14b}$$

$$\sum_{i\in\mathcal{B}} d_{i,f} = 1, \sum_{i\notin B} d_{i,f} = 0, \forall f\in\mathcal{F}, \tag{14c}$$

$$\sum_{f\in\mathcal{F}} r_{i,j}^f \leq R_{i,j}, \forall i\in\mathcal{M}, j\in\mathcal{M}\cup\mathcal{B}, \tag{14d}$$

$$x_{i,j} = 1 - \prod_{f\in\mathcal{F}} (1-x_{i,j}^f), \tag{14e}$$

$$\sum_{k\in\mathcal{K}} p_{i,j}^k \leq P, \forall i\in\mathcal{M}, j\in\mathcal{M}\cup\mathcal{B}, \tag{14f}$$

$$r_f = \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r_{i,j}^f s_{i,f}, \tag{14g}$$

$$d_{i,f}, x_{i,j}^f \in \{0,1\}, \tag{14h}$$

$$0 \leq p_{i,j}^k \leq P, \tag{14i}$$

$$r_{i,j}^f = 0 \Leftrightarrow x_{i,j}^f = 0, r_{i,j}^f > 0 \Leftrightarrow x_{i,j}^f = 1, \tag{14j}$$

where $R_{i,j}$ is given in (8). Constraints (14b) correspond to the flow conservation of each flow. Constraints (14c) imply that each flow path only ends at a SBS. Constraints (14d) are the rate constraints on each link. The constraints in (14e), (14g), and (14j) correspond to equations (3), (10), and (12), respectively. Constraints (14f) are the transmission power constraints for each link.

Note that this is an NP-hard problem due to the binary variables in constraints (14b)-(14e) and the interference term in constraint (14d). In what follows, we tackle the problem through alternating maximization, where resource allocation (RA) and D2D routing (DR) are optimized iteratively. The proposed algorithm is referred to as I-RA-DR algorithm, where "I" stands for "iterative".

The flowchart of the I-RA-DR algorithm is shown in Fig. 2. Given the D2D routing result $(\boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)})$, the resource allocation subproblem can be transformed to an MO problem, where a sub-optimal solution can be obtained by the accelerated polyblock outer approximation method. Likewise, given the resource allocation decision and SBS selection $(\boldsymbol{P}^{(t)}, \boldsymbol{D}^{(t)})$, the D2D routing subproblem can be formulated as a CGP problem, which can be turn into a Geometric Programming (GP) problem [33], [34] by approximating its non-convex constraints. The details of the solution algorithms of the resource allocation and D2D routing subproblems will be discussed in Section IV and V, respectively.

IV. Monotonic Optimization for Resource Allocation

In this section, we first reformulate the resource allocation subproblem as a MO problem, and then propose a RA algo-

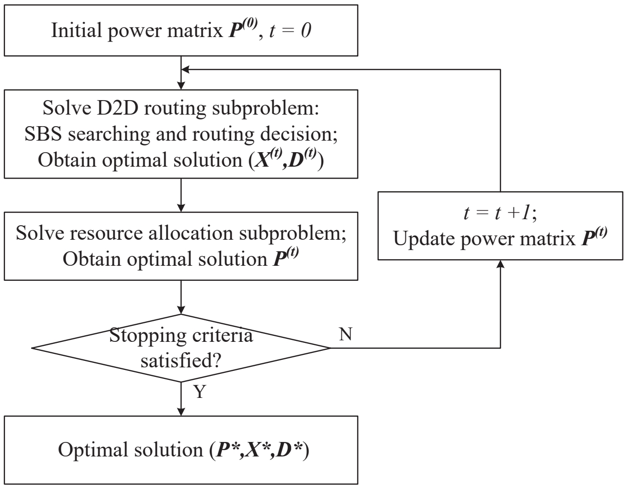

Initial power matrix $P^{(0)}$, $t = 0$
Solve D2D routing subproblem: SBS searching and routing decision; Obtain optimal solution ($X^{(t)}$,$D^{(t)}$)
Solve resource allocation subproblem; Obtain optimal solution $P^{(t)}$
$t = t + 1$; Update power matrix $P^{(t)}$
Stopping criteria satisfied?
N
Y
Optimal solution (P^*,X^*,D^*)

Fig. 2. I-RA-DR algorithm for D2D load balancing.

rithm to solve this problem.

A. Monotonic Optimization Formulation

Suppose that the D2D routing decision, i.e., flow matrix $\boldsymbol{X}$ and destination matrix $\boldsymbol{D}$ is given. Let $\tilde{\gamma}^k_{i,j} = \gamma^k_{i,j} + 1$ and $\tilde{\boldsymbol{\gamma}} = [\tilde{\gamma}^1_{1,1}, \ldots, \tilde{\gamma}^k_{i,j}, \ldots, \tilde{\gamma}^K_{M+B,M+B}]$. Then, problem (P1) can then be rewritten as

$$\begin{aligned} \text{P2:} \quad & \underset{\boldsymbol{P}}{\text{maximize}} \sum_{f\in\mathcal{F}} \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r^f_{i,j} s_{i,f}, \\ s.t. \quad & \text{(14b), (14f), (14g), (14i), and (14j),} \\ & \sum_{f\in\mathcal{F}} r^f_{i,j} \le \sum_{k\in\mathcal{K}} \log_2 \tilde{\gamma}^k_{i,j}. \end{aligned} \tag{15}$$

Due to the coupling interference constraints among all concurrent links, the resource allocation subproblem (P2) is nonconvex. However, the problem exhibits hidden monotonic structures, which can be exploited to obtain global optimal solutions efficiently. In the following, we first introduce the preliminaries of MO [28], [32]. Then, we will show how the resource allocation subproblem can be reformulated as a MO problem.

1) Preliminaries of MO:

Definition 1. *A set $\mathcal{Q} \subset \mathbb{R}^N_+$ is called* normal *if $\forall \boldsymbol{q}_0 \in \mathcal{Q}$, all the points $\boldsymbol{q}$ with $\boldsymbol{0} \preceq \boldsymbol{q} \preceq \boldsymbol{q}_0$ satisfy $\boldsymbol{q} \in \mathcal{Q}$. Here, $\boldsymbol{q} \preceq \boldsymbol{q}_0$ means that $\boldsymbol{q}$ is component-wise less than or equal to $\boldsymbol{q}_0$.*

Definition 2. *An optimization problem can be transformed to a MO if it can be represented by*

$$\begin{aligned} & \underset{\boldsymbol{q}}{\text{maximize}}\, g(\boldsymbol{q}), \\ s.t. \quad & \boldsymbol{q} \in \mathcal{Q}, \end{aligned} \tag{16}$$

where $\mathcal{Q} \subset \mathbb{R}^N_+$ is a non-empty normal set and the function $g(\boldsymbol{q})$ is increasing over $\mathbb{R}^N_+$.

Definition 3. *Let $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^N_+$ and $\boldsymbol{y} \preceq \boldsymbol{z}$. The hyper-rectangle determined by $\boldsymbol{y}$ and $\boldsymbol{z}$ is called a* box *in $\mathbb{R}^N_+$, i.e., $[\boldsymbol{y}, \boldsymbol{z}] = \prod_{n=1}^{N} [y_n, z_n]$, where $[y_n, z_n]$ is the n-th dimension of the box. For example, when $n = 2$, the box is reduced to a rectangle $[y_1, z_1] \times [y_2, z_2]$. The points $\boldsymbol{y}$ and $\boldsymbol{z}$ are called the lower bound vertex and upper bound vertex, respectively. A set is called* polyblock *if it is the union of a finite number of boxes.*

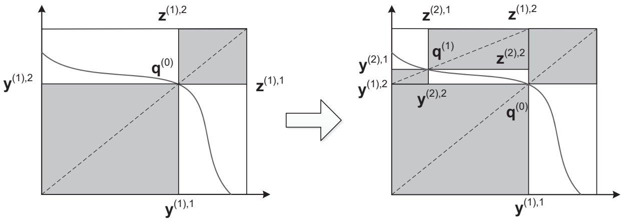


Fig. 3. Illustration of procedure for constructing a new polyblock.

