Device-to-Device Load Balancing for Cellular Networks

Lei Deng, *Member, IEEE*, Yinghui He, Ying Zhang, Minghua Chen, *Senior Member, IEEE*, Zongpeng Li,
Jack Y. B. Lee, *Senior Member, IEEE*, Ying Jun Zhang, *Senior Member, IEEE*,
and Lingyang Song, *Senior Member, IEEE*

Abstract—Small-cell architecture is widely adopted by cellular network operators to increase spectral spatial efficiency. However, this approach suffers from low spectrum temporal efficiency. When a cell becomes smaller and covers fewer users, its total traffic fluctuates significantly due to insufficient traffic aggregation and exhibits a large "peak-to-mean" ratio. As operators customarily provision spectrum for peak traffic, large traffic temporal fluctuation inevitably leads to low spectrum temporal efficiency. To address this issue, in this paper, we advocate device-to-device (D2D) load-balancing as a useful mechanism. The idea is to shift traffic from a congested cell to its adjacent under-utilized cells by leveraging inter-cell D2D communication, so that the traffic can be served without using extra spectrum, effectively improving the spectrum temporal efficiency. We provide theoretical modeling and analysis to characterize the benefit of D2D load balancing, in terms of total spectrum requirements and the corresponding cost, in terms of incurred D2D traffic overhead. We carry out empirical evaluations based on real-world 4G data traces and show that D2D load balancing can reduce the spectrum requirement by 25% as compared to the standard scenario without D2D load balancing, at the expense of negligible 0.7% D2D traffic overhead.

Index Terms—Cellular networks, small-cell architecture, D2D communication, load balancing.

L. Deng is with the School of Electrical Engineering and Intelligentization, Dongguan University of Technology, Dongguan 523808, China (e-mail: denglei@dgut.edu.cn).

Y. He is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: 2014hyh@zju.edu.cn).

Y. Zhang, M. Chen, J. Y. B. Lee, and Y. J. Zhang are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: ying.ie.cuhk@gmail.com; minghua@ie.cuhk.edu.hk; jacklee@computer.org; yjzhang@ie.cuhk.edu.hk).

Z. Li is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: zongpeng@whu.edu.cn).

L. Song is with the School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: lingyang.song@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2018.2890225

I. INTRODUCTION

THE drastic growth in mobile devices and applications has triggered an explosion in cellular data traffic. According to Cisco [2], global cellular data traffic reached 7 exabytes per month in 2016 and will further witness a 7-fold increase in 2016-2021. Meanwhile, radio frequency remains a scarce resource for cellular communication. Supporting the fast-growing data traffic demands has become a central concern of cellular network operators.

There are mainly two lines of efforts to address this concern. The first is to serve cellular traffic by exploring additional spectrum, including offloading cellular traffic to WiFi [3] and the recent 60GHz millimeter-wave communication endeavor [4]. The second is to improve *spectrum spatial efficiency*. A common approach is to adopt a small-cell architecture, such as micro/pico-cell [5]. By reducing cell size, operators can pack more (low-power) base stations in an area and reuse radio frequencies more efficiently to increase network capacity.

While the small-cell architecture improves the *spectrum spatial efficiency*, it comes at a price of degrading the *spectrum temporal efficiency*. When a cell becomes smaller and covers fewer users, there is less traffic aggregation. Consequently, the total traffic of a cell fluctuates significantly, exhibiting a large "peak-to-mean" ratio. As operators customarily provision spectrum to a cell based on peak traffic, high temporal fluctuation in traffic volumes inevitably leads to low spectrum temporal efficiency.

To see this concretely, we carry out a case-study based on 4G cell-traffic traces from Smartone [6] (this complements the study in our conference version [1], which was based on 3G data traces), a major cellular network operator in Hong Kong, a highly-populated metropolis. The detailed analysis and description can be found in our technical report [7]. Based on this case study, we observe that the average cell-capacity utilization is very now and the peak traffic of many pairs of adjacent BSs occurs at different time epochs. This confirms that small-cell architecture indeed causes very low spectrum temporal utilization, and it suggests ample room to do traffic load balancing to improve temporal utilization.

Motivated by the above observations, we advocate *device-to-device (D2D) load-balancing* as a useful mechanism to improve spectrum temporal efficiency. D2D communication [8], [9] is a promising paradigm for improving system

performance in next generation cellular networks that enables direct communication between user devices using cellular frequency. It is conceivable to relay traffic from congested cells to adjacent underutilized cells via inter-cell D2D communication, enabling load-balancing across cells at the expense of incurred inter-cell D2D traffic.

We remark that an idea of this kind was also studied by Liu *et al.* in their recent work [10]. They focus on important aspects of examining the technical feasibility of D2D load balancing and practical algorithm design in three-tier LTE-Advanced networks. This work is complement to their study and focuses on the following two important questions:

- How much spectrum reduction can D2D load balancing bring to a cellular network?
- What is the corresponding D2D traffic overhead for achieving the benefit?

Answers to these questions provide fundamental understanding of the viability of D2D load balancing in cellular networks. In this paper, we answer the questions via both theoretical analysis and empirical evaluations based on real-world traces. We make the following contributions.

▷ In Sec. III, using perhaps the simplest possible example, we illustrate the concept of D2D load balancing and show that it can reduce peak traffic for two adjacent cells by 33%. We also compute the associated D2D traffic overhead.

▷ For general settings beyond the example, we provide tractable models to analyze the performance of D2D load balancing in Sec. IV. We also exploit the optimal solutions without and with D2D load balancing in Sec. V and Sec. VI, respectively.

▷ Theoretically, for arbitrary settings, we derive an upper bound for the benefit of D2D load balancing, in terms of sum peak traffic reduction in Sec. VII-B. We show that the bound is asymptotically tight for a specified network scenario, where we further derive the corresponding overhead, in terms of incurred D2D traffic. Our bound and analysis reveal the insight behind the effectiveness of D2D load balancing: by aggregating traffic among adjacent cells via inter-cell D2D communication, we can leverage statistical multiplexing gains to better serve the overall traffic without requiring extra network capacity.

▷ Empirically, in Sec. X, we use real-world 4G data traces to verify our theoretical analysis and reveal that D2D load balancing can reduce sum peak traffic of individual cells by 25%, at the cost of 0.7% D2D traffic overhead. This implies significant spectrum saving at a negligible system overhead.

Throughout this paper, we assume that time is slotted into intervals of unit length, and each wireless hop incurs one-slot delay. We focus on uplink communication scenarios, while our analysis is also applicable to the downlink communication. In addition, in the rest of this paper, for any two positive integers K_1, K_2 with $K_1 < K_2$, we use notation $[K_1, K_2]$ to denote set $\{K_1, K_1+1, \cdots, K_2\}$, i.e., $[K_1, K_2] \triangleq \{K_1, K_1 + 1, \cdots, K_2\}$. When $K_1 = 1$, we further simplify notation $[1, K_2]$ to be $[K_2]$, i.e., $[K_2] \triangleq \{1, 2, \cdots, K_2\}$. Due to the page limit, we move some of our technical proofs and some discussions into our technical report [7].

II. RELATED WORK

In this paper, we use a dataset from Smartone to show that the peak traffic of different adjacent BSs occurs at different time epochs. Similar observation is also obtained from the measurement studies in [11] and [12]. Hu *et al.* [11] analyze the 3G cellular traffic of three major cities in China during 2010 and 2013 and a city in a Southeast Asian country in 2013. They show that the correlation coefficient of the traffic profiles of different BSs is small (between 0.16 and 0.33). Xu *et al.* [12] analyze the 3G/4G cellular traffic of 9600 BSs in Shanghai, China in 2014. They show that different areas (residential area, business district, transport, entertainment, and comprehensive area) have different traffic patterns, which have different peak epochs. All these traffic measurements motivate us to do load balancing among different BSs so as to reduce the peak demand (spectrum requirement).

In this paper, we propose the D2D load balancing scheme to reduce the peak demand (spectrum requirement) of BSs. There are other load balancing schemes to achieve the goal, including smart user association [13], [14] and mobile offloading [15].

Smart user association [13], [14] dynamically associates users to the BSs so as to balance the traffic demand of all BSs. However, (i) smart user association schemes normally should be operated on large timescale to overcome the large overhead incurred by frequently switching from one BS to another BS (*a.k.a.*, handover) [13]; thus it is not designed for balancing traffic across BSs on small timescale, and (ii) smart user association scheme in [14], where cellular operators globally associate every user to a BS in a centralized manner, incurs high overhead and complexity. Other smart user association schemes through cell breathing [16] or power control methods, where every user locally connects to the BS with strongest signal in a distributed manner, will change the interference levels significantly and thus they may need for spectral re-allocation across the whole networks. Instead, D2D scheme can do load balancing on short timescale since D2D communications often occur locally within short distances and low power and thus D2D scheme has limited impact to the cellular network. Although D2D load balancing may need to switch between the BS mode (connecting to the BS) and the D2D mode (connecting to the device), such a switch happens locally and it is more lightweight than the global handover between different BSs. Therefore, though D2D load balancing scheme will incur some overhead during D2D communications, it has some unique advantages over smart user association schemes. Meanwhile, we also remark that D2D scheme and smart user association schemes are complementary for load balancing in the sense that we might simultaneously use smart user association schemes on large timescale and use D2D scheme on small timescale. Thus, in this paper we advocate the D2D load balancing scheme.

Mobile offloading [3], [15], [17], [18] is another scheme to reduce the cellular traffic demand. It mainly uses WiFi infrastructure. However, mobile offloading and D2D load balancing are technically different schemes: mobile offloading aims to exploit outband spectrum, but our D2D load balancing

scheme targets to increase inband cellular temporal spectrum efficiency. Furthermore in D2D load balancing, the cellular operation can ubiquitously control everything, including both D2D and user-to-BS transmissions. However, mobile offloading usually outsources a portion of traffic to a thirdparty entity, imposing unpleasant unreliability for transmissions. Therefore, our proposed D2D load balancing scheme can ensure better QoS than mobile offloading. Again, our D2D load balancing scheme are orthogonal to the mobile offloading scheme in the sense that the operators can simultaneously use them to reduce the cellular spectrum requirement.

In addition to those traffic load balancing schemes, spectrum reallocation is another effective approach to reduce the spectrum requirement. Instead of moving traffic among different cells, spectrum reallocation *dynamically* allocate the spectrum among different cells to better match the time-varying traffic demands [19]–[22]. However, spectrum allocation incurs high complexity. The state-of-the-art spectrum allocation solution is proposed in [22], which can obtain near-optimal performance for a network with up to 1000 APs and 2500 active users. Furthermore, spectrum reallocation again is operated on large timescale. Hence, the cellular operator can simultaneously do spectrum reallocation on large timescale based on aggregated traffic information [20] and use our proposed D2D load balancing scheme on small timescale based on the fine-grained traffic information to reduce the spectrum requirement.

We further remark that there are some existing works on D2D load balancing. For the three-tier LTE-Advanced heterogenous networks, [10] examines the technical feasibility and designs practical algorithm for D2D load balancing; [23]–[25] propose research allocation strategies to achieve load balancing goal via D2D transmission. In [26], an auction-based mechanism is proposed to incentivize the mobile users to participate in D2D load balancing. However, all existing works do not directly answer the two important questions proposed in Sec. I.

III. AN ILLUSTRATING EXAMPLE

We consider a simple scenario shown in Fig. 1(a), where 4 users are each aiming at transmitting 3 packets to two base stations (BS) subject to a deadline constraint. We compare the peak traffic of both BSs for the case without D2D load balancing (Fig. 1(b)) and for the case with D2D load balancing (Fig. 1(c)). We illustrate the concept of D2D load balancing and show that it can reduce the peak traffic for two adjacent cells by 33%.

