High Throughput Opportunistic Cooperative Device-to-Device Communications With Caching

Binqiang Chen, Chenyang Yang and Gang Wang

***Abstract*—To achieve the potential in providing high throughput for cellular networks by device-to-device (D2D) communications, the interference among D2D links should be carefully managed. In this paper, we propose an opportunistic cooperation strategy for D2D transmission by exploiting the caching capability at the users to control the interference among D2D links. We consider overlay inband D2D, divide the D2D users into clusters, and assign different frequency bands to cooperative and non-cooperative D2D links. To provide high opportunity for cooperative transmission, we introduce a caching policy. To maximize the network throughput, we jointly optimize the cluster size and bandwidth allocation, where the closed-form expression of the bandwidth allocation factor is obtained. Simulation results demonstrate that the proposed strategy can provide $400\% \sim 500\%$ throughput gain over traditional D2D communications when the content popularity distribution is skewed, and can provide $60\% \sim 80\%$ gain even when the content popularity distribution is uniform.**

***Index Terms*—Caching, D2D, Cooperative transmission, Interference, High Throughput**

I. Introduction

Device-to-device (D2D) communications enable direct communications between two user devices without traversing the base station (BS) or core network, and are promising to achieve the high throughput goal of 5th generation (5G) cellular networks [1–7]. The typical use-cases of D2D communications include cellular offloading, content distribution, and relaying, *etc.* [8], where content delivery service has attracted considerable attention recently, since it accounts for the majority of the explosive increasing traffic load.

Motivated by the observation that a large amount of content delivery requests are asynchronous but redundant, i.e., the same content is requested repeatedly at different times, caching has long been studied as a technique to improve performance of wired networks. Due to the rapid reduction in cost of storage device, caching at the wireless edge is also recognized as a promising way for delivering popular contents nowadays, which can improve the network throughput, energy efficiency and the quality of user experience (QoE) [9–17]. However, different from wired networks, the performance of wireless networks is fundamentally limited by the interference, which inevitably limits the throughput gain from local caching.

To take the advantage of the storage capacity at smart phones, cache-enabled D2D communications have been proposed recently, which can offload the content delivery traffic and hence boost the network throughput significantly [18, 19]. Since only the users in proximity communicate to each other, the distance between a user and the undesired transmitters can be close and hence the interference in D2D networks is strong, which needs to be carefully controlled. In an early work of studying cache-enabled D2D communications, the D2D users are divided into clusters. Then, the intra-cluster interference among D2D links is managed by using time division multiple access (TDMA), while the inter-cluster interference between D2D links is simply treated as noise [18]. In [19], only the D2D link from one of the four adjacent clusters is allowed to be active at the same time-frequency resource block, in order to avoid strong inter-cluster interference among adjacent clusters. In [20], interference alignment was employed to mitigate the interference among D2D links, but only three D2D links were coordinated within each cluster, and the interference among clusters was again treated as noise. In [21–23], cooperative relay techniques were proposed to mitigate the interference between cellular and D2D links, which however can not manage the interference among the D2D links.

It is well known that if several transmitters have the required data for some users, they can jointly transmit to the users without generating interference. In fact, if contents have been locally cached, cooperative transmission without data exchange among transmitters becomes possible, which can transform interference into useful signal. Based on such an interesting observation, a BS cooperative transmission strategy was proposed in [24] by exploiting the caches at BSs, where precoding and cache control were optimized to guarantee the QoE of users. Inspired by this work, a natural question is: can we apply cooperative transmission in D2D communications with caching?

Fortunately, cooperative transmission is possible in practice due to the following reasons. (i) In D2D communications, the D2D transmitter (DT) has been proposed to assist other users in additional to transmitting data to its destined D2D receiver (DR), e.g., with cooperative relay [22]. The users have the incentive to do this if their own QoE can be improved or their costs can be compensated by some other rewards [25]. (ii) To facilitate cooperative transmission, the global channel state information (CSI) is required to compute the precoding matrix. The CSI among D2D links can be obtained at DTs and the BS through channel probing and feedback [26]. Then, the precoding vectors can be computed at the BS and sent to the cooperative DTs via multicast. (iii) The synchronization among cooperative DTs is more easier to be implemented than that in Ad-hoc networks, because it can be realized with the assist of the BS [5]. Besides, the synchronization can also be

Binqiang Chen and Chenyang Yang are with the School of Electronics and Information Engineering, Beihang University, Beijing, China, Emails: {chenbq,cyyang}@buaa.edu.cn. Gang Wang is with the NEC labs, China, Email: wang_gang@nec.cn.

realized at users by using the methods proposed in [27].

In this paper, we propose an opportunistic cooperation strategy for cache-enabled D2D communications to manage the interference among D2D links. Different from the BS cooperative transmission strategy [24], the cooperation strategy for D2D communications needs to be optimized in a different way. To control the strong interference among D2D links, we divide the D2D users into virtual clusters. To maximize the opportunity of cooperative transmission via D2D links, we take both redundant caching and diversity caching into account in the users among the clusters, which differs from [24] where all BSs cache the same files. When some users have cached the files requested by other users called DRs, these users act as DTs to jointly transmit the requested files to the DRs. Because only some D2D links can employ cooperative transmission, we assign different frequency bands to cooperative and non-cooperative links to avoid mutual interference. To maximize the average network throughput without compromising the experience of non-cooperative users, we jointly optimize the cluster size and bandwidth allocation under the minimal average user data rate constraint.

The contributions of this paper are summarized as follows:

- We propose an opportunistic cooperation strategy to manage the interference among D2D links, which improve the network throughput remakably.
- We jointly optimize the cluster size and bandwidth allocation and obtain the closed-form expression of optimal bandwidth allocation factor.

The rest of the paper is organized as follows. Section II presents the system model. Section III introduces the cooperation strategy, derives the average network throughput and average user data rate, and jointly optimizes bandwidth allocation and cluster size. Section IV provides numerical and simulation results. Section V concludes the paper.

II. System Model

Consider a cellular network, where M single-antenna users are uniformly located in a square hotspot within a macro cell, where the area is with side length of D_c as shown in Fig. 1. Each user is willing to store N files in its local cache and can act as a helper to share files. When a helper conveys a file in local cache via D2D link to a DR requesting the file, the helper becomes a DT. The BS is aware of the cached files at each user and coordinates the D2D communications.

A. Content Popularity

We consider a static content catalog including N^f files that the users may request. To simplify the analysis, each file is assumed with the same size as in [15–19]. Although in practice the files are with unequal sizes, each file can be divided into chunks of equal size [16], so the same analysis can still be applied. The N^f files are indexed in a descending order of popularity, e.g., the 1st file is the most popular file. The probability that the ith file is requested by a user is assumed to follow a Zipf distribution,

$$P_{Nf}(i) = i^{-\beta} / \sum_{j=1}^{N^f} j^{-\beta}, \tag{1}$$

where $\sum_{i=1}^{N_f} P_{Nf}(i) = 1$, and the parameter β reflects skewness of the popularity distribution, with large β meaning that a few files are requested by majority of users [28].

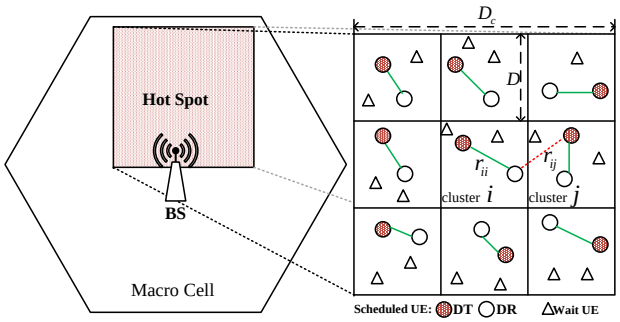

Fig. 1. Cluster division model, "UE" means user equipment.

B. Communication Protocol

D2D links can be established among users in proximity owing to the limited transmit power at each user equipment (UE) and the possible strong interference among UEs. A widely used communication protocol for D2D communications is that two UEs can communicate if their distance is smaller than a given distance [18, 29]. To control the strong interference and make the analysis tractable, the square hotspot area is divided into B smaller square areas called clusters as in [19], where the side length of each cluster is $D = D_c/\sqrt{B}$. For mathematical simplicity, we assume that the number of users per cluster is $K = M/B$ and each user is assumed to transmit with the same power P as in [19].

For the non-cooperative users, only those within the same cluster can establish D2D links in order to control interference. For the cooperative users, the users in different clusters are allowed to establish D2D links to avoid interference and exploit multiplexing gain by jointly transmission. The details will be presented in the next section.

We consider overlay inband D2D [8], and assume that a fixed bandwidth of W is assigned to the D2D links.

III. Opportunistic Cooperation Strategy

In this section, we first introduce a caching policy to provide high opportunity for cache-enabled cooperative D2D transmission. Then, we propose an opportunistic cooperative transmission policy. Finally, we optimize two key parameters in the strategy to maximize the network throughput.

A. Caching Policy

To maximize the probability that a user can fetch files through D2D links, the users within a cluster should cache different files. To maximize the probability of cooperative transmission among DTs in different clusters, the files cached at the users of each cluster should be the same. This suggest that the caching policy needs to balance the diversity of content with the redundancy of the replicas of popular contents. To this end, we consider the following caching policy.

