# Hierarchical Edge Caching in Device-to-Device Aided Mobile Networks: Modeling, Optimization, and Design

Xiuhua Li, *Student Member, IEEE*, Xiaofei Wang, *Senior Member, IEEE*, Peng-Jun Wan, *Fellow, IEEE*, Zhu Han, *Fellow, IEEE*, and Victor C. M. Leung, *Fellow, IEEE*

***Abstract*— The explosive growth of content requests from mobile users is stretching the capability of current mobile networking technologies to satisfy users' demands with acceptable quality of service. An effective approach to address this challenge, which has not yet been thoroughly studied, is to offload network traffic by caching popular content at the edges (e.g., mobile devices and base stations) of mobile networks, thus reducing the massive duplication of content downloads. In this paper, we address the system modeling, large-scale optimization, and framework design of hierarchical edge caching in device-to-device aided mobile networks. In particular, taking into account the analysis of social behavior and preference of mobile users, heterogeneous cache sizes, and the derived system topology, we investigate the maximum capacity of the network infrastructure in terms of offloading network traffic, reducing system costs, and supporting content requests from mobile users locally. Our proposed framework has a low complexity and can be applied in practical engineering implementation. Trace-based simulation results demonstrate the effectiveness of the proposed framework.**

***Index Terms*— Hierarchical edge caching, device-to-device, traffic load, large-scale optimization, time complexity.**

## I. Introduction

DUE to the tremendous popularity of online social communities, requests for content files (e.g., video, photos and audio) from mobile users are explosively growing. However, supporting these requests effectively has become a significant challenge for mobile network operators (MNOs), which is further deteriorated by the scarcity of network resources especially in the current radio access networks (RANs) and backhaul networks [1]. To address those challenges, it is necessary to introduce revolutionary approaches in network architectures and data transmission technologies towards next generation (e.g., 5G) mobile networks [2], [3].

One key approach is to cache popular content at the edges (e.g., mobile devices and base stations (BSs)) of mobile networks, i.e., edge caching. It can bring the requested content closer to mobile users in the routing distance of the network topology, instead of massively downloading duplicated content from service providers (SPs) via backhaul networks. Thus, by satisfying content requests of mobile users locally, edge caching can effectively enhance network performances, such as offloading network traffic [4], reducing system costs [5], and improving quality of service (QoS) or quality of experience (QoE) of mobile users [6]. On the other hand, device-to-device (D2D) communications can significantly improve network spectral efficiency and energy efficiency due to the physical proximity and potential reuse gains [7]. Thus, it is attractive to investigate how the above two techniques can be combined to enhance the performance of mobile networks aided by D2D communications and edge caching. In such a network, content can be cached in mobile devices and BSs, which forms a hierarchical edge caching topology.

There have been a great number of studies focusing on content caching at the edges of mobile networks. For instance, the surveys in [1]–[3] discussed the potentials of deploying the technique of content caching in mobile networks. The proposed cell caching in [4] and uncoded/coded FemtoCaching in [6] and [8] explored cooperative caching in small BSs, aiming to offload network traffic generated by direct content downloads over the Internet. The studies in [9]–[13] proposed the strategies of collaborative caching in BSs to improve users' QoS especially on minimizing the total access delay for satisfying mobile users' content requests. Based on given and fixed content caching policy in BSs, Tao *et al.* [14] and Liu and Lau [15] proposed multicast beamforming schemes for instantaneous wireless transmissions of content delivery

Manuscript received December 11, 2017; revised April 17, 2018; accepted April 18, 2018. Date of publication June 6, 2018; date of current version October 30, 2018. This work was supported in part by the China Scholarship Council Four Year Doctoral Fellowship, in part by the Canadian NSERC under Grants RGPIN-2014-06119 and RGPAS-462031-2014, in part by the China NSFC (Youth) under Grant 61702364, in part by the China NSFC under Grant 61529202, and in part by the U.S. MURI, NSF under Grants CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607, and CNS-1526638. The work of X. Wang was supported by the National Thousand Talents Plan (Youth) of China. *(Corresponding author: Xiaofei Wang.)*

X. Li and V. C. M. Leung are with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: lixiuhua@ece.ubc.ca; vleung@ece.ubc.ca).

