User Preference Aware Caching Deployment for Device-to-Device Caching Networks

Tiankui Zhang, *Senior Member, IEEE,* Hongmei Fan, Jonathan Loo, *Member, IEEE,* and Dantong Liu

***Abstract*—Content Caching in the Device-to-Device (D2D) cellular networks can be utilized to improve the content delivery efficiency and reduce traffic load of cellular networks. In such cache enabled D2D communication networks, how to cache the diversity contents in the multiple cache enabled mobile terminals, namely, the caching deployment, has a substantial impact on network performance since the cache space in a mobile terminal is relatively small compared with the huge amounts of multimedia contents. In this paper, a user preference aware caching deployment algorithm is proposed for D2D caching networks. Firstly, the user preference is defined to measure the interests of users on the data contents, and based on this, the definition of user interest similarity is given. Then a content cache utility of a mobile terminal is defined by taking the communication coverage of this mobile terminal and the user interest similarity of its adjacent mobile terminals into consideration. A general cache utility maximization problem with joint caching deployment and cache space allocation is formulated, where the special logarithmic utility function is integrated. In doing so, the caching deployment and the cache space allocation can be decoupled by equal cache space allocation. Subsequently, we relax the logarithmic utility maximization problem, and obtain a low complexity near-optimal solution via dual decomposition method. Compared with the existing caching placement methods, the proposed algorithm can achieve significant improvement on cache hit ratio, content access delay and traffic offloading gain.**

***Index Terms*—device-to-device communication, content caching, user preference**

I. Introduction

TODAY'S internet traffic internet traffic is dominated by content distribution and retrieval. With the rapid explosion of the data volume and content diversity, it becomes challenging to deliver high quality service to the end user efficiently and securely. In the pioneer work [1], opportunistic multihop transmission had been considered to offload network traffic by exploiting the mobile devices capabilities in the cellular networks. Recently, content caching, a widely adopted content delivery technique in Internet for reducing network traffic load, has been exploited in fifth generation (5G) mobile networks. It has been proven that caching of popular content and pushing them close to consumers can significantly reduce the mobile traffic [2]. [3], [4] have explicitly demonstrated the role of the caching technology in enabling the 5G networks.

With this, in the cellular networks, there have been many works utilizing the cache to improve the system performance of the network. Authors in [5]–[7] have introduced the idea of femto-caching helpers, which are small base stations (BSs) with a low-bandwidth backhaul link and high storage capabilities. Recent work [8]–[11] have shown that one of the most promising approaches for the system performance improvement relies on caching, i.e., storing the video files in the users local caches and/or in dedicated helper nodes distributed in the network coverage area.

Intuitively, caching provides a way to exploit the inherent content reuse while coping with the asynchronism among requests [12]. In addition, caching is appealing since it leverages the wireless devices storage capacity, which can improve the network capacity and mitigate the video stalling [13], minimize the average content access delay [14], and reduce the energy consumption [15]. The content caching can be more efficiency with the help of big data analysis and estimation techniques [16].

Apart from caching, device-to-device (D2D) communication has been regarded as another driving force behind the evolution into 5G, considering D2D communication is able to effectively utilize the air interface resources and offload the cellular network traffic. In the conventional cellular network, a mobile terminal (MT) can only rely on a base station (BS) in the cellular network to acquire the desired content. In the cellular network with D2D, the prospect of cellular communication applications can be extended with direct communication capabilities between devices. For example,if the neighbor MTs have the same content, the content can be directly delivered from his neighbor devices.

D2D caching networks, which take both advantages of caching and D2D communication technologies, have naturally set the stage for the 5G evolution [8], [9]. In the D2D caching networks, the MTs equipped with storage space are used as caching nodes, and the mobile users collaborative download and cache different parts of the same content simultaneously from the serving BS, and then share them by using D2D communications. Although the cache space on each individual MT is not necessarily large, the cache space of multiple devices can form a large virtual unified cache space, which can cache a large amount of multimedia content. By using storage overhead in exchange for transmission efficiency, the D2D caching networks can offload cellular traffic, reduce content access delay, and improve the user experience.

In view of forming a large virtual unified cache space, caching deployment is a problem that looks into how the diversity contents is cached in multiple cache-enabled MTs as it would lead to significant impact to the network performance of D2D caching networks. Some literatures have studied the

This work was supported in part by NSF of China (No. 61461029).

Tiankui Zhang, Hongmei Fan are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: {zhangtiankui, fhm}@bupt.edu.cn).

Jonathan Loo is with the School of Computing and Engineering, University of West London, United Kingdom (e-mail: jonathan.loo@uwl.ac.uk).

Dantong Liu is with the Chief technology and Architecture office, Cisco Systems Inc., CA 95134, USA (e-mail: datliu@cisco.com).

caching deployment optimization problem [17]–[24], which are discussed in the related works section. These works utilize the D2D cache to achieve their goals by taking the channel state information, the popularity of the content, the available bandwidth resources, the data transmission rate, and the distribution of users into account. However, in the context of multimedia content distribution, especially in wireless social networks, the user's preference for content has a great impact on the cache system performance. In [25], the author pointed out that each data object would eventually be sent to interested users. For users, the closer to the storage location of the data object, the less network traffic it will consume to access the data objects. In the selection of caching location of the content replicas, the users interest preference has certain guidance, and the content replicas should be stored in the location much closer to the user who is interested in it. Therefore, the caching deployment strategy can be designed based on the user preference, considering that each user in the D2D network is a relatively independent entity.

In this paper, we propose a user preference aware caching deployment algorithm for the D2D caching networks. We integrate user preference when formulating the content cache utility, establish the optimization problem of cache utility, and implement the near-optimal caching placement algorithm. The outcomes of this research provide the upper bound of caching performance of D2D networks, and also the performance bounds for the follow-up study of distributed and online caching strategies.

The contributions of this paper are shown as below:

1) In order to improve the caching performance, the content cache utility of each MT is defined to measure the caching utilization. Distinguished with the existing research on the caching deployment optimization problem, the proposed utility definition takes both the user preference and the transmission coverage region into consideration. The rationale behind this is that when the content replicas are cached in the nodes near the nodes generating the content requests, the caching utilization can be improved. As such, the MT should cache some specific contents which may be interested by the adjacent MTs with similar user preference to improve the caching efficiency. In addition, more neighbor MTs in the communication coverage region of a MT will lead to the higher possibility of content sharing of the cached content, and the larger content caching utility of this MT.

2) The content cache utility maximization problem is formulated for the caching performance optimization. The existing works on user preference based caching deployment strategy mainly use heuristic method, while the optimal solution of the optimization problem formulated in this work will give the best caching performance, which can be seen as the upper bound of the caching performance obtained by the caching deployment. Firstly, a general cache utility maximization problem is introduced. Then, a specific logarithmic utility function is adopted, which provides a network-wide proportional fairness. With the logarithmic cache utility, the caching deployment and the cache space allocation coupled maximization problem can be reduced to cache utility maximization problem with equal cache space allocation.

3) Finally, for the logarithmic cache utility maximization problem, the single MT caching constraint is relaxed to multiple MTs caching, which converts the intractable combinatorial problem into a convex optimization problem. Then a near-optimal solution is obtained by dual decomposition method. This solution provides a feasible, efficient and low-overhead algorithm for implementation in D2D caching networks. The simulation results show that the proposed algorithm can converge to the maximization solution in a few iterations and achieve significant performance on cache hit ratio, content access delay and traffic offloading gain.

The rest of the paper is organized as follows: In Section II, we review the related work. Section III presents the system model. Section IV defines the user preference, user interest similarity, and the content cache utility. Section V proposes content cache utility optimization problem. Section VI is the proposed dual decomposition algorithm. Section VII evaluates the performance of the proposed algorithm. Finally is the conclusion of this work. The main symbols and variables used in this paper are summarized in Table I.

II. Related Works

The previous works on D2D communication are mainly focused on how D2D communication can run efficiently as an underlay to cellular networks, and the research concentrated on resource allocation and interference avoidance, see [26] and the reference therein.

With a limited amount of storage on each device, the main challenge is how cellular traffic can be maximally offloaded by using D2D communication to satisfy requests for content as well as to share messages between neighboring devices. A carefully designed caching deployment strategy would have a great impact on the network performance of the D2D caching networks. Some of existing contributions include the caching deployment optimization and strategy designing [17]–[24], the design of D2D cache network structure [10], [11], [27], [28], D2D caching clustering [29], [30], and so on.