According to the user cache size N, all files are divided into $K_0 = N^f/N$ groups. The kth *file group* $\mathcal{G}_k$ consists of the $(k-1)N+1$th to the kNth files where $1 \le k \le K_0$, e.g., the 1st file group $\mathcal{G}_1$ contains the most popular N files. Then, the probability that a user requests a file within the kth *file group* $\mathcal{G}_k$ can be obtained as

$$P_k = \sum_{i=(k-1)N+1}^{kN} P_{N^f}(i) = \frac{\sum_{i=(k-1)N+1}^{kN} i^{-\beta}}{\sum_{j=1}^{N^f} j^{-\beta}}. \quad (2)$$

In every cluster, the kth user caches the kth file group $\mathcal{G}_k$. Then, every user in each cluster caches different files, i.e., diversity caching is achieved within each cluster, and the most popular KN files are cached in every cluster, i.e., redundant caching is achieved among clusters. Because each cluster contains K users, the file groups with indices exceeding K (i.e., $\mathcal{G}_k$, $k > K$) are not cached at users.

When $K = K_0$, all the N^f files can be cached at the users in each cluster and all user requests can be served via D2D links, therefore it is not necessary to assign more than K_0 users to each cluster. For this reason, we assume $K \le K_0$.

In practice, the files can be proactively downloaded by the operator from the BS to the cache at each user via multicast during off-peak times according to the user demand statistics.

B. Opportunistic Cooperative Transmission Policy

According to whether a user can find the requested file in the hotspot area, we can classify the users into two types.

D2D users: If the file requested by a user is cached at any UE in the cluster it belongs to (and hence also cached in UEs in other clusters according to the above-mentioned caching policy), then the user can directly obtain the file with D2D communication, either without or with cooperation. Such a user is referred to as a D2D user. Besides, if the file requested by a user is in its local cache, it can retrieve the file immediately with zero delay, but we ignore this case for analysis simplicity as in [19].

Cellular users: If the file requested by a user is not cached in the UEs within the hotspot area, the user fetches the file from the BS and becomes a regular cellular user. The number of cellular users is denoted as N^b.

For easy understanding, we introduce the strategy with the help of an example.

1) Cooperative D2D Users: If there exists at least one user in a cluster requesting the files in $\mathcal{G}_k$, then we say that the cluster *hits the kth file group*. In Fig. 2, the users in the first cluster respectively request the files in $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_4$, and hence the first cluster hits the $\{1,2,4\}$th file groups.

If every cluster hits the same file group $\mathcal{G}_k$, the kth user in each cluster who caches the file group $\mathcal{G}_k$ can act as a DT, and all DTs in these clusters cooperatively transmit files to the DRs requesting the files in $\mathcal{G}_k$.[1] Those DRs are referred to as **cooperative D2D users** (*Coop users* for short), whose number is denoted as N^c.

[1]Though further improvement is possible by allowing cooperation among less than B clusters (called partial cooperation), we only optimize full cooperation among all clusters for mathematical tractability. The impact of partial cooperation is shown via simulation in Section IV.

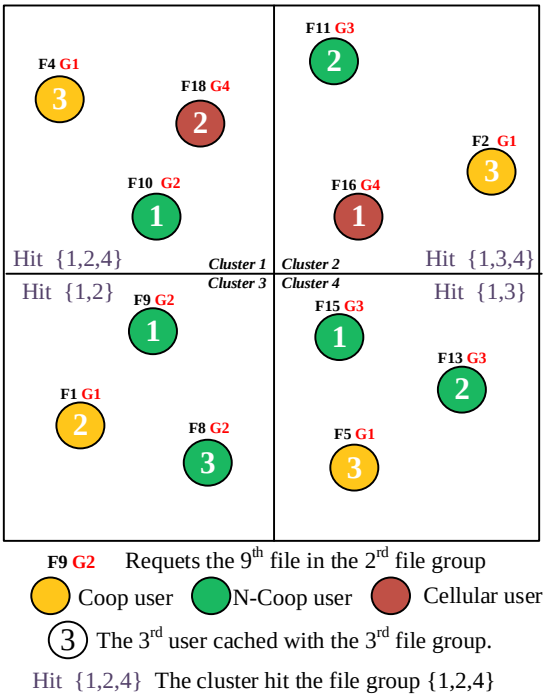

Fig. 2. Illustration of the opportunistic cooperaton strategy. Catalog size $N_f = 20$, $B = 4$ clusters in the hotspot, $K = 3$ users in each cluster and each user caches $N = 5$ files. All 20 files are divided into 4 groups according to content popularity, e.g., $\mathcal{G}_1 = \{1,2,3,4,5\}$.

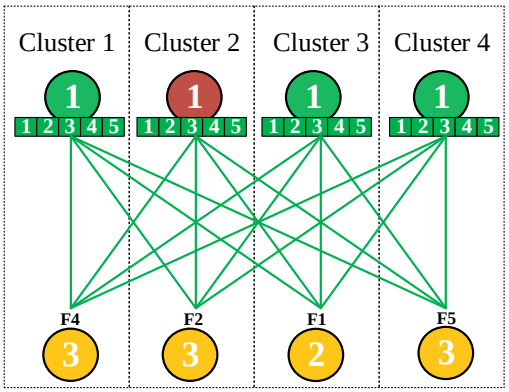

Fig. 3. Illustration for cooperative transmission from multiple DTs to DRs.

In Fig. 2, every cluster hits the 1st file group. Hence, the 1st users in all the four clusters who cache the files in $\mathcal{G}_1$ can act as DTs to cooperatively transmit files with indices $\{4,2,1,5\}$ respectively to the 3rd user in cluster 1, the 3rd user in cluster 2, the 2nd user in cluster 3, and the 3rd user in cluster 4, as shown in Fig. 3.

The remaining users except the cellular and Coop users are **non-cooperative D2D users** (*N-Coop users* for short), whose number is $N^n = M - N^b - N^c$.

2) Interference Control: Due to the random locations of the DTs in proximity, the interference in the network needs to be carefully controlled even with the cooperative transmission.

Inter-type interference: To avoid the mutual interference between *Coop users* and *N-Coop users*, we assign ηW for Coop users and the remaining bandwidth $(1-\eta)W$ for N-

Coop users, where η is the bandwidth allocation factor and $0 \leq \eta \leq 1$. A large value of η means that more bandwidth is allocated to Coop users.

Intra-cluster interference: Considering that the users within each cluster can not cooperate due to caching different files, we randomly select one Coop D2D link and one N-Coop D2D link respectively in each cluster to transmit at the same time to avoid intra-cluster interference as in [18, 19].

Inter-cluster interference: There is no inter-cluster interference among Coop users owing to the joint transmission from multiple DTs. However, there exists inter-cluster interference between N-Coop users, which is weak and can be regarded as noise since N-Coop users only communicate via D2D links within the cluster.

3) Operation Modes: Due to the opportunistic nature of establishing the cooperative D2D links, the network may operate in the following two modes.

- In *Mode* 0, there does not exist any file group hit by every cluster, i.e., all D2D users are N-Coop users. Then, all the DTs transmit independently, and the bandwidth W is assigned to the N-Coop users, i.e., $\eta = 0$.
- In *Mode* 1, there exist file groups hit by every cluster, i.e., there exist Coop users. Then, $0 < \eta \leq 1$.

To become the Coop users, the users are not necessary to request the *same file*, but to request the *files in the same group*. Hence the cooperative probability, i.e., the probability that the network operates in *Mode 1*, is high, which increases with the number of users in each cluster K. To see this, we derive the cooperative probability P^c as follows.

A cluster hits the kth file group if at least one of the K users in the cluster requests a file in $\mathcal{G}_k$, whose probability is denoted as P_k^h. It is the complement of the probability that no user requests any file in the kth file group, which is $(1-P_k)^K$. Then, from (2), P_k^h can be obtained as

$$P_k^h = 1 - (1 - P_k)^K \tag{3}$$

which increases with K, because $0 \leq P_k \leq 1$.

Cooperative probability is the probability that there exists at least one file group hit by all the B clusters. It is the complement of the probability that there is no file group hit by all the B clusters, and hence can be derived as

$$P^c = 1 - \prod_{k=1}^{K} (1 - (P_k^h)^B), \tag{4}$$

where $1-(P_k^h)^B$ is the probability that the number of clusters hitting the kth file group is less than B, which decreases with the growth of K since $B = M/K$. Therefore, for a given value of M, P^c is an increasing function of K.

4) Key Parameters: Since only one Coop D2D link per cluster is allowed to be active each time, B users out of all Coop users can be scheduled simultaneously in *Mode* 1. Therefore, the number of **active Coop users** is $N^a = B$ in *Mode* 1, and is $N^a = 0$ in *Mode* 0.[2] With the cooperative probability, the average number of active Coop users can be obtained as

$$\bar{N^a} = BP^c + 0(1 - P^c) = BP^c. \tag{5}$$

$\bar{N^a}$ characterizes how many *interference-free D2D links* can transmit at the same time-frequency resources in average, which can reflect the multiplexing gain. In general, the number of interference-free D2D links demonstrates the same trend with the network throughput, as to be verified in Section IV. Hence, a large value of $\bar{N^a}$ implies a high network throughput. When the number of users per cluster K is large, the cooperative probability is high, but the number of active Coop users is small since $N^a = B$ and $B = M/K$. This suggests that there is a tradeoff between two counter-running effects: a small value of K leads to more active Coop users if the system operates in Mode 1; a large value of K yields high cooperative probability. In other words, to maximize the network throughput, the cluster size should be optimized, which is reflected by the number of users per cluster K since the number of users in the hotspot M is given.

