Received December 4, 2019, accepted December 31, 2019, date of publication January 6, 2020, date of current version March 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964495

On Consideration of Content and Memory Sizes in 5G D2D-Assisted Caching Networks

YUJING LIN, ZHIJIAN LIN, (Member, IEEE), PINGPING CHEN, (Member, IEEE), ZHIFENG CHEN, (Senior Member, IEEE), AND LINHUANG WU
Department of Electronic & Information Engineering, Fuzhou University, Fuzhou 350116, China

Corresponding author: Zhijian Lin (zlin@fzu.edu.cn)

This work was supported in part by the Natural Science of Foundation of Fujian Province under Grant 2019J05026, in part by the Fuzhou University Fund under Grant XRC-17052, and in part by the Education Scientific Research Project for Young Teachers of Fujian Province under Grant JT180053.

ABSTRACT Device-to-Device (D2D) communication with caching technology has emerged as a promising technique for offloading the traffic and boosting the throughput of the fifth generation (5G) cellular networks. The combined impact of cache memory size of user equipments (UEs) and content sizes, which are two crucial factors in D2D-assisted caching networks, were usually ignored in the existing researches. In this paper, an optimization algorithm is proposed to maximize the cache hit probability and cache-aided throughput, with the consideration of various cache memory sizes of UEs and content sizes. Firstly, users are grouped according to the cache memory sizes and the content sizes. Then the general mathematical expressions for the optimization of cache hit probability and cache-aided throughput with the constraints of cache memory sizes and content sizes are obtained. Subsequently, a Packet Cache Strategy (PCS) algorithm is proposed to obtain the caching probability matrix with the maximum cache hit probability and cache-aided throughout by taking user caching probability of a file as a variable. Finally, numerical results show that the sizes of the requested files affect the caching willingness of users, and the proposed PCS can achieve the highest cache hit probability and the best cache-aided throughput comparing with two other existing methods.

INDEX TERMS D2D-assisted caching, cache hit probability, cache-aid throughput, file size, memory size.

I. INTRODUCTION

With the rapid explosion of data volumes and content diversities, data traffics are increasingly concentrated to hotspot [1]. For limited network capacity and conventional routing protocols, the wireless network is getting congested, especially in the network peak time. Device-to-Device (D2D) communication is proposed to offload the traffic and boost the throughput of cellular networks [2]. D2D-assisted caching networks, which take both advantages of caching and D2D communication technologies, have attracted much interests recently [3]. When the requested contents are stored in local caches, UEs can obtain contents directly from nearby UEs instead of the Base Station (BS) in the D2D-assisted caching networks [4]. It has been proven that effective caching of popular contents can significantly reduce the mobile traffic, ensuring the efficient and fast operation of communication network [5].

Recently, to maximize the total offloading probability of D2D system and reduce the burden of BS, an optimal caching D2D scheme combining the video-file placement and delivery of caching D2D links is proposed for small-cell networks in [6]. The scheme is that closed-form solutoins are derived to solve the optimal caching problem according to the popularity of D2D system. Furthermore, [7] acted unmanned aerial vehicles as small-cell base stations, which are both equipped with caches to provide videos to mobile users in some small cells at off-peak time, and proposed a comprehensive framework for hyper-dense small-cell networks. [8] analyzed YouTube request characteristics as observed at an edge network over a 20 month period, and proposed a workload modelling approach suitable for content delivery applications with ephemeral content. [9] investigated the effect of bursty traffic on the throughput and the delay performance of a wireless caching helper network, considered the case that user

The associate editor coordinating the review of this manuscript and approving it for publication was Jonathan Rodriguez.

requesting for content is not located in its own cache and characterized the performance of the network by throughput and delay. [10] took advantage of a trade-off between video quality and video diversity, and suggested a method to cache videos of different qualities. Meanwhile, a node association algorithm was proposed for coping with the collision that several users request files from the same device at the same time.

In this paper, content preference, which is considered as different users' desires for the same content, is modeled by user request probability of a file. It can be defined that the greater the probability, the more popular the content is. Actually, the UEs' cache memories sizes and the sizes of the requested files will affect the user request probability of files.

Even though researchers have revealed that UEs' cache memory size and the desired file size both affect the content dissemination process, normalized content sizes and normalized UEs' capacities are generally assumed in the existing literatures, and the content popularity is regarded as the only main factor affecting user caching probability, thus ignoring the combined impact of the above-mentioned two factors constraints on the performance of D2D-assisted caching networks. For instance, [11] assumed that the cache memory size of users and the size of the desired files are normalized which simplifies the computation process to obtain the maximum cache hit probability and cache-aided throughput. In [12], the author took account in the capacity restriction and proposed a caching strategy by normalizing the file size to maximize the cache hit probability. Moreover, many studies considered a simple case in which the impact of content size was neglected.

On consideration of the combined impact of both content sizes and capacities restriction, we design a D2D-assisted caching algorithm, in which UEs are classified into various groups according to their cache memory sizes, and the contents are grouped by sizes, then the user caching probability in each group is optimized to maximize the cache hit probability and cache-aided throughput. The main contributions of this paper are outlined as follows.

- We propose a Packet Cache Strategy (PCS) algorithm based on the capacities restriction and the desired content with different sizes to find out the optimal solution for the cache hit probability and the cache-aided throughput.
- Numerical results show that the cache memory size and the content size have impacts on the optimal cache strategy, and the proposed PCS algorithm outperforms two other existing schemes.

The rest of this paper is organized as follows. The system model is described in Section II, and the optimization problem is formulated in Section III. In Section IV, we propose the caching deployment algorithm. The numerical results and analysis are presented in Section V. Finally, we conclude our work in Section VI.

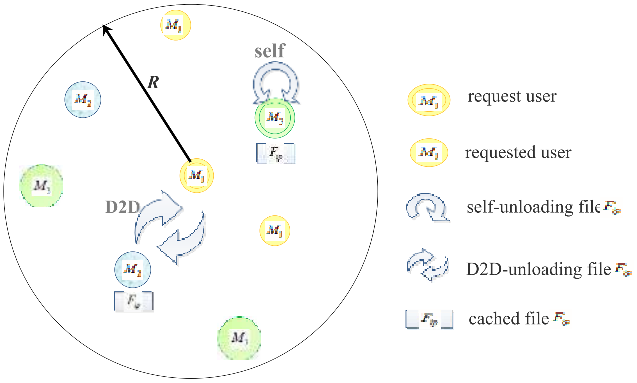

FIGURE 1. D2D caching network schematic under multiple cache memory size.

II. SYSTEM MODEL

A. NETWORK MODEL

The system model of this paper is illustrated in Fig. 1. M_1, M_2, M_3 denote different cache memory sizes. Two offloading methods are considered in the scenario. If a request UE with cache memory size M_1 can obtain the requested file F_{ip} from nearby UE (within D2D communication range R) with cache memory size M_3 in which the requested file has been cached, the method is called as D2D-offloading. Otherwise, if the requested file F_{ip} has already been stored in its own cache memories and the F_{ip} can be obtained directly, the method is called as self-offloading.

To the best of our knowledge, the memories of UEs in our daily life are varied. Therefore, in the system model, U kinds of different cache memory sizes are considered, and the location distribution of all UEs follows homogeneous Poisson Point Process (PPP) with intensity λ. Each UE has cache memory size M_u ($U \in [1, U]$) with a proportion of m_u, $0 < m_u < 1$. Obviously, $\sum_{u=1}^{U} m_u = 1$. Thus, the locations of UEs with the cache memory size M_u can be modeled by PPP Φ_d with intensity $m_u\lambda$.

As far as we know, several existing literatures focused on the content popularity distribution. For example, [13] studied the characteristics of media traffic and found that the requested frequency of multimedia files follow Zipf-like distribution. The distribution of web page requests is studied in [14] which proved that the web page requests approximately follow Zipf distribution. A large number of data researches were made for popular video websites such as YouTube in [15], [16]. The results also showed the request distribution of a video file could approximately follow the Zipf distribution. Assuming a finite content category $F = \{F_i\} = \{F_1, F_2, \cdots, F_N\}$, where F_i is the i-th most popular file library, then the request probability of file library F_i is

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

where i presents popularity order, γ is a popularity parameter which denotes the degree of concentration of popular content.

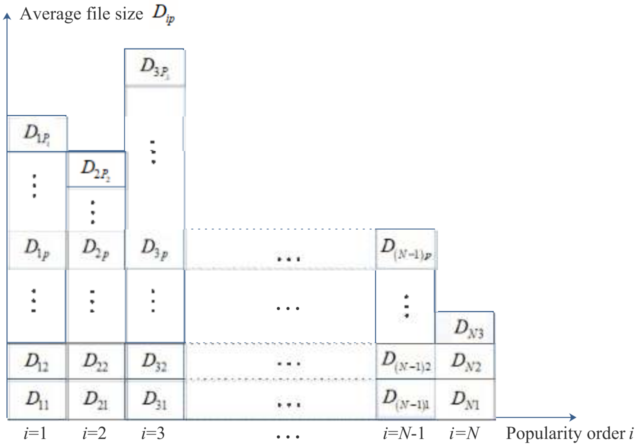

FIGURE 2. Relationship between popularity order i and average file size D_{ip}.

Generally, there are more than one file with different sizes in the same popularity order. Therefore, we give a range to group the files by sizes, and the average value is taken as the average size of files in the group. Consequently, the files can be divided into multiple groups of different sizes in the same popularity order.

