This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

Caching D2D Connections in Small-Cell Networks

Nan Zhao, *Senior Member, IEEE,* Xiaonan Liu, *Student Member, IEEE,* Yunfei Chen, *Senior Member, IEEE,* Shun Zhang, *Member, IEEE,* Zan Li, *Senior Member, IEEE,* Bingcai Chen, *Member, IEEE,* and Mohamed-Slim Alouini, *Fellow, IEEE*

***Abstract*—Small-cell network is a promising solution to high video traffic. However, with the increasing number of devices, it cannot meet the requirements from all users. Thus, we propose a caching device-to-device (D2D) scheme for small-cell networks, in which caching placement and D2D establishment are combined. In this scheme, a limited cache is equipped at each user, and the popular files can be prefetched at the local cache during off-peak period. Thus, dense D2D connections can be established during peak time aided by these cached users, which will reduce the backhaul pressure significantly. To do this, first, an optimal caching scheme is formulated according to the popularity to maximize the total offloading probability of the D2D system. Thus, most edge users can obtain their required video files from the caches at users nearby, instead from the small-cell base station. Then, the sum rate of D2D links is analyzed in different signal-to-noise ratio (SNR) regions. Furthermore, to maximize the throughput of D2D links with low complexity, three D2D-link scheduling schemes are proposed with the help of bipartite graph theory and Kuhn-Munkres algorithm for low, high and medium SNRs, respectively. Simulation results are presented to show the effectiveness of the proposed scheme.**

***Index Terms*—Bipartite graph theory, caching, device-to-device, Kuhn-Munkres algorithm, link scheduling, small-cell networks.**

I. Introduction

Driven by the blooming of mobile devices and their explosive demands for multimedia services, data traffic is expected to increase exponentially in the next decade. To satisfy the ever-increasing demands for video traffic, small-cell networks will be widely deployed in the future wireless systems [1]–[4]. Nonetheless, owing to the dense arrangement of small-cells, the high backhaul cost becomes a fundamental challenge. Recently, it has been reported that most of the video traffic is caused by duplicate downloads of some popular files [5], [6]. Therefore, the problem may be solved by storing popular video files at the caches of small-cell base stations (SBSs) or users during off-peak period, which can be fetched directly from the local caches without backhaul at peak time [7]–[9].

Using local caching, the backhaul congestion and transmission latency can be reduced significantly, and the throughput of small-cell networks can be increased accordingly [10], [11]. In [12], femto-caching was proposed by Shanmugam *et al.* for small-cell networks, where the files cached at the SBSs are optimized in a centralized manner to reduce the transmission delay. Yang *et al.* proposed and analyzed the cache-based content delivery in a three-tier heterogeneous network [13], where base stations (BSs), relays, and device-to-device (D2D) pairs are included. In [14], Taghizadeh *et al.* proposed the cooperative caching policies to minimize the electronic content provisioning cost in social wireless networks. Xu and Tao investigated coded caching in a large-scale small-cell network where the locations of SBSs are modeled by stochastic geometry in [15]. In [16], Han and Ansari investigated the traffic load balancing in backhaul-constrained cache-enabled small-cell networks powered by hybrid energy sources. In our previous works, we have also studied the interference management and power allocation of the caching aided small-cell networks [17]–[20].

On the other hand, as the number of users increases, interference will appear, which will degrade the quality of service (QoS) severely. Consequently, SBSs with caching cannot satisfy the transmission rate of all the devices simultaneously, especially the edge users [1], [21]. Thus, D2D communications can be utilized at the edge of the small cells, which enables two close users to exchange data directly without the help of SBSs [22]–[25]. For the D2D communications of cellular networks, they can be divided into underlay and overlay modes [26]–[28]. In the underlay D2D systems, cellular users and D2D users can share the same spectral resource; while in the overlay D2D systems, D2D links are allocated dedicated spectrum. With the help of D2D, both the spectrum efficiency and energy efficiency of small-cell networks can be improved, the data traffic of SBSs and the transmission delay can be reduced, and the congestion of the backhaul can be alleviated [29]–[31]. In [32], the outage probability of D2D-communication-enabled cellular networks was effectively studied from a general threshold-based perspective by Liu *et al*.

In D2D networks, popular video files can also be cached at users during off-peak period or after watching [33]–[35]. In [33], the redundancy of user requests as well as the storage capacity of devices were exploited by Golrezaei *et al.*, to increase the throughput of video files in cellular networks. In [34], Ji *et al*. proposed a novel caching scheme for a D2D based small-cell network, in which the throughput was improved compared to that of the traditional approach of unicasting from cellular BSs. To dramatically reduce the energy consumption of SBSs and the economical cost of service providers, joint design of transmission and caching was studied by Gregori *et al*. in [35], where caching was performed either at SBSs, or directly at devices. Furthermore, in D2D networks, data placement

N. Zhao, X. Liu and B. Chen are with the School of Inform. and Commun. Eng., Dalian University of Technology, Dalian, 116024, P. R. China (email: zhaonan@dlut.edu.cn, liuxiaonan@mail.dlut.edu.cn, china@dlut.edu.cn).

Y. Chen is with the School of Engineering, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: Yunfei.Chen@warwick.ac.uk).

S. Zhang and Z. Li are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, P. R. China. (Email: zhangshunsdu@xidian.edu.cn, zanli@xidian.edu.cn).

Mohamed-Slim Alouini is with the EE program, King Abdullah University of Science and Technology, Thuwal, Mekkah Province, Saudi Arabia (e-mail: slim.alouini@kaust.edu.sa).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

and delivery are the two main stages of wireless caching, which have attracted great attentions [36]–[39]. Popular files are stored in the caches during off-peak time at the data placement stage; while at the data delivery stage, the users who have cached the required files will perform transmission to the receivers. Particularly, the optimal caching placement via maximizing the density of successful receptions was studied by Malak *et al.* in [36]. Meanwhile, in [37], according to different criteria, some realistic network models were utilized by Zhou *et al*. to describe the stochastic natures of geographic location, and the corresponding optimal caching placements were derived. Moreover, the scaling behavior of the throughput with the number of devices under Zipf distributed request with a different value γ was analyzed by Golrezaei *et al.* in [38]. In addition, in [39], a user preference aware caching deployment algorithm was proposed for D2D caching networks by Zhang *et al.*, to achieve significant improvement on cache hit ratio.

Different from the above-mentioned research works, in this paper, an optimal caching placement scheme is developed for the D2D links of small-cell networks to maximize the total offloading probability of video files. Moreover, to maximize the throughput, three D2D-link establishing schemes are proposed based on the optimal caching scheme for different SNR regions, respectively. The main contributions of this paper are summarized as follows.

- To the best of our knowledge, the video-file placement and delivery of caching D2D links in small-cell networks have not been combined effectively before. In this paper, we propose a caching D2D scheme for small-cell networks, in which the optimal D2D links can be established, with the total offloading probability of the cached video files maximized.
- According to the popularity, an optimal caching problem is first formulated to maximize the total offloading probability of the D2D system, and its closed-form solutions are derived. Thus, edge users can obtain their required video files from the caches at users nearby through D2D links, instead from the SBS.
- To maximize the throughput of the network, the sum rate of D2D links is then analyzed at low SNR, high SNR and medium SNR, respectively, which is the basis for the D2D-link establishing schemes.
- Furthermore, to maximize the throughput with low complexity, three D2D-link scheduling schemes are proposed for different SNR regions. At low SNR, D2D links are established with the help of bipartite graph theory and Kuhn-Munkres (KM) algorithm, assuming that the interference can be ignored. At medium SNR, the bipartite graph theory and KM algorithm are also combined to solve the problem, with the interference assumed to be a constant. At high SNR, a distributed algorithm is designed to construct D2D connections, with the QoS of the current and previous established D2D links guaranteed.

The rest of the paper is organized as follows. In Section II, the system model is presented. Optimal caching scheme to maximize the total offloading probability is proposed in Section III. In Section IV, the sum rate of caching D2D links is analyzed theoretically. In Section V, caching D2D-link scheduling schemes are proposed for different SNR regions. Simulation results are shown in Section VI, followed by the conclusions and future work in Section VII.

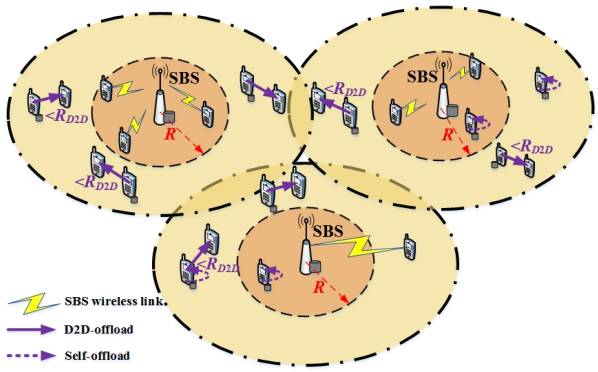

Fig. 1. Dense caching D2D connections in small-cell networks.