Due to the multiplexing gain and interference-free transmission, the average data rate of Coop users usually exceeds that of N-Coop users. As a result, the overall network throughput will be reduced if we simply assign identical bandwidth to these two types of D2D users. Owing to the same reason, simply allocating all the bandwidth to Coop users can maximize the network throughput, but no N-Coop users can be served. This indicates that the bandwidth allocation factor η should be optimized to maximize the throughput of the network under the constraint on the data rate of each user to avoid unfairness.

C. Optimization of Cluster Size and Bandwidth Allocation

In this subsection, we jointly optimize the bandwidth allocation factor η and cluster size K to maximize the average network throughput under a constraint that the average user data rate is larger than a given value, μ (Mbps). Because we assume overlay D2D communications, only D2D users are considered in the network throughput.

In the sequel, we first derive the average network throughput achieved by all D2D users. Then, we derive the average data rate for each of Coop and N-Coop users. Finally, we find the optimal cluster size K and bandwidth allocation factor η.

1) Average Network Throughput: Recall that only one Coop D2D link (if any) and one N-Coop D2D link are scheduled per cluster in each time. Then, the average throughput of the network operating in *Mode* 0 can be obtained as follows,

$$\bar{R_0} = \mathbb{E}\{W \sum_{i=1}^{B} R_i^n\} \overset{(a)}{=} WB\bar{R_i^n}, \tag{6}$$

where the expectation is taken over small scale channel fading and user location, (a) comes from the fact that all users are randomly located and transmit with equal power, and R_i^n and $\bar{R_i^n}$ are the instantaneous and average data rate per unit bandwidth per second of the N-Coop link in the ith cluster,

[2]For the considered strategy, the number of active Coop users is less than the number of Coop users, i.e., $N^a \leq N^c$.

respectively.[3]

Analogically, the average throughput of the network operating in *Mode* 1 can be obtained as

$$\begin{aligned}\bar{R}_1 &= \mathbb{E}\{\eta W \sum_{i=1}^{B} R_i^c + (1-\eta) W \sum_{i=1}^{B} R_i^n\} \\ &= WB(\eta \bar{R}_i^c + (1-\eta)\bar{R}_i^n),\end{aligned} \tag{7}$$

where R_i^c and $\bar{R}_i^c$ are the instantaneous and average data rate per unit bandwidth per second of the Coop link in the ith cluster, respectively.

Further considering the cooperative probability P^c in (4), the average throughput of the network is

$$\begin{aligned}\bar{R} &= P^c \bar{R}_1 + (1-P^c)\bar{R}_0 \\ &= WB(P^c \eta \bar{R}_i^c + (1 - P^c\eta)\bar{R}_i^n).\end{aligned} \tag{8}$$

Proposition 1: The average data rate per unit bandwidth per second of the N-Coop link in the ith cluster is

$$\bar{R}_i^n = \log_2(Q_1(\alpha)) - \log_2(Q_2(\alpha)) - 3, \tag{9}$$

where $Q_1(\alpha) \triangleq \int_0^{\sqrt{2}} r^{-\alpha} g(r) dr + 8\int_0^{\sqrt{5}} r^{-\alpha} f(r) dr$, $Q_2(\alpha) \triangleq \int_0^{\sqrt{5}} r^{-\alpha} f(r) dr$, and $f(r)$ and $g(r)$ are in closed-form expression defined in Appendix A.

Proof: See Appendix A. ∎

In Proposition 1, $Q_1(\alpha)$ and $Q_2(\alpha)$ are easy to be computed numerically. We can see that $\bar{R}_i^n$ only depends on the path loss exponent α.

Proposition 2: The average data rate per unit bandwidth per second of the Coop link in the ith cluster is

$$\bar{R}_i^c = \log_2(1 + \frac{PD^{-\alpha}}{B\sigma^2} Q_1(\alpha)). \tag{10}$$

Proof: See Appendix B. ∎

By substituting (9) and (10) into (8), the average network throughput can be obtained as

$$\begin{aligned}\bar{R} &= WBP^c\eta(\log_2(Q_1(\alpha)) - \log_2(Q_2(\alpha)) - 3) \\ &\quad + WB(1-P^c\eta)\log_2(1 + \frac{PD^{-\alpha}}{B\sigma^2} Q_1(\alpha).\end{aligned} \tag{11}$$

2) Average User Data Rate: Since only one N-Coop user and one Coop user (if exists) are active in a cluster each time, with round robin scheduling, the average data rates of N-Coop and Coop users can be respectively obtained from (9) and (10) as follows

$$\begin{aligned}\bar{R}_u^n &= W(1-\eta)\mathbb{E}\{\frac{BR_i^n}{N^n}\} \overset{(a)}{\approx} \frac{WB(1-\eta)\bar{R}_i^n}{\bar{N}^n} \\ &= \frac{WB(1-\eta)}{\bar{N}^n}\log_2(1 + \frac{PD^{-\alpha}}{B\sigma^2} Q_1(\alpha)), \\ \bar{R}_u^c &= W\eta\mathbb{E}\{\frac{BR_i^c}{N^c}\} \overset{(b)}{\approx} \frac{WB\eta\bar{R}_i^c}{\bar{N}^c} \\ &= \frac{WB\eta(\log_2(Q_1(\alpha)) - \log_2(Q_2(\alpha)) - 3)}{\bar{N}^c},\end{aligned} \tag{12}$$

where (a) and (b) come from the fact that R_i^n and N^n are independent random variables and the same to R_i^c and N^c, thus $\mathbb{E}\{R_i^n/N^n\} = \mathbb{E}\{R_i^n\}\mathbb{E}\{1/N^n\} \approx \mathbb{E}\{R_i^n\}/\mathbb{E}\{N^n\} = \bar{R}_i^n/\bar{N}^n$ according to (A.4) and $\mathbb{E}\{R_i^c/N^c\} \approx \bar{R}_i^c/\bar{N}^c$ analogically. $\bar{N}^c = \mathbb{E}\{N^c\}$ and $\bar{N}^n = \mathbb{E}\{N^n\}$ are the average numbers of Coop users and N-Coop users, respectively.

Proposition 3: The average number of Coop users is

$$\bar{N}^c = \sum_{\Phi_\mathcal{N}} \prod_{i=1}^{B} \frac{K! \prod_{k=1}^{K_0} (P_k)^{n_{ik}}}{\prod_{j=1}^{K_0} n_{ij}!} \sum_{k=1}^{K}\sum_{i=1}^{B} \zeta(k) n_{ik}, \tag{13}$$

which can be approximated as

$$\bar{N}^c \approx \bar{N}_1^c + \bar{N}_2^c, \tag{14}$$

where $\zeta(k)$, n_{ik}, $\bar{N}_1^c$ and $\bar{N}_1^c$ are defined in Appendix C.

Proof: See Appendix C. ∎

Though we can use similar way to derive the average number of N-Coop users $\bar{N}^n$ as for $\bar{N}^c$, the resulting expression is complicated. Considering that $N^n + N^c + N^b = M$, we can obtain $\bar{N}^n$ by deriving the average number of cellular users $\bar{N}^b = \mathbb{E}\{N^b\}$. Since all requests follow a Zipf distribution independently, the number of users that can not fetch files via D2D is a random variable following a Binomial distribution and $N^b \sim B(M, 1 - \sum_{k=1}^{K} P_k)$. Therefore, $\bar{N}^b = M(1 - \sum_{k=1}^{K} P_k)$. Then, the average number of N-Coop users is

$$\bar{N}^n = M - \bar{N}^c - \bar{N}^b. \tag{15}$$

With the average number of Coop and N-Coop users $\bar{N}^c$ and $\bar{N}^n$, we can obtain the corresponding average user data rate $\bar{R}_u^c$ and $\bar{R}_u^n$ using (12).

3) Joint Optimization of η and K: The bandwidth allocation factor and cluster size that maximize the average network throughput under the constraint of average user data rate can be optimized from the following problem

$$\begin{aligned}\max_{\eta,K} \quad & \bar{R} \\ s.t. \quad & \bar{R}_u^c \geq \mu, \quad \bar{R}_u^n \geq \mu, \\ & 0 \leq \eta \leq 1, \quad KB = M.\end{aligned} \tag{16}$$

Since the number of users per cluster K is an integer, we can find the global optimal solution by first finding optimal η for any given K and then enumerating K until the value of $\bar{R}$ computed by (11) achieves the maximum under the two constraints.

By taking the derivative of $\bar{R}$ in (8) with respect to η, we have $\frac{\partial \bar{R}}{\partial \eta} = WBP^c(\bar{R}_i^c - \bar{R}_i^n)$. The constraints $\bar{R}_u^c \geq \mu$ and $\bar{R}_u^n \geq \mu$ can be respectively rewritten as $\eta \leq \frac{WB\bar{R}_i^n - \mu\bar{N}^n}{WB\bar{R}_i^n}$ and

[3]To achieve the data rate, for the non-cooperative D2D users, each DR requires the CSI from its corresponding DT to it for decoding, whereas the DT does not need any CSI. For the cooperative D2D users, each DR requires the CSI from the DTs that jointly transmitting data to it, whereas each DT needs the CSI from itself to its DR. After gathering the CSI from each cooperative DT, the BS computes the precoding matrix and informs the cooperative DTs.