X. Wang is with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: xiaofeiwang@tju.edu.cn).

P.-J. Wan is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: wan@cs.iit.edu).

Z. Han is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 02447, South Korea (e-mail: zhan2@uh.edu).

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

Digital Object Identifier 10.1109/JSAC.2018.2844658

from BSs to users. By satisfying content requests locally via D2D communications, Wang *et al.* [16], Ji *et al.* [17], and Chen and Yang [18] investigated content caching in mobile devices to offload network traffic and analyzed the theoretical performance, respectively, while Zhi *et al.* [19], Wang *et al.* [20], and Bai *et al.* [21] proposed the schemes of content caching in mobile devices for reducing system costs. Yi *et al.* [22] proposed an incentive mechanism for social-aware D2D content sharing and proactive caching by taking power, channel and link management into considerations, aiming to offload downlink cellular traffic. However, these studies only focus on the case of single-level caching in either BSs or mobile devices.

Actually, the idea of hierarchical (or multi-level) caching has been widely applied in web caching systems [23] and IPTV systems [24] to effectively utilize network infrastructures. In this paper, considering the above benefits of D2D communications, edge caching and caching hierarchy, we are motivated to investigate hierarchical edge caching in D2D aided mobile networks in terms of its system modeling, optimization and framework design. Particularly, caches are deployed in both mobile devices and BSs, thereby achieving content sharing among mobile users via D2D communications as well as the caching cooperation among BSs. Our hierarchical caching topology in the network is similar to the widely used hierarchical structures in [23] and [24]. However, the collaborative hierarchical edge caching problem that we address in this paper have serval different features. One important feature that distinguishes our work from similar problems is that the D2D links among mobile users and cellular links between mobile users and BSs are wireless while the user association with the caches is via wired links in [23] and [24]. Another important feature is that the formed hierarchical caching topology in this paper is dynamic due to the mobility of users while it is fixed in [23] and [24]. These features make our problem more challenging to design hierarchical caching strategies in D2D aided mobile networks.

Moreover, in terms of utilizing the idea of hierarchical caching in mobile networks with hierarchical structures, there exist only few works. Bastug *et al.* [25], Yang *et al.* [26], Jiang *et al.* [27], Xu and Tao [28], and Wen *et al.* [29] focused on hierarchical BS caching in heterogeneous networks (HetNets), but did not consider the edge caching in mobile devices, mobile users' social behavior and preference as well as the diversity of content sizes. In [30], a collaborative hierarchical BS caching framework in HetNets was proposed based on the analysis of mobile users' social behavior and preference as well as the diversity of content sizes, but this work did not take the edge caching in mobile devices into consideration as well. Wang *et al.* [31] proposed a hierarchical cooperative caching scheme by dividing the buffer space into three components (i.e., self, friends, and strangers), but it only focused on the caching in mobile devices and did not consider BS caching. Rao *et al.* [32] focused on the stochastic optimization for maximizing the offloading probability of hierarchical content caching in D2D aided mobile networks. But Rao *et al.* [32] only provided the long-run probability of caching each content in each cache (i.e., mobile device and BS), and did not consider the detailed content placement (e.g., which content to cache and how to store content according to the sizes of content and caches) and the caching cooperation (e.g., cooperative content placement and delivery among different levels of caches) among mobile devices and BSs. Wang *et al.* [33] studied edge caching in BSs with D2D offloading where the cached content in mobile devices is fixed, but did not involve the caching design and cooperation of mobile devices. In contrast, our work focuses on hierarchical edge caching in D2D aided mobile networks with practical considerations of network constraints, which can be applied in realistic deployments.

Thus, different from the existing schemes of single-level edge caching in either BSs [1]–[6], [8]–[15] or mobile devices [16]–[19], [21], [22] and hierarchical caching schemes such as in [23]–[33], our problem considers hierarchical edge caching in both mobile devices and BSs, and effectively investigates the caching cooperation among mobile devices and BSs by analyzing the social behavior and preference of mobile users and utilizing the network infrastructures. Besides, our problem is *NP-hard* and focuses on *large-scale* optimization due to the real-world scale of content and mobile users. Accordingly, our problem is quite new and also more challenging.