Remark 1. *A polyblock is normal.*

2) Transformation to MO: Given any feasible resource allocation result $\boldsymbol{P}$ that satisfies constraints (14f) and (14i), the value of $\tilde{\boldsymbol{\gamma}}$ is also determined. We define $\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}(\boldsymbol{P}))$ to denote the global optimal solution of $\boldsymbol{r} = [r^1_{1,1}, \ldots, r^f_{i,j}, \ldots, r^F_{M+B,M+B}]$ by solving problem (P2). In particular, $\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}(\boldsymbol{P}))$ can be obtained by solving the following problem

$$\begin{aligned} \text{P3:} \quad & \underset{\boldsymbol{r}}{\text{maximize}} \sum_{f\in\mathcal{F}} \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r^f_{i,j} s_{i,f}, \\ s.t. \quad & \text{(14b), (14g), and (14j),} \\ & \sum_{f\in\mathcal{F}} r^f_{i,j} \le \sum_{k\in\mathcal{K}} \log_2 \tilde{\gamma}^k_{i,j}. \end{aligned} \tag{17}$$

Note that problem (P3) is a linear programming problem, we can obtain the global optimal solution $\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}(\boldsymbol{P}))$ using a linear programming solver. Correspondingly, we define $g(\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}))$ as the global optimal sum-rate of problem (P2) for given $\tilde{\boldsymbol{\gamma}}$. In this way, the search for resource allocation variable $\boldsymbol{P}$ can be transformed into the search for SINR $\tilde{\boldsymbol{\gamma}}$, and thus, problem (P2) can be rewritten as

$$\begin{aligned} \text{P4:} \quad & \underset{\boldsymbol{\gamma}'}{\text{maximize}}\, g(\boldsymbol{r}^*(\boldsymbol{\gamma}')) = \sum_{f\in\mathcal{F}} \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r^*(\boldsymbol{\gamma}'), \\ s.t. \quad & \boldsymbol{r}^*(\boldsymbol{\gamma}') \text{ is the optimum to problem (P3),} \\ & \boldsymbol{\gamma}' \in \mathcal{Q}, \end{aligned} \tag{18}$$

where the feasible set is defined as

$$\mathcal{Q} = \{\boldsymbol{\gamma}' | \boldsymbol{0} \preceq \ \boldsymbol{\gamma}' \preceq \tilde{\boldsymbol{\gamma}}(\boldsymbol{P}), \text{(14f) and (14i)}\}. \tag{19}$$

The constraints in the problem (P3) indicate that the feasible set of $\boldsymbol{r}$ is enlarged with the increase of $\tilde{\boldsymbol{\gamma}}$. This implies that the function $g(\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}))$ is increasing with $\tilde{\boldsymbol{\gamma}}$. In addition, the feasible set $\mathcal{Q}$ is normal. Thus, problem (P4) is a MO problem as defined in Definition 2.

B. Resource Allocation Algorithm

Now, we describe the solution algorithm for problem (P4). Define $\partial\mathcal{Q}$ as the upper boundary of set $\mathcal{Q}$. The key idea of MO is to successively maximize the increasing objective function $g(\boldsymbol{r}^*(\tilde{\boldsymbol{\gamma}}))$ on a sequence of polyblocks $\mathcal{P}^{(i)}$ enclosing $\partial\mathcal{Q}$ [28]. The most important operation is to calculate the boundary points $\boldsymbol{q}^{(i)}$ on $\partial\mathcal{Q}$ within $\mathcal{P}^{(i)}$.

In the following, we elaborate the detailed steps of the RA algorithm.

Step 1 Initialization: We give an initial polyblock $\mathcal{P}^{(0)}$ that encloses the boundary $\partial\mathcal{Q}$. Without the loss of generality, $\mathcal{P}^{(0)}$ can be simply constructed as $\mathcal{P}^{(0)} = [\boldsymbol{y}^{(0)}, \boldsymbol{z}^{(0)}]$, where $\boldsymbol{y}^{(0)} = \mathbf{1}_{(M+B)^2K\times 1}$ and $\boldsymbol{z} = \overline{\boldsymbol{\gamma}}$ with $\overline{\boldsymbol{\gamma}} = [\overline{\gamma}_{1,1}^{1}, \dots, \overline{\gamma}_{(M+B),(M+B)}^{K}]^{\mathrm{T}}$, and

$$\overline{\gamma}_{i,j}^{k} = 1 + \frac{x_{i,j} P g_{i,j}^{k}}{\sigma^2}, \forall i, j \in \mathcal{M} \cup \mathcal{B}, k \in \mathcal{K}. \qquad (20)$$

Step 2 Shrinking: If the upperbound vertex $\boldsymbol{z}^{(i)}$ over $\mathcal{P}^{(i)}$ are found in the boundary $\partial\mathcal{Q}$, then $\boldsymbol{z}^{(i)}$ is the global optimal solution of problem (P4). Otherwise, we need to construct a smaller polyblock $\mathcal{P}^{(i+1)}$ that still contains the boundary $\partial\mathcal{Q}$ to approximate the global optimal solution. Specifically, in a box $\boldsymbol{B}^{(i)} = [\boldsymbol{y}^{(i)}, \boldsymbol{z}^{(i)}] \in \mathcal{P}^{(i)}$, we need to find a boundary point $\boldsymbol{q}^{(i)}$, and obtain some smaller boxes enclosing the boundary $\partial\mathcal{Q}$, i.e., $\mathcal{B}^{(i)} \backslash ([\boldsymbol{y}^{(i)}, \boldsymbol{q}^{(i)}) \cup (\boldsymbol{q}^{(i)}, \boldsymbol{z}^{(i)}])$, to replace $\mathcal{B}^{(i)}$. Boundary point $\boldsymbol{q}^{(i)}$ is the intersection between the boundary $\partial\mathcal{Q}$ and the line segment from $\boldsymbol{y}^{(i)}$ to $\boldsymbol{z}^{(i)}$, which will be discussed in Subsection IV-C.

Next, we remove all the boxes whose upperbound vertex $\boldsymbol{z}$ satisfies

$$g(\boldsymbol{r}^*(\boldsymbol{z})) \leq \Phi(\mathcal{P}^{(i)}). \qquad (21)$$

Here, $\Phi(\mathcal{P}^{(i)})$ is the current best value on polyblock $\mathcal{P}^{(i)}$, which is equal to

$$\Phi(\mathcal{P}^{(i)}) = \max\{g(\boldsymbol{r}^*(\boldsymbol{q}^{(i)})), \Phi(\mathcal{P}^{(i-1)})\}. \qquad (22)$$

An illustration of this step in a two-dimensional feasible set is given in Fig. 3. The initial polyblock $\mathcal{P}^{(0)}$ is with lower bound vertex $\boldsymbol{y}^{(1)}$ and upper bound vertex $\boldsymbol{z}^{(1)}$. Here, $\boldsymbol{y}^{(i),1}$ and $\boldsymbol{y}^{(i),2}$ are the projections on the first and second dimensions of vertex $\boldsymbol{y}^{(i)}$, respectively. Likewise, $\boldsymbol{z}^{(i),1}$ and $\boldsymbol{z}^{(i),2}$ are the projections on the first and second dimensions of vertex $\boldsymbol{z}^{(i)}$. We obtain the boundary point $\boldsymbol{q}^{(0)}$ by intersecting the boundary and the line from $\boldsymbol{y}^{(1)}$ to $\boldsymbol{z}^{(1)}$, then delete the boxes $[\boldsymbol{y}^{(1)}, \boldsymbol{q}^{(0)}]$ and $[\boldsymbol{q}^{(0)}, \boldsymbol{z}^{(1)}]$ from the initial polyblock $\mathcal{P}^{(0)}$ and form a new polyblock $\mathcal{P}^{(1)}$.

Step 3 Termination: We repeat these steps until the global optimal solution is found. This will lead to a series of polyblocks containing $\partial\mathcal{Q}$: $\mathcal{P}^{(0)} \supset \mathcal{P}^{(1)} \supset \ldots \supset \partial\mathcal{Q}$. The convergence of the RA algorithm is guaranteed only with infinite iterations. For practical implementation, we can terminate this algorithm when either $g(\boldsymbol{r}^*(\boldsymbol{z})) - \Phi(\mathcal{P}^{(i)}) \leq \epsilon$ or $||\boldsymbol{z}^{(i)} - \boldsymbol{y}^{(i)}||_\infty \leq \eta$, where $\epsilon > 0$ and $\eta > 0$ are predefined error tolerance levels.

When the algorithm terminates, we obtain the boundary point in each box as the SINR solution. We define $\boldsymbol{v}^{(i)}$ as the best feasible solution for the SINR at iteration i, where

$$\boldsymbol{v}^{(i)} = \begin{cases} \boldsymbol{v}^{(i-1)}, & g(\boldsymbol{r}^*(\boldsymbol{q}^{(i)})) \leq \Phi(\mathcal{P}^{(i)}); \\ \boldsymbol{q}^{(i)}, & \text{otherwise.} \end{cases} \qquad (23)$$

The flowchart of RA algorithm is provided in Fig. 4.