Notation: $\mathcal{CN}(\mathbf{a}, \mathbf{A})$ stands for the complex Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{A}$. $\mathbb{E}(\cdot)$ denotes expectation.

II. System Model

We consider a heterogeneous network with several small cells, and $\mathcal{K}$ users exist in each small cell[1]. Each SBS is located at the center of its corresponding small cell, and the users are spatially distributed according to homogeneous Poisson Point Process (PPP) with density λ_D, as shown in Fig. 1. For simplicity, we mainly focus on one of the small cells, because the interference among small cells can be neglected due to the long distance and low transmit power. A large cache is equipped at the SBS, which stores N video files. The popularity probability of each video file follows the Zipf's law [40], and can be modeled as

$$q_i = \frac{1/i^{\gamma}}{\sum_{j=1}^{N} 1/j^{\gamma}}, i = 1, 2, ..., N, \tag{1}$$

where i indexes the ith ranked video file according to a descending order of popularity, γ is the coefficient that controls the popularity distribution of video files. Meanwhile, taking the difference of user requirements into consideration, a scalable video coding (SVC) based method is adopted to encode each video file [41], in which each video is divided into a base layer (BL) and an enhancement layer (EL). Basic quality of a video can be guaranteed by the BL, and the additional EL can further improve its quality. Therefore, when users need low-definition video files, BLs can meet their basic requirements. However, for users who request higher video quality, both BLs and ELs should be delivered to satisfy their demands.

Due to the limited SBS transmit power, the large-scale fading from the SBS to users and the high traffic load from the core network, the QoS of users far away from the SBS

[1]Our proposed schemes can also be extended for a macro-cell with caching D2D users.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

may degrade severely. To solve this problem, users who have already cached the required the video files can use D2D transmission to improve the QoS of edge users in this paper. Assume that the effective transmission range of the SBS is R, which means that within R, the users can be served by the SBS directly, and the QoS can be guaranteed when the transmission rate of a specific cellular user is larger than a threshold. In addition, due to the limited transmit power of D2D transmitters and the interference from other D2D users, we assume that a D2D link can be established only if the distance between the D2D transceivers is less than R_{D2D}, which can be derived when the transmission rate of a specific D2D user is larger than a threshold. Due to the fact that the transmit power of the SBS is much higher than that of the D2D transmitters, we can obtain that $R \gg R_{D2D}$.

Assume that all the users are equipped with limited caches equal to μ, which means each device can store at most μ video files, and only m ($N \gg m > \mu$) different kinds of video files can be selected to be stored in the local caches of the users. In addition, $\mathcal{P}_i$ is the proportion of users caching the ith video file, where $0 \leq \mathcal{P}_i \leq 1$. The cache storage constraint for all the users can be denoted as

$$\sum_{i=1}^{N} \mathcal{P}_i \leq m. \tag{2}$$

Thus, the users caching the ith video file follow a PPP with density $\lambda_D \mathcal{P}_i$ as well.

The active D2D users are assumed to perform in the overlay D2D mode with half-duplex transmission [26]. That is, the transmission between the SBS and cellular users shares a specific frequency band, while a different frequency band is allocated for the D2D transmission to avoid interference between the cellular users and D2D users. As indicated in Fig. 1, the video transmission protocols of the network can be summarized as follows.

- *Self-offloading:* When a user requires a video file, it first checks its own cache. The request will be satisfied immediately if the required video exists in its local cache. Meanwhile, the user can still perform transmission to other users if it can be served by itself.
- *D2D-offloading:* If the required video file cannot be found in its local cache, the user will search other users nearby within radius R_{D2D}. If the video exists in at least one users, a proper D2D link can be established to meet the request.
- *SBS-offloading:* If the required video file cannot be found either in the user's own cache or from the other users nearby, the SBS will transmit the file to this user. In this case, even though the channel fading between the SBS and the user may be severe, the SBS has to transmit the requested video in response.

Remark 1: According to the above analysis, for a specific edge user out of the range of R from the SBS in the small-cell network, it should first try to find a potential D2D transmitter nearby that has cached the required file to establish the D2D link with high transmission rate. If no D2D transmitters have cached the required file, the edge user has to obtain the video file from the SBS directly with distance longer than R, and it can only receive the video file with a much lower rate. Thus, an optimal caching scheme exists, which will be discussed in the next section, to maximize the total offloading probability.

III. Optimal Caching for D2D Transmission

In this section, in order to improve the experience of edge users and alleviate the traffic load of small-cell networks, the optimal caching scheme that maximizes the total offloading probability is proposed. First, the offloading probability of the caching D2D links is analyzed. Then, the optimal caching placement problem is formulated and solved.

A. Offloading Probability for Video Streaming

For a homogeneous PPP distribution, the probability that K active users exist in an area with radius r can be denoted as

$$\mathcal{P}_{K,r,\lambda} = \frac{(\pi r^2 \lambda)^K}{K!} e^{-\pi r^2 \lambda}, \tag{3}$$

where λ is the density of the PPP distribution.

For an active edge user, its required video file can be cached at its own memory, at the nearby users or at the SBS, according to the transmission protocols in Section II. Thus, the probability that the ith video file has been stored at at least one another user's caches within R_{D2D} can be written as

$$\mathcal{P}_{i,D2D} = 1 - P_{0,R_{D2D},\lambda_D \mathcal{P}_i} = 1 - e^{-\pi \lambda_D \mathcal{P}_i R_{D2D}^2}, \tag{4}$$

where $P_{0,R_{D2D},\lambda_D P_i}$ is the probability that none of the users have cached the ith video file within R_{D2D}. Thus, the probability that the user can obtain the ith video file directly from its own memory or be served by the users nearby can be presented as

$$\begin{aligned} \mathcal{P}_{i,download} &= \mathcal{P}_i + (1 - \mathcal{P}_i)\mathcal{P}_{i,D2D} \\ &= 1 - e^{-\pi \lambda_D \mathcal{P}_i R_{D2D}^2} + \mathcal{P}_i e^{-\pi \lambda_D \mathcal{P}_i R_{D2D}^2}. \end{aligned} \tag{5}$$

Furthermore, since the ith video file is requested by the user with the popularity q_i and its caching probability is determined by $\mathcal{P}_{i,download}$, the total offloading probability for the caching D2D links can be expressed as

$$\mathcal{P}_{total} = \sum_{i=1}^{N} q_i \mathcal{P}_{i,download}. \tag{6}$$

Using the caching D2D links, the more data are transmitted by D2D users, the less data traffic will need to be delivered via S-BSs. Thus, the traffic load from the backhaul can be alleviated and the QoS of edge users can be improved effectively.

B. Optimal Caching Placement Problem

According to (6), to maximize the offloading probability of D2D links, the optimization problem of caching placement can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

be formulated as

$$(P1) \max_{\mathcal{P}_1,\mathcal{P}_2,\ldots,\mathcal{P}_N} \mathcal{P}_{total} = \sum_{i=1}^{N} q_i \mathcal{P}_{i,download} \quad s.t. \quad 0 \le \mathcal{P}_i \le 1, i = 1, 2.., N, \quad \sum_{i=1}^{N} \mathcal{P}_i \le m. \tag{7}$$

This maximization problem is equivalent to a minimization problem (P1) as (8) (on the next page). To obtain the closed-form solution of (P1), Lemma 1 and Theorem 1 are provided.

Lemma 1: (P1) is a convex optimization problem.

Proof: According to (5), the second-order derivative of $\mathcal{P}_{i,download}$ can be derived as

$$\frac{\partial^2 \mathcal{P}_{i,download}}{\partial^2 \mathcal{P}_i} = -(1-\mathcal{P}_i)a^2 e^{-a\mathcal{P}_i} - 2ae^{-a\mathcal{P}_i} < 0, \tag{9}$$

where $a = \pi\lambda_D R_{D2D}^2 > 0$. We can observe that $\mathcal{P}_{i,download}$ is convex in $\mathcal{P}_i$, and $q_i\mathcal{P}_{i,download}$ is also convex in $\mathcal{P}_i$. Thus, (P1) is a convex optimization problem. ■

Based on Lemma 1, the water-filling method can be utilized to calculate the closed-form solution of (P1), which is given in Theorem 1.