Specifically, we consider a cellular network of two adjacent cells served by BS α and BS β, and four users a, b, c, d. BS α (resp. β) can directly communicate with only users a and b (resp. users c and d). BS α and BS β use orthogonal frequency bands. Due to proximity, users b and c can communicate with each other using frequency band of either BS α or β, creating inter-cell D2D links. Both user a and user b generate 3 packets at the beginning of slot 1, and both user c and user d generate 3 packets at the beginning of slot 3. All packets have the same size and a delay constraint of 2 slots, i.e., a packet must reach BS α or β within 2 slots from its generation time. *We assume that a packet is successfully delivered as long as it reaches*

any BS, since BSs today are connected by a high-speed optical backbone, supported by power clusters, and can coordinate to jointly process/forward packets for users.

In the conventional approach without D2D load balancing, a user only communicates with its own BS. It is straightforward to verify that the minimum peak traffic of both BS α and BS β is 3 (unit: packets), and can be achieved by the scheme in Fig. 1(b). For instance, the minimum peak traffic for BS α is achieved by user a (resp. user b) transmitting all its 3 packets to BS α in slot 1 (resp. slot 2).

With D2D load balancing, we can exploit the inter-cell D2D links between users b and c to perform load balancing and reduce the peak traffic for both BS α and BS β.

- In slot 1, user a transmits two packets a_1 and a_2 to BS α, and user b transmits two packets b_1 and b_2 to user c using the orthogonal frequency band of BS β. The traffic is 2 for both cells. In slot 2, users a and b transmit their remaining packets a_3 and b_3 to BS α, and user c relays the two packets it received in slot 1, i.e., b_1 and b_2, to BS β. The traffic is again 2 for both cells. By the end of slot 2, we deliver 6 packets for users a and b to BSs.
- In slots 3 and 4, note that users c and d have the same traffic pattern as users a and b, but offset by 2 slots. Thus we can also deliver 3 packets for both users c and d in two slots. The traffic of both BSs is 2 per slot.

Overall, with D2D load balancing, we can serve all traffic demands with peak traffic of 2 for both BSs, which is 33% reduced as compared to the case without D2D load balancing.

The intuition behind this example is that the peak traffic for the two cells occurs at different time instances. When users a and b transmit data to BS α in the first two slots, BS β is idle. Meanwhile, BS α is idle when users c and d transmit data to BS β in the last two slots. Therefore, D2D communication can help load balance traffic from the busy BS to the other idle BS, reducing the peak traffic for both BSs. However, D2D load balancing also comes with cost, since it requires transmissions over the inter-cell D2D links. In the example, the total traffic is $8 \times 2 = 16$ packets and the D2D traffic is $2 \times 2 = 4$ packets, yielding an overhead traffic ratio of $\frac{4}{16} = 25\%$. Such D2D traffic is the overhead that we pay in return for peak traffic reduction.

IV. SYSTEM MODEL

In this section, we present the system model for a general network topology and a general traffic demand model beyond the simple example expounded in the previous section. Such models will be used to analyze the benefit of D2D load balancing in general settings, in terms of spectrum reduction ratio, and the cost in terms of D2D traffic overhead ratio.

A. Cellular Network Topology

Consider an uplink wireless cellular network with multiple cells and multiple mobile users. We assume that each cell has one BS and each user is associated with one BS.[1] Define $\mathcal{B}$ as

---

[1] We say that user u is associated with BS b if user u is in the cellular cell covered by BS b. When a user is covered by multiple BSs, we assume that this user has been associated with one of them, e.g., the one with the strongest signal-to-noise ratio. In the rest of this paper, we will also use the terminology, cell b, to represent the cell covered by BS b.
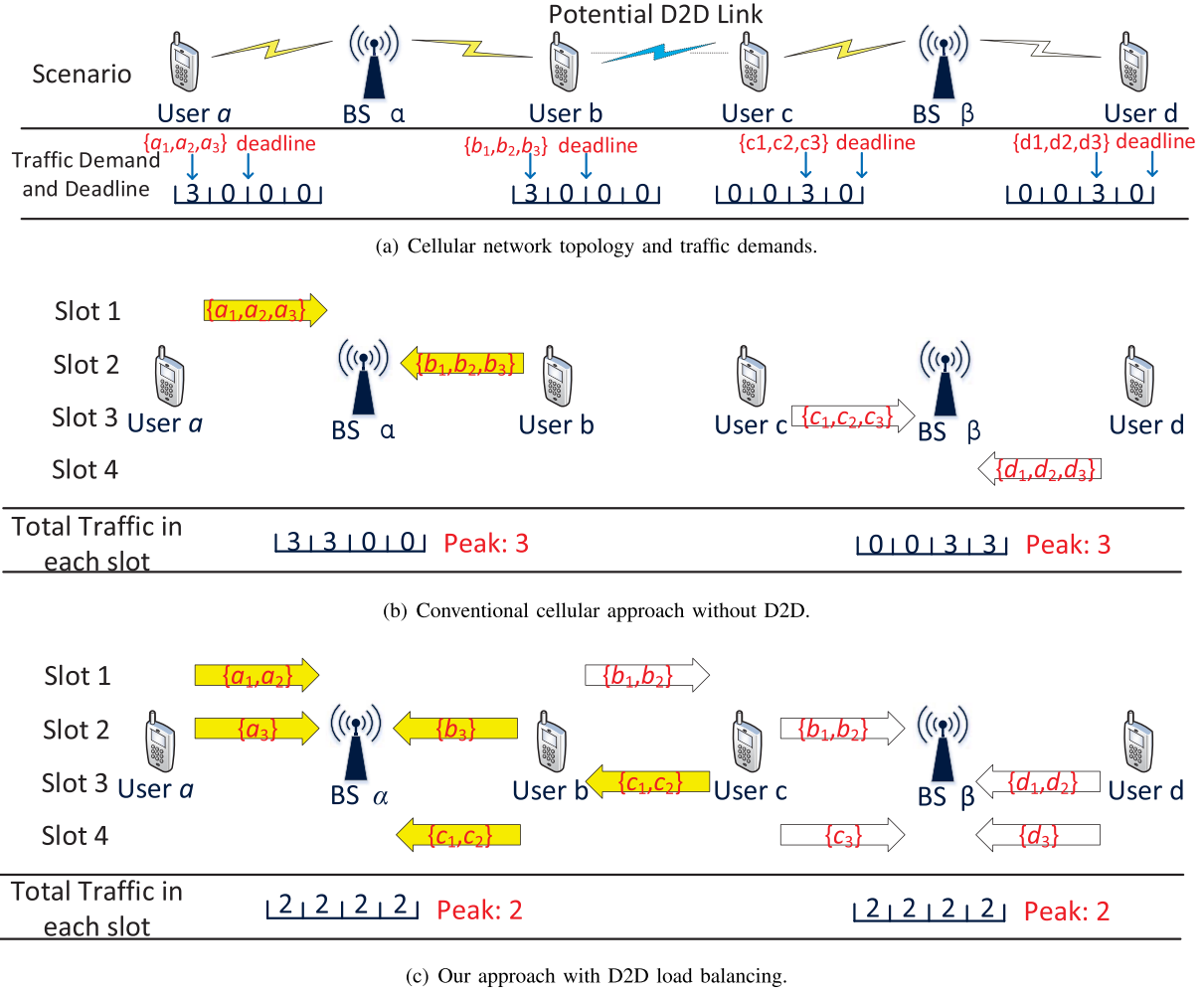
Fig. 1. A simple example for demonstrating the concept of D2D load balancing, and that it can reduce the peak traffic for both cells by 33% (both from 3 to 2) at the cost of 4 extra inter-cell D2D transmissions.

the set of all BSs, $\mathcal{U}_b$ as the set of users belonging to BS $b \in \mathcal{B}$, and $\mathcal{U} = \cup_{b \in \mathcal{B}} \mathcal{U}_b$ as the set of all users in the cellular network. Let $b_u \in \mathcal{B}$ denote the cell (or BS) with which user $u \in \mathcal{U}$ is associated. We model the uplink cellular network topology as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \mathcal{U} \cup \mathcal{B}$ and edge set $\mathcal{E}$ where $(u, v) \in \mathcal{E}$ if there is a wireless link from vertex (user) $u \in \mathcal{U}$ to vertex (BS or user) $v \in \mathcal{V}$.

B. Traffic Model

We consider a time-slotted system with T slots in total, indexed from 1 to T. Each user can generate a delay-constrained traffic demand at the beginning of any slot. We denote $\mathcal{J}$ as the demand set. Each demand $j \in \mathcal{J}$ is characterized by the tuple (u_j, s_j, e_j, r_j) where

- $u_j \in \mathcal{U}$ is the user that generates demand j;
- $s_j \geq 1$ is the starting time/slot of demand j;
- $e_j \in [s_j, T]$ is the ending time/slot (deadline) of demand j;
- $r_j > 0$ is the volume of demand j with unit of bits.

Namely, demand j is generated by user u_j at the beginning of slot s_j with the volume of r_j bits and it must be delivered to BSs before/on the end of slot e_j, implying a *delay* requirement $(e_j - s_j + 1)$. We also call interval $[s_j, e_j]$ the *lifetime* of the

demand j. We further denote $\mathcal{J}_b$ as the set of demands that are generated by the users in BS $b \in \mathcal{B}$, i.e., $\mathcal{J}_b \triangleq \{j \in \mathcal{J} : u_j \in \mathcal{U}_b\}$. Demand j is delivered in time if every bit of demand j reaches a BS before/on the end of slot e_j. Note that different bits in demand j could reach different BSs. Thus, every user can transmit a bit either to its own BS directly in a single hop or to another user via the D2D link between them such that the bit can reach another BS in multiple hops.

C. Wireless Channel/Spectrum Model

For each link $(u, v) \in \mathcal{E}$, we denote its link rate as $R_{u,v}$ (units: bits per slot per Hz), which is the number of bits that can be transmitted in one unit (slot) of time resource and with one unit (Hz) of spectrum resource. Then if we allocate $x \in \mathbb{R}^+$ (unit: Hz) spectrum to link (u, v) at slot t, this link can transmit $x \cdot R_{u,v}$ bits of data from node u to node v in slot t. Note that we simplify the channel model by assuming a linear relationship between the allocated spectrum and the transmitted data. This assumption is reasonable for the high-SNR scenario when we use Shannon capacity as the link rate [27]. In addition, we assume that the total spectrum is not divided into uplink spectrum and downlink spectrum. Instead, our scheme allocates spectrum from a spectrum pool

to mobile users for transmitting or receiving data. Thus, in this paper, we do not consider the switching issue between uplink spectrum and downlink spectrum.

D. Performance Metrics

In this paper, we aim at minimizing the total (amount of) spectrum to deliver all demands in $\mathcal{J}$ in time. In particular, we need to obtain the minimum spectrum/frequency to serve all demands in time without D2D (resp. with D2D), denoted by F^{ND} (resp. F^{D2D}). To evaluate the impact of D2D load balancing, we characterize both the benefit and the cost for D2D load balancing. The benefit is in terms of *spectrum reduction ratio*,

$$\rho \triangleq \frac{F^{\mathsf{ND}} - F^{\mathsf{D2D}}}{F^{\mathsf{ND}}} \in [0, 1). \tag{1}$$

The cost is in terms of *(D2D traffic) overhead ratio*,

$$\eta \triangleq \frac{V^{\mathsf{D2D}}}{V^{\mathsf{D2D}} + V^{\mathsf{BS}}} \in [0, 1), \tag{2}$$

where V^{D2D} is the volume of all D2D traffic and V^{BS} is the volume of all traffic directly sent by cellular users to BSs.

The spectrum reduction ratio ρ evaluates how much spectrum we can save if we apply D2D load balancing. The overhead ratio η evaluates the percentage of D2D traffic among all traffic. D2D traffic incurs cost in the sense that any traffic going through D2D links will consume spectrum and energy of user devices but do not immediately reach any BS. Overall, the spectrum reduction ratio ρ captures the benefit of D2D load balancing and hence larger ρ means larger benefit; the overhead ratio η captures the cost of D2D load balancing and hence smaller η means smaller cost. In the following, we will discuss how to obtain F^{ND} in Sec. V and F^{D2D} in Sec. VI. Then we will show the theoretical upper bounds for ρ and η in Sec. VII.