In [17], the authors considered the distribution of users request when designing the caching strategy to maximize the probability of successful content delivery. With the consideration of content popularity, a cut-off random caching scheme and a segment-based random caching scheme were proposed to improve the cache hitting probability [18]. In [19], an optimization problem was formulated to determine the probability of storing the individual content that could minimize the average caching failure rate, and then a low-complexity search algorithm was proposed for solving the optimization problem. In [20], the authors formulated a continuous time optimization problem to determine the optimal transmission and caching policies that minimize a generic cost function, such as energy, bandwidth or throughput. In [21], a caching allocation scheme was proposed to enhance storage utilization for D2D networks, and the optimal storage assignment achieved tradeoff between static caching and on-demand relaying. In [22], the authors studied the problem of maximizing cellular traffic offloading with D2D communication by selectively caching popular content locally, and exploring maximal matching for sender-receiver pairs. In [23], the authors optimized the content cache

distribution considering the user's geographical location in the D2D network to improve the cache hitting probability. In [24], the authors combined the channel-aware caching and coded multicasting for wireless video delivery.

The above works utilize the D2D cache to achieve their goals by taking the channel state information, the popularity of the content, the available bandwidth resources, the data transmission rate, and the distribution of users into account. However, in the context of multimedia content distribution, the user's preference for content has a great impact on the cache system performance, especially in wireless social networks.

The user preference is a concept from the social networks and the recommendation systems [31]. In the context of D2D caching networks, there are existing some caching strategies that take into account the users interest preferences [3], [32], [33]. The authors in [3] proposed a mechanism to cache popular contents proactively in the mobile users, in which, the files were delivered to some influential mobile users in a social community and then shared in the community by D2D communications. In [32], users were divided into different clusters according to the users' interest preferences. Then, the corresponding cache strategy was obtained by compromising the average download delay of each group. Considering the difference of users preference and the selfishness nature of D2D users, a caching incentive scheme based on backpack theory was proposed in [33].

It is worth mentioning that user preference based caching strategies have been explored in content centric networking (CCN) recently [34], [35]. Nevertheless, the works in CCN pay more attention to the online and on-path caching decision design, which cannot obtain the overall network performance optimization.

Although the works in [32]–[35] laid a good foundation in integrating user's interest preference to the caching strategy design, the effect of the similarity of the users interest preference on the caching strategy design is less well understood. In the caching strategy design, if the MT caches some specific contents which may be interested by the adjacent MTs, the cache space utilization can be improved. Hence, the content caching of a MT should not only consider the user preference itself, but also take account of the user interest similarity on the contents of adjacent MTs. Our work fills the gap by carefully considering the user interest similarity and the D2D transmission coverage region when defining the content caching utility, thereby improving the network performance via the caching deployment problem optimization.

III. System Model

A. Network model

The system model of this paper is illustrated in Fig. 1. A single macrocell is considered, where a macro BS serves N uniformly distributed D2D users. In this paper, we consider the in-band D2D communication, in which, the D2D users can access the licensed spectrum in a dedicated mode (also described as an overlay or orthogonal mode in the literature). In the dedicated mode, the transmission of the cellular users and the D2D users has assigned a non-overlapping orthogonal radio resource, so there is no interference among cellular users and D2D users, nor interference among D2D users [36].

TABLE I: Symbols and variable list

Parameters	Description
B_M	System bandwidth of downlink macrocell
B_D	system bandwidth of D2D communication
p_{BS}^{tx}	Maximum transmit power of BS
p_n^{tx}	Transmit power of TM n
$c_{nn'}$	Data rate from user n' to user n
c_{n_BS}	Data rate from severing BS to user n
σ^2	Additive white Gaussian noise power
N	Number of users
M	Number of contents
M_n	The number of content cached in MT n
S	The cache ability of each user
v	The size of each content
φ_{mn}	The preference of user n to content m
$\varphi_m(n,n')$	the interest similarity of user n and n'
$d(n,n')$	The distance between user n and n'
x_{mn}	user n caching index for content m
y_{mn}	user n cache space allocation for content m
u_{mn}	user n cache utility per unit cache space for content m

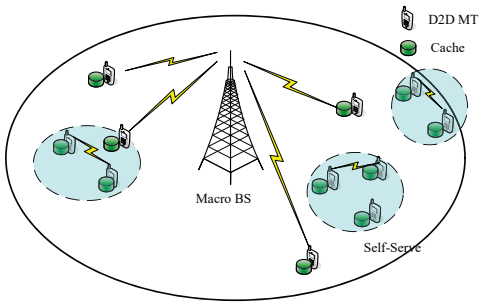

Figure 1. D2D caching networks.

The user n can communicate and share the content directly with his neighbor MTs through D2D communication link if user n cached a content. For a given content, which MT to save the content replica in an overlap region of multiple MTs will be decide by the caching replacement strategy.

The system bandwidth of downlink macrocell is B_M and the system bandwidth of D2D is B_D, we use the time domain Round-Robin scheduling to allocate the radio resource for cellular users and D2D users [36]. When user n is communicated with user n', the data rate from user n' to user n is

$$c_{nn'} = B_D \log\left(1 + \frac{g_{n'n}p_{n'}^{tx}}{\sigma^2}\right), \tag{1}$$

when user n is communicated with BS, the data rate from its severing BS to user n is

$$c_{n\_BS} = B_M \log\left(1 + \frac{g_{BS\_n}p_{BS}^{tx}}{\sigma^2}\right), \tag{2}$$

where σ^2 is additive white Gaussian noise power, $p_{n'}^{tx}$ is the maximum transmit power of user n', p_{BS}^{tx} is the maximum transmit power of BS, and $g_{n'n}$ is the channel gain between user n' and user n , and g_{BS_n} is the channel gain between BS and user n [37].

B. Caching model

Taking into account the diversity of contents within the D2D network, this paper assumes that each cell can cache M contents. More generally, different cells can cache different M

contents, and the value of M in different cells can be various. If considering to relax the assumption, we can further consider that the macro BS (or other functional entity with management function) selects M contents to be cached according to certain criteria, for instance, M contents which having the highest popularity. In addition, due to the limited storage capacity of MT in the actual application scenario, the amount of data in the whole contents is much larger than the cache space available to each MT. The contents are assumed to have the same data size, and the data volume of each content is v. Each MT has a cache space which is able to cache up to S contents.

In our caching model, a user can be a content requester and a content provider. If there exists a complete copy of content m in its own cache, the request is fulfilled with no delay and without the need to establish a communication link. Otherwise, the user broadcasts a request message for the content m to the neighbor MTs within its coverage, if the user can find the requested file from a MT's cache space within its D2D transmission range, then it can establish a D2D communication link and obtain the content. If the user cannot find the requested content neither in its own cache nor its proximity users, it needs to download the file from the BS.

The aforementioned procedure is straightforward and could be implemented via the existing approaches, such as how to establish D2D communication and allocate radio resource, how to measure content popularity and select the most popular contents. This paper mainly focuses on the caching deployment problem in the D2D networks, that is, how to cache M contents among N MTs.

IV. Cache Utility Function

A. User preference and users interest similarity

User preference reflects a user's interest into one content, and can also indirectly reflect the probability that a user requests one content. The users preference for contents are closely related to the type of contents.

We assume that there are K themes for each content in the network, and $\mathrm{W}=\{w_1, w_2 \cdots w_K\}$ represents the set of all themes of each content. The property function of content m under the topic w_k is $Pro\left(m, w_k\right)$, if content m includes the theme w_k , the value of $Pro\left(m, w_k\right)$ is one, otherwise, the value is zero.

The users have their own preference for each type of the themes, and we let the preference function $Pre\left(n, w_k\right)$ representing the preference of the user n for the theme w_k. In this paper, we assume that the user preference function is represented by mutual information [38], which is defined as,

$$Pre\left(n, w_k\right)=I\left(X\left(w_k\right); V_j\right)=\log \frac{p\left(X\left(w_k\right) | V_j\right)}{p\left(X\left(w_k\right)\right)}, \quad (3)$$

where $X\left(w_k\right)$ is the set of all items which contain feature w_k, and $I\left(X\left(w_k\right); V_j\right)$ is the mutual information. $p\left(X\left(w_k\right)\right)$ is the unconditional feature probability representing the probability of contents containing the feature w_k in the whole content set. $p\left(X\left(w_k\right) | V_j\right)$ is the conditional feature probability, i.e., the probability of contents including w_k in the user n history information V_j .