In this paper, we propose an efficient hierarchical edge caching framework in D2D aided mobile networks. In particular, according to some certain considerations of both mobile users and MNOs, uncoded caching is applied in mobile devices to keep the integrity of content while coded caching is used in BSs to investigate the diversity of content. Based on the analysis of social behavior and preference of mobile users, heterogenous cache sizes and the derived system topology, our objective is to investigate the maximum capacity of the network infrastructure on offloading the network traffic, reducing system costs and supporting content requests from mobile users. The contributions of this paper are summarized as follow:

- We integrate the issues of the analysis of social behavior and preference of mobile users and cell association together with hierarchical edge caching in D2D aided mobile networks, for practically offloading duplicated in-network traffic and reducing system costs as our major objective, towards future green mobile networks.
- We decompose the sophisticated hierarchical edge caching optimization problem that is *NP-hard* into some simpler subproblems of collaboration in different levels, and propose the corresponding low-complexity algorithms for solving the formulated *large-scale* integer linear programming (ILP) problems from the perspective of engineering implementations.
- Together with theoretical analysis, numerical simulation and realistic trace-based evaluation, our proposed framework is shown to offload network traffic significantly and reduce system costs effectively while satisfying most content requests of mobile users locally.

The remainder of this paper is organized as follow. Sec. II discusses the system modeling. Sec. III introduces the proposed hierarchical edge caching framework. Sec. IV evaluates

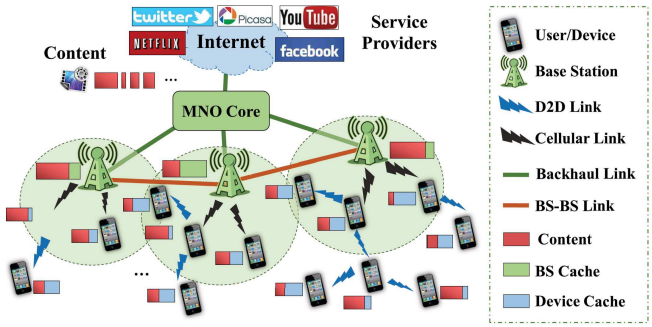



Fig. 1. Illustration of hierarchical edge caching architecture in D2D aided mobile networks.

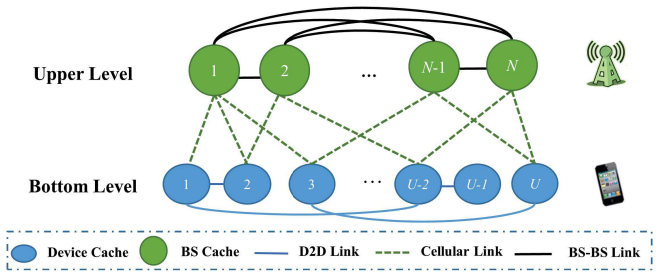



Fig. 2. Topology of hierarchical edge caching in D2D aided mobile networks.

the performance of the proposed framework. Finally, Sec. V concludes the paper.

*Notations:* We use boldfaced letters or numbers to denote real vectors/matrices. The superscript characters "B" and "D" in some notations denote BSs and mobile devices, respectively. $(x_i)_{n\times 1}$ and $(y_{ij})_{n\times m}$ denote a $n$-dimensional vector with elements $\{x_i\}$ and a $n$-by-$m$ matrix with elements $\{y_{ij}\}$, respectively. $\langle \mathbf{x}, \mathbf{y}\rangle$ denotes the sum of all the elements of the Hadamard product $\mathbf{x} \circ \mathbf{y}$. $\lfloor x \rfloor$ and $\lceil y \rceil$ denote the maximum integer that is not greater than $x$ and the minimum integer that is not smaller than $y$, respectively. $[x]^+$ denotes the value of $\max\{0, x\}$. $\mathbf{x} \preceq a$ and $\mathbf{x} \succeq b$ denote that each element of $\mathbf{x}$ is not greater than $a$ and not smaller than $b$, respectively. $\mathbf{x}^{(t)}$ and $\mathbf{x}^*$ denote the $t$-th iteration value and the optimal value of $\mathbf{x}$, respectively. Particularly, we use the letters w.r.t. $(p, q, P, C, E, R)$, $(s, S)$, $\lambda$, and $(x, y, \phi, h)$ for denoting the corresponding probability, storage sizes, average arrival rates of content requests, and caching policy, respectively.