Theorem 1: The approximate optimal probability for caching placement of the ith video file can be expressed as

$$\mathcal{P}_i = \min\left\{\left(\frac{a - \ln\frac{u}{q_i}}{2a}\right)^+, 1\right\}, i = 1, 2, \ldots, N, \tag{10}$$

where $x^+ = \max(x, 0)$. u can be calculated through replacing the expression of $\mathcal{P}_i$ in the following equation using bisection search.

$$\sum_{i=1}^{N} \mathcal{P}_i - m = 0. \tag{11}$$

Proof: According to Lemma 1, we can obtain the following equation using Lagrangian.

$$\mathcal{L} = -\sum_{i=1}^{N} q_i \left(1 - e^{-\pi\lambda_D \mathcal{P}_i R_{D2D}^2} + \mathcal{P}_i e^{-\pi\lambda_D \mathcal{P}_i R_{D2D}^2}\right) + u\left(\sum_{i=1}^{N} \mathcal{P}_i - m\right), \tag{12}$$

where u is the Lagrange multiplier. The Karush-Kuhn-Tucker (KKT) condition for the optimization of caching placement can be denoted as

$$\frac{\partial \mathcal{L}}{\partial \mathcal{P}_i} = -q_i\left(ae^{-a\mathcal{P}_i} + e^{-a\mathcal{P}_i} - a\mathcal{P}_i e^{-a\mathcal{P}_i}\right) + u = 0. \tag{13}$$

According to (13), we can conclude that

$$\ln\frac{u}{q_i} = \ln(a + 1 - a\mathcal{P}_i) - a\mathcal{P}_i. \tag{14}$$

(14) is a transcendental equation, and its closed-form solution is difficult to get. With the help of Taylor expansion, we have

$$\ln(a + 1 - a\mathcal{P}_i) \approx (a - a\mathcal{P}_i) + o(a - a\mathcal{P}_i), \tag{15}$$

where $o(a - a\mathcal{P}_i) \to 0$. Thus, (14) can be approximately rewritten as

$$a - 2a\mathcal{P}_i = \ln\frac{u}{q_i}. \tag{16}$$

According to (16), we have

$$\mathcal{P}_i = \frac{a - \ln\frac{u}{q_i}}{2a}, \tag{17}$$

based on which, we can obtain the expression of $\mathcal{P}_i$ in (10). The multiplier u can be calculated through replacing the expression of $\mathcal{P}_i$ in (11) using bisection search. ■

Using the optimal caching probability derived in Theorem 1, the offloading probability of D2D connections in the network can be maximized. In Sections IV and V, three D2D-link scheduling schemes are proposed to be utilized in different regions of SNR, based on the optimal caching placement.

IV. Sum Rate Analysis of Caching D2D Links

Based on the optimal caching placement scheme, some video files whose $\mathcal{P}_i$ is high are very likely to be cached at the edge users in advance. When users need these videos, most of them can be satisfied by their local caches or the users nearby. Then, D2D communication can be utilized perform transmission. To maximize the throughput of the D2D network, the sum rate of D2D links is first analyzed in this section, which is the basis for the D2D-link establishing schemes.

We assume that the number of users that need video files is L, and $\widehat{\mathcal{L}}$ ($\widehat{\mathcal{L}} \ge L$) users can provide service to them. Thus, the number of D2D links to be established is L. In addition, each D2D transceiver is equipped with only one antenna. Without loss of generality, we assume that the 1st to the Lth D2D transmitters in $\widehat{\mathcal{L}}$ are selected to provide video service to the L D2D receivers, and especially, the video service of the kth D2D receiver is provided by the kth D2D transmitter, $k = 1, \ldots, L$. Thus, the desired signal at the kth D2D receiver from the kth D2D transmitter can be denoted as

$$y_k = h_{k,k}x_k + \sum_{n=1,n\neq k}^{L} h_{k,n}x_n + n_o, \tag{18}$$

where $h_{k,n} = \sqrt{\alpha_p^{[kn]}}\bar{h}_{k,n} = \sqrt{\beta d_{k,n}^{-a}}\bar{h}_{k,n}$ is the channel coefficient between the nth D2D transmitter and the kth D2D receiver, β indicates the channel power gain at the reference distance $d_0 = 1$ from the nth D2D transmitter to the kth D2D receiver, $d_{k,n}$ means the distance between the nth D2D transmitter and the kth D2D receiver, and a is the path-loss exponent. In addition, $\bar{h}_{k,n}$ is identically and independently distributed (i.i.d.) following $\mathcal{CN}(0, 1)$. x_k is the signal transmitted by the kth D2D transmitter for the kth D2D receiver, with transmit power $P = |x_k|^2$. n_o presents the additive white Gaussian noise (AWGN) at the D2D receivers, which follows $\mathcal{CN}(0, N_o)$. Thus, the signal-to-noise-plus-interference (SINR)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

$$(P2)\ \min_{\mathcal{P}_1,\mathcal{P}_2,...,\mathcal{P}_N} -\sum_{i=1}^{N} q_i P_{i,download} = -\sum_{i=1}^{N} q_i \left(1 - e^{-\pi\lambda_D P_i R_{D2D}^2} + P_i e^{-\pi\lambda_D P_i R_{D2D}^2}\right)$$
$$s.t.\quad 0 \le P_i \le 1, i = 1, 2.., N, \tag{8}$$
$$\sum_{i=1}^{N} P_i \le m.$$

at the kth D2D receiver can be denoted as

$$v_k = \frac{Pg(k,k)}{\sum\limits_{n=1,n\neq k}^{L} Pg(k,n) + N_o} = \frac{\Gamma g(k,k)}{\sum\limits_{n=1,n\neq k}^{L} \Gamma g(k,n) + 1}, \tag{19}$$

where

$$\Gamma = P/N_o \tag{20}$$

is the transmit SNR of D2D links, $g(k,n)$ is the channel power between the nth D2D transmitter and the kth D2D receiver, which can be written as

$$g(k,n) = |h_{k,n}|^2 = \alpha_p^{[kn]} \left|\bar{h}_{k,n}\right|^2 = \beta d_{k,n}^{-a} \left|\bar{h}_{k,n}\right|^2. \tag{21}$$

Thus, the sum rate of D2D links can be expressed as

$$R_{sum} = \sum_{k=1}^{L} \log_2(1+v_k) = \sum_{k=1}^{L} \log_2\left(1 + \frac{\Gamma g(k,k)}{\sum\limits_{n=1,n\neq k}^{L} \Gamma g(k,n) + 1}\right). \tag{22}$$

When all the channel gains $g(k,n)$ of the D2D links can be obtained, $\forall k = 1, 2, \ldots, L$ and $\forall n = 1, 2, \ldots, L$, the D2D links with the maximum sum rate can be established via exhaustive searching as follows.

$$R_{sum}^{max} = \arg\max_{s\in\mathcal{S}} R_{sum}^{[s]}, \tag{23}$$

where $\mathcal{S}$ is the set that contains all the available combinations of D2D links.

In fact, due to the fact that the exhaustive searching in (23) is a combinatorial optimization problem, its complexity is extremely high. Thus, we analyze the approximate sum rate for D2D links in different SNR regions in this section, which is the basis for the D2D-link establishing schemes.

A. Low SNR

When Γ is low, i.e., $\Gamma \to 0$, the interference between users $\sum_{n=1,n\neq k}^{L} \Gamma g(k,n)$ in (19) becomes negligible, compared with 1. Thus, the sum rate R_{sum} is only determined by the channel gain between each D2D transceiver, which can be reduced as

$$\widehat{R}_{sum} \approx \sum_{k=1}^{L} \log_2(1 + \Gamma g(k,k)). \tag{24}$$

To obtain the expectation of $\widehat{R}_{sum}$, we present Theorem 2 as follows.

Theorem 2: Due to the fact that $h_{k,k}$ is i.i.d., following $\mathcal{CN}\left(0, \alpha_p^{[kk]}\right)$, $\forall k = 1, 2, \ldots, L$, the expectation of the sum rate in (24) can be expressed as

$$\mathbb{E}\left[\widehat{R}_{sum}\right] = \sum_{k=1}^{L} \frac{\exp\left\{\frac{1}{2\alpha_p^{[kk]}\Gamma}\right\} E_1\left(\frac{1}{2\alpha_p^{[kk]}\Gamma}\right)}{\ln(2)}, \tag{25}$$

where

$$E_1(z) = \int_z^{\infty} t^{-1} \exp^{-t} dt. \tag{26}$$

Proof: Since the channel coefficient of the kth D2D link follows $\mathcal{CN}\left(0, \alpha_p^{[kk]}\right)$, $\forall k = 1, 2, \ldots, L$, the possibility density function (p.d.f) of $g(k,k)$ can be calculated as

$$f_{g(k,k)}(\hat{x}) = \frac{1}{2\alpha_p^{[kk]}} \exp\left\{-\frac{\hat{x}}{2\alpha_p^{[kk]}}\right\}. \tag{27}$$

Thus, the expectation of sum rate in (24) can be derived as

$$\begin{aligned}\mathbb{E}[\widehat{R}_{sum}] &= \sum_{k=1}^{L} \mathbb{E}[\log_2(1+\Gamma g(k,k))] \\ &= \sum_{k=1}^{L} \int_0^{\infty} \frac{1}{2\alpha_p^{[kk]}} \exp\left\{-\frac{\hat{x}}{2\alpha_p^{[kk]}}\right\} \log_2(1+\Gamma\hat{x}) d\hat{x} \\ &= \sum_{k=1}^{L} \frac{\exp\left\{\frac{1}{2\alpha_p^{[kk]}\Gamma}\right\} E_1\left(\frac{1}{2\alpha_p^{[kk]}\Gamma}\right)}{\ln(2)},\end{aligned} \tag{28}$$

where $E_1(z) = \int_z^{\infty} t^{-1} \exp^{-t} dt$ presents the exponential integral. ■

B. High SNR

When Γ is high and the D2D pairs are close to each other, the interference between users will be much larger than 1, i.e., $\sum_{n=1,n\neq k}^{L} \Gamma g(k,n) \gg 1$. Consequently, the sum rate R_{sum} is determined by the channel gain between all the D2D transmitters and receivers, and has no relationship with the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

channel noise. Thus, R_{sum} can be rewritten as

$$\widetilde{R}_{sum} \approx \sum_{k=1}^{L} \log_2 \left(1 + \frac{\Gamma g(k,k)}{\sum\limits_{n=1,n\neq k}^{L} \Gamma g(k,n)} \right) = \sum_{k=1}^{L} \log_2 \left(1 + \frac{g(k,k)}{\sum\limits_{n=1,n\neq k}^{L} g(k,n)} \right). \tag{29}$$

C. Medium SNR

In this case, the sum rate of D2D links can be directly expressed as (22).