The interest of user n to the content m is defined based on cousin theory, that is,

$$\varphi_{mn}=\frac{\sum_{k=1}^{K} Pro\left(m, w_k\right) Pre\left(n, w_k\right)}{\sqrt{\sum_{k=1}^{K}\left[Pro\left(m, w_k\right)\right]^2} \sqrt{\sum_{k=1}^{K}\left[Pre\left(n, w_k\right)\right]^2}}. \quad (4)$$

The more similar $Pro\left(m, w_k\right)$ and $Pre\left(n, w_k\right)$ are, the higher φ_{mn} is, and $0 \leq \varphi_{mn} \leq 1$.

According to the above definition of the user preference, the interest similarity function is further defined to characterize the interest similarity among users. In this paper, we use a simple model to capture the user interest similarity of real social networks [39]. Since φ_m is within the segment $[0,1]$, the interest similarity between user n and user n' is defined as the Euclidean distance on the wrapped segment,

$$\varphi_m\left(n, n'\right)=\min\left\{\left|\varphi_{mn}-\varphi_{mn'}\right|, 1-\left|\varphi_{mn}-\varphi_{mn'}\right|\right\}. \quad (5)$$

The small distance between φ_{mn} and $\varphi_{mn'}$ is, the larger interest similarity of user n and n' on the content m is. A larger interest similarity between two users indicates that the more likely content cached in one user is requested by the another user.

B. Cache utility

In this paper, we define the cache utility function of a user considering both the communication coverage of this user and the user interest similarity of its adjacent users.

As described above, $\varphi\left(n, n'\right)$ represents the interest similarity between user n and user n'. Besides, we let $d\left(n, n'\right)$ represents the physical distance between the two users, and let Φ_n denotes the set of neighbors of user n in its communication range. In the D2D communication coverage region of a user, the more neighbor users, the higher the possibility of content sharing of the cached content, and the larger the caching utility of this user. Therefore, the cache utility per unit cache space of user n caching content m is defined as,

$$u_{mn}=\sum_{n' \in \Phi_n}\left[\varphi_m(n, n')^{-\alpha} \cdot d(n, n')^{-\beta}\right], \quad (6)$$

where α and β are the weighting factors of the user interest similarity and the user physical distance.

In the cache utility function definition, the D2D transmission coverage region is decided by the physical distance of MTs. We used the physical distance, equivalent to the pathloss, in the cache utility definition, the reason is that, i) from the view of the caching management, it is easy to collect the pathloss among MTs in a macrocell, the caching deployment can be implemented in a very short interval; ii) the timescale of content delivery among MTs is larger than that of channel condition varying with fast fading.

Assumption 1: user n can cache a portion of content m. This assumption is practical and necessary, because a user may have multiple interested content to be cached and the cache space in a MT is relative small compared with the data volume of the multiple contents.

we define a caching index $x_{mn} = 1$ of user n and content m, indicating that a portion of (or entire) content m is cached in user n, otherwise $x_{mn} = 0$.

Suppose the cache space of each MT is S, and the cache space allocated by the MT n for caching parts of the content m is y_{mn} , we have $\sum_{m=1}^{M} y_{mn} < S$, it means that all the contents cached in the MT n can not exceed the maximal available cache space.

From the view of the network, the revenue obtained by content m cached in mobile terminal MT n is $x_{mn}y_{mn}u_{mn}$, the total revenue of content m caching in the D2D network is $\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}$, so the cache utility function of content m is $u_m = f\left(\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}\right)$, and the $f(\cdot)$ is a continuously differentiable, monotonically increasing, and strictly concave utility function [40].

In the following section, we will study the problem of maximizing cache utility for the whole network, so as to find the optimal caching deployment and cache space allocation.

V. Problem Formulation

The goal of this paper is to optimize the cache utility of the whole network to obtain the caching deployment algorithm, so as to improve the performance of the network from the backhaul link's service offload, cache hit ratio and content access delay to end users.

A. General utility maximization

1) Unique caching case

Firstly, we consider the scenario that one content only can be cached one MT. In the case of unique caching, the caching deployment strategy has to be combined with the allocation of cache space, because they are interdependent. We construct an optimization problem as the function of caching index x_{mn} and cache space allocation y_{mn}. In the case of general utility function expressions, the utility maximization problem is

$$\begin{aligned}
&\mathbf{P1}: \max_{x,y} \sum_{m=1}^{M}\left(f\left(\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}\right)\right)\\
&\text{s.t. C1}: \sum_{n=1}^{N} x_{mn} = 1,\ \forall m \in \{1,\cdots,M\}\\
&\text{C2}: x_{mn} \in [0,1],\ \forall m \in \{1,\cdots,M\}, \text{and } \forall n \in \{1,\cdots,N\}\\
&\text{C3}: 0 \le y_{mn} \le S,\ \forall m \in \{1,\cdots,M\}, \text{and } \forall n \in \{1,\cdots,N\}\\
&\text{C4}: \sum_{m=1}^{M} y_{mn} \le S,\ \forall n \in \{1,\cdots,N\}.
\end{aligned} \tag{7}$$

In the utility optimization problem $\mathbf{P1}$, x_{mn} can only take 0 or 1, where $\sum_{n=1}^{N} x_{mn} = 1$ represents that content m can only be cached in a single MT. $0 \le y_{mn} \le S$ means that if MT n caches the content m, the size of cache space occupied by content m in the MT n is smaller than the size of the MT cache space. $\sum_{m=1}^{M} y_{mn} \le S$ indicates that the size of all contents cached in MT n cannot exceed the storage capacity of MT n. The above restrictions are all linear. Therefore, the optimization problem is a challenging 0-1 programming problem, it can be proved that the optimization problem is NP-hard problem [41].

2) Multiple caching case

Then, we consider the multiple caching case.

Assumption 2: one content can be cached in multiple MTs simultaneously. This assumption may require more overhead to implement, but it is a practical method in D2D caching networks, since the multiple MTs can collaborative download and cache some large volume contents.

From (7) we can notice that, under the **Assumption 2**, the constraint $\sum_{n=1}^{N} x_{mn} = 1$ can be eliminated, and hence there is no need for x_{mn} as additional indicators for caching. The cache space allocation variable $y_{mn} \in [0,1]$ indicates the state of caching, i.e., $y_{mn} > 0$ means a portion of the content m is cached in MT n, otherwise, $y_{mn} = 0$. In this case, we focus on how the cache space should be allocated to different contents with different u_{mn} so as to maximize the utility of MTs, instead of considering conjunction with caching deployment.

We formulate the optimization problem of multiple caching as follows,

$$\begin{aligned}
&\mathbf{P2}: \max_{y} \sum_{m=1}^{M}\left(f\left(\sum_{n=1}^{N} y_{mn}u_{mn}\right)\right)\\
&\text{s.t. C3, C4.}
\end{aligned} \tag{8}$$

It can be seen that the problem $\mathbf{P2}$ is only related to the cache space allocation of different MTs, without considering the caching deployment. Therefore, the optimization problem is simplified.

In the following sections, we show that with the logarithmic utility function, y_{mn} can be directly found without the **Assumption 2**, and thus there is no need to decouple x_{mn} and y_{mn} in this optimization. However, for general utility maximization, problem $\mathbf{P2}$ can provides an ultimate limit on achievable network performance.

B. Logarithmic utility and cache space allocation

Logarithmic utility function in particular is a very common choice of utility function [40], which naturally achieves some level of utility fairness among the contents. To accomplish this, we use a logarithmic utility function in the cache utility maximization problem. The resulting utility function is

$$f\left(\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}\right) = \log\left(\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}\right). \tag{9}$$

This logarithmic utility function is concave, and hence has diminishing returns. This property encourages cache space allocation balancing.

In the remainder of this paper, we focus on the caching deployment with the logarithmic utility function.