## II. System Modeling

In this section, we introduce the system modeling of hierarchical edge caching in D2D aided mobile networks. Particularly, we introduce the hierarchical edge caching architecture and topology in Sec. II-A. Sec. II-B studies the content popularity and user preference. Sec. II-C and Sec. II-D model the D2D sharing and the association of users and BSs, respectively. Sec. II-E introduces two phases in the scheme design of hierarchical edge caching. Finally, Sec. II-F discusses the content-centric control and management in the network.

### A. Hierarchical Edge Caching Architecture and Topology

An illustration of the hierarchical edge caching architecture in a D2D aided mobile network is shown in Fig. 1. Outside the MNO core, there are some SPs (e.g., YouTube, Facebook, and so on) offering content files over the Internet via backhaul links. Inside the RAN, we consider $N$ BSs (denoted as a set $\mathcal{N} = \{1, 2, \ldots, N\}$) in the whole service area to serve content requests from $U$ geographically distributed mobile users (denoted as a set $\mathcal{U} = \{1, 2, \ldots, U\}$) via cellular links. Besides, BSs are connected to the MNO core via backhaul links and fully connected with each other via BS-BS links through high-capacity cables or optical fibers. Mobile users carrying mobile devices[1] are uniformly distributed in the service area, and can communicate with each other by establishing D2D links via WiFi Direct or Bluetooth when they are in close proximity [7].

All BSs and mobile devices in the network are able to cache some content with limited cache storage capacity, which are denoted by $(S_n^B)_{N\times 1}$ and $(S_u^D)_{U\times 1}$, respectively. Consequently, a hierarchical edge caching topology is formed as shown in Fig. 2, and consists of two levels, i.e., *bottom level* for device caches and *upper level* for BS caches. In the bottom level, content in the device cache can be shared among mobile devices via D2D links. Besides, in the upper level, BSs can provide content to mobile users via cellular links by using BS caches, BS-BS cooperation or direct downloads from SPs via the MNO core.

### B. Content Popularity and User Preference

There is a library of $F$ popular content files (denoted as a set $\mathcal{F} = \{1, 2, \ldots, F\}$) that all mobile users may request in the system. From the practical perspective, the sizes of all the content are assumed to be various, denoted by $(s_f)_{F\times 1}$. The statistics of content requests from all users and from each user are defined below.

*Content popularity*, denoted as $(P_f)_{F\times 1}$, is the probability distribution of content requests from all users in the network, the $f$-th element of which can be calculated as the ratio of the requests of content $f$ to the requests of all the content in the network. Content popularity indicates the common interests of all users in the network and is often modeled by a Mandelbrot-Zipf (MZipf) distribution as [34]

$$P_f = \frac{(R_f + \tau)^{-\beta}}{\sum_{i\in\mathcal{F}} (R_i + \tau)^{-\beta}}, \quad \forall f \in \mathcal{F}, \tag{1}$$

where $R_f$ is the rank of content $f$ in the descending order of content popularity, $\tau \geq 0$ is the plateau factor, and $\beta > 0$ is the skewness factor. Moreover, we assume that the content popularity changes slowly. For instance, short-lifetime popular news with short videos are updated every a few hours, while long-lifetime new movies and new music videos are, respectively, posted weekly and monthly. To reduce the traffic load and avoid possible network traffic congestion especially in busy hours, popular content especially for long-lifetime content can be cached in peak-off hours (e.g., late night). Since the content popularity can be regarded as fixed in a relatively

[1]We use the terms *mobile users* and *mobile devices* interchangeably.