D. Only One D2D Link

When there exists only one active D2D link to perform video transmission in the network, no interference will appear between users, and the corresponding transmission rate can be expressed as

$$R_1 = \log_2(1 + \Gamma g(k,k)). \tag{30}$$

Meanwhile, if several potential transmitters can provide service to this user, a proper D2D transmitter should be selected to perform transmission. When we want to obtain the optimal transmission rate, the D2D transmitter with the highest channel gain should be selected as follows.

$$\widehat{c} = \arg\max_{c\in\mathcal{Q}} g(k,c), \tag{31}$$

where $\mathcal{Q}$ is the set of D2D transmitters that can provide service to the D2D receiver in the network. We assume that the number of the transmitters in set $\mathcal{Q}$ is Q. Then, according to (31), the optimal transmission rate can be achieved. Furthermore, we can obtain the expectation of R_1 according to Theorem 3.

Theorem 3: As for the case of only one D2D link, with its channel i.i.d. and following $\mathcal{CN}\left(0, \alpha_p^{[kk]}\right)$, the expectation of R_1 can be derived as

$$\mathbb{E}[R_1] = \frac{\exp\left\{\frac{1}{2\alpha_p^{[kk]}\Gamma}\right\} E_1\left(\frac{1}{2\alpha_p^{[kk]}\Gamma}\right)}{\ln(2)}. \tag{32}$$

Proof: According to (30) and (31), we have

$$\mathbb{E}[R_1] = \mathbb{E}\left[\log_2\left(1 + \Gamma \max_{c\in\mathcal{Q}} g(k,c)\right)\right]. \tag{33}$$

Due to the optimal selection of D2D transmitters, we assume that $g(k,k)$ is the maximum channel gain among all the possible transceivers and the p.d.f of $g(k,k)$ is expressed as (27). Then, similar to (28), (32) can be calculated. ■

In this section, the sum rate of D2D links are analyzed theoretically in different SNR regions. To establish the D2D connections, we can use exhaustive searching in (23). Nevertheless, the computational complexity of (23) is extremely high, and thus, we will develop three effective D2D-link establishing schemes to construct the D2D links at low SNR, high SNR and medium SNR, respectively.

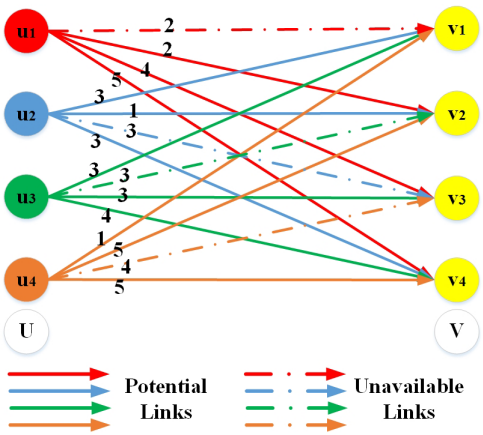

Fig. 2. A concrete example of D2D-link scheduling.

V. Caching D2D Link Scheduling

Due to the high computational complexity of the optimal D2D-link establishing method in (23), in this section, we propose three effective schemes to solve this problem at low SNR, high SNR and medium SNR. Furthermore, the schemes used to establish D2D links can be extended when the network topology changes due to mobility. Nevertheless, the mobility of users will not be further discussed, which is out of the scope of this paper.

Because the caching status of the available D2D transmitters is different, each D2D receiver can be served by several transmitters. We assume that the number of available D2D transmitters that can provide service is $\widehat{\mathcal{L}}$ ($\widehat{\mathcal{L}} \geq L$), and more than one video files are cached at each transmitter. Thus, a proper D2D transmitter should be selected to provide video service to each receiver optimally.

To maximize the sum rate of D2D links, bipartite graph theory is first utilized to include all the potential D2D links. We assume that $G(U,V,E)$ is the set of a potential link graph, where U and V are the vertex sets that present transmitters and receivers, respectively, E presents the links between the D2D transceivers. To avoid the large-scale fading that results from long distance between transceiver, we assume that a D2D link can be established only when the distance between the D2D transmitter and its corresponding receiver is less than R_{D2D}. Thus, E can be denoted as

$$E = \{uv | (u \in U, v \in V), ||u - v|| < R_{D2D}\}. \tag{34}$$

For example, we assume that there are four D2D receivers that require different video files as shown in Fig. 2, while four D2D transmitters can provide service to them. In this figure, the 1st D2D transmitter can perform transmission to the 2nd, 3rd and 4th D2D receivers, the 2nd D2D transmitter can provide video service to the 1st, 2nd and 4th D2D receivers, the 3rd D2D transmitter can provide service to the 1st, 3rd and 4th D2D receivers, and the 4th D2D transmitter can perform transmission to the 1st, 2nd and 4th D2D receivers. Meanwhile, the weights of potential D2D links are marked on each link. Thus, according to the potential D2D links shown in Fig. 2, each D2D receiver should be allocated a proper D2D

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

transmitter to maximize the sum rate of D2D links.

A. Low-SNR Scheme

When Γ is low, we know that the interference can be ignored according to (24), and each potential D2D link can be weighed only by the channel gain and SNR. Thus, to find a proper transmitter for each receiver, the topology of D2D links is first constructed through a weighted bipartite graph.

We consider that L D2D transmitters can provide service to L D2D receivers. Meanwhile, for the kth receiver, assume that B_k $(1 \leq B_k \leq L)$ transmitters can perform service to it. We adopt $\mathbf{Y}$ as a matrix to represent the caching status of the required video files, which can be expressed as

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1L} \\ y_{21} & y_{22} & \cdots & y_{2L} \\ y_{31} & y_{32} & \cdots & y_{3L} \\ \vdots & \vdots & \cdots & \vdots \\ y_{L1} & y_{L2} & \cdots & y_{LL} \end{bmatrix}, \tag{35}$$

where y_{ij} is a binary number. If $y_{ij} = 1$, it means that the requirement of the ith D2D receiver can be satisfied by the jth D2D transmitter; otherwise, $y_{ij} = 0$.

As for a D2D pair, each node works in half-duplex mode, and we define l_{ij} as the link between the jth transmitter to the ith receiver. Thus, the weight of l_{ij} can be expressed as

$$w_{ij} = \Gamma g(i,j). \tag{36}$$

We adopt matrix $\mathbf{W}$ to represent the weight of all the potential D2D links, which can be denoted as

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1L} \\ w_{21} & w_{22} & \cdots & w_{2L} \\ w_{31} & w_{32} & \cdots & w_{3L} \\ \vdots & \vdots & \cdots & \vdots \\ w_{L1} & w_{L2} & \cdots & w_{LL} \end{bmatrix}. \tag{37}$$

To achieve the maximum sum rate of D2D links, the optimization problem can be presented as

$$\begin{aligned} (P3) \quad & \max_{y_{ij}} \sum_{j \in \mathcal{S}_D} \sum_{i=1}^{L} w_{ij} y_{ij} \\ & s.t. \sum_{j \in \mathcal{S}_D} y_{ij} = 1, \forall i, \\ & \sum_{i=1}^{L} y_{ij} = 1, \forall j, \\ & y_{ij} \in (0,1), \end{aligned} \tag{38}$$

where $\mathcal{S}_D$ is the set of D2D transmitters that can provide service to the receivers. Thus, the sum-rate optimization problem is transformed into a maximum weighted matching (MWM) problem as in (38).