$\eta \geq \frac{\bar{N}^c \mu}{WB\bar{R}_i^c}$ according to (12). Therefore, if $\bar{R}_i^c \geq \bar{R}_i^n$, $\bar{R}$ is an increasing function of η, then the optimal solution of problem (16) for any given K is $\eta_K^* = \frac{WB\bar{R}_i^n - \mu\bar{N}^n}{WB\bar{R}_i^n}$; otherwise, $\bar{R}$ is a decreasing function of η, and $\eta_K^* = \frac{\bar{N}^c \mu}{WB\bar{R}_i^c}$.

Proposition 4: If $\bar{I}_i \geq B\sigma^2$, then $\bar{R}_i^c \geq \bar{R}_i^n$.

Proof: See Appendix D. ■

Since the power of interference among N-Coop users $\bar{I}_i$ defined in (A.1) is much larger than noise variance σ^2 in D2D communications and B is finite, the condition $\bar{I}_i \geq B\sigma^2$ in Proposition 4 is easy to be satisfied. Consequently, the optimal value of η for a given K is

$$\eta_K^* = 1 - \frac{\mu\bar{N}^n}{WB\bar{R}_i^n}, \tag{17}$$

which only depends on the average data rate of N-Coop link $\bar{R}_i^n$ and the number of N-Coop users $\bar{N}^n$.

The global optimal solution K and η can be found by one-dimensional searching and η^* is with closed-form expression. Hence the solution is with low complexity. Because K^* and η^* depend on W, μ, B, M, N_f, β and N, which are usually fixed for a long time, they are unnecessary to be updated frequently.

IV. Simulation and Numerical Results

In this section, we validate the analysis by comparing simulation and numerical results, and demonstrate the performance of the proposed opportunistic cooperation strategy by simulation.

In the simulation, we consider a square hotspot area with the side length $D_c = 100$ m, where $M = 180$ users are randomly located. Such a setting reflects relatively high user density as in [30], where more than one user is located within an area of $10\times10\ m^2$. We first consider uniformly distributed non-mobile users and then show the impact of non-uniform distribution and user mobility. The path-loss model is $37.6+36.8\log_{10}(r)$ [19]. Each user is with transmit power $P = 23$ dBm. $W = 20$ MHz, and $\sigma^2 = -100$ dBm. The file catalog size $N^f = 300$, and each user is willing to cache $N = 10$ files. The parameter of Zipf distribution is $\beta = 0 \sim 1$. The user data rate constraint is $\mu = 1 \sim 2$ Mbps. All results are obtained via 100 Monte-Carlo trails, where in each trail the request and location of each user and the Rayleigh fading channel are generated independently. This setup is used in the sequel unless otherwise specified.

A. Impact of Cluster Size and Number of Total Users

In Fig. 4, we provide numerical results of the average number of active Coop users $\bar{N}^a$ obtained from (5) and simulation results for the average sum rate of the active Coop users versus the number of users per cluster K. It is shown from Fig. 4(a) that $\bar{N}^a$ increases with β. This is because with large β, a few popular files are requested by majority of users, which leads to high cooperative probability P^c. With the increase of K, $\bar{N}^a$ first increases and then decreases. It is shown from Fig. 4(b) that with the growth of K, the average sum rate of active Coop users exhibits the same trends with $\bar{N}^a$, which agrees to the analysis in section III.A.

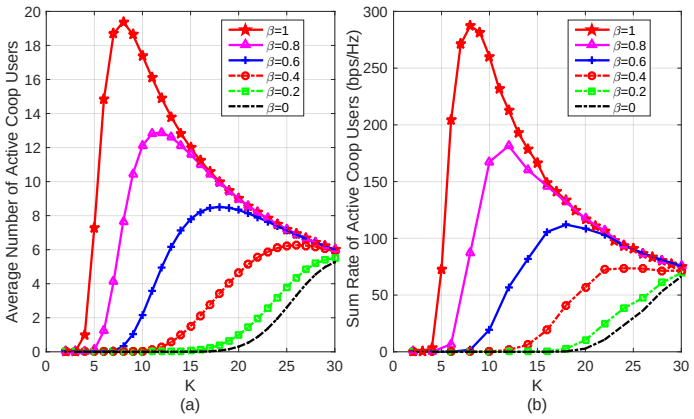

Fig. 4. $\bar{N}^a$ and average sum rate of active Coop users versus K.

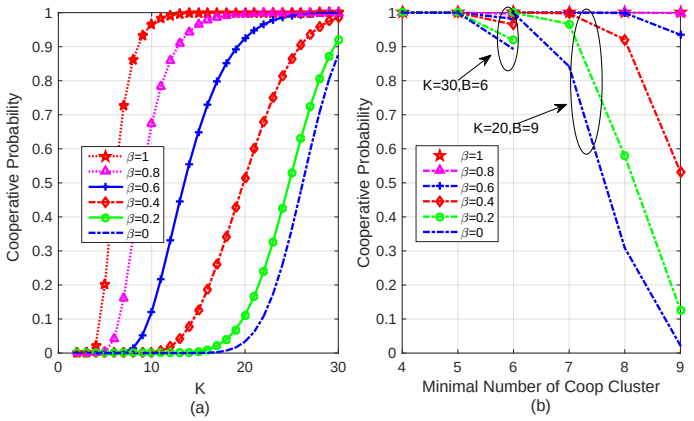

Fig. 5. P^c of (a) full cooperative, and (b) partial cooperation.

In Fig. 5, we simulate the cooperative probability P^c. It is shown from Fig. 5(a) that P^c increases with K, which agrees with the analysis after (4). Moreover, the cooperative probability is high although the full cooperation is allowed only when all clusters hit the same file group, especially when β is not small, say $\beta > 0.4$. In Fig. 5(b), we show the impact of partial cooperation by changing the minimal number of clusters allowed to cooperate.[4] As expected, P^c can be improved if we allow partial cooperation among clusters, but not significant. Therefore, the throughput gain from partial cooperation over full cooperation is marginal for the non-mobile users as to be illustrated later.

In Fig. 6, we provide numerical results of the optimal cluster size K^* and the resulting average number of active Coop users $\bar{N}^a$ for different number of total users in the hotspot area M. When $M = 1000$, there are 10 users in an area of 10×10 m^2, corresponding to a very high traffic load [30]. It is shown from Fig. 6(a) that K^* decreases with β as expected. With the growth of M, K^* first increases and then approaches a constant that equals to $K_0 = N^f/N$. This is because when

[4]We set such a minimal number because when we allow fewer clusters to cooperate, the multiplexing gain will reduce despite that P_c will be higher, and then the resulting throughput will reduce. When this minimal number of clusters is six, it becomes the full cooperation strategy.

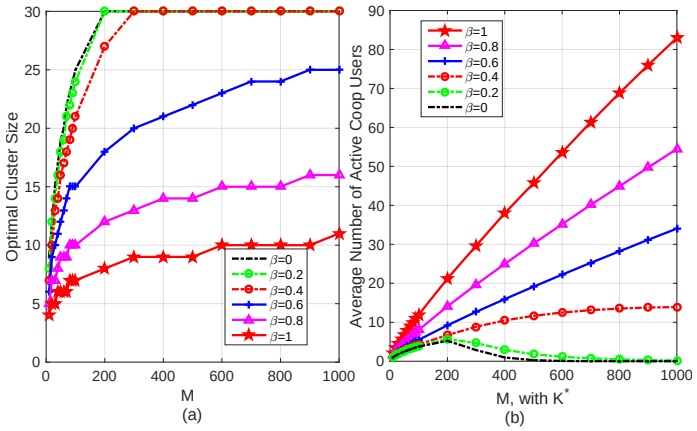

Fig. 6. K^* and $\bar{N}^a$ versus the number of total users in the hotspot M.

$K = K_0$, all files in the catalog can be cached at the users in each cluster. Assigning more than K_0 users to each cluster can not increase the number of Coop users. It is shown from Fig. 6(b) that with the growth of M, the average number of active Coop users monotonously increases when β is large but first increases and then decreases when β is small. This implies that if the user density is large but β is small, the proposed opportunistic cooperation strategy may be not useful.

B. Accuracy of the Approximations

In Fig. 7, we evaluate the accuracy of all the approximations used in deriving the throughput and the number of users by comparing simulation and numerical results. The numerical results of $\bar{R}_i^c$ and $\bar{R}_i^n$ in Fig.7(a) are respectively obtained from (9) and (10) by changing the path loss exponent α from 2 to 4. The numerical results of the number of Coop users $\bar{N}^c$ and the number of N-Coop users $\bar{N}^n$ in Fig. 7(b) are respectively obtained from (14) and (15). We can see from simulation and numerical results that the approximations are accurate, except the average data rate per unit bandwidth per second of the Coop link when α is large, which comes from the first order approximation in (A.5).

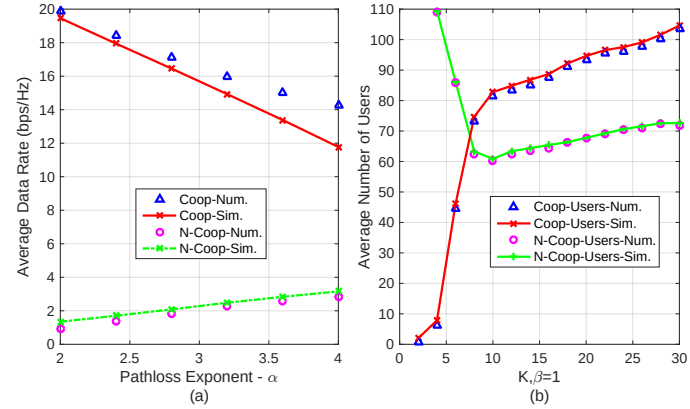

Fig. 7. Average data rate per unit bandwidth per second and number of users. "Num.": Numerical results, "Sim.": Simulation results.