First, we consider the unique caching case. In doing so, the utility maximization problem $\mathbf{P1}$ in (7) is equal to,

$$\begin{aligned}
&\mathbf{P3}: \max_{x,y} \sum_{m=1}^{M}\left(\log\left(\sum_{n=1}^{N} x_{mn}y_{mn}u_{mn}\right)\right)\\
&\text{s.t. C1, C2, C3, C4.}
\end{aligned} \tag{10}$$

In the problem $\mathbf{P3}$, the constraint conditions are the same as that of $\mathbf{P1}$. Based on thus constraint conditions, where the function is a continuously differentiable, monotonically increasing, and strictly concave utility function, we can obtain an equivalent transformation as following,

$$\begin{aligned}&\sum_{m=1}^{M}\left(\log\left(\sum_{n=1}^{N}x_{mn}y_{mn}u_{mn}\right)\right)\\&\Leftrightarrow\sum_{n=1}^{N}\sum_{m\in\{i|x_{in}=1\}}\log\left(y_{mn}u_{mn}\right).\end{aligned}\tag{11}$$

Therefore, the problem $\mathbf{P3}$ can be rewritten as,

$$\begin{aligned}&\mathbf{P3'}:\max_{x,y}\sum_{n=1}^{N}\sum_{m\in\{i|x_{in}=1\}}\log\left(y_{mn}u_{mn}\right)\\&\text{s.t. C1, C2, C3, C4.}\end{aligned}\tag{12}$$

At this point, we first need to consider the optimal allocation of cache space in the MT n, namely

$$\max_{y}\sum_{m\in\{i|x_{in}=1\}}\log\left(y_{mn}u_{mn}\right),\tag{13}$$

which is equal to

$$\max_{y}\sum_{m\in\{i|x_{in}=1\}}\left[\log\left(y_{mn}\right)+\log\left(u_{mn}\right)\right].\tag{14}$$

Where, $\log\left(u_{mn}\right)$ is determined by the unit cache utility u_{mn}, the above optimization goal is converted to

$$\begin{aligned}&\max_{y}\sum_{m\in\{i|x_{in}=1\}}\log\left(y_{mn}\right)\\&\Leftrightarrow\max_{y}\log\prod_{m\in\{i|x_{in}=1\}}y_{mn}\\&\Leftrightarrow\max_{y}\tfrac{1}{M_n}\prod_{m\in\{i|x_{in}=1\}}y_{mn}.\end{aligned}\tag{15}$$

M_n is the number of content cached in MT n, . The content m cached in the MT n, satisfying the condition $m\in\{1,\cdots,M_n\}$.

Due to the geometric mean less than the arithmetic mean, we can obtain

$$\sqrt[M_n]{y_{1n}y_{2n}\cdots y_{M_cn}}\leq\frac{1}{M_n}\left(y_{1n}+y_{2n}+\cdots+y_{M_cn}\right)\tag{16}$$

Proof: See Appendix A.

In this inequality, if and only if $y_{1n}=y_{2n}=\cdots=y_{M_nn}$ the equation holds. Therefore, the optimal solution of (15) is $y_{1n}=y_{2n}=\cdots=y_{M_nn}$, showing that the allocated cache space in the MT n to each cached content is same, namely equal cache space allocation. So the content m cached in the MT n obtain the size of the cache space is $S\left(\sum_{m=1}^{M}x_{mn}\right)^{-1}$.

According to the above analysis, $\mathbf{P3}$ can be rewritten as,

$$\begin{aligned}&\mathbf{P4}:\max_{x}\sum_{m=1}^{M}\left(\sum_{n=1}^{N}x_{mn}\log\left(\frac{Su_{mn}}{\sum_{m=1}^{M}x_{mn}}\right)\right)\\&\text{s.t. C1, C2.}\end{aligned}\tag{17}$$

The problem $\mathbf{P4}$ is combinatorial due to the binary variable x_{mn}, the complexity of the brute force search is $\Theta\left((N)^M\right)$, where N and M denote the number of MTs and number of contents, respectively. The computation is essentially impossible for even a modest-sized cellular network. To overcome this, we again invoke the **Assumption 2** to allow one content to be cached in multiple MTs.

In the following, we provide a physical relaxation of C2 in (17) as $0\leq x_{mn}\leq 1$. With this physical relaxation, the indicators x_{mn} can take on any real value in $[0,1]$, representing that one content can be cached portion in more than one MT, which follows the **Assumption 1** as well. So optimization problem of the multiple caching case is,

$$\begin{aligned}&\mathbf{P5}:\max_{x}\sum_{m=1}^{M}\left(\sum_{n=1}^{N}x_{mn}\log\left(\frac{Su_{mn}}{\sum_{m=1}^{M}x_{mn}}\right)\right)\\&\text{s.t. C1,}\\&\text{C5}:0\leq x_{mn}\leq 1.\end{aligned}\tag{18}$$

This physical relaxation makes the (18) to be convex and decoupled the caching deployment and cache space allocation. The $\mathbf{P5}$ will provide an optimal performance of the caching deployment problem with equal cache space allocation.

Note that the upper bound provided by this physical relaxation is different from that provided by joint caching deployment with a general utility function. The joint caching deployment problem of (7) provides an upper bound without any restriction on cache space allocation with unique caching.

VI. Primal-Dual Algorithm

To solve the convex optimization problem (18), the global network information is necessary, which requires a centralized controller for caching deployment and coordination between mobile users. In this paper, we can assume the macro BSs act as the centralized controllers, which have been updated the functions for D2D caching management. The macro BS controls the D2D users within its covering region.

In this section, we propose a near-optimal algorithm via Lagrangian dual decomposition.

A. Dual decomposition

From (18), we have $\log\left(Su_{mn}\left(\sum_{m=1}^{M}x_{mn}\right)^{-1}\right)=\log\left(Su_{mn}\right)-\log\left(\sum_{m=1}^{M}x_{mn}\right)$. We introduce a new variable $M_n=\sum_{m=1}^{M}x_{mn}$ representing the number of contents cached in the MT n. Then the problem of (18) can be rewritten as,

$$\begin{aligned}&\mathbf{P5'}:\max_{x}\sum_{m=1}^{M}\sum_{n=1}^{N}x_{mn}\log\left(Su_{mn}\right)-\sum_{m=1}^{M}\sum_{n=1}^{N}x_{mn}\log M_n\\&\text{s.t. C1, C5,}\\&\text{C6}:M_n=\sum_{m=1}^{M}x_{mn};\\&\text{C7}:M_n\leq M.\end{aligned}\tag{19}$$

The redundant constraint $M_n \leq M$ is added for the analysis of convergence of the proposed algorithm which represent that the number of contents cached in user n is less than the maximum number of contents cached in user n. The only coupling constraint is $M_n = \sum_{m=1}^{M} x_{mn}$ in problem (19). This motivates us to turn to the Lagrangian dual decomposition method whereby a Lagrange multiplier λ is introduced to relax the coupled constraint. The dual problem is,

$$D: \min_{\lambda} D(\lambda) = f_x(\lambda) + g_{M_n}(\lambda), \tag{20}$$

in which,

$$f_x(\lambda) = \max_{x} \sum_{m=1}^{M}\sum_{n=1}^{N} x_{mn} \log(S u_{mn} - \lambda_n) \quad \text{s.t. C1, C5,} \tag{21}$$

and

$$g_{M_n}(\lambda) = \max_{M_n \leq M} \sum_{n=1}^{M} M_n(\lambda_n - \log M_n). \tag{22}$$

When the optimal value of (19) and (20) is the same, we say that strong duality holds. Slater condition is one of the simple constraint qualifications under which strong duality holds. The constraints in (19) are all linear equalities and inequalities, and thus the Slater condition reduces to feasibility [42]. Therefore, the primal problem (19) can be equivalently solved by the dual problem (20). Denoting $x_{mn}(\lambda^*)$ as the maximizer of the first sub-problem (21) and $M_n(\lambda^*)$ as the maximizer of the second sub-problem (22). There exists a dual optimal λ^* such that $x_{mn}(\lambda^*)$ and $M_n(\lambda^*)$ are the primal optimal. Therefore, given the dual optimal λ^* , we can get the primal optimal solution by solving the decoupled inner maximization problems of (21) and (22) separately.