The optimal method to solve the MWM problem is exhaustive search, whose computational complexity is extremely high when plenty of users exist. For example, when all the required video files are cached at each D2D transmitter, the number of the available solutions to be searched can be calculated as

$$N_c = L! = L(L-1)(L-2)\cdots 1, \tag{39}$$

whose computational complexity can reach $O(L!)$. KM algorithm is effective to solve the MWM problem with much lower complexity of $O(L^3)$ [42], which is thus adopted to solve the problem (P3) in this paper. We present an example in Fig. 2 with $L = 4$ to demonstrate the proposed KM algorithm for D2D-link establishing as follows. The corresponding matrices in the procedure of the proposed KM algorithm are shown in Fig. 3.

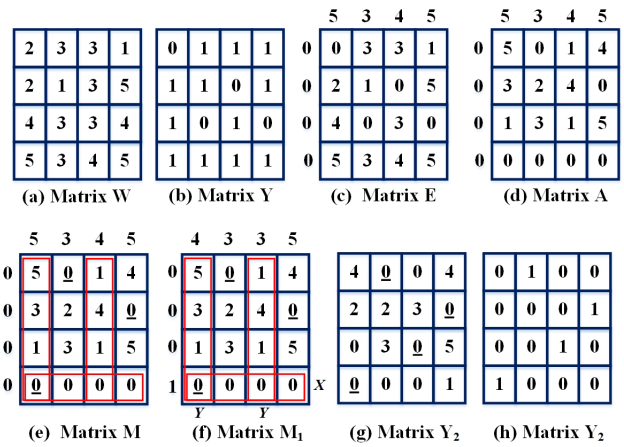

Fig. 3. Matrices in the procedure of the proposed KM algorithm.

Step 1: We use matrix $\mathbf{E} = \{e_{ij}\}_{L \times L}$ to include all the weights of possible D2D links, in which

$$e_{ij} = w_{ij} y_{ij}, \forall i = 1,2,...,L, \forall j = 1,2,...,L. \tag{40}$$

According to Fig. 2, the initial weight matrix $\mathbf{W}$ is shown in Fig. 3(a) and the initial matrix of caching situation $\mathbf{Y}$ is shown in Fig. 3(b). Through (40), we can obtain the weight matrix $\mathbf{E}$ accordingly, which is shown in Fig. 3(c).

Step 2: The KM algorithm begins with the labels of all the columns and rows, which can be calculated as

$$\xi(u_j) = \max_{v_i \in V} e_{ij}, \quad u_j \in U, \tag{41}$$

$$\widetilde{\xi}(v_i) = 0, \quad v_i \in V. \tag{42}$$

As shown in Fig. 3(c), the values of $\xi(u_j)$ and $\widetilde{\xi}(v_i)$ are marked at the top and the left of the matrix $\mathbf{E}$, respectively.

Step 3: With the help of $\xi(u_j)$ and $\widetilde{\xi}(v_i)$, an excess matrix $\mathbf{A} = \{a_{ij}\}_{L \times L}$ can be calculated as

$$a_{ij} = \xi(u_j) + \widetilde{\xi}(v_i) - e_{ij}, \tag{43}$$

which is shown in Fig. 3(d).

Step 4: From the excess matrix $\mathbf{A}$, we can obtain a subgraph with $a_{ij} = 0$ that contains u_j, v_i and the corresponding edge e_{ij}. Then, a maximum matching matrix $\mathbf{M}$ can be obtained from the subgraph as in Fig. 3(e), in which the established maximum matching edges are underlined. If the obtained matrix $\mathbf{M}$ is a perfect one with L underlined edges, the D2D-link scheduled problem is solved and the algorithm comes to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

an end. Otherwise, go to Step 5. In the specific case of Fig. 3(e), we can observe that there are only three underlined edges, and further calculation of Step 5 is needed.

Step 5: The matrix $\mathbf{M}$ is not a perfect one, and the labels should be adjusted. Define set C as the vertexes containing the transmitters that share the same receivers. In addition, we define $X = C \cap V$ and $Y = C \cap U$, as shown in Fig. 3(f). In the specific case of matrix $\mathbf{M}_1$ of Fig. 3(f), C is selected to be the vertexes related to receiver v_4 and transmitters u_1 and u_3. Meanwhile, a coefficient can be calculated as

$$\tau = \min\{a_{ij} | u_j \in Y, v_i \in X\}, \tau \neq 0. \tag{44}$$

Therefore, the labels at the top and the left of matrix $\mathbf{M}_1$ can be updated in Fig. 3(f) according to

$$\xi(u_j) = \begin{cases} \xi(u_j) - \tau, & u_j \in Y, \\ \xi(u_j), & \text{others.} \end{cases} \tag{45}$$

$$\widetilde{\xi}(v_i) = \begin{cases} \widetilde{\xi}(v_i) + \tau, & v_i \in X, \\ \widetilde{\xi}(v_i), & \text{others.} \end{cases} \tag{46}$$

Then, go to Step 3.

After several iterations, the KM algorithm terminates, and the maximum matching matrix $\mathbf{M}_2$ and optimal link-establishing matrix $\mathbf{Y}_2$ can be obtained as shown in Fig. 3(g) and Fig. 3(h), respectively.

B. High-SNR Scheme

When Γ is high, the transmission rate of each D2D link will be seriously influenced by the interference from other D2D transmitters. Based on (29), the weight of link l_{ij} can be expressed as

$$\bar{w}_{ij} = \frac{g(i,j)}{\sum\limits_{s=1, s\neq j}^{L} g(i,s)}. \tag{47}$$

Before establishing D2D links, $\bar{w}_{ij}$ cannot be calculated similarly to the case of low SNR, and thus, the KM algorithm cannot be utilized to schedule the D2D links. Moreover, through (29), we can also obtain that the interference at each D2D link is the sum of channel gains from other D2D transmitters. Thus, we propose a distributed algorithm to establish D2D links at high SNR, which can significantly reduce the signaling overhead caused by the channel state information feedback from the users to the SBS.

At the beginning of the algorithm, a receiver is randomly selected to schedule the first D2D link. Due to the fact that no interference exists for the first link, according to (31), the D2D transmitter with the maximum channel gain can be selected. Then, when constructing the remaining D2D links, interference will appear between D2D users. To guarantee the QoS of D2D links, the following two conditions should be satisfied when constructing the kth D2D link.

$$\frac{g(k,k)}{\sum\limits_{n \in \widehat{S}_{DD}, n \neq k} g(k,n)} \geq \delta_1, \tag{48}$$

$$\frac{g(m,m)}{\sum\limits_{n \in \widetilde{S}_{DD}, n \neq m} g(m,n)} \geq \delta_2, \ \forall m \in \widetilde{S}_{DD}, m \neq k. \tag{49}$$

In (48) and (49), $\widehat{S}_{DD}$ is the set that includes the D2D links established before the kth link, while $\widetilde{S}_{DD}$ contains the kth D2D link to be established and the D2D links already established in $\widehat{S}_{DD}$. δ_1 and δ_2 denote two predetermined thresholds, and the sum rate and the number of scheduled D2D links may vary because of them. The distributed algorithm for scheduling D2D links at high SNR is given in Algorithm 1.

Algorithm 1 Distributed Algorithm for High SNR

1: Randomly select a receiver to construct the first D2D link.
2: As for the first D2D link, choose the transmitter with the maximum channel gain according to (31).
3: **for** $i = 2 : L$ **do**
4: **if** The SNR of the existed $(i-1)$ D2D links and the ith D2D link all satisfy (48) and (49) **then**
5: Select the transmitter with the maximum channel gain to establish the ith D2D link.
6: **else**
7: The ith D2D link cannot be established. Jump to 10.
8: **end if**
9: **end for**
10: Algorithm ends.

Remark 2: Based on (48) and (49), we can guarantee that the D2D link to be established cannot bring great interference to the previous established links according to δ_2, with its own QoS also satisfied according to δ_1. If we want to guarantee the QoS of each D2D link, we should increase δ_1 and δ_2. On the other hand, if we want to increase the number of established D2D links without high QoS requirement, δ_1 and δ_2 should be set lower. Thus, δ_1 and δ_2 should be carefully determined according to the practical requirements of the system. Besides, if several D2D transmitters are able to satisfy (48) and (49) at the same time, the transmitter with maximum channel gain can be selected to serve the current receiver.

C. Medium-SNR Scheme

As for the medium SNR, from (22), the weight of link l_{ij} can be denoted as

$$\widetilde{w}_{ij} = \frac{\Gamma g(i,j)}{\sum\limits_{s=1, s\neq j}^{L} \Gamma g(i,s) + 1}. \tag{50}$$

From (50), we can know that the interference from the other D2D transmitters cannot be ignored at medium SNR. On the other hand, the interference at each D2D receiver is not as serious as that in the case of high SNR. For the dense D2D links at medium SNR, to guarantee that all the D2D receivers can be served by the available transmitters, we assume that the power of interference at each D2D receiver is a constant I. Accordingly, the SINR can be estimated at each D2D receiver, and thus, the weight of link l_{ij} can be rewritten as

$$\widetilde{w}_{ij} = \frac{g(k,k)P}{I + N_o}, \tag{51}$$

and the topology of D2D links can be considered as a weighted bipartite graph as well, similar to the case of low-SNR scheme.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

Therefore, the KM algorithm can also be utilized to construct the D2D links at medium SNR similarly as (38), which will not be demonstrated here for compactness.