Besides, we can see from in Fig.7(a) that with the growth of α, the average data rate per unit bandwidth per second of the Coop link decreases but that for the N-Coop link increases. This is because both signal and interference power decrease when α increases, but the interference power decreases more rapidly due to larger distance of interference link than that of signal link.

As expected, when K increases, the number of Coop users increases due to the decrease of the number of clusters, as shown in Fig. 7(b). However, the number of N-Coop users first decreases and then increases slowly with K. This is because with the growth of K, more files can be cached in each cluster and thus the total number of D2D users increases, and the number of N-Coop users changes as in (15).

C. Optimal Bandwidth Allocation and Network Throughput

In Fig. 8(a), we present the optimal solution of problem (16) η^* versus the Zipf parameter β. We can see that η^* increases with β. This is because the number of interference-free links, $\bar{N}^a$, increases with β, and then allocating more bandwidth to Coop users can increase the network throughput. η^* decreases as μ increases, because more bandwidth is needed for N-Coop users to support higher user data rate.

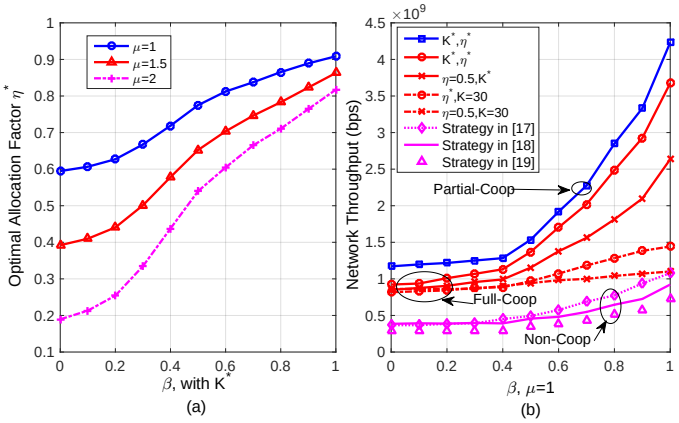

Fig. 8. η^* and average network throughput versus β.

In Fig. 8(b), we provide simulation results for the maximal network throughput. Several relevant strategies for cache-enabled D2D in existing literature serve as the baselines for comparison. The communication protocols used in [18] and [19] are similar to ours, which are also cluster-centric but without cooperation. The caching policy optimized in [17] is user-centric that maximizes cache hit ratio, without considering communication protocol. For a fair comparison, we set the maximal distance allowed to establish D2D communication in [17] equal to the cluster side length in our paper. Then, we simulate the average network throughput achieved by the optimal caching policies in [17], [18] and [19] all without cooperation. In the legends, "$\eta = 0.5, K = 30$" refers to a cooperation strategy without optimizing K and η, which allocates equal bandwidth to Coop users and N-Coop users. "$\eta^*, K = 30$" refers to a cooperation strategy without optimizing K but only optimizing η. "$\eta = 0.5, K^*$" refers to a

cooperation strategy without optimizing η but only optimizing K. We can see that optimizing the bandwidth allocation becomes necessary when $\beta > 0.5$, while optimizing the cluster size is always necessary but the gain from optimization grows with β. With K^* and η^*, the throughput gain over the baseline for $\beta = 1$ is $400\% \sim 500\%$, which demonstrates that the proposed opportunistic cooperation strategy can boost the network throughput remarkably. Even when $\beta = 0$, where file popularity follows a uniform distribution, the throughput gain is still $60\% \sim 80\%$. We also demonstrate the performance of "Partial-Coop", i.e., allowing cooperation among less than B clusters. It is shown that the performance can be improved by allowing partial cooperation, but the gain is marginal.

D. Impact of Non-uniform User Distribution

To show the impact of non-uniform user positions, in this subsection we consider a scenario where the users are distributed according to the Neyman-scott cluster processes, which is a typical Poisson cluster process used to model non-uniform distribution [31]. As illustrated in Fig. 9, the centers of the circle areas follow a homogeneous Poisson process with density $\lambda_c = 3$, and the radius of each circle area D_R is set as 15 or 25 m in simulation. In each circle area, the users are distributed uniformly and the average number of users per cluster is $N_R = 60$. Hence, the average number of users in such a non-uniform hotspot area is 180, which is identical to previous setting.

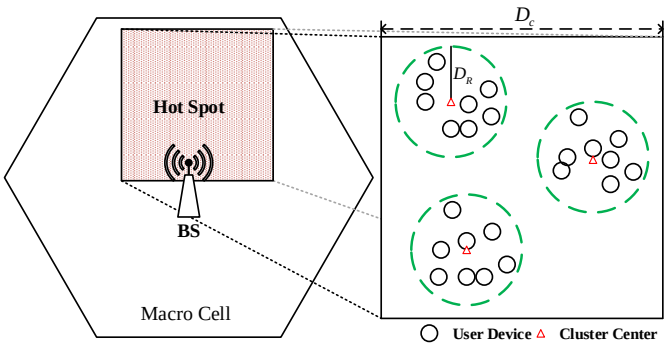

Fig. 9. Illustration for user positions distributed according to the Neyman-scott cluster process.

In Fig. 10, we provide simulation results of the network throughput with the proposed cooperative strategy (with the optimal number of users per cluster K^* and optimal bandwidth allocation factor η^* obtained under the assumption of uniform user's locations) and the non-cooperative strategy in [18] (i.e., with $\eta = 0$) for different user distribution. We do not compare with other related strategies considering that they perform similarly. We can see that compared to uniformly distributed users, the average network throughput achieved by the proposed strategy is higher when the users are non-uniformly distributed, but the gain over non-cooperative strategy is lower. This is because for non-uniform user distribution, the average distance of D2D link is smaller and then the desired signal becomes stronger, whereas the distance between clusters increases and hence the inter-cluster interference becomes

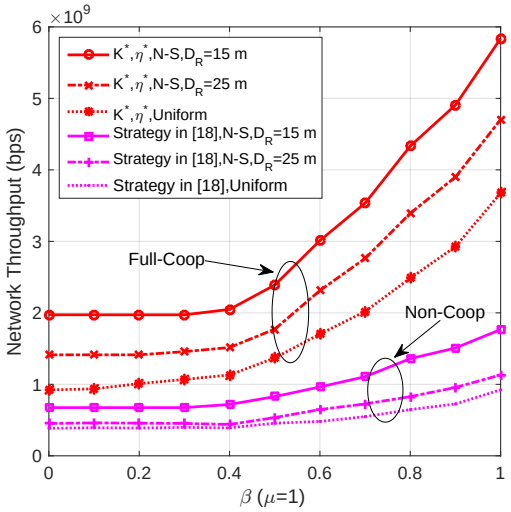

Fig. 10. Average network throughput for different distribution versus β. "N-S" refers to Neyman-scott cluster processes and "Uniform" is for the case where all users are uniformly distributed in the hotspot area.

weaker. In addition, with the decrease of the radius of each cluster D_R, the average network throughput increases for the same reason.

E. Impact of User Mobility

To show the impact of user mobility on the performance of the cooperative strategy, we consider a frequently used model, *random walk mobility* model, where a user moves from its current location to a new location by randomly choosing a direction and speed to travel [32]. In each trail of the simulation, we consider 10 minutes duration, within which the users do not generate new requests, and the speed and direction are uniformly chosen from $[0, 2]$ m/s and $[0, 2\pi]$, respectively. Each user moves 60 seconds before changing direction and speed. The users are initially uniformly distributed. The files cached at each user are determined according to the original uniform user location. Due to the mobility, some users may depart a cluster. Then, the number of users in each cluster is no longer identical, and the distribution of user location may no longer uniform. To keep the overall number of users fixed in the hotspot area, we consider wraparound as in [18], i.e, the leftmost cluster and the rightmost cluster are neighbors and so do the uppermost cluster and lowermost cluster in Fig. 1.

Once the DT or DR of a D2D link departs a cluster due to mobility, the D2D connection is broken and the DR will ask the BS to transmit the remaining data if it can not fetch the file from another D2D link. Then, the BS needs to re-coordinate the cooperative transmission strategy, and update the cached files in each user.[5] To illustrate the necessity of cache replacement for the optimized full cooperation strategy, we consider two caching policies. One is not to update the cached files, which simply ignores the user mobility. The other updates the cached files using the following way: when the 1st user in a cluster (i.e., the user cached with the most popular

[5]We do not consider the overhead required for establishing the D2D links and coordinating the cooperative transmission.

N files) departs the cluster, the cached files at the Kth user will be replaced by the 1st file group (i.e., the most popular N files).