V. OPTIMAL SOLUTION WITHOUT D2D

In this section, we describe how to compute the minimum spectrum without D2D, i.e., F^{ND}. Since there are no D2D links, we can calculate the required minimum spectrum for each BS separately. Let us denote F_b^{ND} as the minimum spectrum of BS b to deliver all its own traffic demands, i.e., $\mathcal{J}_b$. Then the total minimum spectrum without D2D is[2] $F^{\mathsf{ND}} = \sum_{b \in \mathcal{B}} F_b^{\mathsf{ND}}$.

A. Problem Formulation

For each BS $b \in \mathcal{B}$, we formulate the problem of minimizing the spectrum to deliver all demands in cell b without D2D, named as Min-Spectrum-ND$_b$,

$$\min_{x_{u_j,b}^j(t), \gamma_b(t), F_b \in \mathbb{R}^+} F_b \tag{3a}$$

$$\text{s.t.} \sum_{t=s_j}^{e_j} x_{u_j,b}^j(t) R_{u_j,b} = r_j, \quad \forall j \in \mathcal{J}_b \tag{3b}$$

$$\sum_{j \in \mathcal{J}_b : t \in [s_j, e_j]} x_{u_j,b}^j(t) = \gamma_b(t), \quad \forall t \in [T] \tag{3c}$$

$$\gamma_b(t) \le F_b, \quad \forall t \in [T] \tag{3d}$$

$$x_{u_j,b}^j(t) \ge 0, \quad \forall j \in \mathcal{J}_b, \ t \in [s_j, e_j] \tag{3e}$$

where $x_{u_j,b}^j(t)$ is the allocated spectrum (unit: Hz) for transmitting demand j from user u_j to BS b at slot t, the auxiliary variable $\gamma_b(t)$ is the total used spectrum from users to BS b at slot t, and F_b is the allocated (peak) spectrum to BS b,

Our objective is to minimize the total allocated spectrum of BS b, as shown in (3a). Without D2D, users can only be served by its own BS. Equation (3b) shows the volume requirement for any traffic demand j, i.e., the total traffic volume r_j needs to be delivered from user u_j to BS b during its lifetime. Equation (3c) depicts the total needed spectrum of cell b (i.e., $\gamma_b(t)$) in slot t, which is the summation of allocated spectrum for all active jobs in slot t. Inequality (3d) shows that the total needed spectrum of cell b in any slot t cannot exceed the total allocated spectrum of BS b. Finally, inequality (3e) means that the allocated spectrum for a job in any slot is non-negative.

Let us denote $d_{\max} \triangleq \max_{j \in \mathcal{J}}(e_j - s_j + 1)$ as the maximum delay among all demands. Then the number of variables in Min-Spectrum-ND$_b$ is $O(|\mathcal{J}_b| \cdot d_{\max} + T)$ and the number of constraints in Min-Spectrum-ND$_b$ is also $O(|\mathcal{J}_b| \cdot d_{\max} + T)$.

B. Characterizing the Optimal Solution

To solve Min-Spectrum-ND$_b$, we can use standard linear programming (LP) solvers. However, LP solvers cannot exploit the structure of this problem. We next propose a combinatorial algorithm that exploits the problem structure and achieves lower complexity than general LP algorithms.

We note that Min-Spectrum-ND$_b$ resembles a uniprocessor scheduling problem for preemptive tasks with hard deadlines [28]. Indeed, we can attach each task $j \in \mathcal{J}_b$ with an arrival time s_j and a hard deadline e_j and the requested service time $\frac{r_j}{R_{u_j,b}}$. Then for a given amount of allocated spectrum F_b (which resembles the maximum speed of the processor), we can use the earliest-deadline-first (EDF) scheduling algorithm [29] to check its feasibility. Since we can easily get an upper bound for the minimum spectrum, we can use binary search to find the minimum spectrum F_b^{ND}, supported by the EDF feasibility-check subroutine.

More interestingly, we can even get a semi-closed form for F_b^{ND}, inspired by [30, Th. 1]. Specifically, let us define the *intensity* [30] of an interval $I = [z, z']$ to be

$$g_b(I) \triangleq \frac{\sum_{j \in \mathcal{A}_b(I)} \frac{r_j}{R_{u_j,b}}}{z' - z + 1} \tag{4}$$

where $\mathcal{A}_b(I) \triangleq \{j \in \mathcal{J}_b : [s_j, e_j] \subset [z, z']\}$ is the set of all active traffic demands whose lifetime is within the interval $I = [z, z']$. Then we have the following theorem.

Theorem 1: $F_b^{\mathsf{ND}} = \max_{I \subset [T]} g_b(I)$.

Proof: Since the proof of Theorem 1 was omitted in [30] and the theorem is not directly mapped to the minimum spectrum problem, we give a proof in our technical report [7] for completeness. ∎

---

[2]Here for simplicity, we assume that all BSs use orthogonal spectrum. We discuss how to extend our results to the practical case of spectrum reuse in Sec. IX.

Theorem 1 shows that F_b^{ND} is the maximum intensity over all intervals. To obtain the interval with maximum intensity (and hence F_b^{ND}), we adapt the algorithm originally developed for solving the job scheduling problem in [30], which is called YDS algorithm named after the authors, to our spectrum minimization problem. The time complexity of the YDS algorithm is related to the total number of possible intervals. Clearly the optimal interval can only begin from the generation time of a demand and end at the deadline of a demand. So the total number of intervals needed to be checked is $O(|\mathcal{J}_b|^2)$. Thus the time complexity of our adaptive YDS algorithm is $O(|\mathcal{J}_b|^2)$ [30]. But the complexity of general LP algorithms is $O((|\mathcal{J}_b| \cdot d_{\max} + T)^4 L)$ where L is a parameter determined by the coefficients of the LP [31]. Thus, our combinatorial algorithm has much lower complexity than general LP algorithms.

VI. Optimal Solution With D2D

In this section, we formulate the optimization problem to compute the minimum sum spectrum F^{D2D} when D2D communication is enabled. In this case, since the traffic can be directed to other BSs via inter-cell D2D links, all BSs are coupled with each other and need to be considered as a whole. We will first define the traffic scheduling policy with D2D and then formulate the problem as an LP.

A. Traffic Scheduling Policy

Given traffic demand set $\mathcal{J}$, we need to find a routing policy to forward each packet to BSs before the deadline, which is the *traffic scheduling problem*. Since we should consider the traffic flow in each slot, we will use the *time-expanded graph* to model the traffic flow over time [32]. Specifically, denote $x_{u,v}^j(t)$ as the allocated spectrum (unit: Hz) for link (u, v) at slot t for demand $j \in \mathcal{J}$. Then the delivered traffic volume from node u to node v at slot t for demand j is $x_{u,v}^j(t)R_{u,v}$. For ease of formulation, we set the self-link rate to be $R_{u,u} = 1$. Then the self-link traffic i.e., $x_{u,u}^j(t)R_{u,u} = x_{u,u}^j(t)$, is the traffic volume stored in node u at slot t for demand j. But the allocated (virtual) spectrum for self-link traffic, i.e., $x_{u,u}^j(t)$, will not contribute to the spectrum requirements of BSs (see (6c) later). All traffic flows over time are precisely captured by the time-expanded graph and $x_{u,v}^j(t)$. Then we define the *traffic scheduling policy* as follows.

Definition 1: A traffic scheduling policy is the set $\{x_{u,v}^j(t) : (u, v) \in \mathcal{E}, j \in \mathcal{J}, t \in [s_j, e_j]\} \cup \{x_{u,u}^j(t) : u \in \mathcal{V}, j \in \mathcal{J}, t \in [s_j, e_j]\}$ such that

$$\sum_{v \in \text{out}(u_j)} x_{u_j,v}^j(s_j)R_{u_j,v} = r_j, \quad \forall j \in \mathcal{J} \tag{5a}$$

$$\sum_{b \in \mathcal{B}} \sum_{v \in \text{in}(b)} x_{v,b}^j(e_j)R_{v,b} = r_j, \quad \forall j \in \mathcal{J} \tag{5b}$$

$$\sum_{v \in \text{in}(u)} x_{v,u}^j(t)R_{v,u} = \sum_{v \in \text{out}(u)} x_{u,v}^j(t+1)R_{u,v},$$
$$\forall j \in \mathcal{J}, \ u \in \mathcal{V}, \ t \in [s_j, e_j - 1] \tag{5c}$$

$$x_{u,v}^j(t) \geq 0, \quad \forall(u, v) \in \mathcal{E}, \ j \in \mathcal{J},$$
$$t \in [s_j, e_j] \tag{5d}$$

$$x_{u,u}^j(t) \geq 0, \quad \forall u \in \mathcal{V}, \ j \in \mathcal{J}, \ t \in [s_j, e_j] \tag{5e}$$

where $\text{in}(u) = \{v : (v, u) \in \mathcal{E}\} \cup \{u\}$ and $\text{out}(u) = \{v : (u, v) \in \mathcal{E}\} \cup \{u\}$ are the incoming neighbors and outgoing neighbors of node $u \in \mathcal{V}$ in the time-expanded graph.

Constraint (5a) shows the flow balance in the source node while (5b) shows the flow balance in the destination nodes such that all traffic can reach BSs before their deadlines. Equality (5c) is the flow conservation constraint for each intermediate node in the time-expanded graph. Here we assume that all BSs and all users have enough radios such that they can simultaneously transmit data to and receive data from multiple BSs (or users). This is a strong assumption for mobile users because current mobile devices are not equipped with enough radios. However, multi-radio mobile devices could be a trend and there are substantial research work in multi-radio wireless systems (see a survey in [33] and the references therein). We made this assumption here because *wireless scheduling problem* for single-radio users is generally intractable and we want to avoid detracting our attention and focus on how to characterize the benefit of D2D load balancing and get a first-order understanding. We remark that this assumption is also made in recent work [22] on spectrum reallocation in small-cell cellular networks.

B. Problem Formulation

Then we formulate the problem of computing the minimum total spectrum to serve all demands in all cells with D2D, named as Min-Spectrum-D2D,

$$\min_{x_{u,v}^j(t),\alpha_b(t),\beta_b(t),F_b \in \mathbb{R}^+} \sum_{b \in \mathcal{B}} F_b \tag{6a}$$

$$\text{s.t. (5a), (5b), (5c), (5d), (5e)}$$

$$\sum_{v \in \mathcal{U}_b} \sum_{j \in \mathcal{J}:t \in [s_j, e_j]} x_{v,b}^j(t) = \alpha_b(t),$$
$$\forall b \in \mathcal{B}, \ t \in [T] \tag{6b}$$

$$\sum_{u \in \mathcal{U}_b} \sum_{v \in \text{in}(u) \setminus \{u\}} \sum_{j \in \mathcal{J}:t \in [s_j, e_j]} x_{v,u}^j(t) = \beta_b(t),$$
$$\forall b \in \mathcal{B}, \ t \in [T] \tag{6c}$$

$$\alpha_b(t) + \beta_b(t) \leq F_b, \quad \forall b \in \mathcal{B}, \ t \in [T] \tag{6d}$$

where the auxiliary variable $\alpha_b(t)$ is the total used spectrum from users to BS b at slot t, the auxiliary variable $\beta_b(t)$ is the total used spectrum dedicated to all users in BS b at slot t, and F_b is the allocated (peak) spectrum for BS b. Note that in our case with D2D load balancing, a user can adopt *the D2D mode* to transmit to another user via a D2D link (e.g., $\sum_{j \in \mathcal{J}:t \in [s_j, e_j]} x_{v,u}^j(t)$ is the allocated spectrum to the D2D link from user v to user u in slot t) and/or *the cellular mode* to transmit to its BS via a user-to-BS link (e.g., $\sum_{j \in \mathcal{J}:t \in [s_j, e_j]} x_{v,b}^j(t)$ is the allocated spectrum to the user-to-BS link from user v to BS b in slot t). In addition, note that we assume a *receiver-takeover* scheme in the sense that any traffic will consume spectrum resources of the receiver's

BS. Equalities (6b) and (6c) show that BS b is responsible for all traffic dedicated to itself and to its users except self-link (virtual) spectrum (see Sec. VI-A). We also remark that although spectrum sharing is one of the major benefits of D2D communication, in this work we do not model the spectrum sharing among D2D links and user-to-BS links to simplify the analysis. Later in Sec. X, we show that our D2D load balancing scheme can significantly reduce the spectrum requirement even without doing spectrum sharing among D2D links and user-to-BS links. If we further do spectrum sharing, the D2D load balancing has more gains.