D. Only One User

When only one D2D user requires video file and several transmitters can provide the service, according to (31), we can select the transmitter with the maximum channel gain to establish the D2D link.

Remark 3: For low and medium SNRs, the SBS has to obtain the dynamic network topology and caching information through the feedback from users to the SBS via a dedicated frequency band separated from the data service, based on which the D2D links can be established. On the other hand, for the high SNR, a distributed algorithm is proposed in our manuscript to establish the D2D links, in which the SBS does not need to obtain the topology and caching information.

VI. Simulation Results and Discussion

Considering a small cell with dense users [22], [43], all the users are homogeneous PPP distributed in an area with a radius of 100 m, and the effective transmission distance between the SBS and users is set to 50m, i.e., $R = 50$ m. Assume that the D2D communication range is set to no more than 10m, i.e., $R_{D2D} \leq 10$ m. Meanwhile, we assume that the total number of video files is $N = 100$, and the number of different video files to be cached at the D2D transmitters is $m = 15$.

A. Optimal Caching Placement

In this subsection, the total offloading probability of the proposed optimal caching placement scheme in (7) for D2D connections is analyzed. In the simulation, two caching schemes are used for comparison.. The first scheme is the popular m scheme, in which m most popular files are selected to be cached according to the popularity. The caching probability of these m files can be determined according to (7). The second scheme is the equal caching scheme, in which the m files to be cached are selected from the N candidates with equal probability.

First, the total offloading probability of the caching D2D links using different caching schemes is compared with different user density λ in Fig. 4. The Zipf Parameter γ is set to 1. From the results, we can see that, when the user density increases, the offloading probability of the optimal caching scheme and the equal caching scheme both becomes higher, due to the fact that when the D2D users are densely deployed, more opportunity will be provided for each user to access to more different files. In addition, due to the optimization of (7), the offloading probability of the optimal caching is much higher than that of the equal caching scheme. For the popular m scheme, the offloading probability remains almost unchanged with different user density, because only the m most popular files are cached for all the cases.

Then, the total offloading probability of the caching D2D links using different caching schemes is compared with different values of Zipf parameter γ in Fig. 5. The user density

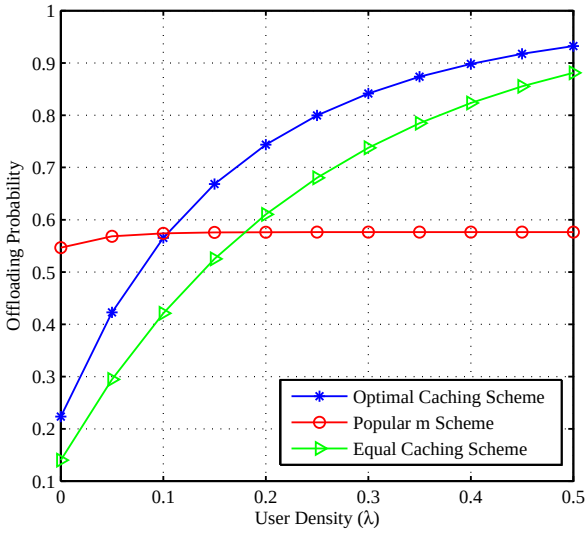

Fig. 4. Total offloading probability comparison of the caching D2D links using different caching schemes with different user density λ. $\gamma = 1$.

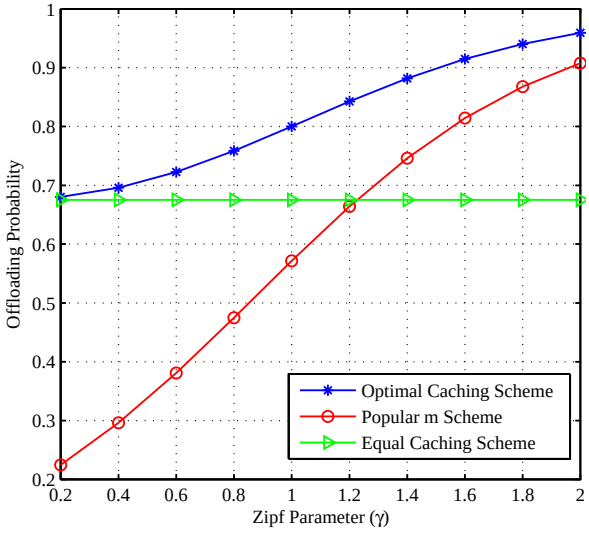

Fig. 5. Total offloading probability comparison of the caching D2D links using different caching schemes with different values of Zipf parameter γ. $\lambda = 0.3$.

λ is set to 0.3. From the results, we can see that when the Zipf parameter γ increases, which means the gap of the popularity of different files becomes larger, the offloading probability of the optimal caching scheme and the popular m scheme becomes higher. This is because the performance of the proposed optimal caching scheme is more effective for the files with more obvious difference in the popularity of different files. In addition, due to the optimization of (7), the offloading probability of the optimal caching is much higher than that of the equal caching scheme. For the equal caching scheme, the offloading probability remains unchanged with different values of γ, because no popularity is considered in this scheme.

B. D2D Link Scheduling

In this subsection, the performance of the D2D link scheduling schemes for different SNRs is analyzed through sim-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

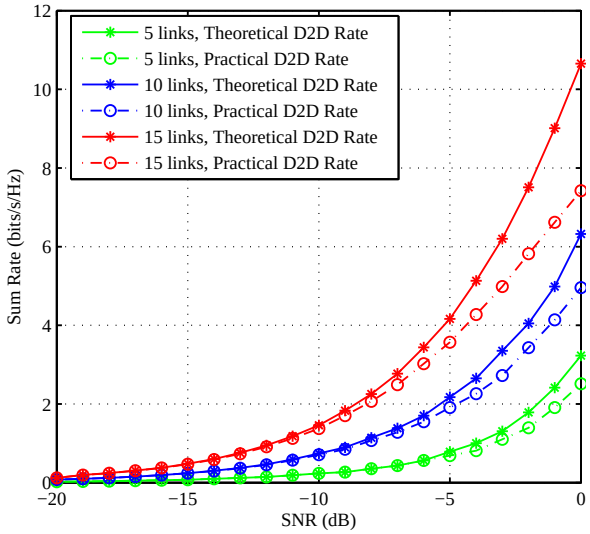

Fig. 6. Theoretical and practical sum rates comparison of D2D links of the low-SNR scheme with different SNRs and different number of D2D links.

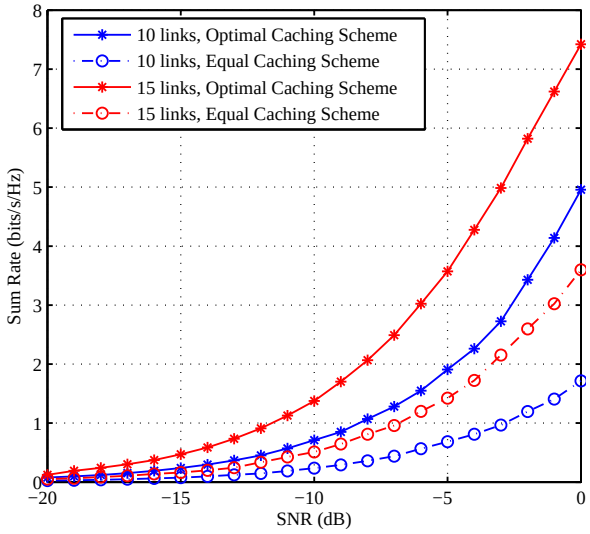

Fig. 7. Sum rate comparison of D2D links of the low-SNR scheme when the optimal caching scheme and equal caching scheme are adopted.

ulation. We assume that the user density λ_D is 0.3 and Zipf parameter γ is 2. The caching capacity of each D2D transmitter is assumed to be $\mu = 4$, i.e., 4 video files can be stored at each local cache. In the simulation, two caching schemes are considered, i.e., the proposed optimal caching scheme and the equal caching scheme. For the optimal caching scheme, some popular video files with high caching probability $\mathcal{P}_i$ can be stored in devices in advance, and $\mathcal{P}_i$ can be calculated through Theorem 1. On the other hand, for the equal caching scheme, the caching probability of each video file can be defined as $m/N = 0.15$. When users can be served by the users nearby, D2D communications can be used to perform transmission. The transmit power of each D2D transmitter is 1W.