Assuming that there are $P_i (i \in [1, N])$ groups of files with different sizes in the i-th popularity order, the average file size of group p ($p \in [1, P_i]$) is D_{ip} and the proportion of files in group p to the total files is denoted by d_{ip}, $0 < d_{ip} < 1$. Then we have $\sum_{p=1}^{P_i} d_{ip} = 1$. As shown in Fig. 2, we can divide the files into P groups, $P = P_1 + P_2 + \cdots + P_N$. Thus, the file library in the i-th popularity order can be subdivided into $F_i = \{f_{ip}\} = \{f_{i1}, f_{i2}, \cdots, f_{iP_i}\}$, $(i \in [1, N])$. We define the probability that the file f_{ip} is cached by the user with memory size M_u is c_{uip} $(0 < c_{uip} < 1)$. The cache strategy for all UEs can be represented as $\mathrm{A} = \left[c_{uip}\right]_{U \times N \times [P_1, P_2, \cdots P_N]^{\mathrm{T}}}$. Moreover, The location distribution of UEs with cache memory size M_u and caching file f_{ip} follows PPP with intensity $m_u \lambda c_{uip}$, each UE has the probability $\rho \in [0, 1]$ to request file f_{ip}. Therefore, the distributions of potential transmitters follow homogeneous PPPs $\Phi_{\mathrm{f}}^{\mathrm{u}}$ with intensity $(1-\rho)\lambda c_{uip}$.

B. CACHE HIT PROBABILITY

When an active UE requests a file in F, the cache hit probability contains the following two parts:

- **Self-request cache hit probability:** the probability that the requested file has been cached in its own memories.
- **D2D cache hit probability:** the probability that the requested file is not cached in its own memories, but in the devices within a certain D2D communication distance R.

While the requested file cannot be found through the above two methods, the file will be downloaded from core network to the nearest base station through backhaul. In this paper, we assume the cross-tier interference between D2D and cellular communications does not exist [11].

III. PROBLEM FORMULATION

Based on the system model in Section II, we consider cache hit probability and cache-aided throughput as the main performance metrics. It is worth noticing that finding the requested file in its own cache memories and the impact of different file sizes on caching willingness of users in the cache-related performance study are both often ignored in the literatures [17], [18], and this is the main innovation of this paper.

The cache hit probability includes two parts, namely, self-request cache hit probability and D2D cache hit probability. Details of each part are given as follows:

1) SELF-REQUEST CACHE HIT PROBABILITY

If the request probability of the i-th popular files follows Zipf distribution and the request probability of file library F_i is q_i, the request probability of file f_{ip} is $q_i \cdot d_{ip}$. Moreover, the self-request cache hit probability of UEs with cache memory size M_u and requesting file f_{ip} can be expressed as $m_u q_i d_{ip} c_{uip}$. Therefore, the self-request cache hit probability of a random user requesting random files is given as

$$p_{hit,self} = \sum_{u=1}^{U} m_u \sum_{i=1}^{N} q_i \sum_{p=1}^{P_i} d_{ip} c_{uip}. \tag{2}$$

2) D2D CACHE HIT PROBABILITY

D2D cache hit probability mainly depends on the popularity degree of the files in the coverage area of D2D communication [19]. According to the Poisson distribution formula, the probability of the file f_{ip} being found in the range of D2D communication radius R can be obtained as

$$p_{hit,f_{ip}}^{d2d} = 1 - \exp\left(-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} \pi R^2 (1-\rho) m_u \lambda c_{uip}\right). \tag{3}$$

The probability that one UE with cache memory size M_u requests file f_{ip} which is not cached in its own memories, is $m_u q_i d_{ip}\left(1 - c_{uip}\right)$. Thus, D2D cache hit probability can be presented as $p_{hit,d2d} = \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} q_i m_u d_{ip}\left(1-c_{uip}\right) p_{hit,f_{ip}}^{d2d}$. Combining with (3), we have

$$\begin{aligned} &p_{hit,d2d} \\ &= \sum_{u=1}^{U} m_u \sum_{i=1}^{N} q_i \sum_{p=1}^{P_i} d_{ip}\left(1-c_{uip}\right) \\ &\quad \cdot \left(1 - \exp\left(-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} (1-\rho) m_u \lambda c_{uip}\right)\right). \end{aligned} \tag{4}$$

Thus, the total cache hit probability $p_{hit} = p_{hit,self} + p_{hit,d2d}$ can be obtained as

$$\begin{aligned} p_{hit} = & \sum_{u=1}^{U} m_u \sum_{i=1}^{N} q_i \sum_{p=1}^{P_i} d_{ip} c_{uip} + \sum_{u=1}^{U} m_u \sum_{i=1}^{N} q_i \sum_{p=1}^{P_i} d_{ip} \\ & \cdot \left(1 - c_{uip}\right) \left(1 - \exp\left(-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} b_u c_{uip}\right)\right), \end{aligned} \quad (5)$$

where $b_u = \pi R^2 (1-\rho) m_u \lambda$. In order to distinguish variables u, i, p in coefficients and exponents, we replace the variables u, i, p with j, n, l, thus

$$\begin{aligned} p_{hit} = & \sum_{j=1}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} d_{nl} c_{jnl} + \sum_{u=1}^{U} m_j \sum_{i=1}^{N} q_n \sum_{p=1}^{P_i} d_{nl} \\ & \cdot \left(1 - c_{jnl}\right)\left(1 - e^{g_u}\right) \\ = & \sum_{j=1}^{U} m_j \sum_{N=1}^{N} q_n \sum_{l=1}^{P_n} d_{nl} \left(1 - e^{g_u} + c_{jnl} e^{g_u}\right) \\ = & 1 + \sum_{j=1}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} e^{g_u} d_{nl} \left(c_{jnl} - 1\right), \end{aligned} \quad (6)$$

where $g_u = \sum_{u=1}^{U} \sum_{i=1}^{N} \sum_{p=1}^{P_i} b_u c_{uip}$.

After simplification, the total cache hit probability is obtained as

$$p_{hit} = 1 + \sum_{j=1}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} e^{g_u} d_{nl} \left(c_{jnl} - 1\right). \quad (7)$$

A. CACHE-AIDED THROUGHPUT

Cache-aided throughput is one of the most important indicators to measure D2D caching network performances [19]. Unlike the traditional cache-aided throughput, the average number of requests that can be successfully and simultaneously handled by local caches per unit area is studied in [20].

Assuming that the transmission time of each file is the same and the impact of transmission time on cache-aided throughput is negligible [21]. In the case of self-request, UE caches the requested file itself, so the probability of finding requested file is 1. Meanwhile, ignoring the influence of channel interference and channel fading, the probability of obtaining files from their own caches and being transmitted is also set as 1. Moreover, in the case of D2D communication, the successful transmission probability of the requested file depends on the received Signal to Interference plus Noise Ratio (SINR) [22]. Therefore, the cache-aided throughput can be presented as

$$\begin{aligned} \mathcal{T} = & \sum_{u=1}^{U} \rho m_u \lambda \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} q_i d_{nl} c_{uip} \cdot 1 \cdot 1 \\ & + \sum_{u=1}^{U} \rho m_u \lambda \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} q_i d_{ip} \left(1 - c_{uip}\right) \\ & \cdot p_{hit,fip}^{d2d} \cdot p_{d2d,fip}^{suc}, \end{aligned} \quad (8)$$

where $\rho m_u \lambda$ is the density of UE with cache memory size M_u per unit area, that is, it is the total average number of UEs with cache memory size M_u in per unit area. $p_{d2d,fip}^{suc}$ is the successful probability of file f_{ip} being transmitted via D2D communication. $q_i d_{ip}$ is defined as the probability of the file being requested in the i-th popularity order with D_{ip} memory size.

Based on the self-request cache hit probability in (2) and the D2D cache hit probability in (4), (8) can be simplified as

$$\begin{aligned} \mathcal{T} = & \rho\lambda \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} m_u q_i d_{nl} c_{uip} \\ & + \rho\lambda \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} m_u q_i d_{ip} \left(1 - c_{uip}\right) \cdot p_{hit,fip}^{d2d} \cdot p_{d2d,fip}^{suc} \\ = & \rho\lambda \left(p_{hit,self} \cdot 1 + p_{hit,d2d} \cdot p_{d2d,fip}^{suc}\right). \end{aligned} \quad (9)$$

In this paper, we denote SINR received by user r as SINR_r, which can be obtained by

$$\mathrm{SINR}_r = \frac{\widetilde{P}_d \left|h_{r,r}\right|^2 s_{fip}^{-\alpha}}{\sigma^2 + \sum_{t \in \Phi_{\mathrm{t}}^{\mathrm{hit}} \setminus \{r\}} \widetilde{P}_d \left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}}, \quad (10)$$

where $\widetilde{P}_d$ refers to the UE's transmission power, $h_{t,r}$ denotes the small-scale channel fading from transmitter t to receiver r, which follows the Rayleigh fading with mean of 1. $s_{t,r}$ denotes the distance from transmitter t to receiver r. s_{fip} is the closest distance from the potential transmitters with file f_{ip} to the request UE. σ^2 is the background thermal noise power. The distribution of transmitter location in the case of D2D cache hit approximately follows PPP $\Phi_{\mathrm{t}}^{\mathrm{hit}}$ with intensity $\lambda_t = \rho m_u \lambda p_{hit,d2d}$.