Given an optimal solution to Min-Spectrum-D2D, we denote F_b^{D2D} as the allocated spectrum for each BS b, and thus the total spectrum is $F^{\mathsf{D2D}} = \sum_{b \in \mathcal{B}} F_b^{\mathsf{D2D}}$. The total D2D traffic and total user-to-BS traffic are

$$V^{\mathsf{D2D}} = \sum_{t=1}^{T} \sum_{j \in \mathcal{J}: t \in [s_j, e_j - 1]} \sum_{u \in \mathcal{U}} \sum_{v: v \in \mathcal{U}, (u,v) \in \mathcal{E}} x_{u,v}^j(t) R_{u,v},$$
(7)

$$V^{\mathsf{BS}} = \sum_{t=1}^{T} \sum_{j \in \mathcal{J}: t \in [s_j, e_j]} \sum_{b \in \mathcal{B}} \sum_{u \in \mathcal{U}_b} x_{u,b}^j(t) R_{u,b},$$
(8)

which are used to calculate the overhead ratio η in (2). We further remark that since all traffic demands must reach any BSs, it is easy to see that the user-to-BS traffic is exactly the total volume of all traffic demands, i.e., $V^{\mathsf{BS}} = \sum_{j \in \mathcal{J}} r_j$.

Given the optimal (minimum) total spectrum, i.e., F^{D2D}, we next minimize the overhead, named Min-Overhead, by solving the following LP,[3]

$$\min_{\substack{x_{u,v}^j(t), \alpha_b(t), \\ \beta_b(t), F_b \in \mathbb{R}^+}} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}: t \in [s_j, e_j - 1]} \sum_{u \in \mathcal{U}} \sum_{\substack{v: v \in \mathcal{U}, \\ (u,v) \in \mathcal{E}}} x_{u,v}^j(t) R_{u,v}$$
(9a)

s.t. (5a), (5b), (5c), (5d), (5e)

$$\sum_{v \in \mathcal{U}_b} \sum_{j \in \mathcal{J}: t \in [s_j, e_j]} x_{v,b}^j(t) = \alpha_b(t), \quad \forall b \in \mathcal{B}, t \in [T]$$
(9b)

$$\sum_{u \in \mathcal{U}_b} \sum_{v \in \text{in}(u) \setminus \{u\}} \sum_{j \in \mathcal{J}: t \in [s_j, e_j]} x_{v,u}^j(t) = \beta_b(t),$$

$$\forall b \in \mathcal{B}, \quad t \in [T]$$
(9c)

$$\alpha_b(t) + \beta_b(t) \le F_b, \quad \forall b \in \mathcal{B}, \ t \in [T]$$
(9d)

$$\sum_{b \in \mathcal{B}} F_b \le F^{\mathsf{D2D}}$$
(9e)

As compared to Min-Spectrum-D2D in (6), Min-Overhead in (9) adds a constraint (9e) for the *given* total spectrum F^{D2D} and changes the objective to be the total D2D traffic defined in (7). Note that even though we write (9e) as an inequality, it must hold as an equality. This is because F^{D2D} is the optimal value of Min-Spectrum-D2D in (6) and any solution in Min-Overhead in (9) is also feasible to Min-Spectrum-D2D in (6).

---

[3]In other words, minimizing the total spectrum is our first-priority objective and minimizing the corresponding D2D traffic overhead (without exceeding the minimum total spectrum) is our second-priority objective.

The number of variables in Min-Spectrum-D2D is $O(|\mathcal{J}| \cdot |\mathcal{E}| \cdot d_{\max} + |\mathcal{B}| \cdot T)$ and the number of constraints in Min-Spectrum-D2D is $O(|\mathcal{J}| \cdot (|\mathcal{V}| + |\mathcal{E}|) \cdot d_{\max} + |\mathcal{B}| \cdot T)$. The problem Min-Overhead has the same complexity as Min-Spectrum-D2D. Solving the problem, even though it is an LP, incurs high complexity. We further discuss how to reduce the complexity without loss of optimality in our technical report [7]. Even with our optimized LP approach, later in our simulation in Sec. X, we show that we cannot solve Min-Spectrum-D2D for practical Smartone network with off-the-shell servers. Thus, we further propose a heuristic algorithm to solve Min-Spectrum-D2D with much lower complexity in Sec. VIII. We also provide performance guarantee for our heuristic algorithm. Before that, we show our theoretical results on the spectrum reduction ratio and the overhead ratio in next section.

VII. THEORETICAL RESULTS

From the two preceding sections, we can compute F^{ND} with the (adaptive) YDS algorithm (Theorem 1) and F^{D2D} by solving the large-scale LP problem Min-Spectrum-D2D (Sec. VI-B). Hence, numerically we can get the spectrum reduction and the overhead ratio. In this section, however, we seek to derive theoretical upper bounds on both spectrum reduction and overhead ratio. Such theoretical upper bounds provide insights for the key factors to achieve large spectrum reduction and thus provide guidance to determine whether it is worthwhile to implement D2D load balancing scheme in real-world cellular systems.

A. A Simple Upper Bound for Spectrum Reduction

We can get a simple upper bound for F^{D2D} by assuming no cost for D2D communication in the sense that any D2D communication will not consume bandwidth and will not incur delays. Then we can construct a virtual grand BS and all users $\mathcal{U}$ are in this BS. Then the system becomes similar to the case without D2D. We can apply the YDS algorithm to compute the minimum peak traffic, which is a lower bound for F^{D2D}, i.e., $\underline{F}^{\mathsf{D2D}} = \max_{I \subset [T]} g(I)$, where

$$g(I) = \frac{\sum_{j \in \mathcal{A}(I)} \frac{r_j}{R_{\max}}}{z' - z + 1}.$$
(10)

Here in (10), $\mathcal{A}(I) = \{j \in \mathcal{J} : [s_j, e_j] \subset [z, z']\}$ is the set of all active traffic demands whose lifetime is within the interval $I = [z, z']$ and $R_{\max} = \max_{s \in \mathcal{U}} R_{s,b_s}$ is the best user-to-BS link. Then we have the following theorem.

Theorem 2: $\rho \le \frac{F^{\mathsf{ND}} - \underline{F}^{\mathsf{D2D}}}{F^{\mathsf{ND}}}$.

Proof: Please see our technical report [7]. ∎

Note that both $\underline{F}^{\mathsf{D2D}}$ and F^{ND} can be computed by the YDS algorithm, much easier than solving the large-scale LP Min-Spectrum-D2D. Therefore, numerically we can get a quick understanding of the maximum benefit that can be achieved by D2D load balancing.

B. A General Upper Bound for Spectrum Reduction

We next describe another general upper bound for any arbitrary topology and any arbitrary traffic demand set. We will begin with some preliminary notations.

TABLE I

DISCREPANCY NOTATIONS

$$r_s = \max_{v:(s,v)\in\mathcal{E},v\in\mathcal{U}_{b_s}} \frac{R_{s,v}}{R_{s,b_s}}, \forall s \in \mathcal{U}$$

$$\tilde{r}_s^b = \max_{v:(s,v)\in\mathcal{E},v\in\mathcal{U}_b} \frac{R_{s,v}}{R_{s,b_s}}, \forall s \in \mathcal{U}, b \in \mathcal{B}$$

$$r_b = \max_{s\in\mathcal{U}_b} r_s, \forall b \in \mathcal{B}$$

$$\tilde{r}_{b,b'} = \max_{s\in\mathcal{U}_b} \tilde{r}_s^{b'}, \forall b \in \mathcal{B}, b' \in \mathcal{B}$$

$$r = \max_{b\in\mathcal{B}} r_b, \tilde{r} = \max_{(b,b')\in\mathcal{E}^{\text{D2D}}} \tilde{r}_{b,b'}$$

We first define some preliminary notations. Let $N = |\mathcal{B}|$ be the number of BSs and we define a directed *D2D communication graph* $\mathcal{G}^{\text{D2D}} = (\mathcal{B}, \mathcal{E}^{\text{D2D}})$ where the vertex set is the BS set $\mathcal{B}$ and $(b, b') \in \mathcal{E}^{\text{D2D}}$ if there exists at least one inter-cell D2D link from user $u \in \mathcal{U}_b$ in BS $b \in \mathcal{B}$ to user $v \in \mathcal{U}_{b'}$ in BS $b' \in \mathcal{B}$. Denote δ_b^- as the in-degree of BS b in the graph $\mathcal{G}^{\text{D2D}}$ and define the maximum in-degree of the graph $\mathcal{G}^{\text{D2D}}$ as $\Delta^- = \max_{b\in\mathcal{B}} \delta_b^-$. In addition, we define some notations in Tab. I to capture the discrepancy of D2D links and non-D2D links for users and BSs. Note that these definitions will be used thoroughly in our technical report [7] to prove Theorem 3.

Now we have the following theorem.

Theorem 3: For an arbitrary network topology $\mathcal{G}$ associated with a D2D communication graph $\mathcal{G}^{\text{D2D}} = (\mathcal{B}, \mathcal{E}^{\text{D2D}})$ and an arbitrary traffic demand set, the spectrum reduction is upper bounded by

$$\rho \leq \frac{\max\{r, 1\} + \tilde{r}\Delta^- - 1}{\max\{r, 1\} + \tilde{r}\Delta^-}. \tag{11}$$

Proof: Please see Appendix A. ∎

Based on this upper bound, we observe that the benefit of D2D load balancing comes from two parts: intra-cell D2D and inter-cell D2D. More interestingly, we can obtain the individual benefit of intra-cell D2D and inter-cell D2D separately, as shown in the following Corollaries 1 and 2. One can go through the proof for Theorem 3 by disabling inter-cell or intra-cell D2D communication and get the proof of these two corollaries.

Corollary 1: If only intra-cell D2D communication is enabled, the spectrum reduction is upper bounded by

$$\rho \leq \frac{\max\{r, 1\} - 1}{\max\{r, 1\}}. \tag{12}$$

This upper bound is quite intuitive. When $r \leq 1$, then for any user s, there does not exist any intra-cell D2D link with better link quality than its direct link to BS b_s. Therefore, using the user-to-BS link is always the optimal choice. Thus the spectrum reduction is 0. When $r > 1$, larger r means more advantages for intra-cell D2D links over the user-to-BS links. Therefore, D2D can exploit more benefit.

Moreover, this upper bound can be achieved by the simple example in Fig. 2. Suppose that user a generates one traffic demand with volume V and delay $D \geq 2$ at slot 1. Suppose link rates $R_1 = 1, R_2 = r, R_3 = (D-1)r$. Then without intra-cell D2D, the (peak) spectrum requirement is $F_1 = \frac{V}{D}$. With intra-cell D2D, user a transmits $\frac{V}{D-1}$ traffic to user b from slot 1 to slot $D-1$ and then user b transmits all V traffic to BS at slot D. The (peak) spectrum requirement
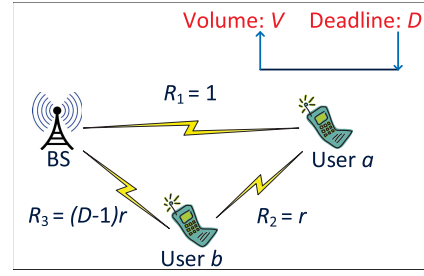

Fig. 2. The benefit of intra-cell D2D communications.

is $F_2 = \max\{\frac{V}{(D-1)R_2}, \frac{V}{R_3}\} = \frac{V}{(D-1)r}$. Then the spectrum reduction is

$$\frac{F_1 - F_2}{F_1} = 1 - \frac{\frac{V}{(D-1)r}}{\frac{V}{D}} \rightarrow \frac{r-1}{r}, \quad \text{as } D \rightarrow \infty. \tag{13}$$

The benefit of intra-cell D2D communication is widely studied (see [8], [9]). However, in this paper, we mainly focus on the benefit of inter-cell D2D load balancing. Indeed, in our simulation settings in Sec. X, the intra-cell D2D brings negligible benefit.