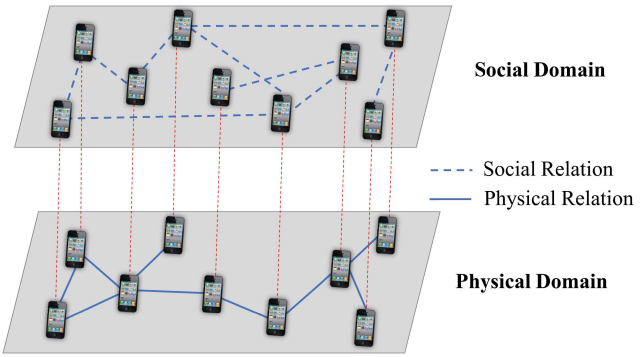



Fig. 3. An illustration of user relation in both physical domain and social domain.

long time, the cost of updating the content in the network can be neglected. In terms of the content popularity, it can be obtained in advance or predicted by system learning and analysis from the social behavior and preference of mobile users [6].

*User preference* denotes the probability distribution of a user's request for each content, denoted as $(q_u^f)_{U\times F}$, where $q_u^f$ is the probability of content $f$ requested by user $u$. Here, each user preference satisfies $\sum_{f\in\mathcal{F}} q_u^f = 1, \forall u \in \mathcal{U}$, which indicates the personal interest of each user. User preference can also be obtained in advance or predicted periodically (e.g., hourly, daily or weekly) by the system learning and analysis from the user social behavior [6], [18].

Besides, we denote the average arrival rate of content requests from user $u$ as $\lambda_u, u \in \mathcal{U}$. Then the corresponding average arrival rate of requests for content $f$ from user $u$ can be calculated as $\lambda_u^f = q_u^f \lambda_u, u \in \mathcal{U}, f \in \mathcal{F}$. Assuming that all content popularity, user preference and average arrival rate of requests are static during a relatively long period, we can get their mathematical relationship as

$$P_f = \frac{\sum_{u\in\mathcal{U}} \lambda_u^f}{\sum_{u\in\mathcal{U}} \lambda_u} = \frac{\sum_{u\in\mathcal{U}} q_u^f \lambda_u}{\sum_{u\in\mathcal{U}} \lambda_u}, \quad \forall f \in \mathcal{F}. \tag{2}$$

### C. D2D Sharing Model

For mobile users in the network, establishing D2D links for content sharing is much cheaper than using cellular links for content downloads. Particularly with a certain probability, a pair of mobile users in close proximity is opportunistic to achieve D2D sharing for their cached content. We assume that instantaneous wireless transmission design of D2D sharing between a pair of mobile users, involving interference, energy status of mobile devices and channel assignment of D2D communications, can be well managed by their local BS [22], which is out of the scope of this paper. As illustrated in Fig. 3, to model D2D sharing of content among mobile users, both physical domain and social domain are considered for characterizing user relation from physical relationship (e.g., geographical distance and encounter) and social relationship (e.g., user preference and relationship), respectively [19]. These two domains will be discussed in details as follow.

*1) Physical Domain:* In the physical domain, when a pair of users are in close proximity, their mobile devices can be connected via WiFi Direct or Bluetooth in a D2D manner. Since mobile devices are carried and operated by humans, users' mobility can affect their physical relationships that are changing over the time. Considering that mobile devices are able to detect neighboring devices within a short certain geographical distance, we use the encounter probability to describe the encounter dynamics of mobile users. In other words, we can record the trace and encounter time period data to predict the strength of physical relationship for each pair of mobile users in the network.

In particular, the encounter probability $E_{uv}^D$ between user $u$ and user $v$ can be approximated as [19]

$$E_{uv}^D = \frac{\sum_i T_{uv}^D(i)}{T_{\text{tot}}}, \quad \forall u \in \mathcal{U}, \ \forall v \in \mathcal{U}, \tag{3}$$

where $T_{uv}^D(i)$ denotes the time period of their $i$-th encounter and $T_{\text{tot}}$ is the chosen total sample time (e.g., a day or a week) according to practical system requirements. This probability in (3) can help predict the encounter probability among users if enough samples are obtained in a trace. That is because longer encounter time means closer physical relationship between a pair of mobile users.