B. Algorithm procedure

The procedure of the dual problem solution is executed by the macro BS. We assume that the all information of caching utility x_{mn} is known by the macro BS.

The outer problem is solved by the gradient projection method [42], where the Lagrange multiplier λ is updated in the opposite direction to the gradient $\nabla D(\lambda)$

$$\frac{\partial D(\lambda)}{\partial(\lambda_n)} = M_n - \sum_{m=1}^{M} x_{mn}. \tag{23}$$

Evaluating the gradient of the dual objective function (20) requires us to solve the inner maximization problem, which has been decomposed into two sub-problems f and g. These sub-problems are solved by **Algorithm 1** as follows.

After iteratively performing the following steps in **Algorithm 1**, the algorithm is guaranteed to converge to a near-optimal solution, as discussed in the Section VI. C.

There is a nice interpretation of λ_n. The multiplier λ_n works as a caching price of MT n in the networks. If we interpret $\sum_{m=1}^{M} x_{mn}$ as the serving demand for the MT n and M_n as the service the MT n can provide, then λ_n is the bridge between demand and supply, and (25) is indeed consistent with the law of supply and demand: if the demand $\sum_{m=1}^{M} x_{mn}$ for the MT n exceeds the supply M_n, the price λ_n will go up; otherwise, the price λ_n will decrease. Thus, when the MT n is over-loaded, it will increase its price λ_n and fewer contents will be cached in it, while other under-loaded MTs will decrease the price so as to attract more contents.

Algorithm 1 Near-optimal algorithm

Set t as the iteration index, and define a caching index matrix $X = \{x_{mn}\}_{M\times N}$. A small positive number ε is predefined as the convergence constant.

Initialization: $t = 0$, $x_{mn} = 0$, and the macro BS generates a random multiplier λ_n for each MT.

Iteration: in the t^{th} iteration of gradient projection algorithm for the content m, the procedure is as following,

Step 1: the macro BS obtain the MT n satisfies $n^* = \arg\max_{n} (\log(S u_{mn}) - \lambda_n(t))$; then set $x_{mn^*} > 0$ and update $M_n^*(t+1) = \sum_{m=1}^{M} x_{mn^*}$;

Step 2: the macro BS updates the values of $M_n(t+1)$ according to the problem (24), we set its gradient to be 0 with the constraint $M_n \leq M$, i.e., $\lambda_n - 1 - \log M_n = 0$, then we have, $M_n = e^{(\lambda_n(t)-1)}$, then the value of M_n is updated by

$$M_n(t+1) = \min\left\{M, e^{(\lambda_n(t)-1)}\right\}. \tag{24}$$

Step 3: the macro BS updates the Lagrange multiplier value $\lambda_n(t+1)$ by the following method,

$$\lambda_n(t+1) = \lambda_n(t) - \delta(t)\left(M_n(t) - \sum_{m} x_{mn}(t)\right), \tag{25}$$

where $\delta(t) > 0$ is a dynamically chosen step size sequence based on some suitable estimates.

Convergence Judgment: when $|D(t+1) - D(t)| \leq \varepsilon$, the iteration will stop and the caching deployment result is given by caching index matrix X; otherwise, it will go to the next iteration.

At each iteration, the complexity of the proposed algorithm is $O(MN)$. The gradient method converges fast generally, especially with the dynamic step size proposed in Sec. VI-C, and thus the number of iterations is a small number (less than 10 in the simulation).

Meanwhile, we assume that the convergence of the proposed algorithm is faster enough compared with the timescale of the cached contents updated and replacement. This can be realized by a specific content selection and maintaining method, for example, the macro BS selects the most popular M contents to be cached for the D2D users in the macrocell. This is reasonable, because the timescale of the popularity changing is relatively long. In the practical implementation scenario, the macro BS acts as a management function entity to update the proposed algorithm procedure when the MT positions are changing.

C. Step Size and convergence

Suppose the step size dynamically updates according to,

$$\delta(t) = \gamma(t)\frac{D(\lambda(t)) - D(t)}{\|\partial D(\lambda(t))\|^2}, 0 < \gamma \leq \gamma(t) \leq \bar{\gamma} < 2, \tag{26}$$

where $D(t)$ is an estimate of the optimal value D^* of problem (20), γ and $\bar{\gamma}$ are some scalars [42].

We consider a procedure for updating $D(t)$, whereby $D(t)$ is given by

$$D(t) = \min_{0 \le \tau \le t} D(\lambda(t)) - \varepsilon(t), \quad (27)$$

and $\varepsilon(t)$ is updated according to

$$\varepsilon(t+1) = \begin{cases} \rho\varepsilon(t) & \text{if } D(\lambda(t+1)) \le D(\lambda(t)) \\ \max\{\theta\varepsilon(t), \varepsilon\} & \text{otherwise} \end{cases}, \quad (28)$$

where ε , θ and ρ are fixed positive constants with $\theta < 1$ and $\rho > 1$ [42].

Thus in this procedure, we want to reach to a target level $D(t)$ that is smaller by $\varepsilon(t)$ over the best value achieved. Whenever the target level is achieved, we increase $\varepsilon(t)$ (i.e., $\rho > 1$) or we keep it at the same value (i.e., $\rho = 1$). If the target level is not attained at a given iteration, $\varepsilon(t)$ is reduced up to a threshold ε , which guarantees that the step size $\delta(t)$ given in (26) is bounded away from zero. As a result, we have the following theorem.

Theorem 1. Assume that the step size $\delta(t)$ is updated by the dynamic step size rule (26) with the adjustment procedure (27) and (28). If $D^* > -\infty$, where D^* denotes the optimal value, then

$$\inf_t D(t) \le D^* + \varepsilon. \quad (29)$$

Proof: The derivative of function $D(\lambda)$(22) is given by

$$\frac{\partial D}{\partial \lambda_n}(\lambda) = M_n(\lambda) - \sum_m x_{mn}(\lambda). \quad (30)$$

In our primal problem $M_n = \sum_{m=1}^{M} x_{mn} \le M$ where N is the total number of MTs. According to (30), when M_n and $\sum_{m=1}^{M} x_{mn}$ are bounded, the sub-gradient of dual objective function ∂D is also bounded,

$$\sup_t \{\|\partial D(\lambda(t))\|\} \le \mu, \quad (31)$$

where μ is some scalar. Thus, our problem satisfies the necessary conditions of Proposition 6.3.6 in [42]. By applying this proposition, the theorem is proved.

VII. Performance Evaluation

A. Simulation assumptions

In the simulation, a macro BS is deployed at the center of the cell and N users are uniformly randomly distributed in the cell, and can communicate with any neighbor users in each user own coverage. We assume the storage capacity of each user are equal and the initial states are empty. The popularity of M contents follows a *Zipf*-like distribution as previous studies [43], and the content size v is set to 1024 bytes. We use φ_{mn} to model the data request of the user n for the content m, varying φ_{mn} characterizes the heterogeneous request among different users. For each user n, we assume that $\sum_{m=1}^{M} \varphi_{mn} = 1$. We also assume that the macro BS is aware of all the users preferences, i.e. $\{\varphi_{mn}\}$ is a common knowledge within the network. Note that, the preference of each user evolves at a timescale much slower than the timescale of content requesting, and it can be learned accurately by monitoring his activity [43].

In the simulation, the parameter setting used for D2D communication is cited from the Technical Report of 3GPP [44], the system bandwidth of D2D communication is 10MHz uplink and 10MHz downlink for FDD, and the indoor to indoor channel model is used as defined in [44], including the pathloss, shadowing and the fast fading. The detailed simulation parameters are given in Table II.

TABLE II: Simulation Parameters

Parameters	Value
Carrier frequency	2 GHz
System bandwidth of downlink Macrocell	10 MHz
System bandwidth of downlink D2D	10 MHz
System bandwidth of uplink D2D	10 MHz
Inter-Cell distance	500 m
Minimum distance between UE and BS	35 m
Minimum distance between UEs	3 m
Maximum transmit power of BS	46 dBm
Transmit power of D2D	30 dBm
Noise power	-174 dBm/Hz

We compare the performance of the following caching schemes:

Preference Aware Caching (PAC): The proposed caching placement algorithm via dual decomposition. In the simulation, the parameters related with PAC are setting as, α= 1, β= 1, ε= 0.1, δ= 2, θ= 0.8, ρ= 1.6 and μ=M.