The performance of the low-SNR D2D-link scheduling scheme is first compared in Fig. 6 and Fig. 7. In Fig. 6, the theoretical and practical sum rates of D2D links at low SNR are compared with different SNRs and different number of D2D links. The optimal caching scheme is adopted. The theoretical sum rate can be calculated according to (25) in Theorem 2 without considering the interference among D2D users, while interference should be considered when calculating the practical sum rate. From the results, we can see that, when SNR or the number of links becomes larger, the sum rate will increase accordingly. In addition, we can see that when SNR is lower, the theoretical sum rate and the practical sum rate is very close to each other. However, as SNR becomes higher, the practical sum rate is becomes lower than that of the theoretical sum rate gradually, due to the fact that interference cannot be ignored when SNR is relatively high. In Fig. 7, the sum rate of D2D links of the low-SNR scheme is compared when the optimal caching scheme and equal caching scheme are adopted. From the results, we can see that the sum rate of D2D links of the optimal caching scheme at low SNR when the low-SNR D2D-link scheduling scheme is adopted is much higher than that of the equal caching scheme, with different number of D2D links. This is because more D2D links can be established by the optimal caching scheme than the equal caching scheme, i.e., the total offloading probability is maximized. The edge users that cannot obtain the required file from D2D transmitters in the equal caching scheme, have to obtain the video file from the SBS directly, with much low transmission rate. Thus, the effectiveness of the proposed optimal caching scheme can be reflected when combined with D2D-link scheduling.

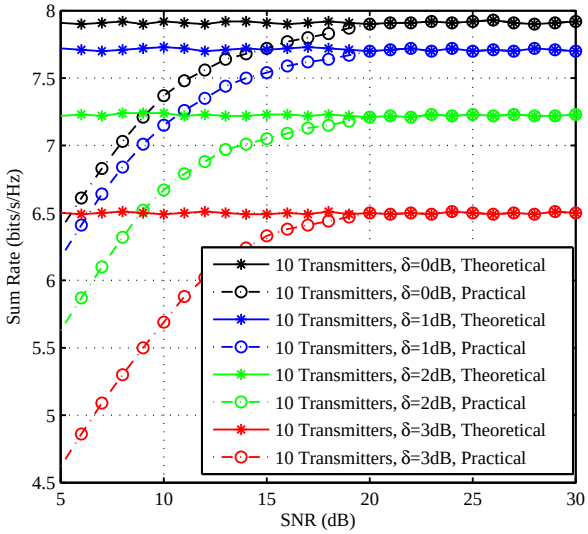

Fig. 8. Theoretical and practical sum rates comparison of D2D links of the high-SNR scheme with different SNRs, different number of D2D links, and different values of $\delta = \delta_1 = \delta_2$.

The performance of the high-SNR D2D-link scheduling scheme is then compared in Fig. 8 and Fig. 9. In Fig. 8, the theoretical and practical sum rates of D2D links of the high-SNR scheme are compared with different SNRs, different number of D2D links, and different values of $\delta = \delta_1 = \delta_2$. The optimal caching scheme is adopted. The theoretical sum rate can be calculated according to (29) without considering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

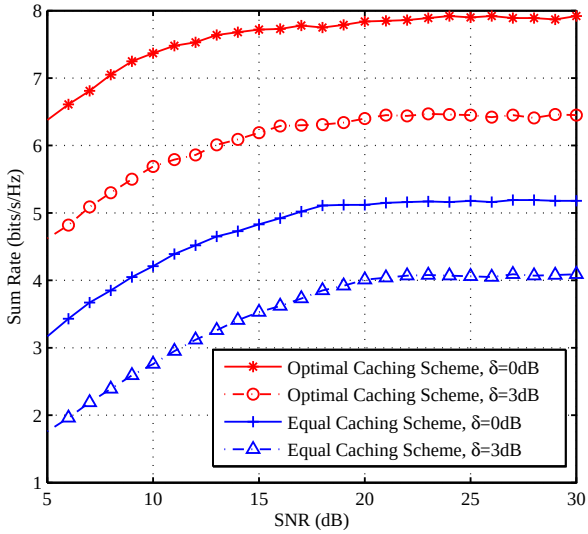

Fig. 9. Sum rate comparison of D2D links of the high-SNR scheme when the optimal caching scheme and equal caching scheme are adopted. There are 10 potential D2D links.

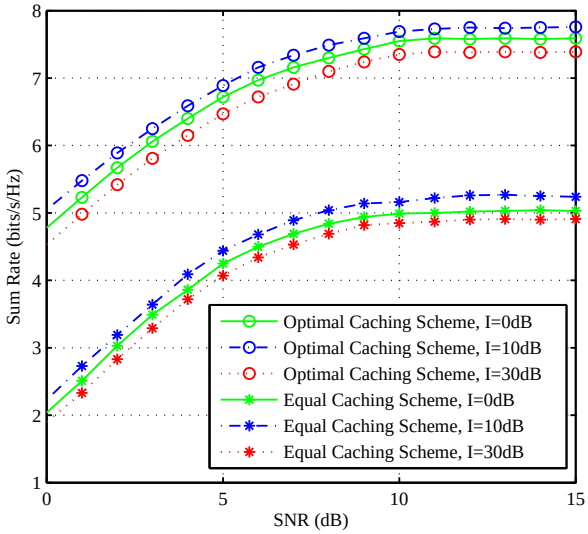

Fig. 10. Sum rate comparison of D2D links of the medium-SNR scheme when the optimal caching scheme and equal caching scheme are adopted, with 10 potential D2D links.

the channel noise, while the noise should be involved when calculating the practical sum rate. From the results, we can see that the theoretical sum rate is close to that of the practical sum rate when SNR becomes higher. However, when the SNR is relatively low, the practical sum rate is lower than the theoretical sum rate, due to the fact that the channel noise can only be ignored when the SNR is enough high. In addition, we can also see that when the threshold δ becomes larger, the sum rate becomes lower accordingly. This is because when δ is larger, the requirement on the performance of the D2D links is stricter, and thus, less D2D links can be established according to (48) and (49). In Fig. 9, the sum rate of D2D links of the high-SNR scheme is compared with 10 potential D2D links, when the optimal caching scheme and equal caching scheme are adopted. From the results, we can see that the sum rate of D2D links of the optimal caching scheme at high SNR when the high-SNR D2D-link scheduling scheme is adopted is much higher than that of the equal caching scheme. Besides, we can also see that when δ is larger, the sum rate becomes lower, which is consistent with the results in Fig. 8. Thus, the effectiveness of the proposed optimal caching scheme can be reflected when combined with D2D-link scheduling.

The performance of D2D-link scheduling of the medium-SNR scheme is compared in Fig. 10 and Fig. 11. From the results in Fig. 10, we can see that the sum rate of D2D links of the optimal caching scheme is much higher than that of the equal caching scheme at medium SNR. Thus, the effectiveness of the proposed optimal caching scheme can be reflected when combined with D2D-link scheduling. In addition, we can observe that the sum rate of D2D links will change slightly with different values of I. Specifically, the sum rate with $I = 10$dB is higher than those with $I = 0$dB and $I = 30$dB. Thus, we should set I properly in the medium-SNR scheme to achieve better performance. To make this point much clearer, in Fig. 11, the sum rate of the D2D links of the medium-SNR scheme is further compared with different values of I

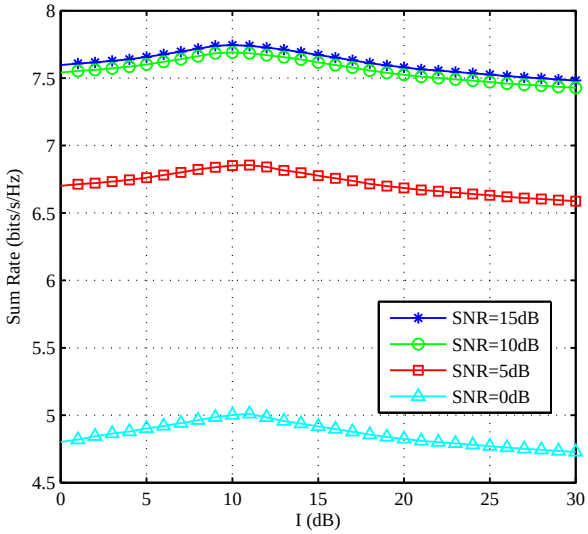

Fig. 11. Sum rate comparison of D2D links of the medium-SNR scheme with different values of I and SNR.

and SNR. The optimal caching scheme is adopted. From the results, we can see that when I=10dB, the sum rate of the medium-SNR scheme is higher than those when I=0dB and I=30dB, which is consistent with the results in Fig. 10.