*2) Social Domain:* In the social domain, considering the selfish nature and security/privacy issues of human beings, mobile users with stronger social relationship are more willing to share their own content directly. To model the social relationship among mobile users, we use the weighted graph $G_s(\mathcal{U}, \mathbf{E}_s)$, where its vertex set is user set $\mathcal{U}$ and element $E_s(u, v) \in [0, 1]$ of $U \times U$ edge matrix $\mathbf{E}_s$ denotes the social strength between users $u$ and $v$. Moreover, we consider that social relationship mainly depends on user preference and relationship types.

Specifically, user preference is the same as discussed in Sec. II-B. To exploit the similarity of the preference of users, we define a *content factor* $C_{uv}$ between users $u$ and $v$, which is covered by calculating the Cosine Similarity of their preference as [35], [36]

$$C_{uv} = \frac{\sum_{f\in\mathcal{F}} q_u^f q_v^f}{\sqrt{\sum_{f\in\mathcal{F}} (q_u^f)^2}\sqrt{\sum_{f\in\mathcal{F}} (q_v^f)^2}}, \ \forall u \in \mathcal{U}, \ \forall v \in \mathcal{U}. \tag{4}$$

Larger value of $C_{uv}$ indicates that user $u$ and user $v$ have more common interests in content.

We divide users into four types based on their social relationship for each user: *self*, *close friends*, *normal friends* and *strangers*. The detailed social type of each pair of users can be measured by using social clustering methods such as the $k$-means grouping method in [37] and [38]. To exploit and simplify the effects of relationship types on social relationship among users, we define a *relationship factor* $R_{uv}$ between users $u$ and $v$ as

$$R_{uv} = \begin{cases} 1, & \text{self } (u = v), \\ \alpha_1, & \text{close friends}, \\ \alpha_2, & \text{normal friends}, \\ \alpha_3, & \text{strangers}, \end{cases} \tag{5}$$

where $\{\alpha_i\}_{i=1}^3$ are constant parameters and satisfy $0 \leq \alpha_3 < \alpha_2 < \alpha_1 < 1$. In particular, the parameters $\{\alpha_i\}_{i=1}^3$ can be achieved according to the large-scale practical measurement and analysis [37].

Based on the idea of Jaccard Similarity [19], we define the social strength between users $u$ and $v$ as the weighted sum of the above two factors as

$$E_s(u,v) = \omega_0 \cdot C_{uv} + (1-\omega_0) \cdot R_{uv}, \quad \forall u \in \mathcal{U},\ \forall v \in \mathcal{U}, \tag{6}$$

where $\omega_0 \in [0,1]$ is a weighted parameter. Note that $E_s(u,u) = 1$ holds for $\forall u \in \mathcal{U}$.

Based on the considerations of the physical domain and social domain, we can calculate the probability $p_{uv}^D$ of D2D sharing between users $u$ and $v$ as

$$p_{uv}^D = E_{uv}^D \cdot E_s(u,v), \quad \forall u \in \mathcal{U},\ \forall v \in \mathcal{U}. \tag{7}$$

The joint consideration in (7) shows that only when a pair of users $u$ and $v$ have higher encounter probability $E_{uv}^D$ or stronger social strength $E_s(u,v)$, they can achieve D2D sharing with higher probability $p_{uv}^D$. Note that the sum of probability of D2D sharing between each user and other users is not greater than 1, i.e., $\sum_{v\in\mathcal{U}} p_{uv}^D \leq 1, \forall u \in \mathcal{U}$.