For a given SINR target ε of successful D2D transmission, the success probability of D2D transmission is given by

$$\begin{aligned} & p_{d2d,fip}^{suc} \\ & = \mathbb{P}\left[\frac{\widetilde{P}_{\mathrm{d}} \left|h_{r,r}\right|^2 s_{fip}^{-\alpha}}{\sigma^2 + \sum_{t \in \Phi_{\mathrm{t}}^{\mathrm{hit}} \setminus \{r\}} \widetilde{P}_{\mathrm{d}} \left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}} > \varepsilon\right] \\ & = \mathbb{P}\left[\left|h_{r,r}\right|^2 > \frac{\varepsilon s_{fip}^{\alpha}}{\widetilde{P}_{\mathrm{d}}}\left(\sigma^2 + \sum_{t \in \Phi_{\mathrm{t}}^{\mathrm{hit}} \setminus \{r\}} \widetilde{P}_{\mathrm{d}} \left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}\right)\right] \\ & = \mathbb{P}\left[\left|h_{r,r}\right|^2 > \frac{\varepsilon s_{fip}^{\alpha} \sigma^2}{\widetilde{P}_{\mathrm{d}}} + \varepsilon s_{fip}^{\alpha} \sum_{t \in \Phi_{\mathrm{t}}^{\mathrm{hit}} \setminus \{r\}} \left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}\right] \\ & = \mathbb{E}_{s_{fip}}\left[\exp\left(-\frac{\varepsilon s_{fip}^{\alpha} \sigma^2}{\widetilde{P}_{\mathrm{d}}} - \varepsilon s_{fip}^{\alpha} \sum_{t \in \Phi_{\mathrm{t}}^{\mathrm{hit}} \setminus \{r\}} \left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}\right)\right] \\ & = \mathbb{E}_{s_{fip}}\left[\exp\left(-\frac{\varepsilon s_{fip}^{\alpha} \sigma^2}{\widetilde{P}_{\mathrm{d}}}\right) \cdot \mathcal{L}_{I_d}\left(\varepsilon s_{fip}^{\alpha}\right)\right], \end{aligned} \quad (11)$$

where

$$\begin{aligned}\mathcal{L}_{I_d}\left(\varepsilon s_{fip}^{\alpha}\right) &= \exp\left(-\varepsilon s_{fip}^{\alpha}\sum_{t\in\Phi_{\mathrm{t}}^{\mathrm{hit}}\backslash\{r\}}\left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}\right)\\ &= \prod_{t\in\Phi_{\mathrm{t}}^{\mathrm{hit}}\backslash\{r\}}\exp\left(-\varepsilon s_{fip}^{\alpha}\cdot\left|h_{t,r}\right|^2 s_{t,r}^{-\alpha}\right)\\ &= \exp\left(-2\pi\rho m_u\lambda p_{hit,d2d}\cdot\mathcal{I}\right)\\ &= \exp\left(-\frac{\pi\rho m_u\lambda p_{hit,d2d}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}\cdot\left(\varepsilon s_{fip}^{\alpha}\right)^{\frac{2}{\alpha}}\right)\\ &= \exp\left(-\frac{\pi\rho m_u\lambda p_{hit,d2d}\cdot s_{fip}^2\cdot\varepsilon^{\frac{2}{\alpha}}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}\right), \qquad (12)\end{aligned}$$

and $\mathcal{I}=\int_0^\infty\left(1-\exp\left(-\varepsilon s_{fip}^{\alpha}\left|h_{t,r}\right|^2 r^{-\alpha}\right)\right)rdr$.

Substituting (12) into (11), we obtain the success probability of D2D transmission as

$$p_{d2d,fip}^{suc}=\int_0^\infty f_{s_{fip}}(r)\cdot e^{-\frac{\varepsilon r^{\alpha}\sigma^2}{\widetilde{P}_{\mathrm{d}}}-\frac{\pi\rho m_u\lambda r^2\varepsilon^{\frac{2}{\alpha}}p_{hit,d2d}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}}dr, \qquad (13)$$

where $f_{s_{fip}}(r)$ is the probability density function (PDF) of the distance s_{fip} between the request UE and the requested UE. $f_{s_{fip}}(r)$ can be obtained by the corresponding probability distribution function $F_{s_{fip}}(r)$.

Setting event A that the requested users are in the circle which the request user is in the center, and R (maximum D2D communication distance) is the radius. The event B is that the requested users with file f_{ip} locate in the circle, which the request user is in the center, and r (the shortest distance between the receiver and the transmitter) is the radius, $r<R$. According to conditional probability, we have

$$F_{s_{fip}}(r)=P(B|A)=\frac{P(B)}{P(A)}. \qquad (14)$$

Let $\mathcal{S}(R)=\left(-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}\right)$. Then, we have $P(A)=1-\mathcal{S}(R)$, $P(B)=1-\mathcal{S}(r)$. Substituting $P(A)$ and $P(B)$ into (14), the probability distribution function can be rewritten as

$$F_{s_{fip}}(r)=\begin{cases}\dfrac{1-e^{-\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi r^2(1-\rho)m_u\lambda c_{uip}}}{1-e^{-\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}}, & \text{if } r\leq R\\ 0, & \text{otherwise}\end{cases} \qquad (15)$$

The corresponding PDF is given by

$$\begin{aligned}f_{s_{fip}}(r)&=F_{s_{fip}}{}'(r)\\ &=\begin{cases}\dfrac{\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}2\pi r(1-\rho)m_u\lambda c_{uip}e^{-g_u}}{1-e^{-g_u}}, & \text{if } r\leq R\\ 0, & \text{otherwise}\end{cases} \qquad (16)\end{aligned}$$

IV. SOLUTION ANALYSIS FOR PACKET CACHING STRATEGY

In this section, we propose a Packet Caching Strategy (PCS) algorithm and study the optimal caching probability by maximizing the cache hit probability and cache-aided throughput.

A. CACHE HIT PROBABILITY

The aim of this part is to discuss the value of c_{uip}, which is the probability of UEs with cache memory size M_u and caching file f_{ip} in the 3-dimensional (3D) matrix $\mathrm{A}=\left[c_{uip}\right]_{U\times N\times[P_1,P_2,\cdots P_N]^{\mathrm{T}}}$. The cache hit probability is given by (7), accordingly, the optimization problem can be expressed as

$$\begin{aligned}\mathcal{P}1:\max_{\mathrm{A}}p_{hit}&=1+\sum_{j=1}^{U}m_j\sum_{n=1}^{N}q_n\sum_{l=1}^{P_n}d_{nl}\left(c_{jnl}-1\right)e^{-g_u}.\\ s.t.\ &\sum_{i=1}^{N}\sum_{p=1}^{P_i}c_{uip}\cdot D_{ip}\leq M_u,\quad u\in[1,U],\\ &0\leq c_{uip}\leq 1. \qquad (17)\end{aligned}$$

wherein, the constraint conditions mean that the total average size of caching files cannot exceed the UEs' cache memory size.

In the 3D matrix $\mathrm{A}=\left[c_{uip}\right]_{U\times N\times[P_1,P_2,\cdots P_N]^{\mathrm{T}}}$, assuming that all the caching probabilities $c_{jnl}(j\in[1,U],j\neq u,n\in[1,N],n\neq i,l\in[1,2,\cdots P_n],l\neq p)$ are constant except the variable c_{uip}. Thus, equation (7) can be rewritten as

$$\begin{aligned}p_{hit}&=1+\underbrace{m_u d_{ip}q_i\left(c_{uip}-1\right)e^{-g_u}}_{(i).j=u,n=i,l=p}\\ &\quad+\underbrace{m_u\sum_{l=1,l\neq p}^{P_i}d_{il}q_i\left(c_{uil}-1\right)e^{-g_u}}_{(ii).j=u,n=i,l\neq p}\\ &\quad+\underbrace{m_u\sum_{n=1,n\neq i}^{N}q_n\sum_{p=1}^{P_n}d_{np}\left(c_{unp}-1\right)e^{-g_u}}_{(iii).j=u,n\neq i}\\ &\quad+\underbrace{\sum_{j=1,j\neq u}^{U}m_j\sum_{n=1}^{N}q_n\sum_{l=1}^{P_n}d_{nl}\left(c_{jnp}-1\right)e^{-g_u}}_{(iv).j\neq u}. \qquad (18)\end{aligned}$$

Firstly, the solvability of problem $\mathcal{P}1$ is revealed by Lemma 1. Then, we apply the relaxation variable μ_u ($u\in 1,U$) in Karush-Kuhn-Tucher (KKT) conditions to find out the optimal result, the corresponding solution process of relaxation variable is given in Corollary 1. Finally, based on Lambert W Function, the general expression of the optimal caching probability $\widehat{c}_{uip}$ is summarized in Corollary 2.

Lemma 1: When $u\in[1,U]$, $i\in[1,N]$, $p\in[1,P_i]$, the cache hit probability p_{hit} has the maximum value on the UE's caching probability c_{uip}, that is, $\frac{\partial^2 p_{hit}}{\partial c_{uip}{}^2}<0$.

Proof: Refer to Appendix A for the detailed proof.

To find the optimal solution, we utilize Lemma 1 to prove the solvability of the optimization problem $\mathcal{P}1$. As Appendix A shows, the optimization function p_{hit} is a concave function about c_{uip}, that is, the function p_{hit} exists a maximum value. Thus, problem $\mathcal{P}1$ can be rewritten as

$$\begin{aligned}\min_{\mathrm{A}} p_{hit} &= -1-\sum_{j=1}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} d_{nl}\left(c_{jnl}-1\right)e^{-g_u}\\ s.t. &\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{uip}\cdot D_{ip}\le M_u,\, u\in[1,U],\\ &0\le c_{uip}\le 1. \end{aligned} \tag{19}$$

Considering relaxation variable $\mu_u(u \in [1,U])$, the Lagrange function of (19) is as follows

$$\begin{aligned}\mathcal{L}(\mathrm{A},\mu) = &-1-\sum_{j=1}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} d_{np}\left(c_{jnl}-1\right)e^{-g_u}\\ &+\sum_{k=1}^{U}\mu_k\left(\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{kip}\cdot D_{ip}-M_k\right). \end{aligned} \tag{20}$$

Meanwhile, according to KKT, the following conditions must be satisfied to solve the above optimization problem.

$$\begin{cases}\dfrac{\delta\mathcal{L}(\mathrm{A},\mu)}{\delta c_{uip}}=0, & (22.i)\\ \mu_k\ge 0,\ k\in[1,U], & (21.ii)\\ \displaystyle\sum_{k=1}^{U}\mu_k\left(\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{kip}\cdot D_{ip}-M_k\right)=0, & (22.iii)\end{cases} \tag{21}$$

where $(21.i)$ denotes a necessary condition for acquiring extreme value by Lagrange function. $(21.ii)$ is the coefficient constraints. $(21.iii)$ denotes the constraint condition to guarantee the establishment of the equation. The solution of the optimization problem is obtained by condition $(21.i)$.

Furthermore, proof of Lemma 1 leads to the following two corollaries.