C. Boundary Point $\boldsymbol{q}^{(i)}$ Calculation

According to the aforementioned algorithm, at iteration i, a critical step is to calculate the boundary point $\boldsymbol{q}^{(i)}$. The boundary point can be obtained by the intersection between $\partial\mathcal{Q}$ and the line segment from $\boldsymbol{y}^{(i)}$ to $\boldsymbol{z}^{(i)}$, which can be given by $\boldsymbol{q}^{(i)} = (1-\kappa^*)\boldsymbol{y}^{(i)} + \kappa^*\boldsymbol{z}^{(i)}$. Since $\boldsymbol{y}^{(i)}$ is an inner vertex of Q and $\boldsymbol{z}^{(i)}$ is an outer vertex of Q, the larger κ with $\boldsymbol{u}^{(i)}(\kappa) \in \mathcal{Q}$ implies that it gets closer to the boundary $\partial\mathcal{Q}$. Therefore, κ^* is the solution of the following problem.

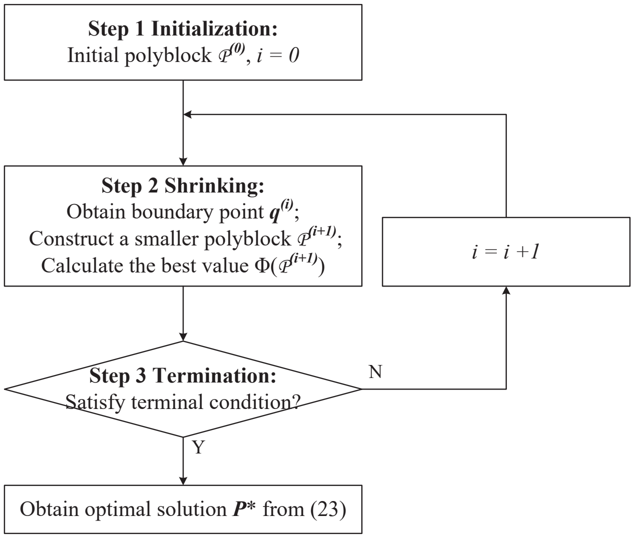

Step 1 Initialization:
Initial polyblock $\mathcal{P}^{(0)}$, $i = 0$
Step 2 Shrinking:
Obtain boundary point $q^{(i)}$;
Construct a smaller polyblock $\mathcal{P}^{(i+1)}$;
Calculate the best value $\Phi(\mathcal{P}^{(i+1)})$
$i = i + 1$
Step 3 Termination:
Satisfy terminal condition?
N
Y
Obtain optimal solution $\boldsymbol{P}$* from (23)

Fig. 4. RA algorithm for resource allocation.

$$\begin{aligned} \text{P5:} \quad & \text{maximize } \kappa, \\ s.t. \quad & \boldsymbol{u}^{(i)}(\kappa) = (1-\kappa)\boldsymbol{y}^{(i)} + \kappa\boldsymbol{z}^{(i)} \in \mathcal{Q}, \\ & 0 \leq \kappa \leq 1. \end{aligned} \qquad (24)$$

The bisection search starts with the initial interval $[\overline{\kappa}, \underline{\kappa}] = [0, 1]$. The algorithm first calculates $\kappa = (\overline{\kappa} + \underline{\kappa})/2$, and then checks the feasibility of $\boldsymbol{u}^{(i)}(\kappa)$. If we can find a feasible solution of problem (24), $\underline{\kappa}$ is replaced by κ. Otherwise, $\overline{\kappa}$ is replaced.

The feasibility check of $\boldsymbol{u}^{(i)}(\kappa)$ works as follows. Since $\boldsymbol{u}^{(i)}(\kappa) = \tilde{\boldsymbol{\gamma}}(\kappa)$, we will check whether $\tilde{\boldsymbol{\gamma}}(\kappa)$ is feasible instead. Define interference matrix $\boldsymbol{A}$ with entries

$$a_{i,j}^{k} = \begin{cases} \dfrac{g_{i,j}^{k}}{g_{j,j}^{k}}, & i \neq j, \\ 0, & \text{otherwise,} \end{cases} \qquad (25)$$

and a vector

$$\boldsymbol{\nu} = (\frac{\sigma^2}{g_{1,1}^{1}}, \dots, \frac{\sigma^2}{g_{(M+B),(M+B)}^{1}}, \frac{\sigma^2}{g_{1,1}^{2}}, \dots, \frac{\sigma^2}{g_{(M+B),(M+B)}^{K}})^{\mathrm{T}}. \qquad (26)$$

Note that $\boldsymbol{A}$ is block diagonal due to synchronous transmission. According to the SINR definition in (7), we have

$$\tilde{\gamma}_{i,j}^{k}(\kappa) - 1 = \frac{p_{i,j}^{k}}{(\boldsymbol{A}\boldsymbol{p} + \boldsymbol{\nu})_{i,j}^{k}}, i, j \in \mathcal{B} \cup \mathcal{M}. \qquad (27)$$

By rearranging the terms, (27) can be rewritten as

$$\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{\nu} = (\boldsymbol{I} - \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})\boldsymbol{p}. \qquad (28)$$

Denote the spectral radius of matrix $\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}$ by $\rho(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})$. The following property helps to check whether $\tilde{\boldsymbol{\gamma}}(\kappa)$ is feasible.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2819648, IEEE Transactions on Wireless Communications

Theorem 1. *For any $\tilde{\boldsymbol{\gamma}}(\kappa) \succeq 1$, $\tilde{\boldsymbol{\gamma}}(\kappa)$ is feasible if $\rho(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}) < 1$ and $\sum_{k\in\mathcal{K}} p_{i,j}^k \le P$ with $\boldsymbol{p} = (\boldsymbol{I} - \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})^{-1}\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{\nu}$.*

Proof: The proof of Theorem 1 is given in Appendix A. ■

Based on Theorem 1, the steps of feasibility check are summarized as follows:

1) Construct A according to (25).
2) If $\rho(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}) \ge 1$, then $\tilde{\boldsymbol{\gamma}}(\kappa)$ and thus $\boldsymbol{u}^{(i)}(\kappa)$ is infeasible. Otherwise, set $\boldsymbol{p} = \sigma^2(\boldsymbol{I} - \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})^{-1}\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})$.
3) If $\sum_{k\in\mathcal{K}} p_{i,j}^k \le P$, then $\boldsymbol{u}^{(i)}(\kappa)$ is feasible. Otherwise, it is infeasible.

Remark 2. *Denote $\mathcal{N} = \{1, \ldots, 2(M+B)^2K\}$. In each iteration, from computational point of view, we will not further partition the box along n-th axis for any $n \in \mathcal{N}$ when $z_n - y_n \le \eta$. Let $\mathcal{I} = \{n \in \mathcal{N} : z_n - y_n \le \eta\}$. We can only partition the boxes in $\mathcal{N} \setminus \mathcal{I}$, and thus, the number of new boxes is $2|\mathcal{N} \setminus \mathcal{I}| - 2$.*

D. An Accelerated Method

According to Remark 2, there are $2|\mathcal{N} \setminus \mathcal{I}| - 2$ new boxes are generated at iteration i. Thus, the size of the box set grows quickly at each iteration when the number of small cell users M is large. To tackle this problem, we present an accelerated algorithm that prevents the size of box set from growing too fast, so as to expedite the convergence. The solution is sub-optimal when the accelerated method is applied. The accelerated method is to skip searching the neighboring boxes to achieve a trade-off between the optimality and the complexity.

The partition scheme in RA algorithm suggests that partitions over the boxes with symmetric upper bound and lower bound vertices $[\boldsymbol{y}_1, \boldsymbol{z}_1]$ and $[\boldsymbol{y}_2, \boldsymbol{z}_2]$ yield nearly equally lower and upper bounds $(g(\boldsymbol{r}^*(\boldsymbol{q}_1)), \boldsymbol{r}^*(\boldsymbol{z}_1))$ and $(g(\boldsymbol{r}^*(\boldsymbol{q}_2)), \boldsymbol{r}^*(\boldsymbol{z}_2))$. Such symmetry implies that there could exist more than one equally optimal box at each iteration. Therefore, selecting any one of these boxes and eliminating the others would not severely affect the optimality of the algorithm.

One difficulty of this idea is the identification of symmetric boxes. To address this issue, we first give two definitions.