### D. Association of Users and BSs

If a content request from a mobile user cannot be satisfied via D2D sharing, then the mobile user has to be served by the associated local BS for wireless content delivery. Denote $p_u^B$ as the *cellular serving ratio* for user $u$, which is the average probability that the content requests of user $u$ have to be served by BSs via cellular links rather than D2D sharing. Based on the above analysis in Sec. II-C, we can get $p_u^B = 1 - \sum_{v\in\mathcal{U}} p_{uv}^D, \forall u \in \mathcal{U}$. Moreover, considering the user mobility, the associated local BS is not fixed in the long run of time and may change with the dynamics of geographical locations of mobile users. We assume that during the content delivery between a pair of a user and the local BS, the user is only served by the local BS, and the local BS can satisfy the content request by using its own cache, fetching the requested content from other BSs via BS-BS links or downloading it over the Internet via backhaul links. On the condition that user $u$ has to be served by a BS, we denote $p^B\{u|\text{BS } n\}$ as the conditional probability that BS $n$ serves user $u$. Here, this conditional probability satisfies $\sum_{n\in\mathcal{N}} p^B\{u|\text{BS } n\} = 1, \forall u \in \mathcal{U}$, and can also be approximated as

$$p^B\{u|\text{BS } n\} = \frac{\sum_i T_{un}^B(i)}{\sum_{n\in\mathcal{N}}\sum_i T_{un}^B(i)}, \quad \forall u \in \mathcal{U},\ \forall n \in \mathcal{N}, \tag{8}$$

where $T_{un}^B(i)$ denotes the time period of the $i$-th cellular serving from BS $n$ to user $u$ during the total sample time $T_{\text{tot}}$. Thus, from the perspective of the whole network, the probability $p_{un}^B$ that user $u$ is served by BS $n$ is calculated as

$$p_{un}^B = p_u^B \cdot p^B\{u|\text{BS } n\}, \quad \forall u \in \mathcal{U},\ \forall n \in \mathcal{N}. \tag{9}$$

Note that $\sum_{n\in\mathcal{N}} p_{un}^B + \sum_{v\in\mathcal{U}} p_{uv}^D = 1$ holds for $\forall u \in \mathcal{U}$.

### E. Content Placement and Delivery

After achieving the above mentioned statistical information of mobile users by system learning and analysis from their social behavior and preference, designing caching schemes needs to consider two phases, i.e., content placement phase and content delivery phase.

In the content placement phase, the content cached in BSs are assumed to be coded with maximum distance separable (MDS) coding, in order to increase content diversity and improve the cooperation of content delivery among BSs [28], [39]. Based on the theory of MDS coding, content $f$ in the library is split into $M_f$ original segments, and then the $M_f$ original segments are encoded into a large number of packets.[2] Each packet is assumed to have an equal size of $B_o$ bits. Besides, content $f$ can be decoded successfully from any $K_f$ different packets with MDS coding. Denote $y_n^f \in \{0,1,\ldots,K_f\}$ as the number of different packets of content $f$ that are cached in BS $n$. Different from the coded caching in BSs, we employ uncoded caching (that keeps the integrity of each content without coding) in D2D sharing according to the following considerations:

- *Sharing Convenience Perspective*: Mobile devices store the entire data of the content that the mobile users are interested in, then the content can be entirely shared with a certain probability from a mobile user that has the content to a nearby user. In such an uncoded caching way, fetching different parts of the interested content from multiple mobile users and decoding them can be avoided.
- *User QoS/QoE Perspective*: Note that D2D sharing between a pair of mobile users is operated with a certain probability, which depends on their physical/socical relations and certain content availability/request. If a mobile user cannot fetch the entire interested content from a nearby mobile user via D2D sharing, then the mobile user is not willing to wait for another possible mobile user that has the content and moves closer since such D2D sharing may take a longer time and have a bigger uncertainty than those the mobile user can tolerate. Instead, the mobile user tends to download the entire content directly from the associated BS.
- *Complexity Perspective*: Uncoded caching in D2D sharing has much lower complexity than coded caching in terms of caching design and practical implementations.

Denote $x_u^f \in \{0,1\}$ for whether user $u$ caches content $f$ entirely or not, where $x_u^f = 1$ means caching while $x_u^f = 0$ means no caching. Moreover, considering content is large-scale in the real-world systems, we assume that each BS and each mobile device can only cache a small portion of content due to the limited cache storage, i.e., $\sum_{f\in\mathcal{F}} K_f > \lfloor S_n^B/B_0 \rfloor, \forall n \in \mathcal{N}$ and $\sum_{f\in\mathcal{F}} s_f > S_u^D, \forall u \in \mathcal{U}$.