Corollary 1: When $u \in [1,U]$, the relaxation variable μ_u of Lagrange function to solve the optimization problems can be given as

$$\mu_u = \frac{e^{-g_j}}{\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}}\left[a-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} m_j q_n d_{nl}\left(1-c_{jnl}\right)\right], \tag{22}$$

where $a = m_u d_{ip} q_i$, $g_j = \sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_i} b_j c_{jnl}$, $b_j = \pi R^2(1-\rho)m_j\lambda$ is the average number of all UEs in the circle which the request user with cache memory size M_j is in the center and R is the radius.

Proof: Refer to Appendix B for the detailed proof.

According to Corollary 1, we can find the highest cache hit probability by relaxation variable.

Corollary 2: If the relaxation variable $\mu_u(u \in [1,U])$ is determined, the solution of problem $\mathcal{P}1$ can be achieved by Lambert W Function. Thus, we have

$$\begin{aligned}\widehat{c}_{uip} = \frac{d_{ip}q_i\widehat{b}}{b_m b_u}+\frac{\widehat{b}}{b_u}-\frac{\mathcal{W}\left(\dfrac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right)}{b_u}\\ -\sum_{l=1,l\neq p}^{P_i} c_{uil}-\chi, \end{aligned} \tag{23}$$

where $b_m = \sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} b_j q_n d_{nl}$, $\widehat{b} = \sum_{j=1}^{U} b_j = \sum_{j=1}^{U}\pi R^2(1-\rho)m_j\lambda$, $\mathcal{W}()$ is Lambert W Function, $\chi = \sum_{n=1,n\neq i}^{N}\sum_{l=1}^{P_n} c_{unl} + \sum_{j=1,j\neq u}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\frac{b_j}{b_u}c_{jnl}$.

Proof: Refer to Appendix C for the detailed proof.

In Corollary 2, the optimal caching probability $\widehat{c}_{uip}$ can be obtained. Based on the general expression of $\widehat{c}_{uip}$ which is the probability of UE with cache memory size M_u and caching file f_{ip}. Thus, the optimal user caching probability of the rest UEs can be obtained.

B. CACHE-AIDED THROUGHPUT

Similar to the study of the optimal cache hit probability, the UE caching probability c_{uip} is an unknown variable, the cache strategy $\mathrm{A} = \left[c_{uip}\right]_{U\times N\times[P_1,P_2,\cdots P_N]^{\mathrm{T}}}$ is also discussed to maximize the cache-aided throughput. The optimization problem can be given as follows:

$\mathcal{P}2$:

$$\begin{aligned}\max_{\mathrm{A}}\mathcal{T} = &\sum_{u=1}^{U}\rho m_u\lambda\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} q_i d_{ip} c_{uip}\cdot 1\cdot 1\\ &+\sum_{u=1}^{U}\rho m_u\lambda\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i} q_i d_{ip}\left(1-c_{uip}\right)\cdot p_{hit,f_{ip}}^{d2d}\cdot p_{d2d,f_{ip}}^{suc}\\ s.t. &\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{uip}\cdot D_{ip}\le M_u,\, u\in[1,U],\\ &0\le c_{uip}\le 1. \end{aligned} \tag{24}$$

The calculating process of D2D successful transmission probability in the section III is particularly complex and hard to simulate, in order to simplify the process, the result of (11) is estimated by (25).

$$\mathbb{E}_x\left[\exp\left(-ax^b\right)\right]\approx\exp\left[-a\mathbb{E}\left[x^2\right]^{\frac{b}{2}}\right]. \tag{25}$$

Lemma 2: The cache-aided throughput can be simplified by (25) as

$$\mathcal{T} = \sum_{u=1}^{U} \rho m_u \lambda \sum_{u=1}^{U} \sum_{i=1}^{N} \sum_{p=1}^{P_i} q_i d_{ip} c_{uip} + \sum_{u=1}^{U} \rho m_u \lambda \sum_{u=1}^{U} \sum_{i=1}^{N} \sum_{p=1}^{P_i} q_i d_{ip} \left(1 - c_{uip}\right) \cdot p_{hit,fip}^{d2d} \cdot \widehat{p}_{d2d,fip}^{suc}, \tag{26}$$

where $\widehat{p}_{d2d,fip}^{suc}$ is the probability of the estimated D2D successful transmission, which is obtained as:

$$\widehat{p}_{d2d,fip}^{suc} \approx e^{\left(-\frac{\varepsilon \mathbb{E}\left[s_{fip}^2\right]^{\frac{\alpha}{2}} \sigma^2}{\widetilde{P}_{\mathrm{d}}} - \frac{\pi \rho m_u \lambda p_{hit,d2d} \cdot \mathbb{E}\left[s_{fip}^2\right] \cdot \varepsilon^{\frac{2}{\alpha}}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}\right)}, \tag{27}$$

where $\mathbb{E}\left[s_{fip}^2\right]$ is the result of simplification combining with the PDF in (16), which is calculated as

$$\mathbb{E}\left[s_{fip}^2\right] = \frac{R^2 \cdot e^{-g_u}}{1 - e^{-g_u}} - \frac{R^2}{g_u}. \tag{28}$$

Proof: Refer to Appendix D for the detailed proof.

From Lemma 2, the complex expression of cache-aided throughput $\mathcal{T}$ is simplified, which is used to solve problem $\mathcal{P}2$. And the proof of Lemma 2 is also carried out by KKT conditions and Lambert W Function, which is similar to the proof in the Lemma 1, Corollary 1 and Corollary 2. Details are omitted for brevity.

According to the simplification results of $\mathcal{T}$, it can be found that the unit area throughput in the specified area is affected by the UE distribution density λ and the probability q_i that UE requests the i-th order popular files F_i. We can infer that the greater λ, the more distribution density the UE is, which will lead the greater the throughput per unit area. The greater q_i, the higher probability of F_i being requested is and the greater throughput per unit area is.

The algorithm of Problem $\mathcal{P}1$ and Problem $\mathcal{P}2$ is given above.

Algorithm 1 Implementation algorithm of PCs

```
Initialization: Set the number of different UE cache memory $U$ and their corresponding size $M_u$ ($u \in [1, U]$), the total file popularity order $N$ and the size of each file $S_{ic}$ ($i \in [1, N]$, $c$ refers to the number of files at $i$-th popularity order), D2D communication radius $R$, Poisson distribution intensity $\lambda$, user request probability of a file $\rho$, background thermal noise power $\sigma^2$, etc.
Set initial caching strategy matrix A $= \left[c_{uip}\right]_{U \times N \times [P_1, P_2, \cdots P_N]^{\mathrm{T}}}$, $(0 < c_uip < 1)$.
Grouping: Grouping all files of each popularity order by the size, get the number of the groups $P_i$ and the average file size of each group $D_{ip}(p \in [1, P_i])$. i.e., the 1st order popularity files can be divided into $P_1$ groups at intervals of 10M, the average file size of each group is $D_{1p}(p \in [1, P_1])$.
Optimization: Matching the dimension of the matrix by adding zero.
$flag = 1$;
while $flag == 1$ do
    for $u = 1 : U$ do
        for $i = 1 : N$ do
            for $p = 1 : P_i\ (i)_{max}$ do
                Using KKT conditions and Lambert W Function to calculate the UE caching probability $\widehat{c}_{uip}$ as (23).
                if $c_{uip}$==0 then
                    Restraining the corresponding element of the optimal cache strategy matrix $\widehat{c}_{uip}$=0.
                end if
                Update matrix A and get the new matrix A*.
            end for
        end for
    end for
    if A*==A then
        flag=0;
    end if
end while
Calculating the best cache hit probability $p_{hit,\max}$ according to (20)
Calculating the cache-aided throughput $\mathcal{T}$ according to (26)
```

V. NUMERICAL RESULTS

In this section, we make simulations to evaluate the impact of the cache memory size constraint and the requested file size on the caching performance. We consider a single cell scenario where UEs density λ of each group is set to be 10^{-3} per m^2. The D2D communication range is $R = 10$ m.

A. CACHE HIT PROBABILITY

Assuming that the caching network has $U = 2$ different cache memory size M_u ($u \in [1, U]$), where $M_1 = 1$, $M_2 = 2$. UEs request a random file in F with probability $\rho = 0.5$. In the file library F, there are $N = 2$ popularity orders. The 1-st order popular content library F_1 is divided into $P_1 = 2$ groups, and the size matrix is $D_{1p} = [1, 2]$, whose corresponding proportion is $d_{1p} = [0.5, 0.5]$. The 2-nd order popular content library F_2 is divided into $P_2 = 3$ groups, and the size is $D_{2p} = [1, 2, 3]$, whose corresponding proportion is $d_{2p} = [1/3, 1/3, 1/3]$. The aforementioned assumptions are referred to as initial hypothesis (IH).

In Fig. 3, the numerical and convergence of the optimal cache hit probability are compared between value init $A_0 = \{0.5\}_{2\times2\times[2,3]^{\mathrm{T}}}$ and zero init $A_0 = \left\{10^{-3}\right\}_{2\times2\times[2,3]^{\mathrm{T}}}$. It is obvious that two initial strategies have no effect on the optimal cache hit probability and its convergence under the same caching network. Moreover, three different caching networks settings are given in Fig. 3, in which the points of the pentagram denote IH. Compared with the IH, the UEs in

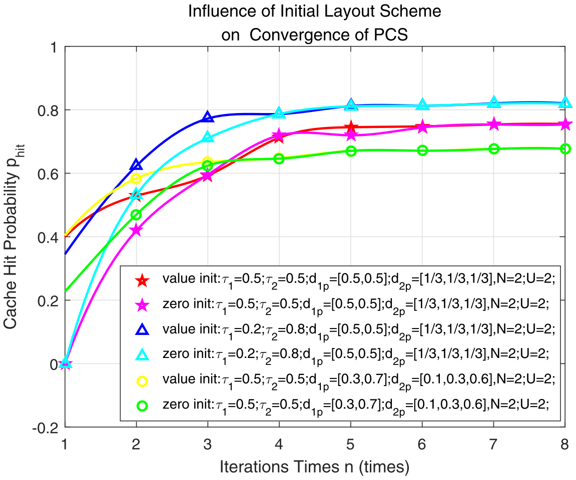

FIGURE 3. Impact of initial layout scheme on convergence of PCS.