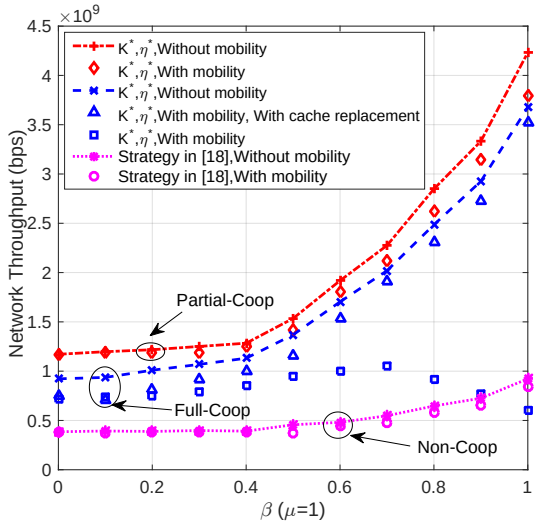

Fig. 11. Average network throughput versus β with or without user mobility.

In Fig. 11, we present simulation results of the network throughput versus β for the scenarios with and without user mobility. The impact of mobility on the performance of the strategy in [18] is negligible, which agrees to the simulation results in [18]. For the "Full-Coop" strategy without cache update, when β is high, the number of clusters is large and thus the cooperative probability reduces significantly due to mobility. For the "Full-Coop" strategy with cache replacement, however, even though K^* and η^* are optimized under the assumption of fixed position of users, the performance loss due to user mobility is small, especially for large value of β. Besides, the negative impact of mobility on the performance of "Partial-Coop" is also small despite that K^* and η^* are obtained with the assumption of fixed user location. This suggests that the proposed strategy is resilient to mobility by introducing cache replacement or partial cooperation.

V. Conclusions

In this paper, we proposed an opportunistic cooperation strategy for cache-enabled D2D communications. We jointly optimized the cluster size and the bandwidth allocated to Coop and N-Coop users to maximize the network throughput with minimal user data rate constraint. Simulation results showed that the proposed strategy can boost the throughput even when the content popularity follows a uniform distribution, and the gain over existing strategies is remarkable when the popularity distribution is more skewed.

Appendix A
Proof of Proposition 1

Without cooperation, the closest DT in the same cluster of its corresponding DR delivers the requested file to the DR, which treats the inter-cluster interference as noise when decoding the desired signal. The signal to interference plus noise ratio (SINR) at the DR of the active N-Coop link in the ith cluster can be expressed as

$$\gamma_i^n = \frac{P|h_{ii}|^2 r_{ii}^{-\alpha}}{I_i + \sigma^2}, \tag{A.1}$$

where P is the transmit power, σ^2 is the variance of white Gaussian noise, $I_i = P\sum_{j=1,j\neq i}^{B} r_{ij}^{-\alpha}|h_{ij}|^2$ is the total power of inter-cluster interference, h_{ij} and r_{ij} are respectively the channel coefficient and distance between the DT and the DR (h_{ii} in (A.1) is the specially case for h_{ij} when $j = i$), α is the path loss exponent, and both the interference channel coefficient h_{ij} $(i \neq j)$ and the desired channel coefficient h_{ii} follow a complex Gaussian distribution with zero mean and unit variance.

Due to the short distance between D2D links, the D2D network is interference-limited and hence the noise can be ignored, i.e., $I_i \gg \sigma^2$.

Then, from (A.1) the data rate per unit bandwidth per second for the N-Coop link in the ith cluster is $R_i^n = \log_2(1 + \frac{P|h_{ii}|^2 r_{ii}^{-\alpha}}{I_i})$.

Considering that $|h_{ij}|^2$ follows an Exponential distribution, which is a special case of the Gamma distribution, the interference power I_i (which is a sum of random variables following a Gamma distribution) can be approximated as a Gamma distribution [33]. Further consider that for a Gamma distributed random variable X with parameters k and θ, $\mathbb{E}\{\ln(X)\} = \psi(k) + \ln(\theta)$, where $\psi(k)$ is the Digamma function [34]. Then, the average data rate per unit bandwidth per second of the N-Coop link can be obtained according to Proposition 9 in [33] as

$$\mathbb{E}_{\mathbf{h}}\{R_i^n\} \approx \log_2(1 + \frac{P r_{ii}^{-\alpha}}{\bar{I}_i}), \tag{A.2}$$

where $\mathbb{E}_{\mathbf{h}}\{\cdot\}$ represents the expectation taken over small scale channel fading, $\bar{I}_i = P\sum_{j=1,j\neq i}^{B} r_{ij}^{-\alpha}$ is the average total power of the inter-cluster interference.

Since channel fading and user location are distributed independently, the average data rate per unit bandwidth per second of the N-Coop link taken over both channel fading and user location can be obtained as

$$\bar{R}_i^n = \mathbb{E}_{\mathbf{p}}\{\log_2(1 + \frac{P r_{ii}^{-\alpha}}{\bar{I}_i})\}, \tag{A.3}$$

where $\mathbb{E}_{\mathbf{p}}\{\cdot\}$ denotes the expectation taken over user location.

Because the joint probability density function (pdf) of the distances among D2D users is hard to obtain, we introduce the first order approximation to derive the expression of $\bar{R}_i^n$. Specifically, for a random variable X, the expectation of a function of X, $\varphi(X)$, can be approximated as [35]

$$\begin{aligned}\mathbb{E}\{\varphi(X)\} &= \mathbb{E}\{\varphi(\mu_x + X - \mu_x)\} \\ &\approx \mathbb{E}\{\varphi(\mu_x) + \varphi'(\mu_x)(X - \mu_x)\} \\ &= \varphi(\mu_x),\end{aligned} \tag{A.4}$$

where $\mu_x = \mathbb{E}\{X\}$ and the approximation is accurate when the variance of X is small. With this approximation, $\bar{R}_i^n$ in (A.3) can be approximated as

$$\bar{R}_i^n \approx \log_2(\mathbb{E}_{\mathbf{p}}\{P r_{ii}^{-\alpha} + \bar{I}_i\}) - \log_2(\mathbb{E}_{\mathbf{p}}\{\bar{I}_i\}). \tag{A.5}$$

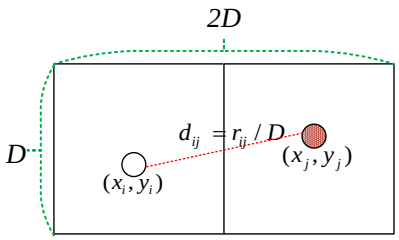

Fig. 12. The distance of users in two adjacent cluster

The pdf of the signal link distance r_{ii} can be obtained from [36] by variable substitution $r = r_{ii}/D$ as

$$g(r) = \frac{1}{D}\begin{cases} 2r(r^2-4r+\pi), & 0 \le r < 1 \\ 8r\epsilon - 2r(r^2+2) & \\ +4r(\arcsin(\frac{1}{r}) & \\ -\arccos(\frac{1}{r})), & 1 \le r < \sqrt{2} \end{cases}. \tag{A.6}$$

To simplify the analysis, the interference link distance r_{ij} is assumed to have the same distribution $f(r)$, where $r = r_{ij}/D$. We can derive the pdf of the interference link distance r_{ij} as follows.

Denote the position of the DR in the ith cluster as (x_i, y_i) and the position of the DT in the jth cluster as (x_j, y_j), as illustrated in Fig. 12. The link distance between them normalized by the cluster side D can be expressed as $d_{ij} = \frac{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}}{D} = \sqrt{(\Delta x)^2+(\Delta y)^2}$, where $\Delta x = \frac{x_i-x_j}{D}$ and $\Delta y = \frac{y_i-y_j}{D}$. The pdf of $|\Delta y|$ can be obtained according to [36] as

$$p_{\Delta y}(v) = \begin{cases} 2(1-v), & 0 \le v \le 1 \\ 0, & \text{otherwise} \end{cases}. \tag{A.7}$$

Analogically, the pdf of $|\Delta x|$ can be obtained as

$$p_{\Delta x}(u) = \begin{cases} 1-|1-u|, & 0 \le u \le 2 \\ 0, & \text{otherwise} \end{cases}. \tag{A.8}$$

Then, the cumulative probability distribution function (cdf) of d_{ij} is

$$\begin{aligned} F_d(r) &= \mathbb{P}\{d_{ij} \le r\} = \mathbb{P}\{\sqrt{|\Delta x|^2+|\Delta y|^2} \le r\} \\ &= \iint dudv p_{\Delta x,\Delta y}(u,v) \\ &\overset{(a)}{=} \iint dudv p_{\Delta x}(u) p_{\Delta y}(v), \end{aligned} \tag{A.9}$$

where (a) comes from the fact that $|\Delta x|$ and $|\Delta y|$ are independent random variables.

When $0 \le r \le 1$, from (A.8) and (A.7), we can obtain $p_{\Delta x,\Delta y}(u,v) = 2u(1-v)$, where $0 \le u, v \le 1$. Then, the cdf in (A.9) can be derived as

$$\begin{aligned} F_d(r) &= \iint dudv p_{\Delta x,\Delta y}(u,v) \\ &= \int_0^r dv \int_0^{\sqrt{r^2-v^2}} 2u(1-v)du \\ &= \frac{2}{3}r^3 - \frac{1}{4}r^4, \end{aligned} \tag{A.10}$$

When $1 \le r \le \sqrt{2}$, we have

$$p_{\Delta x,\Delta y}(u,v) = \begin{cases} 2u(1-v), & 0 \le u \le 1, 0 \le v \le 1 \\ 2(2-u)(1-v), & 1 \le u \le 2, 0 \le v \le 1 \\ 0, & \text{otherwise} \end{cases},$$

and the corresponding cdf can be derived as

$$\begin{aligned} &F_d(r) = \\ &\int_0^{\sqrt{r^2-1}} dv \left(\int_0^1 2u(1-v)du + \int_1^{\sqrt{r^2-v^2}} 2(2-u)(1-v)du \right) \\ &+ \int_{\sqrt{r^2-1}}^1 dv \int_0^{\sqrt{r^2-v^2}} 2u(1-v)du \\ &= \frac{5}{4} + \epsilon\left(-2r^2 + (1+r^2)\epsilon + \frac{2}{3}\epsilon^2 - \frac{1}{2}\epsilon^2\right) \\ &+ r^2\left(2\sin^{-1}(\frac{3}{r}) + \frac{1}{2} - \frac{4}{3}r\right), \end{aligned} \tag{A.11}$$

where $\epsilon \triangleq \sqrt{r^2-1}$.