Definition 4. *The distance between two boxes $\mathcal{B}_1 = [\boldsymbol{y}_1, \boldsymbol{z}_1]$ and $\mathcal{B}_2 = [\boldsymbol{y}_2, \boldsymbol{z}_2]$ is defined as $d(\mathcal{B}_1, \mathcal{B}_2) = \max\{||\boldsymbol{y}_1 - \boldsymbol{y}_2||, ||\boldsymbol{z}_1 - \boldsymbol{z}_2||\}$.*

Definition 5. *Two boxes $\mathcal{B}_1$ and $\mathcal{B}_2$ are said to be symmetric if and only if $d(\mathcal{B}_1, \mathcal{B}_2) \le \delta$, where $\delta \ge 0$ is predefined.*

Based on these two definitions, the symmetric box set of the best box $\overline{\mathcal{B}}$ is defined as

$$\mathcal{J}^{(i)} = \{\mathcal{B} : d(\mathcal{B}, \overline{\mathcal{B}}) \le \delta, \mathcal{B} \in \mathcal{P}^{(i)}\}. \tag{29}$$

Therefore, RA algorithm can be modified as follows. Find the symmetric box set $\mathcal{J}^{(i)}$ of $\overline{\mathcal{B}}^{(i)}$ according to (29), and skip all the boxes in $\mathcal{J}^{(i)}$ in the shrinking process.

V. Complementary Geometric Programming for D2D Routing

In this section, we first reformulate the D2D routing subproblem as a CGP problem given the SBS decision, and then propose a DR algorithm to solve this problem.

A. Complementary Geometric Programming Formulation

Given the resource allocation result $\boldsymbol{p}$, problem (P1) can be rewritten as

$$\begin{aligned} \text{P6:} \quad & \underset{\boldsymbol{x},\boldsymbol{d}}{\text{maximize}} \sum_{f\in\mathcal{F}} \sum_{i\in\mathcal{M}} \sum_{j\in N_i} r_{i,j}^f s_{i,f}, \\ s.t. \quad & \text{(14b), (14c),(14d),(14e), (14h), and (14j).} \end{aligned} \tag{30}$$

A close look at these constraints suggests that all the constraints are convex except constraint (14d). Note that

$$1 + \gamma_{i,j}^k = \frac{\sigma^2 + \sum_{i,j} x_{i,j} p_{i,j}^k g_{i,j}^k}{\sigma^2 + \sum_{i'\neq i, j'\neq j} x_{i',j'} p_{i',j'}^k g_{i',j}^k} \tag{31}$$

in constraint (14d) is nonconvex of variable $\boldsymbol{x}$. However, $1 + \gamma_{i,j}^k$ is a ratio between two posynomials of $\boldsymbol{x}$. Minimizing a ratio between two posynomials is a class of nonconvex problems known as CGP. In the following, we first briefly introduce some preliminaries of CGP [29]. Then, given the destination matrix $\boldsymbol{d}$, problem (P6) will be reformulated as a CGP problem .

1) Preliminaries of CGP:

Definition 6. *A* CGP *is an optimization problem of the form*

$$\begin{aligned} & \min f_0(\boldsymbol{x}) \\ s.t. \quad & g_l(\boldsymbol{x}) \le 1, \forall l = 1, \ldots, n_1, \\ & h_l(\boldsymbol{x}) = 1, \forall l = 1, \ldots, n_2, \\ & \tfrac{w_{l1}(\boldsymbol{x})}{w_{l2}(\boldsymbol{x})} \le 1, \forall l = 1, \ldots, n_3, \end{aligned} \tag{32}$$

where $h_l(\boldsymbol{x})$ are monomial functions and $f_0(\boldsymbol{x})$, $g_l(\boldsymbol{x})$, $w_{l1}(\boldsymbol{x})$, as well as $w_{l2}(\boldsymbol{x})$ are posynomial functions.

2) Transformation to CGP: First, we show that relaxing the binary constraints on $\boldsymbol{x}$ does not affect the optimality of the solution.

Theorem 2. *By approximating $x_{i,j}^f$ with $x_{i,j}^f = 1 - \exp(-\theta r_{i,j}^f)$, the relaxation of constraint (14e), i.e.,*

$$x_{i,j} + \prod_{f\in\mathcal{F}} (\tilde{r}_{i,j}^f)^{(-\theta)} \le 1, x_{i,j} \ge 0, \tag{33}$$

is exact when θ is sufficiently large.

Proof: The proof of Theorem 2 is given in Appendix B. ■

Let $\tilde{r}_{i,j}^f = \exp(r_{i,j}^f)$. Problem (P6) can then be rewritten as follows, given the destination $\boldsymbol{d}$.

$$\text{P7:} \underset{\boldsymbol{x},\boldsymbol{r}}{\text{maximize}} \prod_{f\in\mathcal{F}}\prod_{i\in\mathcal{M}}\prod_{j\in N_i} (\tilde{r}_{i,j}^f)^{s_{i,f}}, \tag{34a}$$

$$s.t. \ \frac{\prod\limits_{j\in N_i} \tilde{r}_{j,i}^f}{\prod\limits_{j\in N_i} \tilde{r}_{i,j}^f \prod\limits_{u\in\mathcal{M}} \prod\limits_{v\in N_u} (\tilde{r}_{u,v}^f)^{s_{u,f}(d_{i,f}-s_{i,f})}} = 1, \tag{34b}$$

$$\prod_{f\in\mathcal{F}} \tilde{r}_{i,j}^f \prod_{k\in\mathcal{K}} \frac{\sigma^2 + \sum\limits_{i'\neq i, j'\neq j} x_{i',j'} p_{i',j'}^k g_{i',j'}^k}{\sigma^2 + \sum\limits_{i,j} x_{i,j} p_{i,j}^k g_{i,j}^k} \leq 1, \tag{34c}$$

$$x_{i,j} + \prod_{f\in\mathcal{F}} (\tilde{r}_{i,j}^f)^{(-\theta)} \leq 1, \tag{34d}$$

$$0 \leq x_{i,j}. \tag{34e}$$

Note the fact that the left hand side (LHS) of constraint (34b) is monomial, the LHS of constraint (34c) is the ratio of two posynomial functions, and the LHS of constraint (34d) is posynomial. Therefore, problem (34) is a CGP problem when $\boldsymbol{d}$ is given. In the following, we will design an iterative algorithm to solve the D2D routing subproblem efficiently.

B. Algorithm Design for D2D Routing

In the last subsection, we have reformulated the D2D routing subproblem as a CGP problem (34) given the destination matrix $\boldsymbol{d}$. This subsection first presents an efficient algorithm to solve the problem. Then, we present an efficient algorithm for the SBS searching.

1) Posynomial Approximation: We first present an algorithm to solve problem (34). Note that the LHS of constraint (34c) is nonconvex. Here, we simplify the problem by approximating it by a monomial approximation. Then, the problem can be turned into a GP problem that can be solved efficiently [33]. Denote the LHS of constraint (34c) by $W(\tilde{\boldsymbol{r}}, \boldsymbol{x})$ and its approximation by $\hat{W}(\tilde{\boldsymbol{r}}, \boldsymbol{x})$. In addition, denote $h_k(\boldsymbol{x}) = \sigma^2 + \sum\limits_{i,j} x_{i,j} p_{i,j}^k g_{i,j}^k$, and the monomial approximation $\hat{h}_k(\boldsymbol{x}) = c_k \prod\limits_{i,j} x_{i,j}^{a_{i,j}^k}$.