Corollary 2: If only inter-cell D2D communication is enabled, the spectrum reduction is upper bounded by $\rho \leq \frac{\tilde{r}\Delta^-}{1+\tilde{r}\Delta^-}$.

The intuition behind the parameter $\tilde{r}$ is similar to the effect of parameter r in the intra-cell D2D case. In what follows, we will only discuss the effect of parameter Δ^-, which actually reveals the insight of our advocated D2D load balancing scheme. Now suppose that all the links have the same quality and *w.l.o.g.* let $R_{u,v} = 1, \forall (u,v) \in \mathcal{E}$. Then $r = \tilde{r} = 1$, meaning that no intra-cell D2D benefit exists. And the benefit of inter-cell D2D is reduced to the following upper bound

$$\rho \leq \frac{\Delta^-}{1+\Delta^-}. \tag{14}$$

The rationale to understand this upper bound is as follows. On a high level of understanding, the main idea for load balancing is traffic aggregation. If each BS can aggregate more traffic from other BSs, it can exploit more statistical multiplexing gains to serve more traffic with the same amount of spectrum. Since the in-degree for each BS indeed measures its capacity of traffic aggregation, it is not surprising that the upper bound for ρ is related to maximum in-degree Δ^-.

To evaluate how good the upper bound in (14) is, two natural questions can be asked. The first is: *Is this upper bound tight?* Another observation is that if we want to achieve unbounded benefit, i.e., $\rho \rightarrow 1$, it is necessary to let $\frac{\Delta^-}{\Delta^-+1} \rightarrow 1$, which means that $\Delta^- \rightarrow \infty$. Then the second question is: *Can ρ indeed approach 100% as $\Delta^- \rightarrow \infty$?*

In the rest of this subsection, we will answer these two questions by constructing a specified network and traffic demand set. Specifically, we consider $N = |\mathcal{B}|$ BSs each serving one user only. To facilitate analysis, let b_i be the i-th BS and u_i be the user in BS i, for all $i \in [N]$. We consider a *singleton-decoupled* traffic demand set as follows. Each user has one and only one traffic demand with the same volume

V and the same delay $D \geq 2$. Let $T = ND$ and the traffic generation time of user i be slot $D(i-1)+1$. Therefore, the lifetime of user u_i's traffic demand is $[D(i-1)+1, Di]$, during which there are no other demands.

Under such settings, we will vary the user-connection pattern such that the D2D communication graph is different. Specifically, we will prove that this upper bound is asymptotically tight in the ring topology for $\Delta^- = 2$ in Fact 1, and $\rho \to 100\%$ in the complete topology as the number of BSs $N \to \infty$ in Fact 2. Moreover, we will also discuss the overhead ratio for these two special topologies.

Fact 1: If $N = 2D-1$ and the D2D communication graph forms a bidirectional ring graph, then there exists a traffic scheduling policy such that the spectrum reduction is

$$\rho = \frac{2(D-1)}{3D-2} \to \frac{2}{3} = \frac{\Delta^-}{\Delta^- + 1}, \quad \text{as } D \to \infty. \quad (15)$$

Besides, the overhead ratio in this case is

$$\eta = \frac{D(D-1)}{D^2 + 2D - 2}. \quad (16)$$

Proof: Please see our technical report [7]. ∎

Fact 2: If the D2D communication graph forms a bidirectional complete graph, then there exists a traffic scheduling policy such that the spectrum reduction is

$$\rho = \frac{N-1}{N+1} \to 100\%, \quad \text{as } N \to \infty. \quad (17)$$

Besides, the overhead ratio in this case is

$$\eta = \frac{N-1}{2N}. \quad (18)$$

Proof: Please see our technical report [7]. ∎

Remark: (i) Fact 1 shows the tightness of the upper bound in (14) for the ring-graph topology when $\Delta^- = 2$. (ii) Fact 2 shows that ρ can indeed approach 100%, implying that in the best case, ρ goes to 100%. This gives us strong motivation to investigate D2D load balancing scheme both theoretically and practically. (iii) For the complete-graph topology, the upper bound $\frac{\Delta^-}{\Delta^- + 1}$ is not tight. Indeed, since $\Delta^- = N-1$ in the complete-graph topology, we have

$$\frac{\Delta^-}{\Delta^- + 1} = \frac{N-1}{N} > \frac{N-1}{N+1}. \quad (19)$$

(iv) Let us revisit the toy example in Fig. 1 which forms a complete-graph topology with $N = 2$. It verifies the spectrum reduction and overhead ratio in Fact 2, i.e., $\rho = \frac{1}{3} = \frac{N-1}{N+1}$ and $\eta = \frac{1}{4} = \frac{N-1}{2N}$. (v) We also highlight the tradeoff between the benefit ρ and the cost η, as illustrated in Fig. 3. Furthermore, Fig. 3 shows that the complete-graph topology outperforms the ring-graph topology asymptotically because $\rho \to \frac{2}{3}$ and $\eta \to 1$ for the ring-graph topology but $\rho \to 1 > \frac{2}{3}$ (larger benefit) and $\eta \to \frac{1}{2} < 1$ (smaller cost) for the complete-graph topology.

C. An Upper Bound for Overhead Ratio

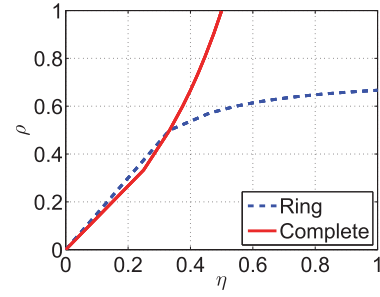Previously we study upper bounds for the spectrum reduction. Now we instead propose an upper bound for overhead

Fig. 3.   Tradeoff between ρ and η.

ratio. Recall that $d_{\max}$ is the maximum demand delay. We then have the following result.

Theorem 4: $\eta \leq \frac{d_{\max}-1}{d_{\max}}$.

Proof: Please our technical report [7]. ∎

The upper bound in Theorem 4 increases when the maximum demand delay $d_{\max}$ increases. This is reasonable because a traffic demand can travel more D2D links (and thus incurs more D2D traffic overhead) if its delay is large. For our toy example in Fig. 1, we have $d_{\max} = 2$ and thus the upper bound for the overhead ratio is $\frac{d_{\max}-1}{d_{\max}} = 50\%$, which is in line with our actual overhead ratio 25%.

VIII. A LOW-COMPLEXITY HEURISTIC ALGORITHM FOR Min-Spectrum-D2D

Our proposed LP formulation for Min-Spectrum-D2D has high complexity due to the size of input traffic demand and cellular network. To reduce the complexity, in this section, we propose a heuristic algorithm which can significantly reduce the number of traffic demands that is needed to be considered. Moreover, our algorithm has a parameter (which is λ defined shortly) such that we can balance the complexity and the performance.

Our proposed algorithm has three steps.

Step I: We solve Min-Spectrum-ND$_b$ for each BS $b \in \mathcal{B}$, and get the optimal solution $\{x_{u_j,b}^j(t), \gamma_b(t), F_b\}$.

Step II: For each BS b with the spectrum profile $\gamma_b(t)$, we consider the following set,

$$T_b(\lambda) \triangleq \{t \in [T] : \gamma_b(t) > \lambda F_b\}, \quad (20)$$

where parameter $\lambda \in [0,1]$ controls the split level. Now we divide all cell-b traffic demands $\mathcal{J}_b$ into two demand sets

$$\mathcal{J}_b^{\text{D2D}}(\lambda) \triangleq \{j \in \mathcal{J}_b : \exists t \in [s_j, e_j] \cap T_b(\lambda) \text{ s.t. } x_{u_j,b}^j(t) > 0\}, \quad (21)$$

and

$$\mathcal{J}_b^{\text{ND}}(\lambda) \triangleq \{j \in \mathcal{J}_b : x_{u_j,b}^j(t) = 0, \forall t \in [s_j, e_j] \cap T_b(\lambda)\}. \quad (22)$$

For all traffic demand in $\mathcal{J}_b^{\text{ND}}(\lambda)$, we schedule them according to $\{x_{u_j,b}^j(t)\}$ without D2D, which results in at most $\gamma_b(t)$ spectrum requirement for BS b at slot t. Note that no demand in $\mathcal{J}_b^{\text{ND}}(\lambda)$ is served in slot set $T_b(\lambda)$. We thud denote $\tilde{\gamma}_b(t)$ as the already allocated spectrum spectrum for demand set $\mathcal{J}_b^{\text{ND}}(\lambda)$ for BS b at slot b, which satisfies $\tilde{\gamma}_b(t) \leq \gamma_b(t)$ when $t \notin T_b(\lambda)$ and $\tilde{\gamma}_b(t) = 0$ when $t \in T_b(\lambda)$.

Step III: We solve the D2D load balancing problem with traffic demands $\mathcal{J}^{\text{D2D}}(\lambda) \triangleq \{\mathcal{J}_b^{\text{D2D}}(\lambda) : b \in \mathcal{B}\}$, according to the following LP, which adapts Min-Spectrum-D2D in (6) by considering the already allocated spectrum $\{\tilde{\gamma}_b(t)\}$,

$$\min_{x_{u,v}^j(t), \alpha_b(t), \beta_b(t), F_b \in \mathbb{R}^+} \sum_{b \in \mathcal{B}} F_b \tag{23a}$$

s.t. (5a), (5b), (5c), (5d), (5e)

$$\sum_{v \in \mathcal{U}_b} \sum_{j \in \mathcal{J}^{\text{D2D}}(\lambda):t\in[s_j,e_j]} x_{v,b}^j(t) = \alpha_b(t),$$

$$\forall b \in \mathcal{B}, \quad t \in [T] \tag{23b}$$

$$\sum_{u \in \mathcal{U}_b} \sum_{v \in \text{in}(u) \setminus \{u\}} \sum_{j \in \mathcal{J}^{\text{D2D}}(\lambda):t\in[s_j,e_j]}$$

$$\times x_{v,u}^j(t) = \beta_b(t), \quad \forall b \in \mathcal{B}, \ t \in [T] \tag{23c}$$

$$\alpha_b(t) + \beta_b(t) + \tilde{\gamma}_b(t) \leq F_b, \quad \forall b \in \mathcal{B},$$

$$t \in [T] \tag{23d}$$

Similar to the overhead minimization problem Min-Overhead in (9), given the optimal spectrum requirement of (23), denoted as, $F^{\text{Heuristic}}(\lambda)$, we next minimize the overhead by solving the following LP,

$$\min_{\substack{x_{u,v}^j(t), \alpha_b(t), \\ \beta_b(t), F_b \in \mathbb{R}^+}} \sum_{t=1}^{T} \sum_{j \in \mathcal{J}^{\text{D2D}}:t\in[s_j,e_j-1]} \sum_{u \in \mathcal{U}} \sum_{\substack{v:v \in \mathcal{U}, \\ (u,v) \in \mathcal{E}}} x_{u,v}^j(t) R_{u,v}$$

$$\tag{24a}$$

s.t. (5a), (5b), (5c), (5d), (5e)

$$\sum_{v \in \mathcal{U}_b} \sum_{j \in \mathcal{J}^{\text{D2D}}(\lambda):t\in[s_j,e_j]} x_{v,b}^j(t) = \alpha_b(t),$$

$$\forall b \in \mathcal{B}, \quad t \in [T] \tag{24b}$$

$$\sum_{u \in \mathcal{U}_b} \sum_{v \in \text{in}(u) \setminus \{u\}} \sum_{j \in \mathcal{J}^{\text{D2D}}(\lambda):t\in[s_j,e_j]} x_{v,u}^j(t) = \beta_b(t),$$

$$\forall b \in \mathcal{B}, \quad t \in [T] \tag{24c}$$

$$\alpha_b(t) + \beta_b(t) + \tilde{\gamma}_b(t) \leq F_b, \quad \forall b \in \mathcal{B}, \ t \in [T] \tag{24d}$$

$$\sum_{b \in \mathcal{B}} F_b \leq F^{\text{Heuristic}}(\lambda) \tag{24e}$$

Note that in (23)/(24), all variables, $x_{u,v}^j(t), \alpha_b(t), \beta_b(t), F_b$, have the same meanings of those in (6)/(9). There are two differences between (23)/(24) and (6)/(9). First, the traffic demand set in (23)/(24) is $\mathcal{J}^{\text{D2D}}(\lambda)$ while that in (6)/(9) is $\mathcal{J}$. Likewise, the traffic scheduling policy characterized by (5a), (5b), (5c), (5d), (5e) in (23)/(24) is for the traffic demand set $\mathcal{J}^{\text{D2D}}(\lambda)$ while that in (6)/(9) is for the traffic demand set $\mathcal{J}$. Second, constraint (23d)/(24d) is different from constraint (6d)/(9d) in that (23d)/(24d) considers the already allocated spectrum $\{\tilde{\gamma}_b(t)\}$. Namely, the spectrum requirement for BS b at slot t includes the already allocated spectrum $\tilde{\gamma}_b(t)$ to serve the traffic demand $\mathcal{J}_b^{\text{ND}}$ and the new allocated spectrum $(\alpha_b(t) + \beta_b(t))$ to serve the traffic demand $\mathcal{J}_b^{\text{D2D}}(\lambda)$.