Analogically, when $\sqrt{2} \le r \le 2$, the cdf can be derived as

$$\begin{aligned} F_d(r) &= \int_0^1 dv \int_0^1 2u(1-v)du \\ &+ \int_0^1 dv \int_1^{\sqrt{r^2-v^2}} 2(2-u)(1-v)du \\ &= -\frac{11}{12} + 2\epsilon - \frac{r^2}{2} + 2r^2\sin^{-1}(\frac{1}{r}) + \frac{4}{3}(\epsilon^3 - r^3). \end{aligned} \tag{A.12}$$

When $2 \le r \le \sqrt{5}$, the cdf can be derived as

$$\begin{aligned} F_d(r) &= 1 - \int_{\sqrt{r^2-4}}^1 dv \int_{\sqrt{r^2-v^2}}^2 2u(1-v)du \\ &= -\frac{45}{12} - \frac{r^2}{2} - \frac{r^4}{2} - 2r^2\left(\sin^{-1}(\frac{\xi}{r}) - \sin^{-1}(\frac{1}{r})\right) \\ &+ r^2\xi + 2\epsilon - \frac{1}{3}\xi^3 + \frac{4}{3}\epsilon^3 + \frac{1}{4}\xi, \end{aligned} \tag{A.13}$$

where $\xi \triangleq \sqrt{r^2-4}$.

Finally, by combining (A.10) (A.11) (A.12) and (A.13), and considering the pdf $f(r) = \frac{dF(r)}{dr}$, the pdf of the interference link distance r_{ij} is

$$f(r) = \frac{1}{D}\begin{cases} 2r^2 - r^3, & 0 \le r < 1 \\ 2r - 4r^2 + 2r^3 & \\ -2r\epsilon + \frac{2r}{\epsilon} & \\ -\frac{2r^3}{\epsilon} + 4r\arcsin(\frac{\epsilon}{r}), & 1 \le r < \sqrt{2} \\ 4r\epsilon + 4r\arcsin(\frac{1}{r}) & \\ -r - 4r^2, & \sqrt{2} \le r < 2 \\ -5r - r^3 + 4r\epsilon & \\ -4r\arcsin(\frac{\xi}{r}) - \arcsin(\frac{1}{r}) & \\ -\frac{4r}{\xi} + r\xi + \frac{r^3}{\xi}, & 2 \le r < \sqrt{5} \end{cases} \tag{A.14}$$

where $r = r_{ij}/D$, $\epsilon \triangleq \sqrt{r^2-1}$, and $\xi \triangleq \sqrt{r^2-4}$.

Since the interference generated by DTs far away from the DR can be ignored due to pathloss, we only consider dominant

interference generated from the nearest eight clusters around the ith cluster as shown in Fig. 1. Then, by substituting the pdf of interference and signal link distance into (A.5), we can obtain the average data rate per unit bandwidth per second of the N-Coop link in the ith cluster as

$$\begin{aligned}\bar{R}_i^n &\approx \log_2\left(\int_0^{\sqrt{2}D} Pr_{ii}^{-\alpha} g(\frac{r_{ii}}{D})dr_{ii}\right.\\ &\left.+8\int_0^{\sqrt{5}D} Pr_{ij}^{-\alpha} f(\frac{r_{ij}}{D})dr_{ij}\right)\\ &-\log_2\left(8\int_0^{\sqrt{5}D} Pr_{ij}^{-\alpha} f(\frac{r_{ij}}{D})dr_{ij}\right)\\ &=\log_2\left(\int_0^{\sqrt{2}} Pr^{-\alpha} g(r)dr + 8\int_0^{\sqrt{5}} Pr^{-\alpha} f(r)dr\right)\\ &-\log_2\left(8\int_0^{\sqrt{5}} Pr^{-\alpha} f(r)dr\right)\\ &=\log_2(Q_1(\alpha)) - \log_2(Q_2(\alpha)) - 3,\end{aligned} \tag{A.15}$$

This proves Proposition 1.

Appendix B
Proof of Proposition 2

In *Mode 1*, the cooperative DTs jointly transmit the requested files to the Coop users with zero-forcing beamforming, which is of low complexity and hence practical. Then, the SINR at the DR of the active Coop link in the ith cluster can be expressed as

$$\begin{aligned}\gamma_i^c &= \frac{P\|\mathbf{h}_i\|^2\delta_i}{\sigma^2}\\ &\approx \frac{P\sum_{j=1}^B r_{ij}^{-\alpha}|h_{ij}|^2}{B\sigma^2},\end{aligned} \tag{B.1}$$

where $\mathbf{h}_i = [\sqrt{r_{i1}^{-\alpha}}h_{i1}, \sqrt{r_{i2}^{-\alpha}}h_{i2}, ..., \sqrt{r_{iB}^{-\alpha}}h_{iB}]$ is the composite channel vector between all DTs and the DR, $0 \le \delta_i \le 1$, and a larger value of δ_i indicates a better orthogonality between $\mathbf{h}_i$ and $\mathbf{h}_j$ for $i \ne j$. The approximation comes from the fact $\delta_i \approx (BN^t - B + 1)/B = 1/B$ [37], where N^t is the number of antennas per DT and $N^t = 1$ in this paper. This approximation is accurate when the variance of δ_i is small.

Using the same approximation as deriving (A.2), the average data rate per unit bandwidth per second of the Coop link is obtained as

$$\bar{R}_i^c \approx \mathbb{E}_{\mathbf{p}}\{\log_2(1 + \frac{P\sum_{j=1}^B r_{ij}^{-\alpha}}{B\sigma^2})\}. \tag{B.2}$$

By applying the first-order approximation in (A.4), using (A.6) and (A.14), and only considering dominant signal, we can obtain the average data rate per unit bandwidth per second for the Coop link as

$$\begin{aligned}\bar{R}_i^c &\approx \log_2\left(B\sigma^2 + 8\int_0^{\sqrt{5}D} Pr_{ij}^{-\alpha} f(\frac{r_{ij}}{D})dr_{ij}\right.\\ &\left.+\int_0^{\sqrt{2}D} Pr_{ii}^{-\alpha} g(\frac{r_{ii}}{D})dr_{ii}\right) - \log_2\left(B\sigma^2\right)\\ &=\log_2(1 + \frac{PD^{-\alpha}}{B\sigma^2}Q_1(\alpha)).\end{aligned} \tag{B.3}$$

This proves Proposition 2.

Appendix C
Proof of Proposition 3

Denote the number of users in the ith cluster who request files in $\mathcal{G}_k$ as n_{ik}, $1 \le k \le K_0$, $1 \le i \le B$.

Since the users request files independently, the probability that the combination of the numbers of users in the ith cluster who request files in each file group is $\{n_{i1}, n_{i2}, ..., n_{iK_0}\} \triangleq \mathcal{N}_i$ can be derived as

$$\begin{aligned}p_{\mathcal{N}_i} &= \prod_{m=1}^{K_0} C_{K-\sum_{j=1}^{m-1} n_{ij}}^{n_{im}} \prod_{k=1}^{K_0}(P_k)^{n_{ik}}\\ &\overset{(a)}{=} \frac{K!\prod_{k=1}^{K_0}(P_k)^{n_{ik}}}{\prod_{j=1}^{K_0} n_{ij}!},\end{aligned} \tag{C.1}$$

where (a) comes from $C_n^m C_{n-m}^k = \frac{n!}{m!k!(n-m-k)!}$.

Only when all B clusters hit a file group $\mathcal{G}_k$ and $k \le K$ (i.e., $n_{ik} > 0$ is satisfied for $k \le K$ and any i, $1 \le i \le B$), the users requesting the files within $\mathcal{G}_k$ are Coop users, and we call $\mathcal{G}_k$ a *hit file group*.

The number of Coop users for all hit file groups can be obtained as

$$N^c = \sum_{k=1}^{K}\sum_{i=1}^{B}\zeta(k)n_{ik}, \tag{C.2}$$

where $\zeta(k) = \lceil\sum_{i=1}^B u(n_{ik}) - B + 1)\rceil^+$ $(k \le K)$ indicates that whether $\mathcal{G}_k$ is a *hit file group*, the function $u(x) = 1$ when $x > 0$, otherwise $u(x) = 0$, and $\lceil\Lambda\rceil^+ = \max(\Lambda, 0)$.