TABLE 1. Optimal cache strategy for IH.

Cache Memory size M_1=1		Cache Memory size M_1=2	
0.5 0.5 0	0.9327 0.5251 0.0000	0.5 0.5 0	0.9953 0.9975 0.0000
1/3 1/3 1/3	0.0087 0.0033 0.0025	1/3 1/3 1/3	0.9463 0.0264 0.1704
The optimal cache hit probability: 0.7537			

the second caching network (noted by triangle) have larger cache memory sizes. However, in the third caching network (noted by circular), popular file size is larger. In the above three caching networks, the second caching network has the highest cache hit probability while the third caching network is the lowest. Therefore, it can be obtained that UEs with larger cache memory sizes have higher cache hit probability, and the size of file will also affect the caching intention of UEs, that is, compared to more popular but larger size files, the users may prefer to the smaller size but less popular ones.

Furthermore, we can obtain the optimal cache strategy A corresponding to the maximum cache hit probability. Taking the IH condition as an example, the optimal caching strategy is given in TABLE 1, where the rank denotes the popularity order and the column denotes file group. Moreover, if the caching probability of the initial strategy is set to be 0, it indicates that the file does not exist. From the strategy, we can prioritize caching more desired contents, to maximize the use of limited cache memory size.

We compare the proposed caching strategy with two other exsiting ones as follows:

(1) Packet Caching Strategy (PCS), which is the proposed strategy, is to consider the combined impact of different UE cache memory sizes and requested file sizes.

(2) Separate Cache Placement (SCP), is the strategy that the users are assumed to have equal cache memory size and the requested file sizes are normalized [11].

(3) Joint Cache Placement (JCP), is the strategy that the different UE cache memory sizes are considered, but the requested file sizes are normalized [12].

In Fig. 4, Fig. 5, Fig. 6 and Fig. 7, four effecting parameters (the content popularity parameter γ, user request probability ρ of a file, D2D communication radius R and the user distribution density λ) of the cache hit probability are studied, separately. The same color denotes the same

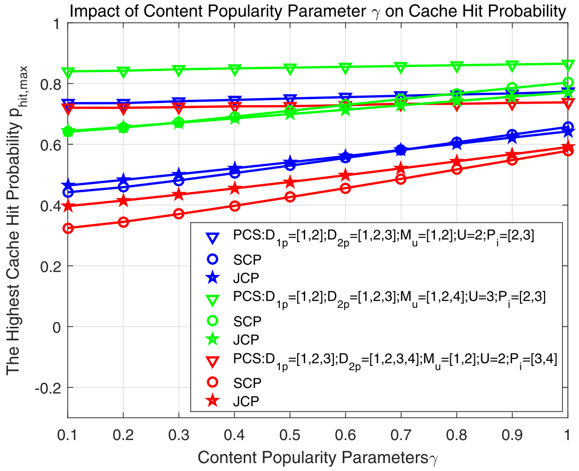

FIGURE 4. Impact of content popularity parameter γ on cache hit probability.

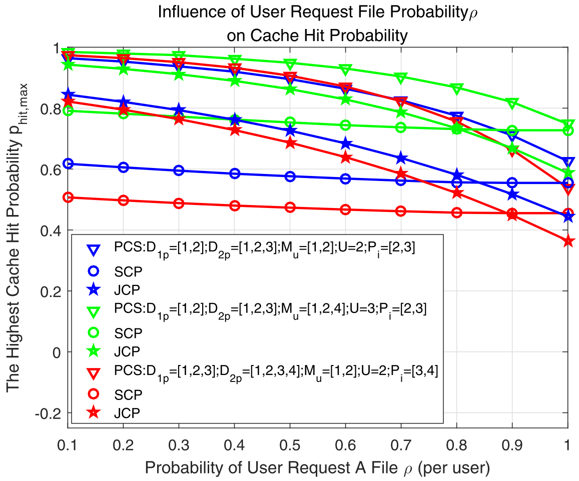

FIGURE 5. Impact of user request probability of a file ρ on cache hit probability.

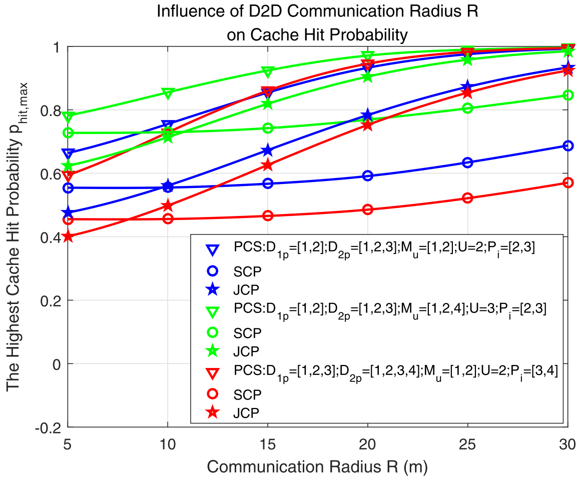

FIGURE 6. Impact of D2D communication radius R on cache hit probability.

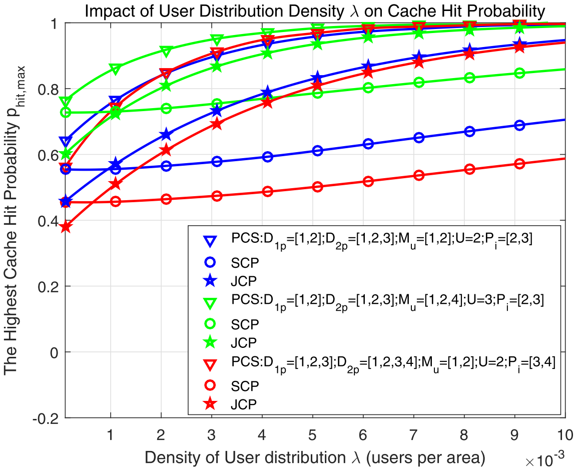

FIGURE 7. Impact of user distribution density λ on cache hit probability.

caching network, different point shapes represent different cache strategies. Fig. 4 shows the impact of popularity parameter γ on $p_{hit,\max}$. From Fig. 4, we can observe that SCP and JCP have an obvious upward trend while the content popularity parameter γ increases. However, PCS is relatively flat. Comparing PCS with JCP, we know that the growth rate of PCS is obviously lower than that of JCP when the probability q_i of the i-th order popular files being requested increases, that is, the requested file size may prevent the increase of cache hit probability. With the normalization of file size, that is, the file size is small and the same, the more popular the content is, the more users will request. While the file size is large, the users have to concern about whether if their cache memory size can support, not always more users will cache more popular files. The cache hit probability may almost the same while the file size is too large for the cache memory size, even though the file is the most popular of all. This proves again that the larger the requested file size, the lower the cache hit probability is.

In Fig. 5, all curves show that the optimal cache hit probability decreases as user request probability of a file increases. User request probability of a file can also be understood as the proportion of request users to all users in the caching network. When the number of request users increases, the number of potential D2D transmitters in the network decreases. Therefore, some UEs cannot make full use of adjacent resources, only rely on the files cached by their own memories, this will lead the fact that the cache hit probability of the system become relatively reduced. In addition, the curve of PCS and JCP is closer, while their interval increases as ρ increases. This is because the file size is normalized in JCP algorithm and request probability of each popularity order represents that of a file. Under the same number of files, user request probability of a file in JCP is relatively low so that the caching probability of file is also reduced. Compared with PCS, when the proportion of request UEs increases, the probability of request UEs obtaining a file is also reduced. Thus, the caching strategy PCS proposed in this paper has a higher cache hit probability.

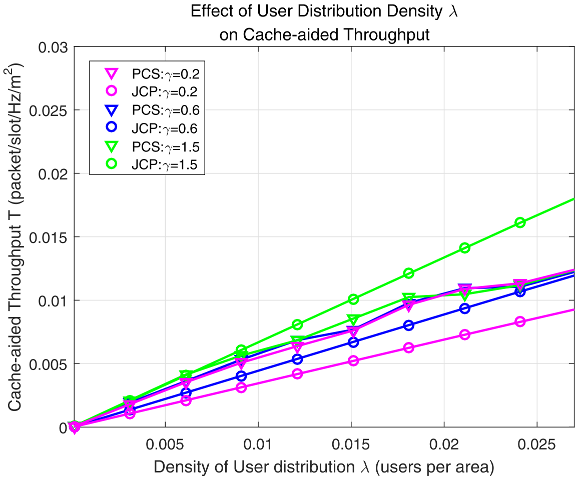

FIGURE 8. Impact of user distribution density λ on cache-aided throughput.

In Fig. 6, the cache hit probability affected by different D2D communication radius is analyzed. In SCP, each popularity order corresponds to a file of size 1, we called it file library in this paper, while PCS supposes few files of different sizes with the same popularity order. That means, within a certain popularity order, PCS considers more files in the system, which is closer to reality. That is why PCS is superior to SCP. With the expansion of communication distance, more devices are included, then the possibility of finding the requested file also increases.

In Fig. 7, the number of users become dense in the D2D networks when the distribution density λ increases, more UEs can provide D2D services for nearby UEs, which leads the increase of the cache hit probability. Compared with the other two existing strategy, PCS still has the highest cache hit probability, which also demonstrates the effectiveness of PCS.

B. CACHE-AIDED THROUGHPUT

In this section, we mainly study the impact of UE distribution density λ and the probability of UE requesting the i-th order popular file library q_i on cache-aided throughput. Set UE transmission power $\widetilde{P}_d = 0.1$ mW, the background noise power $\sigma^2 = -140$ dB and the target SINR of successful D2D transmissions $\varepsilon = 0$ dB.