Obviously, if the number of traffic demand in $\mathcal{J}^{\text{D2D}}(\lambda)$ is much less than the total number of traffic demands in $\mathcal{J}$, which is indeed the case according to our empirical study in Sec. X, we can significantly reduce the number of variables and constraints in (23)/(24) in Step III as compared to the

LP problem Min-Spectrum-D2D/Min-Overhead in (6)/(9). After these three steps, the total spectrum is given by the objective value of (23) and the corresponding overhead is given by the objective value of (24). An example of our heuristic algorithm is shown in our technical report [7].

We denote the spectrum reduction of our heuristic algorithm as

$$\rho^{\text{Heuristic}}(\lambda) \triangleq \frac{F^{\text{ND}} - F^{\text{Heuristic}}(\lambda)}{F^{\text{ND}}}. \tag{25}$$

Similarly, we denote $\eta^{\text{Heuristic}}(\lambda)$ as the overhead ratio of our heuristic algorithm. We next show that the performance guarantee of our heuristic algorithm.

First, for the spectrum we reduction, we have,

Theorem 5: $(1 - \lambda)\rho \leq \rho^{\text{Heuristic}}(\lambda) \leq \rho$.

Proof: Please see Appendix B. ∎

Theorem 5 shows that when $\lambda = 0$, we have $\rho^{\text{Heuristic}}(0) = \rho$. This is because when $\lambda = 0$, we have $\mathcal{J}^{\text{D2D}}(0) = \mathcal{J}$, i.e., all demands participate in D2D load balancing in our heuristic algorithm when $\lambda = 0$ and thus the objective value of (23) when $\lambda = 0$ is exactly F^{D2D}. When $\lambda = 1$, since $\mathcal{J}^{\text{D2D}}(1) = \emptyset$, all traffic demands are served locally without D2D and therefore the objective value of (23) when $\lambda = 1$ is exactly F^{ND}. Thus, the lower bound $(1 - \lambda)\rho = 0$ is tight. Further, the lower bound $(1 - \lambda)\rho$, decreases as λ increases, but the computational complexity decreases as λ increases. Thus, this lower bound illustrates the tradeoff between the performance and the complexity of our heuristic algorithm.

Second, we give an upper bound for the overhead ratio.[4]

Theorem 6: $\eta^{\text{Heuristic}}(\lambda) \leq \dfrac{(d_{\max}-1)\sum\limits_{j \in \mathcal{J}^{\text{D2D}}(\lambda)} r_j}{(d_{\max}-1)\sum\limits_{j \in \mathcal{J}^{\text{D2D}}(\lambda)} r_j + \sum\limits_{j \in \mathcal{J}} r_j} \leq \dfrac{d_{\max}-1}{d_{\max}}.$

Proof: Please see our technical report [7]. ∎

We can see that the upper bound of the overhead ratio is 0 when $\lambda = 1$ because $\mathcal{J}^{\text{D2D}}(1) = \emptyset$, i.e., all traffic demands are served locally without D2D. Moreover, when λ increases, the upper bound decreases because less traffic demands participate in D2D load balancing.

Overall, our heuristic algorithm reduce the complexity of our global LP approach and has performance guarantee. Moreover, our proposed heuristic algorithm has a controllable parameter λ to balance the benefit in terms of spectrum reduction, the cost in terms of overhead ratio, and the computational complexity for our D2D load balancing scheme.

IX. TOWARDS SPECTRUM REDUCTION WITH FREQUENCY REUSE

In this paper, we use the sum spectrum to describe how many resources are needed to serve all users' traffic demands in cellular networks. This may not directly reflect the total required spectrum for cellular operators, because the same spectrum can be spatially reused by multiple BSs who are sufficiently far away from each other. The benefit of spectrum spatial reuse is characterized by the frequency reuse factor K, which represents the proportion of the total spectrum that

---

[4]Recall that $d_{\max}$ is the maximum demand delay.

one cell can utilize. For instance, $K = 1$ means that any cell can use all spectrum, and $K = 1/7$ means that one cell can only utilize 1/7 of the total spectrum, to avoid excessive interference among adjacent cells. A *back-of-the-envelope* calculation suggests that, if the total number of required channels for all N BSs is C, then $\frac{C/N}{K}$ distinct radio channels are needed to serve the entire cellular network.

In the case without D2D, the sum spectrum of all BSs is F^{ND}, which corresponds to the total number of channels for all cells. Thus, with frequency reuse factor K, $\frac{F^{\mathsf{ND}}}{NK}$ distinct channels are needed without D2D.

In the case with D2D, D2D communication can degrade the original frequency reuse pattern if they are sharing the same spectrum with cellular users (which is called underlay D2D [8]). Given the new frequency reuse factor $K^{\mathsf{D2D}} (\leq K)$. A back-of-the-envelope analysis suggests that $\frac{F^{\mathsf{D2D}}}{NK^{\mathsf{D2D}}}$ distinct radio channels are needed with D2D load balancing. Consequently, the spectrum reduction can be estimated as

$$\frac{\frac{F^{\mathsf{ND}}}{NK} - \frac{F^{\mathsf{D2D}}}{NK^{\mathsf{D2D}}}}{\frac{F^{\mathsf{ND}}}{NK}} = 1 - \frac{K}{K^{\mathsf{D2D}}} \times \frac{F^{\mathsf{D2D}}}{F^{\mathsf{ND}}} = 1 - \frac{K}{K^{\mathsf{D2D}}}(1 - \rho). \tag{26}$$

Eq. (26) suggests that our calculation of ρ without frequency reuse gives us a first-order understanding of how much spectrum reduction can be achieved by D2D load balancing with frequency reuse.

X. EMPIRICAL EVALUATIONS

In this section, we use real-world 4G uplink traffic traces from Smartone, a major cellular network operator in Hong Kong, to evaluate the performance of our proposed D2D load balancing scheme.

Our objectives are three-fold: (i) to evaluate the performance and complexity of our proposed low-complexity heuristic algorithm in Sec. VIII, (ii) to evaluate the benefit in terms of spectrum reduction and the cost in terms of D2D traffic overhead ratio of D2D load balancing scheme, and (iii) to measure the impact of different system parameters.

A. Methodology

1) Dataset: Our Smartone dataset contains 510 cell sectors covering a highly-populated area of 22 km² in Hong Kong. We merge them based on their unique site locations and get 152 BSs/cells. The data traffic traces are sampled every 15 minutes, spanning a 29-day period from 2015/01/05 to 2015/02/02.

2) Network Topology: Each BS's location is its corresponding site location. Each BS covers a circle area with radius 300m centered around its location. In each BS, 40 users are uniformly distributed in the coverage circle. Assume that the communication range for all user-to-BS links is 300m and the communication range for all D2D links is 30m. Then we can construct the cellular network topology $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. For each link $(u, v) \in \mathcal{E}$ with distance $d_{u,v}$, we use Shannon capacity to be the link rate, i.e., $R_{u,v} = \log_2(1 + P_t d_{u,v}^{-3.5}/N)$, where $P_t = 21$dBm is the transmit power and $N = -102$dBm is the noise power.

3) Traffic Model: We let each slot last for 2 seconds and thus we have $T = 24 \times 3600/2 = 43200$ slots in each day. Each data point in the raw traffic trace is the aggregate traffic volume of 15 minutes. To get fine-granularity traffic demands, we randomly[5] generate 120 positive real numbers in $(0, 1]$ and then divide the aggregate traffic volume on a pro-rata basis according to the values of such 120 numbers. Thus, we get 120 traffic demands of different volumes for each data point. For each generated traffic demand j, we randomly assign it to a user u_j from the total 40 users, randomly set its start time s_j from the total $15 \times 60/2 = 450$ slots, and randomly set its delay $(e_j - s_j + 1)$ from the range $\{3, 4, 5\}$.

4) Tools: We use the state-of-the-art LP solver Gurobi [34] and implement all evaluations with Python language. All evaluations are running in a cluster of 30 computers, each of which has a 8-core Intel Core-i7 3770 3.4Ghz CPU with 30GB memory, running CentOS 6.4.

B. Performance and Complexity of the Heuristic Algorithm

As seen soon, our global LP approach cannot be applied to the whole cellular network due to its high complexity. Instead, we should apply our low-complexity heuristic algorithm. In this section, we show the performance and complexity of our heuristic algorithm and hence justify why we can apply it to the whole cellular network.

The global LP approach is the benchmark to evaluate the heuristic algorithm but we cannot use it for large-scale networks. We thus evaluate them for small-scale networks. More specifically, we divide the entire 22km² region of 152 BSs into 22 small regions of 3 to 10 BSs. For each small region and each day, we use the global LP approach and the heuristic algorithm with different λ values to solve the problem Min-Spectrum-D2D and get the spectrum reduction and the overhead ratio. We then get the average spectrum reduction and average overhead ratio of both algorithms over all 22 small regions and all 29 days, as shown in Fig. 4(a). Similarly, we show the normalized time/space complexity of our heuristic algorithm with different λ values in Fig. 4(b).

From Fig. 4(a) and Fig. 4(b), we can see the tradeoff between performance (in terms of spectrum reduction) and the time/space complexity controlled by parameter λ. Increasing λ reduces the complexity but degrades the performance. However, our heuristic algorithm achieves close-to-optimal performance when λ is in $[0, 0.5]$ and we can achieve 100x complexity reduction when we use $\lambda = 0.5$. Since our results in Fig. 4(a) and Fig. 4(b) consider all 22 small regions of the entire region and all 29-day traffic traces, it is reasonable to apply our heuristic algorithm with $\lambda = 0.5$ to the whole cellular network. Thus, in the rest of this section, we set $\lambda = 0.5$ for our heuristic algorithm

In Fig. 4(a), we also show our spectrum reduction lower bound $(1 - \lambda)\rho$ proposed in Theorem 5 and our overhead ratio upper bound $\frac{(d_{\max}-1) \sum_{j \in \mathcal{J}^{\mathsf{D2D}}(\lambda)} r_j}{(d_{\max}-1) \sum_{j \in \mathcal{J}^{\mathsf{D2D}}(\lambda)} r_j + \sum_{j \in \mathcal{J}} r_j}$ proposed in Theorem 6. As we can see, we verify the correctness of both bounds. More importantly, our empirical overhead ratio is

---

[5]When we say "randomly", we draw a number from its range uniformly.