Random Complete Caching (RCC): In this caching scheme MTs cache data items randomly and then other MTs choose to get the replica from the nearest cache MT or from the BS. The computational complexity of RCC is $O(MN)$.

In this paper, the performance criteria considered are the average content access delay, cache hit ratio, offloading ratio, and the content caching utility. The detailed definitions are given as following.

Average content access delay is defined as the average service delay of M content request within the simulation period of N users, which reflects the effect of the caching strategy on the user experience. The data volume of the content m queried by the user n is $v\varphi_{mn}$, and the content m can be transferred from the BS if $x_{mn} = 0$, or partial from BS and partial from D2D users if $x_{mn} > 0$. So the average content access delay is calculated as,

$$\tau = \frac{1}{MN} \sum_{n=1}^{N} \sum_{m=1}^{M} \left(\frac{v\varphi_{mn}}{c_{nn'} x_{mn} + c_{n\_BS}(1 - x_{mn})} \right), \quad (32)$$

where, $c_{nn'}$ and c_{n_BS} are the data rate from D2D user and BS calculated according to (1) and (2).

Cache hit ratio is the percentage that the interested content item of a user transferred by its neighbor D2D users within its transmission coverage region or its own cache space. The cache hit ratio reflects the effect of caching strategy on reducing the traffic of BS. The higher the cache hit ratio is, the better performance of the caching strategy we can achieve. The cache hit ratio is calculated as,

$$\eta = \frac{\sum_{n}^{N} \sum_{m}^{M} \left(1 - e^{-d^2}\right) \varphi_{mn} x_{mn}}{\sum_{n}^{N} \sum_{m}^{M} \varphi_{mn}}, \quad (33)$$

where, the denominator is the total data requests of N users for M contents, and the numerator is total data requests responded by D2D transmissions.

Offloading ratio is the ratio of the amount of data offloaded by the D2D content delivery to the total amount of the request data in the cell, which reflects the ability of offloading the traffic of BS. Considering the actual application scenario, the data volume of all the contents is much larger than the cache space available to each MT, so the expression of the offloading ratio is,

$$h = \frac{\sum_{n}^{N}\sum_{m}^{M}\left(\left(1-e^{-d^2}\right)\varphi_{mn}x_{mn}S\left(\sum_{m=1}^{M}x_{mn}\right)^{-1}\right)}{v\sum_{n}^{N}\sum_{m}^{M}\varphi_{mn}}. \quad (34)$$

where, the denominator is the data volume of all the M contents requested by all the N users, and the numerator is the data volume of requested contents delivered by the D2D transmissions.

B. Simulation results

i) Convergence validation of PAC

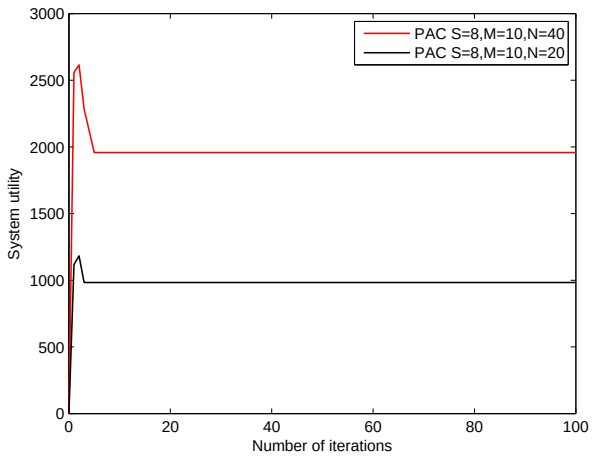

Figure 2. Convergence validation of the PAC.

Firstly, the convergence of the proposed PAC algorithm is verified in Fig. 2. In the simulation, the cache space size of each MT is set as S = 8, and the number of contents is M = 10. It can be observed that the PAC can converge to the approximate optimal solution within 10 times iterations both in the cases of 20 and 40 MTs. When N = 40, the fluctuations before convergence are relatively larger compared with the case of N=20. This is because when there are the more users, there will be more the state of the caching, consequently the proposed algorithm needs to take more time to choose the best fitted MT to cache the contents before reaching a steady state.

ii) Impact of the user number and the cache space size

In this part, we give the total system utility, average content access delay, offloading ratio and cache hit ratio of the proposed PAC against the RCC when the cache space is S=4 and S=8 respectively. Here we set the number of contents as M=10, and the number of user is changing from 20 to 100.

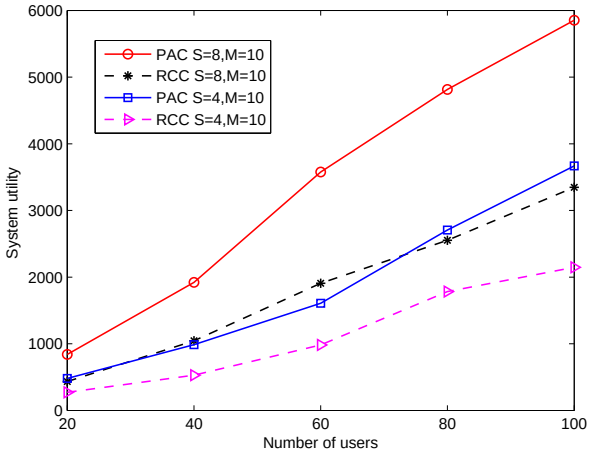

Figure 3. System utility comparison with varying number of users.

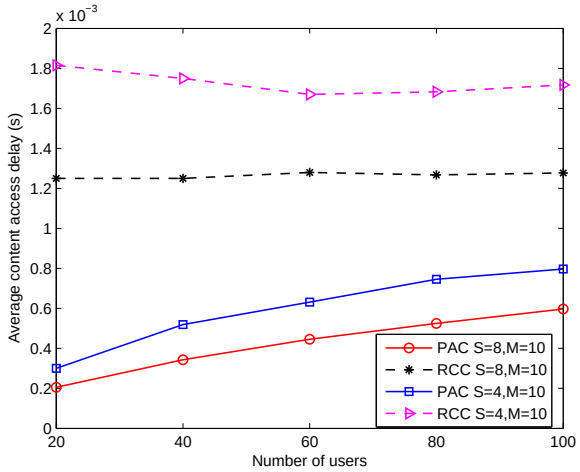

Figure 4. Average content access delay comparison with varying number of users.

Fig. 3 shows the total system utility increases as the number of users increasing. This is due to the fact that the higher the total users number, the higher the neighbor users number, leads to the higher of the unit cache utility and the number of the cached users. Meanwhile, since we take the user preference into account, the increasing of the total system utility of the proposed PAC is faster than that of the RCC, especially in the case S=8 and M=10.

From the Fig. 4, we can see that the proposed PAC achieves significant gains in terms of lower average content access delay compared to RCC with varying number of users. We should notice that, when the number of users is increasing, the average content access delay of PAC increases gradually, this is because that the number of users that cache content does not increase as fast as the total number of users. On the other hand, the average content access delay of RCC almost remains invariable value as the user number increasing. From the result, we can understand that even a very small cache space can obtain a considerable average content access delay gain when the caching strategy is designed elaborated.

Fig. 5 is the cache hit ratio performance of PAC and RCC. The cache hit ratio of the proposed PAC decreases as the number of users increases, while that of the RCC

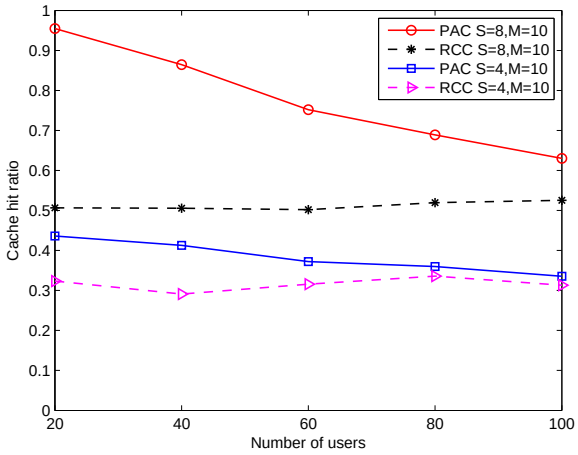

Figure 5. Cache hit ratio comparison with varying number of users.