Denote $\mathcal{N} = \{\mathcal{N}_1, \mathcal{N}_2, ..., \mathcal{N}_B\}$, and $\Phi_{\mathcal{N}} = \{\mathcal{N}|n_{ik} \ge 0, \sum_{k=1}^{K_0} n_{ik} = K\}$ represents a set of all possible combinations of $\mathcal{N}$. Then, by taking average of N^c in (C.2) over $\Phi_{\mathcal{N}}$, we can derive the average number of Coop users as

$$\begin{aligned}\bar{N^c} &= \sum_{\Phi_{\mathcal{N}}}\prod_{i=1}^{B} p_{\mathcal{N}_i} N^c\\ &= \sum_{\Phi_{\mathcal{N}}}\prod_{i=1}^{B}\frac{K!\prod_{k=1}^{K_0}(P_k)^{n_{ik}}}{\prod_{j=1}^{K_0} n_{ij}!}\sum_{k=1}^{K}\sum_{i=1}^{B}\zeta(k)n_{ik}.\end{aligned} \tag{C.3}$$

The number of Coop users can be computed accurately using (C.3). However, the cardinality of $\Phi_{\mathcal{N}}$ is $K_0^{(K-1)B}$, and hence the computational complexity exponentially increases with K. For example, when $K = K_0 = 15$ and $B = 9$, we obtain $|\Phi_{\mathcal{N}}| = 1.35 \times 10^{140}$.

In the sequel, we seek an alternative solution.

Noticing that the probability that multiple *hit file groups* exist simultaneously decreases with the growth of the number

of hit file groups, we only consider the case where only one or two hit file groups exist to approximate the average number of Coop users as follows. This approximation is accurate when the number of clusters is large.

The probability that a *hit file group* only contains $\mathcal{G}_k$ can be obtained from (3) as $\prod_{j=1,j\neq k}^{K}(1-(P_j^h)^B)(P_k^h)^B$. As a result, the probability of $n_{ik}=m$ $(1\leq m\leq K)$ when a *hit file group* only contains $\mathcal{G}_k$ is $C_K^m(P_k)^m(1-P_k)^{(K-m)}/P_k^h$. Then, the average number of Coop users in the cases where only one hit file group exists can be derived as

$$\begin{aligned}\bar{N}_1^c=&\sum_{k=1}^{K}\prod_{j=1,j\neq k}^{K}(1-(P_j^h)^B)(P_k^h)^B\cdot\\ &B\sum_{m=1}^{K}\frac{C_K^m(P_k)^m(1-P_k)^{(K-m)}}{P_k^h}m\\ =&\sum_{k=1}^{K}\prod_{j=1,j\neq k}^{K}(1-(P_j^h)^B)(P_k^h)^{B-1}BP_kK.\end{aligned}\tag{C.4}$$

The probability that there only exist two *hit file groups*, $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$, can be obtained from (3) as $\prod_{j=1,j\neq k_1,k_2}^{K}(1-(P_j^h)^B)(P_{k_1}^h)^B(P_{k_2}^h)^B$. Consequently, the probability of $n_{ik_1}=m_1$ and $n_{ik_2}=m_2$ (where $1\leq m_1\leq K$, $1\leq m_2\leq K$ and $2\leq m_1+m_2=m\leq K$) when the *hit file groups* only contain $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$ is $C_m^{m_1}C_{m-m_1}^{m_2}\frac{(p_{k_1})^{m_1}(p_{k_2})^{m_2}}{P_{k_1}^hP_{k_2}^h}$. Then, the average number of Coop users in the case where only two hit file groups exist can be derived as

$$\begin{aligned}\bar{N}_2^c=&\sum_{\Phi_{k_1k_2}}\prod_{j=1,j\neq k_1,k_2}^{K}\left(1-(P_j^h)^B\right)(P_{k_1}^h)^B(P_{k_2}^h)^B\cdot\\ &B\sum_{m=2}^{K}\sum_{\Phi_{m_1m_2}}C_m^{m_1}C_{m-m_1}^{m_2}\frac{(p_{k_1})^{m_1}(p_{k_2})^{m_2}}{P_{k_1}^hP_{k_2}^h}m\\ =&\sum_{\Phi_{k_1k_2}}\prod_{k=1,k\neq i,j}^{K}\left(1-(P_k^h)^B\right)(P_i^hP_j^h)^{B-1}\cdot\\ &B\sum_{m=2}^{K}\sum_{\Phi_{m_1m_2}}\frac{mm!}{m_1!m_2!}(p_{k_1})^{m_1}(p_{k_2})^{m_2},\end{aligned}\tag{C.5}$$

where $\Phi_{k_1k_2}=\{k_1,k_2|0\leq k_1,k_2\leq K,i\neq j\}$, $\Phi_{m_1m_2}=\{m_1,m_2|m_1+m_2=m\}$, whose cardinality are K^2 and m^2 $(m\leq K)$, respectively.

By combining (C.4) and (C.5), the average number of Coop users is approximated as

$$\bar{N}^c\approx\bar{N}_1^c+\bar{N}_2^c,\tag{C.6}$$

which can be obtained much easier than (C.3).

This proves Proposition 3.

Appendix D
Proof of Proposition 4

By simply subtracting $\bar{R}_i^n$ from $\bar{R}_i^c$ and considering $\bar{I}_i\gg\sigma^2$, we can obtain that

$$\begin{aligned}\bar{R}_i^c-\bar{R}_i^n=&\mathbb{E}\{\log_2(1+\frac{P\sum_{j=1}^{B}r_{ij}^{-\alpha}}{B\sigma^2})\}\\ &-\mathbb{E}\{\log_2(1+\frac{Pr_{ii}^{-\alpha}}{\bar{I}_i})\}\\ =&\mathbb{E}\{\log_2(\frac{\bar{I}_i}{B\sigma^2})\}.\end{aligned}$$

When $\log_2(\frac{\bar{I}_i}{B\sigma^2})\geq 0$ (i.e., $\bar{I}_i\geq B\sigma^2$), $\bar{R}_i^c-\bar{R}_i^n\geq 0$.

This proves Proposition 4.

References

[1] 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, 2009.

[2] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: a survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1923–1940, 2014.

[3] A. T. Gamage, H. Liang, R. Zhang, and X. Shen, "Device-to-device communication underlaying converged heterogeneous networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 98–107, 2014.

[4] 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.*, vol. 21, no. 2, pp. 57–65, 2014.

[5] X. Lin, J. 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, 2014.

[6] Y. Zhang, L. Song, W. Saad, Z. Dawy, and Z. Han, "Exploring social ties for enhanced device-to-device communications in wireless networks," in *IEEE GLOBECOM*, 2013.

[7] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 67–80, 2015.

[8] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, 2014.

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

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

[11] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[12] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, 2016.

[13] B. D. Higgins, J. Flinn, T. J. Giuli, B. Noble, C. Peplin, and D. Watson, "Informed mobile prefetching," in *ACM MobiSys*, 2012.

[14] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2894–2905, 2014.

[15] B. Chen and C. Yang, "Performance gain of precaching at users in small cell networks," in *IEEE PIMRC*, 2015.

[16] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE ICC*, 2015.

[17] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *IEEE ICC*, 2016.

[18] N. Golrezaei, P. Mansourifard, A. Molisch, and A. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, 2014.

[19] M. Ji, G. Caire, and A. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, 2016.

[20] H. Elkotby, K. Elsayed, and M. Ismail, "Exploiting interference alignment for sum rate enhancement in D2D-enabled cellular networks," in *IEEE WCNC*, 2012.

[21] Y. Cao, T. Jiang, and C. Wang, "Cooperative device-to-device communications in cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 124–129, 2015.

[22] C. Ma, G. Sun, X. Tian, K. Ying, Y. Hui, and X. Wang, "Cooperative relaying schemes for device-to-device communication underlaying cellular networks," in *IEEE GLOBECOM*, 2013.

[23] Y. Pei and Y. Liang, "Resource allocation for device-to-device communications overlaying two-way cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3611–3621, 2013.

[24] A. Liu and V. K. Lau, "Mixed-timescale precoding and cache control in cached MIMO interference network," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6320–6332, 2013.

[25] Q. Sun, L. Tian, Y. Zhou, J. Shi, and X. Wang, "Energy efficient incentive resource allocation in D2D cooperative communications," in *IEEE ICC*, 2015.

[26] K. Doppler, C.-H. Yu, C. Ribeiro, and P. Janis, "Mode selection for device-to-device communication underlaying an LTE-advanced network," in *IEEE WCNC*, 2010.

[27] Y. Jiang and X. You, "Research of synchronization and training sequence design for cooperative D2D communications underlaying hyper-cellular networks," in *IEEE ICC*, 2013.

[28] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *IEEE INFOCOM*, 1999.

[29] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, 2000.

[30] 3GPP, "TR 22.891 feasibility study on new services and markets technology enablers," 2016.

[31] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[32] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wiley Online Library Wireless commun. and mobile comput.*, vol. 2, no. 5, pp. 483–502, 2002.

[33] R. W. Heath, T. Wu, Y. H. Kwon, and A. C. Soong, "Multiuser MIMO in distributed antenna systems with out-of-cell interference," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4885–4899, 2011.

[34] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1964, vol. 55.

[35] A. Papanicolaou, *Taylor approximation and the delta method*, 2009. [Online]. Available: http://web.stanford.edu/class/cme308/OldWebsite/notes/TaylorAppDeltaMethod.pdf

[36] L. E. Miller, "Distribution of link distances in a wireless network," *J. Res. Natl. Inst. Stand. Technol.*, vol. 106, no. 2, pp. 401–412, 2001.

[37] Q. Zhang and C. Yang, "Transmission mode selection for downlink coordinated multipoint systems," *IEEE Trans. Veh. Technol.*, vol. 62, no. 1, pp. 465–471, 2013.