Define logarithm transformations $t_{i,j} = \log x_{i,j}$, $g_k(\boldsymbol{t}) = \log h_k(\boldsymbol{t}) = \log(\sum\limits_{i,j} e^{t_{i,j}} g_{i,j}^k p_{i,j}^k + \sigma^2)$, and $\hat{g}_k(\boldsymbol{t}) = \log \hat{h}_k(\boldsymbol{t}) = \log c_k + \sum\limits_{i,j} a_{i,j}^k t_{i,j}$. Equating the first-order Taylor expansion of $g_k(\boldsymbol{t})$ at $\boldsymbol{t}'$ with $\hat{g}_k(\boldsymbol{t})$, we have

$$g_k(\boldsymbol{t}') + \sum_{i,j} \frac{\partial g_k(\boldsymbol{t})}{\partial t_{i,j}} (t_{i,j} - t'_{i,j}) = \log c_k + \sum_{i,j} a_{i,j}^k t_{i,j}, \tag{35}$$

which implies that

$$a_{i,j}^k = \frac{\partial g_k(\boldsymbol{t})}{\partial t_{i,j}} = \frac{x_{i,j}}{h_k(\boldsymbol{x})} \left.\frac{\partial h_k(\boldsymbol{x})}{\partial x_{i,j}}\right|_{x_{i,j}=x'_{i,j}}, \tag{36}$$

and

$$c_k = \exp(g_k(\boldsymbol{t}') - \sum_{i,j} \frac{\partial g_k(\boldsymbol{t})}{\partial t_{i,j}} t'_{i,j}) = \left.\frac{h_k(\boldsymbol{x})}{\prod\limits_{i,j} x_{i,j}^{a_{i,j}^k}}\right|_{x_{i,j}=x'_{i,j}}. \tag{37}$$

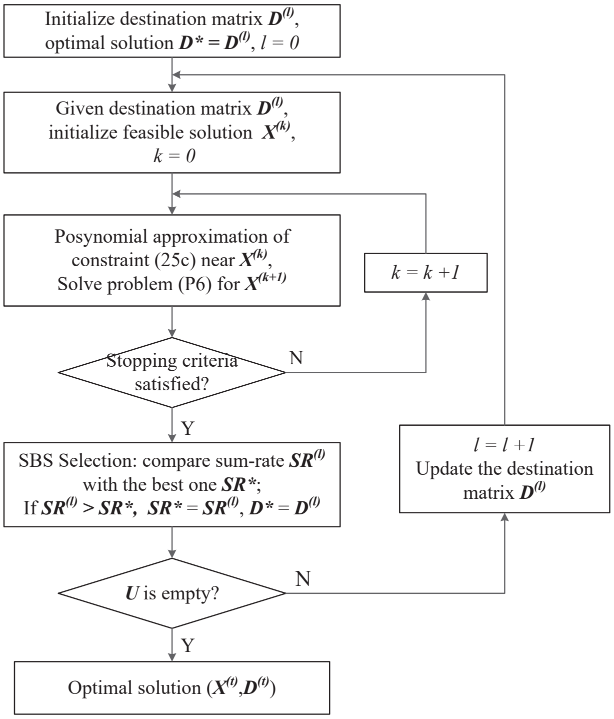

Initialize destination matrix $\boldsymbol{D}^{(l)}$, optimal solution $\boldsymbol{D}^* = \boldsymbol{D}^{(l)}$, $l = 0$
Given destination matrix $\boldsymbol{D}^{(l)}$, initialize feasible solution $\boldsymbol{X}^{(k)}$, $k = 0$
Posynomial approximation of constraint (25c) near $\boldsymbol{X}^{(k)}$, Solve problem (P6) for $\boldsymbol{X}^{(k+1)}$
$k = k + 1$
Stopping criteria satisfied?
N
Y
SBS Selection: compare sum-rate $\boldsymbol{SR}^{(l)}$ with the best one $\boldsymbol{SR}^*$; If $\boldsymbol{SR}^{(l)} > \boldsymbol{SR}^*$, $\boldsymbol{SR}^* = \boldsymbol{SR}^{(l)}$, $\boldsymbol{D}^* = \boldsymbol{D}^{(l)}$
$l = l + 1$ Update the destination matrix $\boldsymbol{D}^{(l)}$
$\boldsymbol{U}$ is empty?
N
Y
Optimal solution $(\boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)})$

Fig. 5. Iterative GP algorithm for the D2D routing subproblem.

Once $a_{i,j}^k$ and c_k are computed, the LHS of the approximated constraint can be expressed by

$$W(\tilde{\boldsymbol{r}}, \boldsymbol{x}) = \prod_{f\in\mathcal{F}} \tilde{r}_{i,j}^f \prod_{k\in\mathcal{K}} \frac{\sigma^2 + \sum\limits_{i'\neq i, j'\neq j} x_{i',j'} p_{i',j'}^k g_{i',j'}^k}{c_k \prod\limits_{i,j} x_{i,j}^{a_{i,j}^k}}, \tag{38}$$

which is monomial. Using the monomial approximation, the problem (P7) is converted into GP that can be solved efficiently using convex optimization techniques [34].

2) Iterative GP Algorithm: In this subsection, we present the iterative GP algorithm to solve problem (P7) through a sequence of posynomial approximation. At the beginning of each iteration, we have an initial value for variables $(\tilde{\boldsymbol{r}}, \boldsymbol{x})$. Then, we replace constraint (34c) using the aforementioned monomial approximation approach, and convert this problem into a GP problem. The solution of the GP problem is taken as the initial value of the monomial approximation in the next iteration.

Theorem 3. *The solutions obtained by the iterative GP algorithm converges to the KKT conditions of problem (P7).*

Proof: Please refer to [29]. ■

3) SBS Selection: In this subsection, we will discuss how to search the solution for destination matrix $\boldsymbol{d}$. Since destination variables are binary, the feasible solution for each flow is finite. Thus, we can search these feasible solutions for the global optimal one. Note that the exhaustive search for the global optimal solution requires solving B^F CGP problems, whose computational complexity grows exponentially as the number of flows. Thus, we propose a local search method to reduce the computational complexity.

Denote $\mathcal{U}$ by the set of flows that have not found the local optimal destination. Initially, $\mathcal{U} = \mathcal{F}$. SBS selection follows three searching steps as below:

Step 1: Give an initial feasible solution for the destination matrix $\boldsymbol{d}$, and then solve problem (P7) to obtain the D2D routing result $\boldsymbol{x}$;

Step 2: Choose the flow f_l with the lowest data rate from set $\mathcal{U}$, and try each SBS $b \in \mathcal{B}$, i.e., $d_{f_l,b} = 1, d_{f_l,i} = 0, i \in \mathcal{B}\backslash\{b\}$. In the enumeration for SBS b, recall the iterative GP algorithm to obtain the total sum-rate. Among the enumerations, the SBS with the maximum sum-rate b_m will be selected as the destination for flow f_l, i.e., $d_{f_l,b_m} = 1, d_{f_l,i} = 0, i \in \mathcal{B}\backslash\{b\}$. Then, remove flow f_l from the set $\mathcal{U}$;

Step 3: Repeat Step 2 until set $\mathcal{U}$ is empty.

Using this local search method, we can find the solution for the destination matrix $\boldsymbol{d}$ by solving BF CGP problems. The flowchart of the iterative GP algorithm for the D2D routing subproblem is shown in Fig. 5.

VI. Overall Convergence and Complexity Analysis

In this subsection, we will first prove the convergence of the I-RA-DR algorithm. Then, we provide some analyses on the computational complexity and the signaling cost.

A. Convergence Analysis

Define $g(\boldsymbol{P}, \boldsymbol{X}, \boldsymbol{D}) = \sum\limits_{f \in \mathcal{F}} r_f$. First, in the resource allocation subproblem, we obtain the solution given $\boldsymbol{X}^{(t)}$ and $\boldsymbol{D}^{(t)}$. Therefore, we have

$$g(\boldsymbol{P}^{(t+1)}, \boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)}) \geq g(\boldsymbol{P}^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)}). \qquad (39)$$

Second, in the D2D routing subproblem, we solve a local optimal solution given $\boldsymbol{P}^{(t+1)}$, thus, the following inequality holds:

$$g(\boldsymbol{P}^{(t+1)}, \boldsymbol{X}^{(t+1)}, \boldsymbol{D}^{(t+1)}) \geq g(\boldsymbol{P}^{(t+1)}, \boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)}). \quad (40)$$

Based on the above two inequalities (39) and (40), we can obtain

$$g(\boldsymbol{P}^{(t+1)}, \boldsymbol{X}^{(t+1)}, \boldsymbol{D}^{(t+1)}) \geq g(\boldsymbol{P}^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{D}^{(t)}), \qquad (41)$$

which implies that the objective value of problem (P1) is non-decreasing after each iteration of the I-RA-DR Algorithm. Since the objective value of problem (P1) is upper bounded by a finite value, the proposed I-RA-DR algorithm is guaranteed to converge.

B. Computational Complexity Analysis

In the resource allocation subproblem, we transform it as MO problem and obtain its solution by polyblock outer approximation method. However, the computational complexity of MO is still an open problem. To accelerate the convergence of the polyblock outer approximation method, we skip those boxes which are symmetric, at least $\frac{2\delta}{2\delta+||\boldsymbol{z}-\boldsymbol{y}||}$ of boxes will be skipped in the RA algorithm, and the proportion of the skipped boxes is positively related to the dimension of $\boldsymbol{z}$.