(a) Spectrum reduction v.s. λ.   (b) Time/Space complexity v.s. λ.   (c) Solving time.   (d) Memory usage.
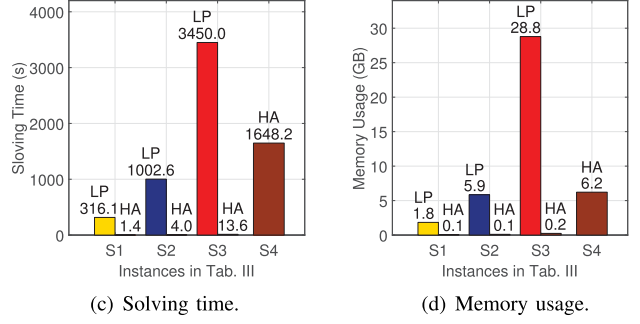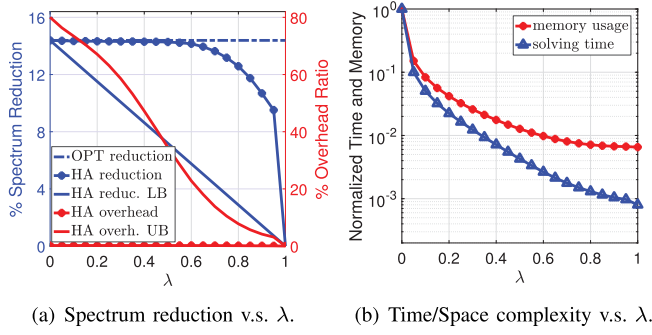
Fig. 4. Performance and complexity of our heuristic algorithm. Here, (a) and (b) show the performance and the complexity of the heuristic algorithm with different λ values; (c) and (d) compare the solving time and memory usage of the global LP approach (LP) and the heuristic algorithm (HA) with $\lambda = 0.5$.

TABLE II

FOUR DIFFERENT PROBLEM INSTANCES

Instance	$|\mathcal{B}|$	$|\mathcal{U}|$	$|\mathcal{E}|$	$|\mathcal{J}|$	$\sum_{b\in\mathcal{B}}|\mathcal{J}_b^{\mathrm{D2D}}(\lambda)|$	T
S1	3	120	155	34080	182	43200
S2	6	240	351	65520	377	43200
S3	9	360	674	103680	632	43200
S4	152	6080	11794	1647480	11960	43200

much lower than the upper bound, almost close to 0, meaning that we can achieve the spectrum reduction with very low overhead.

To more concretely compare our heuristic algorithm (with $\lambda = 0.5$) and our global LP approach, we consider four different problem instances as shown in Tab. II. They have different number of BSs, users, links, and demands. Instance S4 is our whole cellular network. We show their computational cost in Fig. 4(c) and Fig. 4(d). From instances S1-S3, we can see that our heuristic algorithm has much lower time/space complexity than our global LP approach. For our whole cellular network, i.e., instance S4, we cannot apply our global LP approach with our computational resources, but our heuristic algorithm takes less than 30 minutes of time and consumes less than 6GB of memory. The reason that we can get substantial complexity reduction is because the number of demands participating in D2D load balancing in our heuristic algorithm, i.e., $\sum_{b\in\mathcal{B}}|\mathcal{J}_b^{\mathrm{D2D}}(\lambda)|$, is much smaller than the total number of demand, i.e., $|\mathcal{J}|$. As we can see from Tab. II, $\sum_{b\in\mathcal{B}}|\mathcal{J}_b^{\mathrm{D2D}}(\lambda)|$ is only about 0.7% of $|\mathcal{J}|$ for instance S4.

C. Spectrum Reduction and Overhead Ratio of D2D Load Balancing

As justified in the previous subsection, we apply our heuristic algorithm with $\lambda = 0.5$ to the whole cellular network of all 152 BSs in the area of 22km^2. We show the 29-day spectrum reduction and overhead ratio in Fig. 5. On average our proposed D2D load balancing scheme can reduce spectrum by 25% and the overhead ratio is only 0.7%. Thus, to serve the same set of traffic demands, cellular network operators like Smartone could reduce its spectrum requirement by 25% at the cost of negligible 0.7% more D2D traffic by using our D2D load balancing scheme. Fig. 5 also verifies the upper bound, represented in Theorem 2 and Theorem 3. The average value of the upper bound of spectrum reduction is 68.69%.
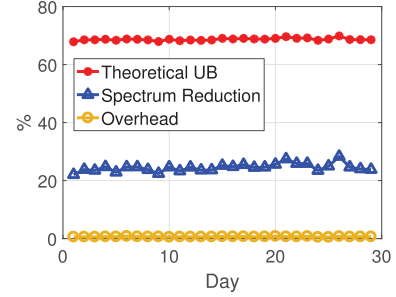

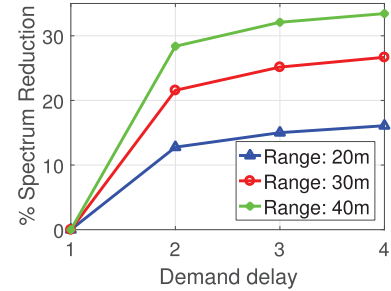Fig. 5. Spectrum reduction (ant its upper bound) and overhead ratio in 29 days.

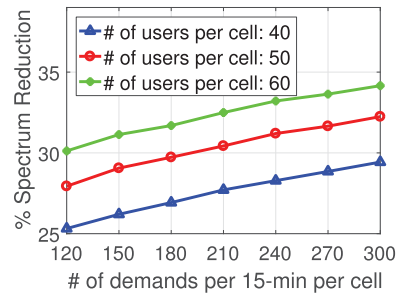Fig. 6. Impact of demand delay and D2D communication range.

Fig. 7. Impact of user density and demand intensity.

D. Impact of System Parameters

In this subsection, we evaluate the impact of four system parameters: the demand delay, the D2D communication range, the number of users per cell (user density), and the number of demands per cell per 15 minutes (demand intensity). The results are shown in Fig. 6 and Fig. 7. We observe that our D2D load balancing scheme brings more spectrum reduction

with larger demand delay, larger D2D communication range, larger user density, or larger demand intensity. The reason is as follows. Larger demand delay and larger demand intensity imply that traffic demands can be balanced with more freedom, and larger D2D communication range and larger user density result in better network connectivity, both of which enable D2D load balancing scheme to exploit more benefit.

XI. CONCLUSION AND FUTURE WORK

To the best of our knowledge, this is the first work to characterize the system-level benefit and cost of D2D load balancing, through both theoretical analysis and empirical evaluations. We show that D2D load balancing can substantially reduce the spectrum requirement at low cost, which provides strong support to standardize D2D in the coming cellular systems. This work aims to provide performance metrics/benchmarks and call for participation on the D2D load balancing scheme. In the future, it is important and interesting to jointly consider D2D load balancing and spectrum reuse/sharing among different cells and/or among different links, design online and/or distributed traffic scheduling algorithms, incorporate more realistic considerations such as transmission outage and user mobility, and eventually implement the D2D load balancing scheme in practical systems.

APPENDIX A
PROOF OF THEOREM 3

The proof logic is to construct a feasible solution to Min-Spectrum-ND$_b$ based on the optimal solution with D2D. Let us denote the optimal traffic scheduling policy for Min-Spectrum-D2D as $x_{u,v}^j(t)$ and the optimal spectrum amount for each BS b as F_b^{D2D}. Then consider BS $b \in \mathcal{B}$. For each traffic demand $j \in \mathcal{J}_b$, user $s \in \mathcal{U}_b$ must transmit all volume r_j either to BS b directly and/or to any other neighbour users via D2D links. Thus $\forall j \in \mathcal{J}_b$, the following equality holds,

$$r_j = \sum_{t=s_j}^{e_j} [x_{u_j,b}^j(t)R_{u_j,b} + \sum_{v:v\in\mathcal{U}_b,(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)R_{u_j,v}$$
$$+ \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} \sum_{v:v\in\mathcal{U}_{b'},(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)R_{u_j,v}],$$

In addition, the (peak) spectrum requirement should be satisfied,

$$\sum_{j\in\mathcal{J}_b} x_{u_j,b}^j(t) + \sum_{u\in\mathcal{U}_b}\sum_{v\in\text{in}(u)\setminus\{u\}}\sum_{j\in\mathcal{J}:t\in[s_j,e_j]} x_{v,u}^j(t) \le F_b^{\text{D2D}},$$

Now we construct a feasible solution to Min-Spectrum-ND$_b$, i.e., for any $j \in \mathcal{J}_b$,

$$\bar{x}_{u_j,b}^j(t) = x_{u_j,b}^t(t) + \sum_{v:v\in\mathcal{U}_b,(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)\frac{R_{u_j,v}}{R_{u_j,b}}$$
$$+ \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} \sum_{v:v\in\mathcal{U}_{b'},(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)\frac{R_{u_j,v}}{R_{u_j,b}},$$
$$\text{(27)}$$

Thus we have

$$\gamma_b(t)$$
$$= \sum_{j\in\mathcal{J}_b:t\in[s_j,e_j]} \bar{x}_{u_j,b}^j(t) = \sum_{j\in\mathcal{J}_b:t\in[s_j,e_j]} [x_{u_j,b}^t(t)$$
$$+ \sum_{v:v\in\mathcal{U}_b,(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)\frac{R_{u_j,v}}{R_{u_j,b}}$$
$$+ \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} \sum_{v:v\in\mathcal{U}_{b'},(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)\frac{R_{u_j,v}}{R_{u_j,b}}]$$
$$\le \sum_{j\in\mathcal{J}_b:t\in[s_j,e_j]} [x_{u_j,b}^t(t) + r_{u_j} \sum_{v:v\in\mathcal{U}_b,(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)$$
$$+ \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} \tilde{r}_{u_j}^{b'} \sum_{v:v\in\mathcal{U}_{b'},(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)]$$
$$\overset{(a)}{\le} \max\{r,1\} \sum_{j\in\mathcal{J}_b:t\in[s_j,e_j]} [x_{u_j,b}^t(t) + \sum_{v:v\in\mathcal{U}_b,(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)]$$
$$+ \tilde{r} \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} \sum_{j\in\mathcal{J}_b:t\in[s_j,e_j]} \sum_{v:v\in\mathcal{U}_{b'},(u_j,v)\in\mathcal{E}} x_{u_j,v}^j(t)$$
$$\le \max\{r,1\}F_b^{\text{D2D}} + \tilde{r} \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} F_{b'}^{\text{D2D}}, \quad \text{(28)}$$

where (a) trivially holds for $r > 1$ and also holds for $r \le 1$ by noting that there is no intra-cell D2D traffic when $r \le 1$. Therefore, $\max\{r,1\}F_b^{\text{D2D}} + \tilde{r}\sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} F_{b'}^{\text{D2D}}$ is a feasible spectrum amount for BS b without D2D. Thus we must have

$$F_b^{\text{ND}} \le \max\{r,1\}F_b^{\text{D2D}} + \tilde{r} \sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} F_{b'}^{\text{D2D}}. \quad \text{(29)}$$

Then we do summation over all BSs and get

$$F^{\text{ND}} = \sum_{b\in\mathcal{B}} F_b^{\text{ND}}$$
$$\le \sum_{b\in\mathcal{B}} \max\{r,1\}F_b^{\text{D2D}} + \tilde{r} \sum_{b\in\mathcal{B}}\sum_{b':(b,b')\in\mathcal{E}^{\text{D2D}}} F_{b'}^{\text{D2D}}$$
$$\overset{(b)}{=} \max\{r,1\}\sum_{b\in\mathcal{B}} F_b^{\text{D2D}} + \tilde{r} \sum_{b'\in\mathcal{B}}\sum_{b:(b,b')\in\mathcal{E}^{\text{D2D}}} F_{b'}^{\text{D2D}}$$
$$= \max\{r,1\}\sum_{b\in\mathcal{B}} F_b^{\text{D2D}} + \tilde{r} \sum_{b'\in\mathcal{B}} \delta_{b'}^- F_{b'}^{D2D}$$
$$\le \max\{r,1\}\sum_{b\in\mathcal{B}} F_b^{\text{D2D}} + \tilde{r} \sum_{b'\in\mathcal{B}} \Delta^- F_{b'}^{\text{D2D}}$$
$$= [\max\{r,1\} + \tilde{r}\Delta^-]F^{\text{D2D}}, \quad \text{(30)}$$

where (b) holds because any $(b,b') \in \mathcal{E}^{\text{D2D}}$ contributes one $\tilde{r}F_{b'}^{\text{D2D}}$ on both sides. Thus, we conclude that

$$\rho = \frac{F^{\text{ND}} - F^{\text{D2D}}}{F^{\text{ND}}} \le \frac{\max\{r,1\} + \tilde{r}\Delta^- - 1}{\max\{r,1\} + \tilde{r}\Delta^-}. \quad \text{(31)}$$

APPENDIX B
PROOF OF THEOREM 5

It is obvious that $\rho^{\text{Heuristic}} \le \rho$. To show $\rho^{\text{Heuristic}} \ge (1-\lambda)\rho$, we need to show that

$$F^{\text{Heuristic}} \le (1-\lambda)F^{\text{D2D}} + \lambda F^{\text{ND}}. \quad \text{(32)}$$

In the following, we construct a feasible solution to (23) whose total spectrum requirement is at most $(1 - \lambda)F^{\text{D2D}} + \lambda F^{\text{ND}}$. Then since $F^{\text{Heuristic}}$ is the optimal value of (23), we clearly have that $F^{\text{Heuristic}} \leq (1 - \lambda)F^{\text{D2D}} + \lambda F^{\text{ND}}$.