In Fig. 8, the cache-aided throughput $\mathcal{T}$ increases as user distribution density λ increases when γ is constant, which indicates that the number of D2D successful transmission per unit area increases in the intensive environment. The reason is that the distances of UEs are relatively shorter in the intensive environment, the probability of D2D successful transmission is higher, and hence the cache-aided throughput also increases. In addition, if λ is constant, the higher the user request probability of a file, the higher the D2D cache hit probability is, the probability of successful D2D transmission increases, and hence the cache-aided throughput increases. Compared with JCP, when $\lambda = 0.2$ and $\lambda = 0.6$, PCS performs better in the cache-aided throughput, while $\lambda = 1.5$,

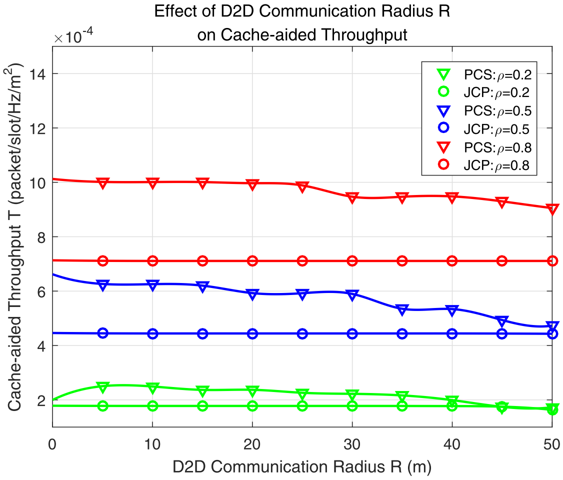

FIGURE 9. Impact of D2D communication radius *R* on cache-aided throughput.

PCS cannot be better than JCP. The reason is that JCP set the file size as 1, and set the users memory size unequal and lager than 1, that means, in density environment, when file become more popular, the cache hit probability will highly increase, which leads the increase of the cache-aided throughput. To some extent, that is unreasonable. In PCS, though in the density environment, and the file is much more popular, the cache hit probability is still limited by the combine effect of file size and cache memory size.

Fig. 9 shows the impact of D2D communication radius R and the proportion of request UEs ρ on cache-aided throughput. From Fig. 9, a higher cache-aided throughput can be obtained in the environment of more request users. The reason is that, when the request users increase, more files will be requested, so that the cache-aided throughput will also increase. when the proportion of request users is fixed, the cache-aided throughput decreases with the increase of D2D communication distance. Because of the effect of interference and fading of signals, some requests cannot be successfully and simultaneously handled by local caches. Although the requested content has been cached in the network, the D2D connect is failed, which leads the cache-aided throughput not as high as small-scale range.

VI. CONCLUSION

In this paper, we propose a cache algorithm PCS based on different cache memory sizes and file sizes to maximize the cache hit probability and cache-aided throughput. Users are grouped according to cache memory sizes and file sizes, and the optimal caching strategy is solved by KKT conditions, and the corresponding cache-aided throughput is finally obtained. The superiority and effectiveness of the algorithm PCS are verified by numerical simulation. Specifically, the following insights are observed.

- The requested file size will affect the user's willingness to cache the file.
- In the intensive environment, more users cached popular files, the probability that users find requested files from nearby UEs is higher, that means D2D service will be more popular.
- In areas of popular files, the willingness of users caching files will increase, the cache hit probability and the probability of D2D successful transmission will also increase.

When the proposed caching algorithm is applied to the mobile scenario, a coupling effect caused by the interference and fading of signals will be considered in the proposed strategy. Besides, the coupling relationship between D2D in small-cell networks collaborative caching and the combining effect of file size and cache memory size is the focus of our next research.

APPENDIXES

APPENDIX A

PROOF OF LEMMA 1

Considering (18), two order partial derivatives wich variable c_{uip} is conducted as follows

$$\frac{\partial^2 p_{hit}}{\partial {c_{uip}}^2} = \frac{\partial^2 p_{hit,j=u,n=i,l=p}}{\partial {c_{uip}}^2} + \frac{\partial^2 p_{hit,j=u,n=i,l\neq p}}{\partial {c_{uip}}^2} + \frac{\partial^2 p_{hit,j=u,n\neq i}}{\partial {c_{uip}}^2} + \frac{\partial^2 p_{hit,j\neq u}}{\partial {c_{uip}}^2}. \quad \text{(A.1)}$$

From (18), the first order partial derivative of c_{uip} in (i) is given as

$$\frac{\partial p_{hit,j=u,n=i,l=p}}{\partial c_{uip}} = m_u d_{ip} q_i e^{-g_u} \left(1 - b_u c_{uip} + b_u\right). \quad \text{(A.2)}$$

The second order partial derivative of c_{uip} in (i) is given as

$$\begin{aligned} \frac{\partial^2 p_{hit,j=u,l=p}}{\partial {c_{uip}}^2} &= m_u d_{ip} q_i b_u e^{-g_u} \left(1 - b_u c_{uip} + b_u\right) \\ &\quad + m_u d_{ip} q_i e^{-g_u} \cdot \left(-b_u\right) \\ &= m_u d_{ip} q_i b_u e^{-g_u} \left(-2 + b_u c_{uip} - b_u\right). \end{aligned} \quad \text{(A.3)}$$

The positive and negative characteristics of (A.3) can be judged as follows:

$$\begin{aligned} &\because b_u = \pi R^2 (1-\rho) m_u \lambda > 0, 0 \leq c_{uip} \leq 1 \\ &\therefore b_u c_{uip} \leq b_u \\ &\therefore -2 + b_u c_{uip} - b_u < 0 \\ &\because m_u d_{ip} q_i b_u e^{-g_u} > 0 \\ &\therefore \frac{\partial^2 p_{hit,j=u,l=p}}{\partial {c_{uip}}^2} = m_u d_{ip} q_i b_u e^{-g_u} \left(-2 + b_u c_{uip} - b_u\right) < 0. \end{aligned} \quad \text{(A.4)}$$

The first order partial derivative of c_{uip} in (18) (ii) is given as

$$\frac{\partial p_{hit,j=u,n=i,l\neq p}}{\partial c_{uip}} = \sum_{l=1,l\neq p}^{P_i} m_u d_{il} q_i \cdot \left(-b_u\right) \cdot \left(c_{jil} - 1\right) e^{-g_u}. \quad \text{(A.5)}$$

The second order partial derivative of c_{uip} in (18) (*ii*) is given as

$$\frac{\partial^2 p_{hit,j=u,n=i,l\neq p}}{\partial {c_{uip}}^2} = \sum_{l=1,l\neq p}^{P_i} m_u d_{il} q_i {b_u}^2 \left(c_{jil}-1\right) e^{-g_u}. \quad \text{(A.6)}$$

The same as the analysis in (A.4), the positive and negative characteristics of (A.6) can be obtained as

$$\frac{\partial^2 p_{hit,j=u,n=i,l\neq p}}{\partial {c_{uip}}^2} = \sum_{l=1,l\neq p}^{P_i} m_u d_{il} q_i {b_u}^2 \left(c_{jil}-1\right) e^{-g_u} < 0. \quad \text{(A.7)}$$

Similarly, the positive and negative characteristics of the second order partial derivative of c_{uip} in (18) (*iii*) and (*iv*) can be observed, respectively.

For (*iii*) in (18):

$$\frac{\partial p_{hit,j=u,n\neq i}}{\partial c_{uip}} = m_u \sum_{n=1,n\neq i}^{N} q_n \sum_{l=1}^{P_n} d_{nl}\cdot\left(-b_u\right)\cdot\left(c_{unl}-1\right) e^{-g_u}. \quad \text{(A.8)}$$

$$\frac{\partial^2 p_{hit,j=u,n\neq i}}{\partial {c_{uip}}^2} = m_u \sum_{n=1,n\neq i}^{N} q_n \sum_{l=1}^{P_n} d_{nl} {b_u}^2 \left(c_{unl}-1\right) e^{-g_u} < 0. \quad \text{(A.9)}$$

For (*iv*) in (18):

$$\frac{\partial p_{hit,j\neq u}}{\partial c_{uip}} = \sum_{j=1,j\neq u}^{U} \sum_{n=1}^{N} \sum_{l=1}^{P_n} m_u d_{nl} q_n \cdot\left(-b_u\right)\cdot\left(c_{jnl}-1\right) e^{-g_u}. \quad \text{(A.10)}$$

$$\frac{\partial^2 p_{hit,j\neq u}}{\partial {c_{uip}}^2} = \sum_{j=1,j\neq u}^{U} \sum_{n=1}^{N} \sum_{l=1}^{P_n} m_j q_n d_{np} {b_u}^2 \left(c_{jnl}-1\right) e^{-g_u} < 0. \quad \text{(A.11)}$$

Substituting (A.4), (A.7), (A.9), (A.11) into (A.1), we can obtain that (A.1) is negative, that is, $\frac{\partial^2 p_{hit}}{\partial {c_{uip}}^2} < 0$.