TABLE I
Parameters for Simulation

Network layout	50m-by-50m area
Number of small cells	3
Small cell radius	10 m
User's Transmit Power P	20 dBm
Transmission Bandwidth	20 MHz
Carrier Frequency	1.9 GHz
Noise Figure	5 dB
Decay Factor of the Path Loss α	3.5
Power Gains Factor κ	-31.5 dB
Error Tolerance Level ϵ	0.1
Error Tolerance Level η	3
Error Tolerance Level δ	20

In the D2D routing subproblem, we reformulate it as a CGP problem given the destination and solve it by successive GP approximation. According to the results in [35], the GP problem can be solved in polynomial time and GP approximation can terminate within few iterations in practice. Besides, the SBS selection requires to solve BF CGP problems, therefore, the D2D routing subproblem can be solved in polynomial time and the computational complexity of one iteration mainly comes from the resource allocation subproblem.

C. Signaling Cost Analysis

To describe the signaling cost of the proposed I-RA-DR algorithm, we assume that μ messages are required for a user to report its location, υ messages are necessary for the subchannel estimation, ζ messages are needed for the BS to notify a user about the allocated subchannels, transmission power, and routing results, and ξ messages for two SBSs to exchange information such as location and resource utilization. In the D2D routing subproblem, each user $i \in \mathcal{M}$ needs to upload its location to the associated SBS, and thus, this requires at most μM messages. Besides, each SBS also needs to exchange information with the others, therefore, at most $B(B-1)\xi$ messages are needed for the information interaction among SBSs. Note that the data traffic can only be offloaded to the neighboring small cells, and thus, the number of involved SBS B can be restricted.

In the resource allocation subproblem, each active link needs to inform the associated SBS of the estimated channel information, and then the SBS will response to the requested link. Therefore, $(\mu + \upsilon) \sum\limits_{i,j} x_{i,j}$ messages are required for the resource allocation. Due to the limited subchannel resources, the number of active links is constrained, and thus, the signaling cost of the resource allocation can be also restricted to a tolerable level.

VII. Simulation Results

In this section, we evaluate the performance of the I-RA-DR algorithm for D2D load balancing. We consider a three-cell scenario, in which the users are uniformly distributed, and the data traffic generated in different cells are imbalanced to show the proposed algorithms can achieve efficient load balancing. In addition, we show that the D2D load balancing can improve the system performance on the data rate. In different small

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2819648, IEEE Transactions on Wireless Communications

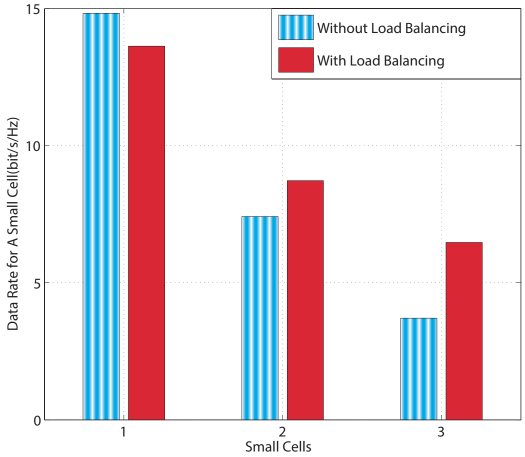


Fig. 6. Distribution of the data rate for a small cell with $K = 4$ and $N = 30$.

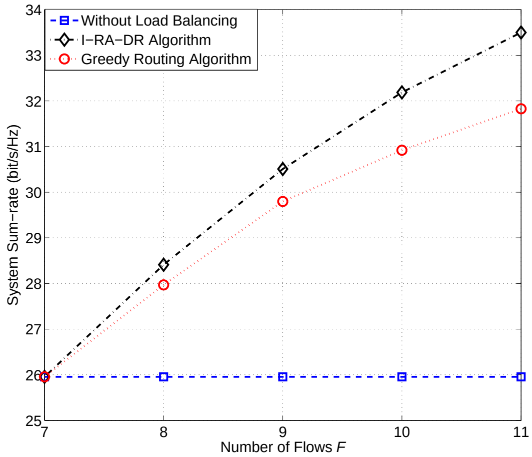


Fig. 7. Sum-rate with the number of flows F for $N = 30$ and $K = 4$.

cells, the same set of available subchannels are deployed. For simplicity, we assume that the numbers of users are also identical in different small cells. The simulation parameters are listed in Table I based on the existing LTE/LTE-Advanced specifications [36].

In Fig. 6, we plot the distribution of the data rate for a small cell with $K = 4$ subchannels and $N = 30$ users. Suppose that there are $[8, 2, 1]$ users generating data traffic in respective small cells. Without load balancing, users get access to subchannels orthogonally, and each small cell can only support at most 4 users. On the contrary, with D2D load balancing, some users can set up D2D communications and detour the load traffic to the neighboring small cells. In Fig. 6, the data rate in small cell 2 and 3 increase by 18% and 75%, respectively. However, due to the interference from D2D communications, the data rate in small cell 1 decreases by 8%. It can be easily observed that the increased sum-rate caused by traffic detouring can make up the decreased sum-rate due to the intra-channel interference, and hence the total sum-rate grows.

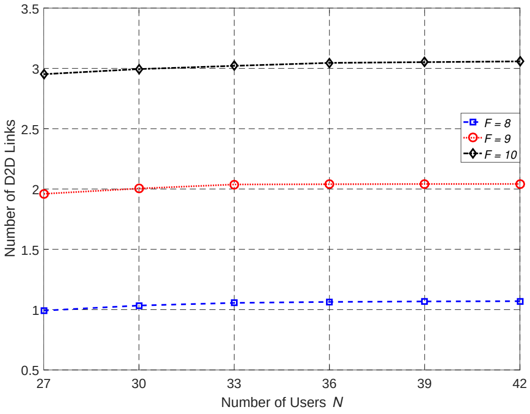


Fig. 8. Number of D2D links with the number of users N for $K = 4$.

In Fig. 7, we plot the system sum-rate as a function of the number of flows F with $N = 30$ users and $K = 4$ channels. There are $[6, 2, 1]$ users generating data flows in small cell 1, 2, and 3, respectively, and the increased flows are all from small cell 1. In Fig. 7, we compare the system performance with the scheme without load balancing, and the scheme using Greedy routing algorithm with power control. Greedy algorithm always selects the nearest SBS as the destination, and the node closest to the destination SBS in the node candidate set as the D2D relay node. Without load balancing, the concurrent transmitting users in each small cell cannot exceed the number of subchannels. Thus, even the number of active users increases, the total sum-rate remains the same. It can be also observed that the system sum-rate obtained by the I-RA-DR algorithm is always higher than that using the greedy algorithm. This is because the greedy algorithm only considers interference, and the target data rate is not maximized.

In Fig. 8, we plot the number of D2D links as a function of the number of users N with $K = 4$ channels. There are $[7, 2, 1]$ users generating data flows in small cell 1, 2, and 3, respectively, and the increased flows are all from small cell 1. In this figure, we can learn that most flows can be offloaded to the neighboring SBS by one-hop D2D communications due to the ultra-dense SBS deployment. In addition, we can also observe that the number of D2D links increases slightly with the number of users N. This is because more users can provide more options in D2D routing.

In Fig. 9, we plot the sum data rate as a function of the number of users N with $F = 9$ flows, i.e., there are $[6, 2, 1]$ users generating data flows in small cell 1, 2, and 3, respectively, and $K = 4$ channels. Without load balancing, the concurrent transmission users need to be less than K, thus, there are 4 users communicating with SBS directly. For this reason, the sum data rate without load balancing keeps a constant even N grows. As for the I-RA-DR and greedy algorithms, more D2D relay users provide a diversity for relay selection. Thus, the system sum-rate increases as the number of users grows. The simulation results show that the system sum-rate increases 0.2 bit/s/Hz using the I-RA-DR algorithm for there more D2D relays.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2018.2819648, IEEE Transactions on Wireless Communications

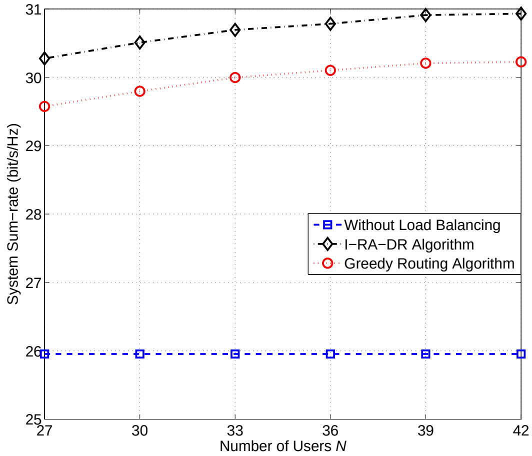


Fig. 9. Sum-rate with the number of users N for $F = 9$ and $K = 4$.