All jobs in $\mathcal{J}_b^{\text{ND}}(\lambda)$ are served locally according to the results in Step I, i.e., $\{x_{u_j,b}^j(t)\}$. Each job j in the demand set $\mathcal{J}_b^{\text{D2D}}(\lambda)$ is served as follows. For all slots not in $T_b(\lambda)$, we serve it according to the results in Step I, i.e., $\{x_{u_j,b}^j(t)\}$. Thus, the resulting total spectrum for any slot $t \notin T_b(\lambda)$ is $\gamma_b(t) \leq \lambda F_b$, where F_b is the optimal value of Min-Spectrum-ND$_b$. At any slot $t \in T_b(\lambda)$, if job j is delivered at the volume of v at slot t when solving Min-Spectrum-ND$_b$ in Step I of our heuristic algorithm, we serve job j at the volume of λv locally without D2D, i.e., directly sending λv amount to BSs. Thus, any job $j \in \mathcal{J}_b^{\text{D2D}}(\lambda)$ will be served *at least* at the volume of λr_j locally. And the resulting total used spectrum from users to BS b at slot $t \in T_b(\lambda)$ is at most λF_b. In summary, the resulting total spectrum for any slot $t \in [T]$ is at most $\sum_{b \in \mathcal{B}} \lambda F_b = \lambda F^{\text{ND}}$.

After that, every job $j \in \mathcal{J}_b^{\text{D2D}}(\lambda)$ has a remaining volume of at most $(1-\lambda)r_j$, i.e., scaling with a factor $(1-\lambda)$. We then serve all those jobs in $\mathcal{J}_b^{\text{D2D}}(\lambda)$ with the remaining volume with D2D by solving the problem Min-Spectrum-D2D. The resulting total spectrum for any slot $t \in [T]$ is at most $(1-\lambda)F^{\text{D2D}}$.

Thus, the total spectrum of this constructed solution is at most $(1 - \lambda)F^{\text{D2D}} + \lambda F^{\text{ND}}$, which completes the proof.

REFERENCES

[1] L. Deng *et al.*, "Device-to-device load balancing for cellular networks," in *Proc. IEEE MASS*, Oct. 2015, pp. 19–27.

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

[3] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, "Mobile data offloading: How much can WiFi deliver?" in *Proc. ACM CoNEXT*, Apr. 2010, pp. 536–550.

[4] R. C. Daniels, J. N. Murdock, T. S. Rappaport, and R. W. Heath, Jr., "60 GHz wireless: Up close and personal," *IEEE Microw. Mag.*, vol. 11, no. 7, pp. 44–50, Dec. 2010.

[5] A. Ghosh *et al.*, "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 54–64, Jun. 2012.

[6] *SmarTone*. Accessed: Dec. 4, 2018. [Online]. Available: http://www.smartone.com

[7] L. Deng *et al.* (2018). "Device-to-device load balancing for cellular networks." [Online]. Available: https://arxiv.org/abs/1710.02636

[8] 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.

[9] G. Fodor *et al.*, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[10] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-device communications achieve efficient load balancing in LTE-advanced networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 2, pp. 57–65, Feb. 2014.

[11] Z. Hu *et al.*, "An in-depth analysis of 3G traffic and performance," in *Proc. ACM AllThingsCellular*, Aug. 2015, pp. 1–6.

[12] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 1147–1161, Apr. 2017.

[13] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[14] Q. Ye, B. Rong, Y. Chen, M. Al-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.

[15] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of WiFi offloading: Trading delay for cellular capacity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1540–1554, Mar. 2014.

[16] A. Jalali, "On cell breathing in CDMA networks," in *Proc. IEEE Int. Conf. Commun. Conf. Rec. (ICC)*, vol. 2, Jun. 1998, pp. 985–988.

[17] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li, "Cellular traffic offloading through WiFi networks," in *Proc. IEEE MASS*, Oct. 2011, pp. 192–201.

[18] B. Han, P. Hui, and A. Srinivasan, "Mobile data offloading in metropolitan area networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 14, no. 4, pp. 28–30, 2011.

[19] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Scalable spectrum allocation and user association in networks with many small cells," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2931–2942, Jul. 2017.

[20] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, Oct. 2015.

[21] Z. Zhou, D. Guo, and M. L. Honig, "Licensed and unlicensed spectrum allocation in heterogeneous networks," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1815–1827, Apr. 2017.

[22] Z. Zhou and D. Guo, "1000-cell global spectrum management," in *Proc. ACM MobiHoc*, Jul. 2017, p. 20.

[23] Z. Chen, H. Zhao, Y. Cao, and T. Jiang, "Load balancing for D2D-based relay communications in heterogeneous network," in *Proc. WiOpt*, 2015, pp. 23–29.

[24] C. Vlachos and V. Friderikos, "Optimal device-to-device cell association and load balancing," in *Proc. IEEE ICC*, Jun. 2015, pp. 5441–5447.

[25] F. Jiang, Y. Liu, B. Wang, and X. Wang, "A relay-aided device-to-device-based load balancing scheme for multitier heterogeneous networks," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1537–1551, Oct. 2017.

[26] 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.

[27] *Shannon–Hartley Theorem*. [Online]. Available: https://en.wikipedia.org/wiki/Shannon%E2%80%93Hartley_theorem

[28] G. Buttazzo, *Hard Real-Time Computing Systems: Predictable Scheduling Algorithms and Applications*. New York, NY, USA: Springer, 2011.

[29] C. L. Liu and J. W. Layland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," *J. ACM*, vol. 20, no. 1, pp. 46–61, 1973.

[30] F. Yao, A. Demers, and S. Shenker, "A scheduling model for reduced CPU energy," in *Proc. IEEE FOCS*, Oct. 1995, pp. 374–382.

[31] L. G. Khachiyan, "Polynomial algorithms in linear programming," *USSR Comput. Math. Math. Phys.*, vol. 20, no. 1, pp. 53–72, 1980.

[32] M. Skutella, "An introduction to network flows over time," in *Proc. Res. Trends Combinat. Optim.*, pp. 451–482, 2009.

[33] W. Si, S. Selvakennedy, and A. Y. Zomaya, "An overview of channel assignment methods for multi-radio multi-channel wireless mesh networks," *J. Parallel Distrib. Comput.*, vol. 70, no. 5, pp. 505–524, 2010.

[34] *Gurobi*. Accessed: Dec. 4, 2018. [Online]. Available: http://www.gurobi.com
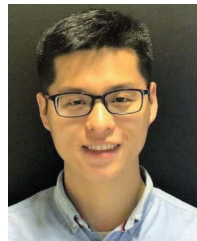
Lei Deng (S'14–M'18) received the B.Eng. degree from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2012, and the Ph.D. degree from the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, in 2017. In 2015, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. He is currently an Assistant Professor with the School of Electrical Engineering and Intelligentization, Dongguan University of Technology. His research interests are timely network communications, intelligent transportation system, and spectral-energy efficiency in wireless networks.

Yinghui He received the B.S.E. degree in information engineering from Zhejiang University, Hangzhou, China, in 2018, where he is currently pursuing the master's degree with the College of Information Science and Electronic Engineering. His research interests mainly include mobile edge computing and device-to-device communications.

Ying Zhang received the B.Eng. degree from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2013, and the Ph.D. degree from the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, in 2017. His research interests include energy system operation and optimization, machine learning, statistical arbitrage, and algorithmic trading.

Minghua Chen (S'04–M'06–SM'13) received the B.Eng. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, in 1999 and 2001, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in 2006. He spent one year visiting Microsoft Research Redmond as a Post-Doctoral Researcher. He joined the Department of Information Engineering, The Chinese University of Hong Kong, in 2007, where he is currently an Associate Professor. He is also an Adjunct Associate Professor with the Institute of Interdisciplinary Information Sciences, Tsinghua University. His recent research interests include energy systems (e.g., smart power grids and energy-efficient data centers), intelligent transportation systems, distributed optimization, multimedia networking, wireless networking, delay-constrained networking, and characterizing the benefit of data-driven prediction in algorithm/system design. He received the Eli Jury Award from UC Berkeley in 2007 (presented to a graduate student or recent alumnus for outstanding achievement in the area of systems, communications, control, or signal processing) and The Chinese University of Hong Kong Young Researcher Award in 2013. He also received several best paper awards, including the IEEE ICME Best Paper Award in 2009, the IEEE Transactions on Multimedia Prize Paper Award in 2009, and the ACM Multimedia Best Paper Award in 2012. He received the ACM Recognition of Service Award in 2017 for service contribution to the community. He is currently the Steering Committee Chair of ACM e-Energy. He serves as the TPC Co-Chair of ACM e-Energy 2016 and the General Chair of ACM e-Energy 2017. He also serves as an Associate Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING from 2014 to 2018.

Zongpeng Li received the B.Sc. degree in computer science from Tsinghua University in 1999, and the Ph.D. degree from the University of Toronto in 2005. He was with the University of Calgary. He is currently a Professor at the School of Computer Science, Wuhan University. His research interests include computer networks, cloud computing, and IoT.

Jack Y. B. Lee (M'95–SM'03) received the B.Eng. and Ph.D. degrees in information engineering from The Chinese University of Hong Kong, Hong Kong, in 1993 and 1997, respectively. He is currently an Associate Professor with the Department of Information, The Chinese University of Hong Kong. His research group focuses on research in multimedia communications systems, mobile communications, protocols, and applications. He specializes in tackling research challenges arising from real-world systems. He works closely with the industry to uncover new research challenges and opportunities for new services and applications. Several of the systems research from his lab have been adopted and deployed by the industry.
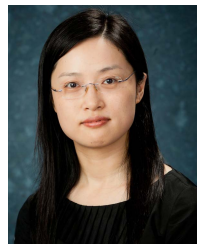
Ying Jun (Angela) Zhang (S'00–M'05–SM'10) received the Ph.D. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 2004. Since 2005, she has been with Department of Information Engineering, The Chinese University of Hong Kong, where she 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 a fellow of IET. She was a co-recipient of the 2011 IEEE Marconi Prize Paper Award on Wireless Communications, the 2013 IEEE SmartgridComm Best Paper Award, and the 2014 IEEE ComSoc APB Outstanding Paper Award. She was a recipient of the Young Researcher Award from The Chinese University of Hong Kong in 2011. As the only winner from engineering science, she has received the Hong Kong Young Scientist Award 2006, conferred by the Hong Kong Institution of Science. She serves as the Chair for the Executive Editor Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Previously, she served many years as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON COMMUNICATIONS, *Security and Communications Networks* (Wiley), and a Feature Topic in the *IEEE Communications Magazine*. She has served on the organizing committee for major IEEE conferences, including ICC, GLOBECOM, SmartgridComm, VTC, CCNC, ICCC, and MASS. She was a Co-Chair of the IEEE ComSoc Multimedia Communications Technical Committee and the IEEE Communication Society GOLD Coordinator. She is currently the Chair of IEEE ComSoc Emerging Technical Committee on Smart Grid. She is a Distinguished Lecturer of IEEE ComSoc.

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