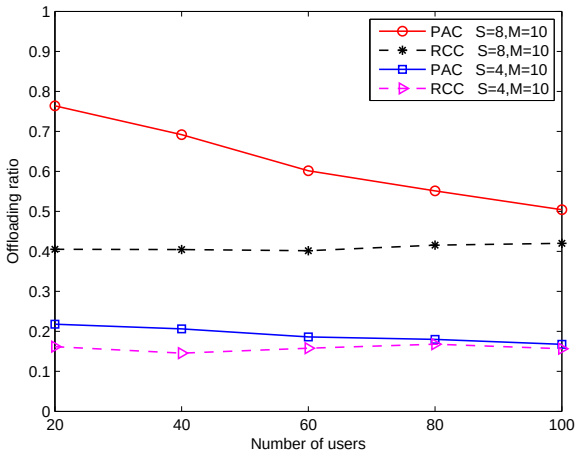

Figure 6. Offloading ratio comparison with varying number of users.

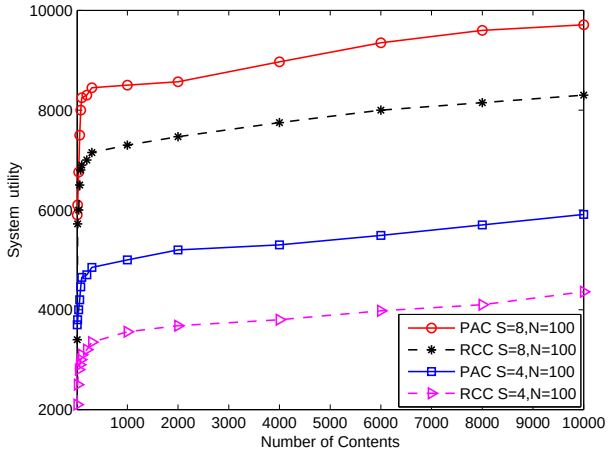

Figure 7. System utility with varying number of contents.

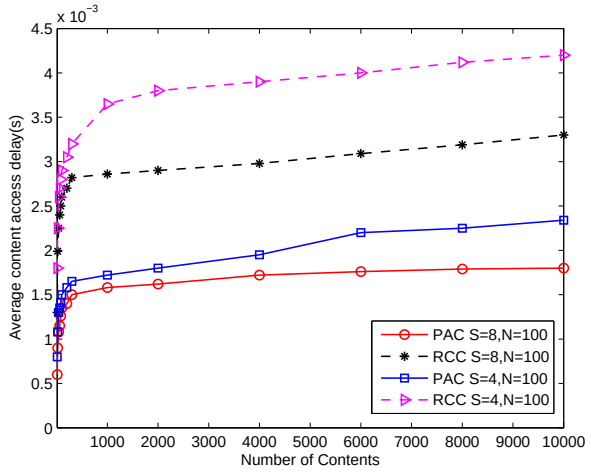

Figure 8. Average content access delay with varying number of contents.

has no significant changes. The reason is that the proposed PAC is preference aware and consider the limitation of D2D communication distance between users, the higher the number of neighbor users, the higher the possibility of content sharing, and therefore the greater the cache hit ratio. The performance advantage of PAC reaches up to 98% and 18.9% relative to the RCC when S=8 and S=4, respectively. The simulation results indicate that, the cache space has a large impact on the cache hit ratio and the content delivery performance.

Fig. 6 compares the offloading ratio for different caching schemes, the results shows that this performance advantage reaches up to 78% and 41%, respectively. The results show that the proposed PAC achieves significant offloading ratio by considering users similarity preference, especially in the case S=8 and M=10. The results also show that, the offloading ratio decreases with the increasing number of users, as more requested contents cannot be fulfilled.

From the simulation results above, we can see that all the performances of the proposed *PAC* are significantly improved as compared to RCC. Meanwhile, we note that with a given user number, the larger the cache space will lead to the higher the system's total utility, better offloading ratio gain, higher the cache hit ratio, and the smaller the average content access delay. The reason is obvious, the higher cache capacity allows the cached *MT*s to cache more popular contents.

iii) Impact of the content number and the cache space size

In this part, we consider a special but practical case in D2D caching networks, that is, there are a large number of contents in the network and a small cache space of each MT. We verify the performance of the caching schemes with varying number of contents. Here we set the number of users is N=100, the number of contents changes from 10 to 10000, the number of cache space S = 4 and$S$ = 8. Fig. 7-Fig. 10 show the system total utility, average content access delay, offloading ratio and cache hit ratio of the proposed PAC against the RCC.

Fig. 7 shows that the total system utility increases as the number of users increasing, and the utility value of the proposed PAC is always higher than that of the RCC. We also note that the utility value of the PAC and RCC do not change obviously when the number of contents is around 500. This is due to the fact that with the increase of the number of contents, the distribution of popular content is more and more dispersed, and the user preference for contents are also increasingly scattered. RCC randomly select contents without considering the user preference, whereas the proposed PAC considers the user preference to maximize the caching utility.

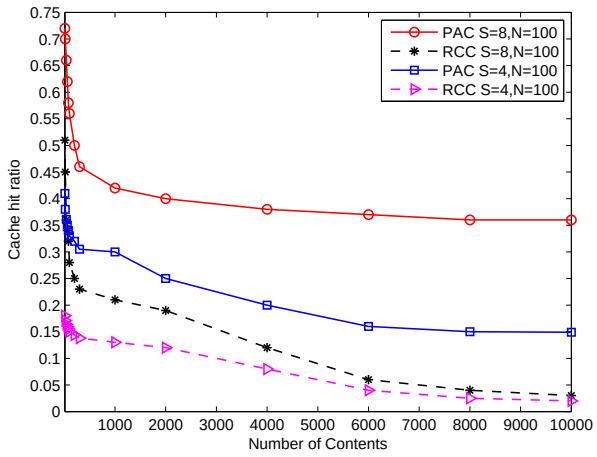

Figure 9. Cache hit ratio with varying number of contents.

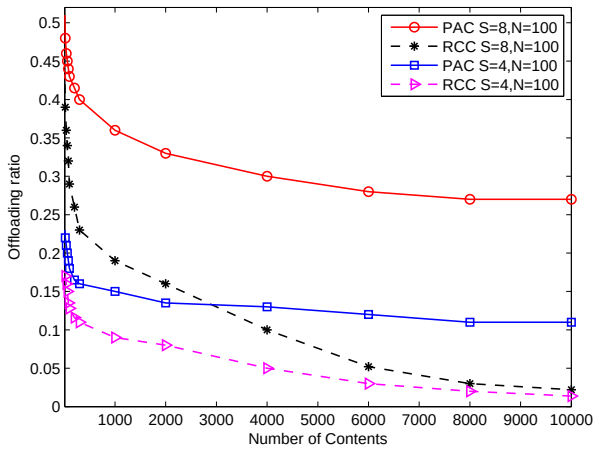

Figure 10. offloading ratio with varying number of contents.

The comparison of the average content access delay with varying number of contents is given in Fig. 8. We can see that the proposed PAC achieves much lower content access delay compared to the RCC. Meanwhile, the average content access delay increases with the increasing of the number of contents. This is due to the fact that with the increase of the number of contents, the MTs with limited cache space cannot cache all the contents which requested by the neighbor users, which leads that macro BS will provide more contents to the users.

Fig. 9 compares the cache hit ratio of the PAC and RCC for different numbers of contents with S=8 and S=4. The cache hit ratio of the proposed PAC decreases as the number of contents increasing. The reason is that with the increasing of the number of contents, the users cannot cache all contents with limited cache space. The cache hit ratio of the PAC outperforms that of the RCC. The performance gain increases as the storage capacity increases, since higher capacity allows the MTs to cache more popular contents following their caching strategies.

We can see from the Fig. 10 that the offloading ratio of the proposed PAC has great improvement compared that of the RCC, especially when the case space size is large. Meanwhile, we can also observe that the offloading ratio of the proposed PAC with S=4 is higher than that of the RCC with S=8 even when the number of contents is larger than 3000. This is due to the fact that with the increase in the number of contents, the distribution of popular contents is more and more dispersed, and the user preference for contents is also increasingly scattered. RCC randomly selects contents without considering the user preference, whereas the proposed PAC considers the user preference to cache the most suitable contents.