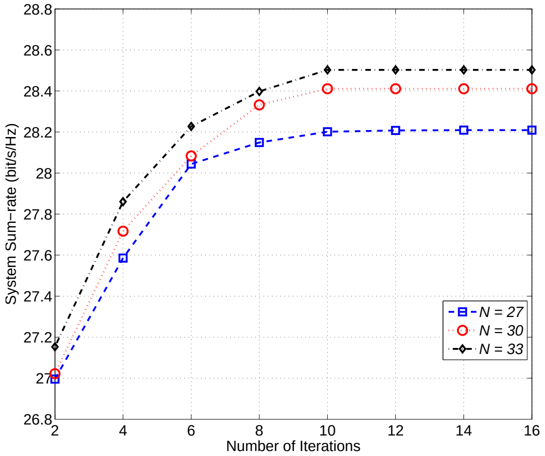


Fig. 11. Convergence speed of I-RA-DR algorithm.

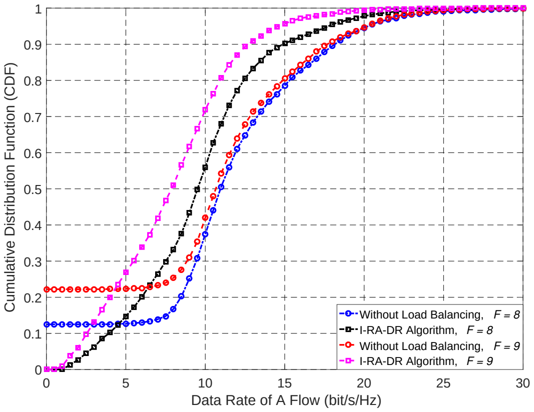


Fig. 10. Cumulative distribution function of data rate of a flow with $K = 4$ and $N = 30$.

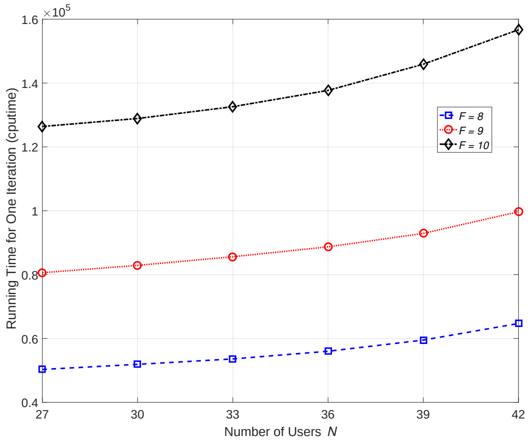


Fig. 12. Running time for one iteration of the I-RA-DR algorithm.

In Fig. 10, we plot the Cumulative Distribution Function (CDF) of the data rate of a flow with $K = 4$ and $N = 30$. There are $[5, 2, 1]$ users generating data flows in small cell 1, 2, and 3, respectively, and the increased flows are all from small cell 1. Without load balancing, the concurrent transmission users need to be less than K. Thus, there are 12.5% of edge users in outage when $F = 8$ and 22.2% of edge users in outage when $F = 9$. While in the I-RA-DR algorithm, the data generated from the cell edge users can be offloaded to the neighboring small cells by D2D communications, and thus, more users are active. For this reason, the D2D load balancing can effectively improve the data rate of the cell edge users. However, due to the spectrum sharing of the D2D links, the average data rate for the active flows decreases.

In Fig. 11, we demonstrate the convergence speed of the proposed I-RA-DR algorithm. We plot the sum data rate versus the number of iterations when $F = 8$ flows, i.e., there are $[5, 2, 1]$ users generating data flows in small cell 1, 2, and 3, respectively, and $K = 4$ channels for different values of N. It can be seen that around 10 iterations are needed for convergence. In addition, we can observed that the number of iterations does not increase as the number of users N grows.

In Fig. 12, we plot the running time per each iteration versus the number of users with $K = 4$ channels for different values of F. We can learn that the larger number of users or flows will lead to a longer running time for one iteration. It can also be observed that F affects the running time more significantly than N, because F will affect the running time of the both two subproblems while N only affects that of the D2D routing subproblem. This is also consistent with the analysis in Section VI-B.

VIII. Conclusions

In this paper, we studied the D2D load balancing in an OFDMA ultra-dense small cell network, in which users can set up direct cellular links with SBS or the data traffic was detoured to the neighboring SBS by D2D communications. We solved the D2D load balancing problem by an iterative method, where resource allocation and D2D routing subproblems were optimized iteratively. In particular, we reformulated the resource allocation subproblem as an MO problem. Likewise, the D2D routing subproblem was solved as a CGP. Simulation results showed that the D2D load balancing can not only balance the traffic effectively, but also improve the system performance in terms of data rate. Compared to the scheme without load balancing, the performance on system sum-rate using the I-RA-DR algorithm for D2D load balancing was improved by 12%.

APPENDIX A
PROOF OF THEOREM 1

First, the variable $\boldsymbol{p}$ needs to satisfy the power constraint $\sum\limits_{k\in\mathcal{K}} p_{i,j}^k \leq P$. Second, $\tilde{\boldsymbol{\gamma}}(\kappa)$ is feasible when $\boldsymbol{p}$ in this equation has a unique solution $\boldsymbol{p} = (\boldsymbol{I} - \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})^{-1}\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{\nu}$.

On the other hand, $\boldsymbol{p}$ in this equation can be solved in an iterative method,

$$\boldsymbol{p}^{(i+1)} = \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{\nu} + \text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})\boldsymbol{p}^{(i)}, \quad (42)$$

and the necessary and sufficient condition of convergence is $\rho(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}) < 1$. Let $\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A} = X^{-1}JX$, with

$$J = \begin{bmatrix} J_1(\lambda_1) & \cdots & 0 \\ \vdots & J_n(\lambda_n) & \vdots \\ 0 & \cdots & J_N(\lambda_N) \end{bmatrix} \quad (43)$$

where $J_n(\lambda_n)$ is a Jordan matrix, and λ_n is the n-th eigenvalue of $\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}$. Thus, we have $(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A})^i = X^{-1}J^iX$. The iterative method converges only when $\lim\limits_{i\to\infty} J_n^i = 0$, that is, $\forall n, |\lambda_n| < 1$. Therefore, $\rho(\text{diag}(\tilde{\boldsymbol{\gamma}}(\kappa) - \mathbf{1})\boldsymbol{A}) < 1$. ■

APPENDIX B
PROOF OF THEOREM 2

We approximate $x_{i,j}^f$ by $x_{i,j}^f = 1 - \exp(-\theta r_{i,j}^f)$ in constraint (14j). When θ is sufficiently large, the approximation will not affect the optimal solution, because the value of $x_{i,j}^f$ is always near 0 or 1, which indicates whether the link $i-j$ is active. By substituting this approximation to constraint (14e), we can get $x_{i,j} + \prod\limits_{f\in\mathcal{F}} (\exp(-\theta r_{i,j}^f)) = 1$.

Now, we need to prove that the optimal solution can be obtained by replacing the equality constraints with the inequality constraints.

Case 1: There does not exist a flow f on link $i-j$, i.e., $r_{i,j}^f = 0$ for all f. According to the inequality constraints, $x_{i,j}$ only has the solution $x_{i,j} = 0$, which is equivalent to the equality constraint.

Case 2: There exists a flow f on link $i-j$, i.e., $r_{i,j}^f > 0$. The value of $x_{i,j}$ can be any number between 0 and 1. Note that the capacity of link $i-j$ grows as the value of $x_{i,j}$. Since our objective is to maximize the system sum-rate, the value of $x_{i,j}$ will reach the maximum, i.e., $x_{i,j} = 1$. Therefore, the optimal solution can also be obtained by the inequality constraints. ■

REFERENCES

[1] H. Zhang, L. Song, and Y. Zhang, "Load Balancing for Cellular Networks using Device-to-Device Communications," in *Proc. IEEE VTC-Spring*, Sydney, Australia, Jun. 2017.

[2] Ciso, "Cisco Virtual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020 White Paper," Feb. 2016.

[3] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Ultra Dense Small Cell Networks: Turning Density Into Energy Efficiency," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1267-1280, May 2016.

[4] A. Ghost, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo, H. S. Dhillon, and T. D. Nolvan, "Heterogeneous Cellular Networks: From Theory to Practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54-64, Jun. 2012.

[5] T. Zhou, Z. Liu, D. Qin, N. Jiang, and C. Li, "User association with maximizing weighted sum energy efficiency for massive MIMO-enabled heterogeneous cellular networks," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2250-2253, Jul. 2017.