APPENDIX B
PROOF OF COROLLARY 2

Assuming that c_{uip} is an unknown variable, and the rest caching probability c_{jnl} ($j \in [1, 2, \cdots U]$, $n \in [1, 2, \cdots N]$, $l \in [1, 2, \cdots P_n]$) are constants. Expanding (20), the following equation is obtained.

$$\begin{aligned}
\mathcal{L}\left(\mathrm{A}, \mu\right) = &-1 - \underbrace{m_u d_{ip} q_i \left(c_{uip}-1\right) e^{-g_u}}_{j=u,n=i,l=p} \\
&- \underbrace{m_u \sum_{l=1,l\neq p}^{P_i} d_{il} q_i \left(c_{uil}-1\right) e^{-g_u}}_{j=u,n=i,l\neq p} \\
&- \underbrace{m_u \sum_{n=1,n\neq i}^{N} q_n \sum_{p=1}^{P_n} d_{np} \left(c_{unp}-1\right) e^{-g_u}}_{j=u,n\neq i} \\
&- \underbrace{\sum_{j=1,j\neq u}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} d_{nl} \left(c_{jnp}-1\right) e^{-g_u}}_{j\neq u} \\
&+ \sum_{k=1}^{U} \mu_k \left(\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{kip}\cdot D_{ip} - M_k\right). \qquad \text{(B.1)}
\end{aligned}$$

Let $\frac{\delta\mathcal{L}(\mathrm{A},\mu)}{\delta c_{uip}} = 0$:

$$\begin{aligned}
\frac{\delta\mathcal{L}\left(\mathrm{A}, \mu\right)}{\delta c_{uip}}
= &\left(-1 - \underbrace{m_u d_{ip} q_i \left(c_{uip}-1\right) e^{-g_u}}_{j=u,n=i,l=p}\right)' \\
&- \left(\underbrace{m_u \sum_{l=1,l\neq p}^{P_i} d_{il} q_i \left(c_{uil}-1\right) e^{-g_u}}_{j=u,n=i,l\neq p}\right)' \\
&- \left(\underbrace{m_u \sum_{n=1,n\neq i}^{N} q_n \sum_{p=1}^{P_n} d_{np} \left(c_{unp}-1\right) e^{-g_u}}_{j=u,n\neq i}\right)' \\
&- \left(\underbrace{\sum_{j=1,j\neq u}^{U} m_j \sum_{n=1}^{N} q_n \sum_{l=1}^{P_n} d_{nl} \left(c_{jnp}-1\right) e^{-g_u}}_{j\neq u}\right)' \\
&+ \left(\mu_u \left(\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{uip}\cdot D_{ip} - M_u\right)\right)' \\
&+ \left(\sum_{k=1,k\neq u}^{U} \mu_k \left(\sum_{i=1}^{N}\sum_{p=1}^{P_i} c_{uip}\cdot D_{ip} - M_k\right)\right)'. \qquad \text{(B.2)}
\end{aligned}$$

Substituting (A.3), (A.6), (A.9), (A.11) into (B.2), we have

$$\begin{aligned}
\frac{\delta\mathcal{L}\left(\mathrm{A}, \mu\right)}{\delta c_{uip}}
= &-\underbrace{m_u d_{ip} q_i e^{-g_u} \left(1 - b_u c_{uip} + b_u\right)}_{j=u,n=i,l=p} \\
&- \underbrace{\sum_{l=1,l\neq p}^{P_i} m_u d_{il} q_i \cdot\left(-b_u\right)\cdot\left(c_{jil}-1\right) e^{-g_u}}_{j=u,n=i,l\neq p} \\
&- \underbrace{m_u \sum_{n=1,n\neq i}^{N} q_n \sum_{l=1}^{P_n} d_{nl}\cdot\left(-b_u\right)\cdot\left(c_{unl}-1\right) e^{-g_u}}_{j=u,n\neq i}
\end{aligned}$$

$$\begin{aligned}
&-\underbrace{\sum_{j=1,j\neq u}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} m_j d_{nl} q_n \cdot \left(-b_u\right)\cdot\left(c_{jnl}-1\right)e^{-g_u}}_{j\neq u}\\
&+\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}\\
&= -e^{-g_u}\left[\upsilon_u-\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\left(-\upsilon_j b_u c_{jnl}+\upsilon_j b_u\right)\right]\\
&+\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}\\
&= -e^{-g_u}\left[\upsilon_u-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}m_j q_n d_{nl}\left(1-c_{jnl}\right)\right]\\
&+\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}\\
&= -e^{-x}\left[m_u d_{ip} q_i-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}m_j q_n d_{nl}\left(1-c_{jnl}\right)\right]\\
&+\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}, \qquad \text{(B.3)}
\end{aligned}$$

where $\upsilon_u = m_u d_{ip} q_i$, $\upsilon_j = m_j d_{nl} q_n$, $x = -\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} b_j c_{jnl}$.

$$\begin{aligned}
&\because \frac{\delta\mathcal{L}\left(\mathrm{A},\mu\right)}{\delta c_{uip}}=0\\
&\therefore -e^{-x}\left[\upsilon_u-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\upsilon_j\left(1-c_{jnl}\right)\right]\\
&\quad+\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}=0. \qquad \text{(B.4)}\\
&\therefore e^{-x}\left[\upsilon_u-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\upsilon_j\left(1-c_{jnl}\right)\right]\\
&\quad=\mu_u\sum_{i=1}^{N}\sum_{p=1}^{P_i}D_{ip}.
\end{aligned}$$

$$\mu_u=\frac{e^{-x}}{\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}}\left[\upsilon_u-b_u\sum_{j=1}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\upsilon_u\left(1-c_{jnl}\right)\right]. \qquad \text{(B.5)}$$

Finally, the relaxation variable μ_u ($u \in [1, 2, \cdots U]$) is obtained in the similar way.

APPENDIX C
PROOF OF COROLLARY 3

Based on the assumptions in Appendix B, the optimal caching probabilities can be obtained according to (B.5). Expanding (B.5).

$$\begin{aligned}
\because b_u &= \pi R^2\left(1-\rho\right)m_u\lambda\\
\therefore \mu_u &= \frac{b_u}{m_u}m_j = b_j = \pi R^2\left(1-\rho\right)m_j\lambda. \qquad \text{(C.1)}\\
\therefore \mu_u &= \frac{e^x\upsilon_u+e^x\left[\sum\limits_{j=1}^{U}\sum\limits_{n=1}^{N}\sum\limits_{l=1}^{P_n}v-\sum\limits_{j=1}^{U}\sum\limits_{n=1}^{N}\sum\limits_{l=1}^{P_n}vc_{jnl}\right]}{\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}}. \qquad \text{(C.2)}
\end{aligned}$$

where $v = b_j q_n d_{nl}$.

Let $b_m = \sum\limits_{j=1}^{U}\sum\limits_{n=1}^{N}\sum\limits_{l=1}^{P_n} b_j q_n d_{nl}$, $\overset{\frown}{b} = \sum\limits_{j=1}^{U} b_j = \sum\limits_{j=1}^{U}\pi R^2\left(1-\rho\right)m_j\lambda$

$$\begin{aligned}
&\because \sum_{j=1}^{U} m_j = 1\\
&\therefore \overset{\frown}{b} = \pi R^2\left(1-\rho\right)\lambda. \qquad \text{(C.3)}
\end{aligned}$$

Substituting b_m, $\overset{\frown}{b}$ into (C.2), we obtain

$$\begin{aligned}
&\mu_u = \frac{m_u}{\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}}\cdot e^x\left[d_{ip}q_i+b_m+\frac{b_m x}{\overset{\frown}{b}}\right]\\
&\therefore \frac{b_m}{\overset{\frown}{b}}\cdot e^x\left[\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right]=\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}}{m_u}\\
&e^x\left[\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right]=\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}\overset{\frown}{b}}{m_u b_m}\\
&e^{\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right)-\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}\right)}\cdot\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right)\\
&\quad=\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}\overset{\frown}{b}}{m_u b_m}\\
&e^{\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right)}\cdot\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}+x\right)\\
&\quad=\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}D_{ip}\overset{\frown}{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\overset{\frown}{b}}{b_m}+\overset{\frown}{b}\right)}. \qquad \text{(C.4)}
\end{aligned}$$

Let $\varpi = \frac{d_{ip}q_i\widehat{b}}{b_m} + \widehat{b} + x$,

$$e^{\varpi}\cdot\varpi = \frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}. \tag{C.5}$$

According to Lambert W Function, (C.5) can be rewritten as

$$\begin{aligned}
&\frac{d_{ip}q_i\widehat{b}}{b_m} + \widehat{b} + x \\
&= \mathcal{W}\left(\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right) \\
&x = \mathcal{W}\left(\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right) - \frac{d_{ip}q_i\widehat{b}}{b_m} - \widehat{b}.
\end{aligned} \tag{C.6}$$

$$\begin{aligned}
&\because x = -\sum_{n=1}^{N}\sum_{j=1}^{U}\sum_{l=1}^{P_n} b_j c_{jnl} \\
&\quad \therefore -\sum_{n=1}^{N}\sum_{j=1}^{U}\sum_{l=1}^{P_n} b_j c_{jnl} \\
&= \mathcal{W}\left(\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right) - \frac{d_{ip}q_i\widehat{b}}{b_m} - \widehat{b}.
\end{aligned} \tag{C.7}$$

Expanding $\sum\limits_{n=1}^{N}\sum\limits_{j=1}^{U}\sum\limits_{l=1}^{P_n} b_j c_{jnl}$ in (C.7)

$$\begin{aligned}
&b_u\left(c_{uip} + \sum_{l=1,l\neq p}^{P_i} c_{uil} + \sum_{n=1,n\neq i}^{N}\sum_{l=1}^{P_n} b_j c_{unl}\right) \\
&\quad + \sum_{j=1,j\neq u}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} c_{jnl} \\
&\quad = \frac{d_{ip}q_i\widehat{b}}{b_m} + \widehat{b} - \mathcal{W}\left(\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right)
\end{aligned} \tag{C.8}$$

$$\begin{aligned}
b_u c_{uip} &= \frac{d_{ip}q_i\widehat{b}}{b_m} + \widehat{b} - \mathcal{W}(\bullet) \\
&\quad - b_u\left(\sum_{l=1,l\neq p}^{P_i} c_{uil} + \sum_{n=1,n\neq i}^{N}\sum_{l=1}^{P_n} c_{unl}\right) \\
&\quad - \sum_{j=1,j\neq u}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n} b_j c_{jnl}.
\end{aligned} \tag{C.9}$$

After simplification, the general expression of optimal caching probability $\widehat{c}_{uip}$ is obtained as

$$\begin{aligned}
\widehat{c}_{uip} &= \frac{d_{ip}q_i\widehat{b}}{b_m b_u} + \frac{\mathcal{W}\widehat{b}}{b_u} - \frac{\left(\frac{\mu_u\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} D_{ip}\widehat{b}}{m_u b_m}\cdot e^{\left(\frac{d_{ip}q_i\widehat{b}}{b_m}+\widehat{b}\right)}\right)}{b_u} \\
&\quad - \sum_{l=1,l\neq p}^{P_i} c_{uil} - \sum_{n=1,n\neq i}^{N}\sum_{l=1}^{P_n} c_{unl} - \sum_{j=1,j\neq u}^{U}\sum_{n=1}^{N}\sum_{l=1}^{P_n}\frac{b_j}{b_u}c_{jnl}.
\end{aligned} \tag{C.10}$$