From Fig. 7-Fig. 10, we can obtain the conclusion that the performance of the proposed PAC is affected by the number of contents when the number of contents is small, e.g., less than 1000 in the simulation, and the performance of PAC is almost unaffected when the content number is large enough. This is because that the preferred contents of each user will be affected by the *Zipf*-like distribution. Even when the number of contents is very large, the contents interested by users gathered some most popular contents in a *Zipf*-like distribution.

It is worth mentioning that the performance of the proposed PAC algorithm can be improved if we consider the multi-hop D2D communication. In this paper we consider the single-hop D2D communication, in which, each MT can only communicate with others within its own communication coverage in one hop. When the user's request cannot be responded by the D2D caching within its transmission coverage, the contents can only be obtained from the BS directly. If we consider a multi-hop communication in the D2D networks, the contents can be fetched via a MT far away from one hop transmission coverage or the BS which has larger content caching utility.

In short, we prove the effectiveness and efficiency of our proposed PAC by changing the number of users, the size of cache space, and the number of contents. Simulation results show that compared to the existing RCC scheme, our proposed PAC can reduce the content access delay, increase the cache hit ratio, offloading ratio, and the system utility. With this, it comes to the conclusion that the user preference aware caching scheme is superior to other caching schemes.

VIII. Conclusion

In this paper, we have proposed a user preference aware caching deployment algorithm. The proposed algorithm measures the content caching utilization taking account of both the user preference and the transmission coverage region with the aim to replicate the content of interest near the nodes generating the content requests. By doing so, the proposed algorithm would be able to cache specific contents that match the user preference that may also be interested by the adjacent nodes at unpopulated region. Beyond that, we have introduced a caching utility function with the aim to maximize caching utility in order to enhance the possibility of content sharing of among the multiple MTs. The proposed centralized algorithm has obtained the near-optimal performance of the caching deployment, which can be used as the benchmark for the online caching strategy design.

APPENDIX A
PROOF OF THE EQUATION(16)

$$\begin{aligned} \sqrt[M_n]{y_{1n}y_{2n}...y_{M_{cn}}} &\leq \frac{1}{M_n}\left(y_{1n}+y_{2n}+y_{3n}+...y_{M_{cn}}\right) \\ \sqrt[M_n]{y_{1n}y_{2n}...y_{M_{cn}}} &= e^{\frac{\sum_{m=1}^{M_c}\ln y_{mn}}{M_n}} \\ \frac{y_{1n}+y_{2n}+...y_{M_{cn}}}{M_n} &= \frac{\sum_{m=1}^{M_c} e^{\ln y_{mn}}}{M_n} \end{aligned} . \tag{35}$$

Let$f\left(x\right)=e^{x}$, by the nature of the convex function we can obtain that

$$f\left(\frac{\sum_{m=1}^{M_c}\ln y_{mn}}{M_n}\right) \leq \frac{\sum_{m=1}^{M_c} f(\ln y_{mn})}{M_n}. \tag{36}$$

So, we can prove that

$$\sqrt[M_n]{y_{1n}y_{2n}...y_{M_{cn}}} \leq \frac{1}{M_n}\left(y_{1n}+y_{2n}+y_{3n}+...y_{M_{cn}}\right). \tag{37}$$

REFERENCES

[1] B. Coll-Perales, J. Gozalvez, O. Lazaro, and M. Sepulcre, "Opportunistic multihopping for energy efficiency: Opportunistic multihop cellular networking for energy-efficient provision of mobile delay-tolerant services," *IEEE Vehicular Technology Magazine*, vol. 10, no. 2, pp. 93–101, Jun. 2015.

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

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

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

[5] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, March 2012, pp. 1107–1115.

[6] ——, "Wireless video content delivery through coded distributed caching," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 2467–2472.

[7] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, Aug. 2014, pp. 649–653.

[8] 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 Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, July 2014.

[9] 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 Communications Magazine*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[10] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching," in *2012 IEEE International Symposium on Information Theory Proceedings*, July 2012, pp. 2781–2785.

[11] ——, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on information Theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

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

[13] H. Ahlehagh and S. Dey, "Video caching in radio access network: Impact on delay and capacity," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2012, pp. 2276–2281.

[14] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 936–944.

[15] Y. Bao, X. Wang, S. Zhou, and Z. Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," in *2012 18th Asia-Pacific Conference on Communications (APCC)*, Oct. 2012, pp. 566–571.

[16] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: moving from cloud to edge," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 36–42, Sep. 2016.

[17] 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 Transactions on Communications*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[18] X. Huang, G. Zhao, and Z. Chen, "Segment-based random caching in device-to-device (D2D) caching networks," in *2015 International Symposium on Wireless Communication Systems (ISWCS)*, Aug. 2015, pp. 731–735.

[19] H. J. Kang, K. Y. Park, K. Cho, and C. G. Kang, "Mobile caching policies for device-to-device (D2D) content delivery networking," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2014, pp. 299–304.

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

[21] W. Wang, X. Wu, L. Xie, and S. Lu, "Joint storage assignment for D2D offloading systems," *Computer Communications*, vol. 83, pp. 45–55, 2016.

[22] J. Jiang, S. Zhang, B. Li, and B. Li, "Maximized cellular traffic offloading via device-to-device content sharing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 82–91, Jan. 2016.

[23] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing the spatial content caching distribution for device-to-device communications," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 280–284.

[24] A. S. Cacciapuoti, M. Caleffi, M. Ji, J. Llorca, and A. M. Tulino, "Speeding up future video distribution via channel-aware caching-aided coded multicast," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2207–2218, Aug. 2016.

[25] K. C. Chen, M. Chiang, and H. V. Poor, "From technological networks to social networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 548–572, Sep. 2013.

[26] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in LTE-advanced networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1923–1940, Fourthquarter 2015.

[27] K. Wang, F. R. Yu, and H. Li, "Information-centric virtualized cellular networks with device-to-device communications," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9319–9329, Nov. 2016.

[28] K. Wang, H. Li, F. R. Yu, and W. Wei, "Virtual resource allocation in software-defined information-centric cellular networks with device-to-device communications and imperfect CSI," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 10 011–10 021, Dec. 2016.

[29] B. Chen, C. Yang, and G. Wang, "Cooperative device-to-device communications with caching," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[30] Y. Shen, C. Jiang, T. Q. S. Quek, and Y. Ren, "Device-to-device-assisted communications in cellular networks: An energy efficient approach in downlink video sharing scenario," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1575–1587, Feb. 2016.

[31] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015.

[32] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching and file sharing under heterogeneous file preferences," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[33] Z. Chen, Y. Liu, B. Zhou, and M. Tao, "Caching incentive design in wireless D2D networks: A stackelberg game approach," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[34] J. Iqbal and P. Giaccone, "Interest-based cooperative caching in multi-hop wireless networks," in *2013 IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 617–622.
[35] L. Wu, T. Zhang, X. Xu, Z. Zeng, and Y. Liu, "Grey relational analysis based cross-layer caching for content centric networking," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, Nov. 2015, pp. 1–6.
[36] P. Mach, Z. Becvar, and T. Vanek, "In-band device-to-device communication in OFDMA cellular networks: A survey and challenges," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1885–1922, Fourthquarter 2015.
[37] S. Shalmashi and S. B. Slimane, "Cooperative device-to-device communications in the downlink of cellular networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Apr. 2014, pp. 2265–2270.
[38] S. Y. Jung, J.-H. Hong, and T.-S. Kim, "A statistical model for user preference," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 834–843, June 2005.
[39] M. Boguñá, R. Pastor-Satorras, A. Díaz-Guilera, and A. Arenas, "Models of social networks based on social distance attachment," *Physical review E*, vol. 70, no. 5, p. 056122, 2004.
[40] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, June 2013.
[41] L. A. Wolsey, *Integer programming*. Wiley New York, 1998, vol. 42.
[42] D. P. Bertsekas, "Convex optimization theory athena scientific, 2009," *Cited on*, p. 9, 2014.
[43] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 1–14.
[44] 3GPP, "Technical specification group radio access network; study on lte device to device proximity services," *TR 36.843, Release 12*, pp. 33–46, 2014.