[6] Q. We, B. Rong, Y. Chen, M. A.-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706-2716, Jun. 2013.

[7] Y. Zhao, X. Fang, R. Huang, and Y. Fang, "Joint Interference Coordination and Load Balancing for OFDMA Multihop Cellular Networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 89-101, Jan. 2014.

[8] H. Jiang and S. Rappaport, "CBWL: A New Channel Assignment and Sharing Method for Cellular Communication Systems," *IEEE Trans. Veh. Technol.*, vol. 43, no. 2, pp. 313-322, May 1994.

[9] S. K. Das, S. K. Sen, and R. Jayaram, "A Dynamic Load Balancing Strategy for Channel Assignment Using Selective Borrowing in Cellular Mobile Environment," *Wireless Networks*, vol. 3, no. 5, pp. 333-347, Oct. 1997.

[10] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated Cellular and Ad-hoc Relaying Systems: iCAR," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 10, pp. 2105-2115, Oct. 2001.

[11] E. Yanmaz and O. K. Tonguz, "Dynamic Load Balancing and Sharing Performance of Integrated Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 5, pp. 862-872, Jun. 2004.

[12] C. K. Ho, D. Yuan, and S. Sun, "Data Offloading in Load Coupled Networks: A Utility Maximization Framework," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1921-1931, Apr. 2014.

[13] M. Madhavan, H. Ganapathy, M. Chetlur, and S. Kalyanaraman, "Adapting Cellular Networks to Whitespaces Spectrum," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 383-397, Apr. 2015.

[14] L. Song, D. Niyato, Z. Han, and E. Hossain, *Wireless Device-to-Device Communications and Networks*, Cambridge University Press, UK, 2015.

[15] L. Wang and H. Wu, "Fast Pairing of Device-to-Device Link Underlay for Spectrum Sharing with Cellular Users," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1803-1806, Oct. 2014.

[16] H. Zhang, L. Song, and Z. Han, "Radio Resource Allocation for Device-to-Device Underlay Communication Using Hypergraph Theory," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4852-4861, Jul. 2016.

[17] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-Device Communication as an Underlay to LTE-advanced Networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42-49, Dec. 2009.

[18] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao, "Efficiency Resource Allocation for Device-to-Device Underlay Communication Systems: A Reverse Iterative Combinatorial Auction Based Approach," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 348-358, Sep. 2013.

[19] L. Wang, H. Tang, H. Wu, and G. L. Stüber, "Resource Allocation for D2D Communications Underlay in Rayleigh Fading Channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1159-1170, Apr. 2016.

[20] T. Zhou, Y. Huang, and L. Yang, "Joint User Association and Interference Mitigation for D2D-Enabled Heterogeneous Cellular Networks," *Mobile Networks Applicat.*, vol. 21, no. 4, pp. 589-602, Aug. 2016.

[21] X. Lin, J. G. Andrews, A. Ghosh, and R. Ratasuk, "An Overview of 3GPP Device-to-Device Proximity Services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40-48, Apr. 2014.

[22] K. W. Choi and Z. Han, "Device-to-Device Discovery for Proximity-Based Service in LTE-Advanced System," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 55-66, Jan. 2015.

[23] H. Zhang, Y. Liao, and L. Song, "D2D-U: Device-to-Device Communications in Unlicensed Bands for 5G System," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3507-3519, Jun. 2017.

[24] J. Liu, Y. Kawamto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-Device Communications Achieve Efficient Load Balancing in LTE-advanced Networks," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 57-65, Apr. 2014.

[25] L. Deng, Y. Zhang, M. Chen, Z. Li, J. Lee, Y. Zhang, and L. Song, "Device-to-Device Load Balancing for Cellular Networks," in *Proc. IEEE MASS*, Dallas, USA, Oct. 2015.

[26] Z. Chen, H. Zhao, Y. Chao, and T. Jiang, "Load Balancing for D2D-based relay Communications in Heterogeneous Network," in *Proc. WiOpt*, Mumbai, India, May 2015.

[27] M. H. Hajiesmaili, L. Deng, M. Chen, and Z. Li, "Incentivizing Device-to-Device Load Balancing for Cellular Networks: An Online Auction Design", *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 265-279, Feb. 2017.

[28] Y. J. Zhang, L. Qian, and J. Huang, "Monotonic Optimization in Communication and Networking Systems," *Found. Trends Netw.*, vol. 7, no. 1, pp. 1-75, Oct. 2013.

[29] M. Chiang, "Nonconvex Optimization for Communication Systems," *Advances in Mechanics and Mathematics*, vol. 3, pp. 137-196, 2008.

[30] Q. Chen, G. Yu, R. Yin, and G. Y. Li, "Energy-Efficient User Association and Resource Allocation for Multistream Carrier Aggregation," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6366-6376, Aug. 2016.

[31] F. Liu, K. Zheng, W. Xiang, and H. Zhao, "Design and Performance Analysis of An Energy-Efficient Uplink Carrier Aggregation Scheme," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 197-207, Feb. 2014.

[32] J. Li, L. P. Qian, Y. J. Zhang, and L. Shen, "Global Optimial Rate Control and Scheduling for Spectrum-Sharing Multi-Hop Networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6462-6473, Sep. 2016.

[33] M. Chiang, "Geometric Programming for Communication Systems," *Found. Trends Commun. Inform. Theory*, vol. 2, no. 1-2, pp. 1-156, Aug. 2005.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

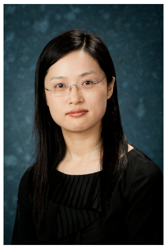
[35] P. C. Weeraddana, M. Codreanu, M. Latva-aho, A. Ephremides, and C. Fischione, "Weighted Sum-Rate Maximization in Wireless Networks: A Review," *Found. Trends Netw.*, vol. 6, no. 1-2, pp. 1-163, Sep. 2012.

[36] 3GPP TR 36.931, "LTE; Evolved universal terrestrial radio sccess (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B," Release 9, May 2011.

Hongliang Zhang (S'15) received the B. S. degree in Electronic Engineering from Peking University, Beijing, China, in 2014. He is currently pursuing his PhD's degree at School of Electrical Engineering and Computer Science in Peking University.

His current research interest includes wireless communications, hypergraph theory, and optimization theory.

Lingyang Song (S'03-M'06-SM'12) received his Ph.D. from the the University of York, United Kingdom, in 2007, where he received the K. M. Stott Prize for excellent research. He worked as a research fellow at the University of Oslo, Norway, until rejoining Philips Research UK in March 2008. In May 2009, he joined the School of Electronics Engineering and Computer Science, Peking University, and is now a Boya Distinguished Professor. His main research interests include wireless communication and networks, signal processing, and machine learning. He was the recipient of the IEEE Leonard G. Abraham Prize in 2016 and the IEEE Asia Pacific (AP) Young Researcher Award in 2012. He has been an IEEE Distinguished Lecturer since 2015.



Ying-Jun Angela Zhang (S'00-M'05-SM'10) received her PhD degree in Electrical and Electronic Engineering from the Hong Kong University of Science and Technology, Hong Kong in 2004. Since 2005, she has been with Department of Information Engineering, The Chinese University of Hong Kong, where she is currently an Associate Professor. Her research interests include mainly wireless communications systems and smart power systems, in particular optimization techniques for such systems.

She is an Executive Editor of the IEEE Transactions on Wireless Communications. She is also an Associate Editor of the IEEE Transactions on Communications. Previously, she served many years as an Associate Editor of the IEEE Transactions on Wireless Communications, Security and Communications Networks (Wiley), and a Feature Topic in the IEEE Communications Magazine. She has served on the organizing committee of major IEEE conferences including ICC, GLOBECOM, SmartgridComm, VTC, CCNC, ICCC, MASS, etc.. She is now the Chair of IEEE ComSoc Emerging Technical Committee on Smart Grid. She was a Co-Chair of the IEEE ComSoc Multimedia Communications Technical Committee and the IEEE Communication Society GOLD Coordinator.

She was the co-recipient of the 2014 IEEE ComSoc APB Outstanding Paper Award, the 2013 IEEE SmartgridComm Best Paper Award, and the 2011 IEEE Marconi Prize Paper Award on Wireless Communications. She was the recipient of the Young Researcher Award from the Chinese University of Hong Kong in 2011. As the only winner from engineering science, she has won the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science. Dr. Zhang is a Fellow of IET and a Distinguished Lecturer of IEEE ComSoc.