APPENDIX D
PROOF OF LEMMA 4

Considering $p_{d2d,fip}^{suc}$ in (11) and the approximation in (25), $p_{d2d,fip}^{suc}$ can be rewritten as

$$\begin{aligned}
&\widehat{p}_{d2d,fip}^{suc} \\
&\quad = \mathbb{E}_{s_{fip}}\left[\exp\left(-\frac{\varepsilon s_{fip}^{\alpha}\sigma^2}{\widetilde{P}_{\mathrm{d}}}\right)\right] \\
&\quad\quad \cdot\mathbb{E}_{s_{fip}}\left[\exp\left(-\frac{\pi\rho m_u\lambda p_{hit,d2d}\cdot s_{fip}^2\cdot\varepsilon^{\frac{2}{\alpha}}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}\right)\right]
\end{aligned} \tag{D.1}$$

$$\begin{aligned}
&\approx \exp\left(-\frac{\varepsilon\mathbb{E}\left[s_{fip}^2\right]^{\frac{\alpha}{2}}\sigma^2}{\widetilde{P}_{\mathrm{d}}}\right) \\
&\quad\cdot\exp\left(-\frac{\pi\rho m_u\lambda p_{hit,d2d}\cdot\mathbb{E}\left[s_{fip}^2\right]\cdot\varepsilon^{\frac{2}{\alpha}}}{\operatorname{sinc}\left(\frac{2}{\alpha}\right)}\right).
\end{aligned} \tag{D.2}$$

Combining the PDF in (16), $\mathbb{E}\left[s_{fip}^2\right]$ is given as

$$\begin{aligned}
&\mathbb{E}\left[s_{fip}^2\right] \\
&= \int_0^{\infty} r^2 f_{s_{fip}}(r)\,dr \\
&= \int_0^{R} r^2\frac{\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i} 2\pi r(1-\rho)m_u\lambda c_{uip}e^{-R_r}}{1-e^{-R_R}}dr \\
&= \frac{\int_0^R R_r e^{-R_r}dr^2}{1-e^{-R_R}} \\
&= \frac{\int_0^R R_r e^{-R_r}d\left[\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi r^2(1-\rho)m_u\lambda c_{uip}\right]}{\left(1-e^{-R_R}\right)\cdot\left(\sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi(1-\rho)m_u\lambda c_{uip}\right)},
\end{aligned} \tag{D.3}$$

where $R_r = \sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi r^2(1-\rho)m_u\lambda c_{uip}$, $R_R = \sum\limits_{u=1}^{U}\sum\limits_{i=1}^{N}\sum\limits_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}$.

Let $\kappa = \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi r^2(1-\rho)m_u\lambda c_{uip}$, and $\kappa \in \left[0, \sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}\right]$

$$
\begin{aligned}
&\mathbb{E}\left[s_{fip}^2\right]\\
&= \frac{\int_0^{\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}} \kappa e^{-\kappa}d\kappa}{\left(1-e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}\right)\frac{\kappa}{r^2}}\\
&= \frac{-\int_0^{\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}} \kappa de^{-\kappa}}{\left(1-e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}\right)\cdot\frac{\kappa}{r^2}}\\
&= \frac{-\kappa e^{-\kappa}\Big|_0^{\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}}{\left(1-e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}\right)\cdot\frac{\kappa}{r^2}}\\
&\quad + \frac{\int_0^{\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}} e^{-\kappa}d\kappa}{\left(1-e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}\right)\cdot\frac{\kappa}{r^2}}\\
&= \frac{R_R e^{-R_R} + e^{-\kappa}\Big|_0^{R_R}}{\left(1-e^{-R_R}\right)\cdot\left(\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi(1-\rho)m_u\lambda c_{uip}\right)}\\
&= \frac{R^2\cdot e^{-R_R}}{1-e^{-R_R}} - \frac{1}{\left(\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi(1-\rho)m_u\lambda c_{uip}\right)}
\end{aligned}
\tag{D.4}
$$

$$
\begin{aligned}
&\therefore \mathbb{E}\left[s_{fip}^2\right]\\
&= \frac{R^2\cdot e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}}{1-e^{-\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi R^2(1-\rho)m_u\lambda c_{uip}}}\\
&\quad - \frac{1}{\left(\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}\pi(1-\rho)m_u\lambda c_{uip}\right)}.
\end{aligned}
\tag{D.5}
$$

Combining (D.5) with (8), the approximated cache-aided throughput is obtained as follows.

$$
\begin{aligned}
\mathcal{T} &= \sum_{u=1}^{U}\rho m_u\lambda\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}q_i d_{nl}c_{uip}\cdot 1\cdot 1\\
&\quad + \sum_{u=1}^{U}\rho m_u\lambda\sum_{u=1}^{U}\sum_{i=1}^{N}\sum_{p=1}^{P_i}q_i d_{ip}\left(1-c_{uip}\right)\\
&\quad \cdot p_{hit,fip}^{d2d}\cdot\widehat{p}_{d2d,fip}^{suc},
\end{aligned}
\tag{D.6}
$$

where $\widehat{p}_{d2d,fip}^{suc}$ is given in (D.2), the result of $\mathbb{E}\left[s_{fip}^2\right]$ is given in (D.5).

REFERENCES

[1] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching under heterogeneous file preferences," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 444–457, Oct. 2017.

[2] 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, 4th Quart., 2014.

[3] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[4] H. Xu, W. Xu, Z. Yang, J. Shi, and M. Chen, "Pilot reuse among D2D users in D2D underlaid massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 467–482, Jan. 2018.

[5] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27–34, Mar. 2011.

[6] N. Zhao, X. Liu, Y. Chen, S. Zhang, Z. Li, B. Chen, and M.-S. Alouini, "Caching D2D connections in small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12326–12338, Dec. 2018.

[7] 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.

[8] N. Carlsson and D. Eager, "Ephemeral content popularity at the edge and implications for on-demand caching," *IEEE Trans. Parallel Distrib. Syst.*, vol. 28, no. 6, pp. 1621–1634, Jun. 2017.

[9] N. Pappas, Z. Chen, and I. Dimitriou, "Throughput and delay analysis of wireless caching helper systems with random availability," *IEEE Access*, vol. 6, pp. 9667–9678, 2018.

[10] M. Choi, J. Kim, and J. Moon, "Wireless video caching and dynamic streaming under differentiated quality requirements," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1245–1257, Jun. 2018.

[11] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Nov. 2017.

[12] Y. Long, W. U. Dan, and Y. Cai, "Cache placement optimization scheme in D2D networks with heterogeneous cache capacity," *J. Comput. Appl.*, no. 4, pp. 1–5, May 2018.

[13] M. Hajimirsadeghi, N. B. Mandayam, and A. Reznik, "Joint caching and pricing strategies for popular content in information centric networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 3, pp. 654–667, Mar. 2017.

[14] K. Takeda, "USENIX symposium on Internet technologies and systems proceedings," *J. Periodontol.*, vol. 19, no. 4, pp. 697–713, 1997.

[15] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM Conf. Comput. Commun., 18th Annu. Joint Conf. IEEE Comput. Commun. Societies*, Mar. 1999.

[16] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.

[17] H. J. Kang and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework," *J. Commun. Netw.*, vol. 16, no. 5, pp. 568–577, Oct. 2014.

[18] S. H. Chae and W. Choi, "Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6626–6637, Oct. 2016.

[19] Y. Zhang, E. Pan, L. Song, W. Saad, Z. Dawy, and Z. Han, "Social network aware device-to-device communication in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 177–190, Jan. 2015.

[20] Y. Qiu, C. Tan, and M. Zheng, "Throughput maximization for D2D model selection in 5G network," *Inf. Technol.*, vol. 4, pp. 102–105, Apr. 2017.

[21] M. Ni, L. Zheng, F. Tong, J. Pan, and L. Cai, "A geometrical-based throughput bound analysis for device-to-device communications in cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 100–110, Jan. 2015.

[22] W. B. Tang and S. Q. Li, "Linear precoding based on maximum SJNR in multi-user MIMO downlink," *Acta Electron. Sinica*, vol. 35, no. 1, pp. 157–160, 2007.

YUJING LIN was born in Fuzhou, China, in 1995. She received the B.S. degree in electronic information engineering from Fuzhou University, Fuzhou, China, in 2018. She is currently pursuing the master's degree with the Department of Communication Engineering, Xiamen University, China. Her current research interests are D2D communications and D2D caching.

ZHIJIAN LIN received the B.S., M.S., and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2008, 2011, and 2017, respectively. In 2016, he was a Visiting Scholar with the University of NC State, USA. He is currently working as an Assistant Professor with the Department of Electronic and Information Engineering, Fuzhou University, China. His current research interests include D2D communications, D2D caching, and NOMA.

PINGPING CHEN received the Ph.D. degree in electronic engineering, Xiamen University, China, in 2013. From May 2012 to September 2012, he was a Research Assistant in electronic and information engineering with The Hong Kong Polytechnic University, Hong Kong. From January 2013 to January 2015, he was a Postdoctoral Fellow with the Institute of Network Coding, The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with Fuzhou University, China. His primary research interests include channel coding, joint source and channel coding, network coding, and UWB communications.

ZHIFENG CHEN received the Ph.D. degree in electrical and computer engineering from the University of Florida, USA, in 2010. He is currently a Professor with the College of Physics and Information Engineering, Fuzhou University, China. His research interests include video coding, video transmission, computer vision, and machine learning.

LINHUANG WU received the Ph.D. degree in electronic engineering from Xiamen University, China, in 2013. From May 2012 to September 2012, he was a Research Assistant in electronic and information engineering with The Hong Kong Polytechnic University, Hong Kong. From January 2013 to January 2015, he was a Postdoctoral Fellow with the Institute of Network Coding, The Chinese University of Hong Kong, Hong Kong. He is currently a Professor with Fuzhou University, China. His primary research interests include channel coding, joint source and channel coding, network coding, and UWB communications.