Finally, the sum rate of all the three proposed schemes for low SNR, medium SNR and high SNR is compared with SNR from -20dB to 30dB and with 10 potential links in Fig. 12. The optimal caching scheme is adopted. We set $\delta_1 = \delta_2 = 0$dB for the high-SNR scheme, and $I = 10$dB for the medium-SNR scheme. From the results, we can see that when SNR is low, i.e., from -20dB to -3dB, the low-SNR scheme can achieve the highest sum rate. When medium SNR is considered, i.e., from -3dB to 15dB, the performance of the medium-SNR scheme is optimal. For the high-SNR scheme, it can achieve the optimal performance when SNR is relatively high, i.e., above 15dB.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2877645, IEEE Transactions on Vehicular Technology

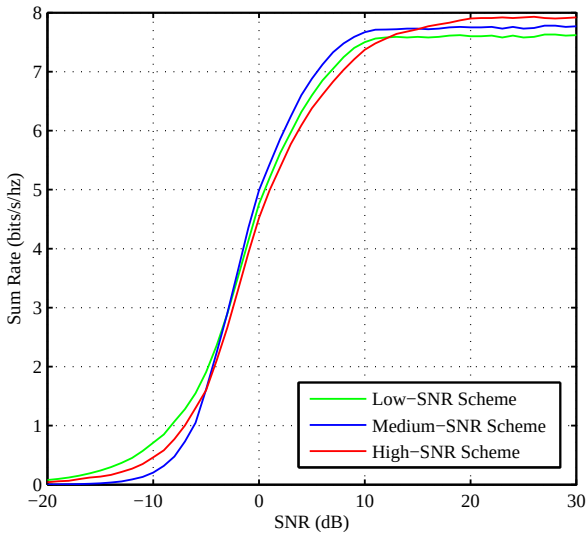

Fig. 12. Sum rate comparison of D2D links of the low-SNR, medium-SNR and high-SNR schemes with SNR from -20dB to 30dB and with 10 potential links. The optimal caching scheme is adopted. We set $\delta_1 = \delta_2 = 0$dB for the high-SNR scheme, and $I = 10$dB for the medium-SNR scheme.

Thus, we should choose the proper scheme to establish D2D links at different transmit SNRs.

VII. Conclusions and Future Work

In this paper, the caching D2D scheme has been proposed for D2D links in small-cell networks, by jointly designing the caching placement and D2D-link establishing. We first proposed an optimal caching scheme to maximize the total offloading probability according to the popularity, and thus, edge users can obtain their required video files from the caches at users nearby through D2D transmission, instead of the SBS. Then, the sum rate of D2D links was analyzed theoretically in different SNR regions, which is the basis for link establishing. In addition, to maximize the throughput of D2D links with low computational complexity, three effective D2D link scheduling schemes were proposed for low SNR, high SNR and medium SNR, respectively, with the help of bipartite graph theory and KM algorithm. Plenty of simulation results were presented to show the effectiveness of the proposed caching D2D scheme. In our future work, the mobility of users and multi-tier heterogeneous small cells will be further considered.

Acknowledgment

We thank the editor and reviewers for their detailed reviews and constructive comments, which have greatly improved the quality of this paper.

References

[1] I. Hwang, B. Song, and S. S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[2] W. Feng, Y. Wang, N. Ge, J. Lu, and J. Zhang, "Virtual MIMO in multi-cell distributed antenna systems: Coordinated transmissions with large-scale CSIT," *IEEE J. Select. Areas Commun.*, vol. 31, no. 10, pp. 2067–2081, Oct. 2013.

[3] Z. Li, L. Guan, C. Li, and A. Radwan, "A secure intelligent spectrum control strategy for future THz mobile heterogeneous networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 116–123, Jun. 2018.

[4] D. Wang, Z. Li, N. Zhang, H. Wu, and X. Shen, "Channel state classification in cognitive small-cell networks with multiple transmission powers," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6023–6036, Jul. 2018.

[5] H. Liu, Z. Chen, X. Tian, X. Wang, and M. Tao, "On content-centric wireless delivery networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 118–125, Dec. 2014.

[6] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[7] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sept. 2016.

[8] W. Han, A. Liu, and V. K. N. Lau, "PHY-caching in 5G wireless networks: Design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[9] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[10] T. Liu, J. Li, F. Shu, M. Tao, W. Chen, and Z. Han, "Design of contract-based trading mechanism for a small-cell caching system," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6602–6617, Oct. 2017.

[11] Y. Guo, Q. Yang, F. R. Yu, and V. C. M. Leung, "Cache-enabled adaptive video streaming over vehicular networks: A dynamic approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5445–5459, Jun. 2018.

[12] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[13] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless. Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.

[14] M. Taghizadeh, K. Micinski, S. Biswas, C. Ofria, and E. Torng, "Distributed cooperative caching in social wireless networks," *IEEE Trans. Mobile. Comput.*, vol. 12, no. 6, pp. 1037–1053, Jun. 2013.

[15] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[16] T. Han and N. Ansari, "Network utility aware traffic load balancing in backhaul-constrained cache-enabled small cell networks with hybrid power supplies," *IEEE. Trans. Mob. Comput.*, vol. 16, no. 10, pp. 2819–2832, Oct. 2017.

[17] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, "Communications, caching, and computing oriented small cell networks with interference alignment," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 29–35, 2016.

[18] F. Cheng, Y. Yu, Z. Zhao, N. Zhao, Y. Chen, and H. Lin, "Power allocation for cache-aided small-cell networks with limited backhaul," *IEEE Access*, vol. 5, pp. 1272–1283, Jan. 2017.

[19] X. Liu, N. Zhao, F. R. Yu, Y. Chen, and V. C. M. Leung, "A cooperative video-streaming transmission strategy in information-centric networks," in *Proc. IEEE SPAWC'17*, pp. 1–5, Hokkaido, Japan, Jul. 2017.

[20] N. Zhao, F. Cheng, F. R. Yu, J. Tang, Y. Chen, G. Gui, and H. Sari, "Caching UAV assisted secure transmission in hyper-dense networks based on interference alignment," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2281–2294, May 2018.

[21] W. Feng, Y. Wang, D. Lin, N. Ge, J. Lu, and S. Li, "When mmWave communications meet network densification: A scalable interference coordination perspective," *IEEE J. Select. Areas Commun.*, vol. 35, no. 7, pp. 1459–1471, Jul. 2017.

[22] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and Z. Han, "Distributed interference and energy-aware power control for ultra-dense D2D networks: A mean field game," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1205–1217, Feb. 2017.

[23] Y. He, F. R. Yu, N. Zhao, and H. Yin, "Secure social networks in 5G systems with mobile edge computing, caching, and device-to-device communications," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 103–109, Jun. 2018.

[24] F. Tang, Z. M. Fadlullah, N. Kato, F. Ono, and R. Miura, "AC-POCA: Anticoordination game based partially overlapping channels assignment in combined UAV and D2D-based networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1672–1683, Feb. 2018.

[25] 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, Apr. 2014.

[26] 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, 4th Quart. 2015.

[27] C. Xu, L. Song, Z. Han, Q. Zhao, X. Wang, X. Cheng, and B. Jiao, "Efficiency resource allocation for device-to-device underlay communication systems: A Reverse iterative combinatorial auction based approach," *IEEE J. Select. Areas Commun.*, vol. 31, no. 9, pp. 348–358, Sept. 2013.

[28] H. Zhang, Y. Liao, and L. Song, "D2D-U: Device-to-device communications in unlicensed bands for 5G system," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3507–3519, Jun. 2017.

[29] L. Wei, R. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: Challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90–96, Jun. 2014.

[30] H. Nishiyama, M. Ito, and N. Kato, "Relay-by-smartphone: realizing multihop device-to-device communications," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 56–65, Apr. 2014.

[31] 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, 4th Quart. 2015.

[32] J. Liu, H. Nishiyama, N. Kato, and J. Guo, "On the outage probability of device-to-device-communication-enabled multichannel cellular networks: An RSS-threshold-based perspective," *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 163–175, Jan. 2016.

[33] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. 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, Jul. 2014.

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

[35] M. Gregori, J. Gmez-Vilardeb, J. Matamoros, and D. Gndz, "Wireless content caching for small cell and D2D networks," *IEEE J. Select. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May. 2016.

[36] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[37] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE ICC'15*, pp. 3358–3363, London, UK, Jun. 2015.

[38] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Device-to-Device collaboration through distributed storage," in *Proc. IEEE Globecom'12*, pp. 2397–2402, Anaheim, CA, Dec. 2012.

[39] T. Zhang, H. Fan, J. Loo, and D. Liu, "User preference aware caching deployment for device-to-device caching networks," *IEEE Syst. J.*, to be published, DOI: 10.1109/JSYST.2017.2773580. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8125781.

[40] S. W. Jeon, S. N. Hong, M. Ji, G. Caire, and A. F. Molisch, "Wireless multihop device-to-device caching networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1662–1676, Mar. 2017.

[41] X. Zhang, H. Gao, and T. Lv, "Multicast beamforming for scalable videos in cache-enabled heterogeneous networks," in *Proc. IEEE WCNC'17*, pp. 1–6, San Francisco, CA, Mar. 2017.

[42] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Res. Logist.*, vol. 52, no. 1, pp. 7–21, Feb. 2005.

[43] D. Lpez-Prez, M. Ding, H. Claussen, and A. H. Jafari, "Towards 1 Gbps/UE in cellular systems: Understanding ultra-dense small cell deployments," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2078–2101, 4th Quart. 2015.