In the content delivery phase, each user requests content based on its own preference. If a user's requested content is locally cached in its mobile device, then the request can be satisfied locally; Otherwise, the user can first find the content in caches of other users in close proximity, then establishes a D2D link with a user where the content is

[2] In this paper, we call the coded segments as packets.

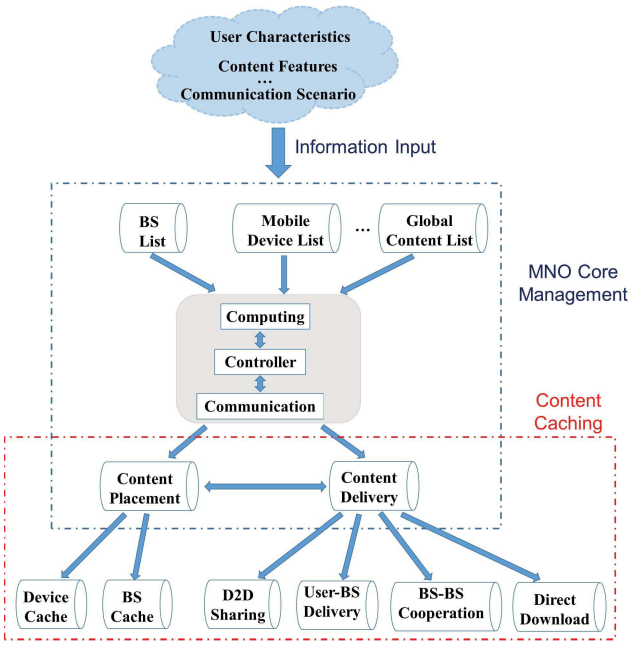



Fig. 4. Brief process of content-centric control and management in D2D-aided mobile networks.

available and finally fetches it in a D2D manner. Both fetching locally and via D2D links are called D2D sharing in this paper. If the content request cannot be satisfied via D2D sharing, the user has to be served by the associated local BS using cellular links. Moreover, if the packets of a requested content cached in the associated local BS are not enough to decode successfully, the associated local BS needs to first fetch a required number of packets from other BSs via BS-BS links or finally downloads the required packets from SPs over the Internet to the associated BS via backhaul links. We assume that the capacity of BS-BS links are large to support the packet delivery of content among BSs, and denote $C_{kn}$ as the average available link capacity from BS $k$ to BS $n$. Denote $\phi_{kn}^f \in \{0, 1, \ldots, K_f\}$ and $h_n^f \in \{0, 1, \ldots, K_f\}$ as the numbers of different packets of content $f$ that BS $n$ fetches from BS $k$ and downloads from SPs, respectively. Here, we set $\phi_{nn}^f = 0, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$.

### *F. Content-Centric Control and Management*

Based on the above modeling and analysis, Fig. 4 shows the brief process of content-centric control and management in the considered D2D-aided mobile networks. From the perspective of content caching, we respectively discuss the functionality of the MNO core, BSs and mobile devices as well as their communications as follow:

*1) MNO Core:* The MNO core manages various resources such as computing and communication for making decisions on how to provide services to mobile users. In particular, the MNO core's main function is to play a role of central management in deciding the content placement in mobile devices and BSs, as well as controlling content delivery process. Besides, the MNO core can provide content delivery service by pushing content to BSs. Moreover, the MNO core in the network can achieve global information, including the following lists:

- BS List: a list for recording all BSs that are connected to and managed by the MNO core. Particularly, this list records BS ID, content ID in each BS, mobile device ID associated with each BS, and so on.
- Mobile Device List: a list for recording mobile device ID, historical information of associated BS ID and mobile device ID, association time, geographic mobility pattern, and so on.
- Global Content List: a list for recording all cached or incoming content in the BSs and mobile devices. Specifically, this list records content ID, content size, content popularity, BS ID and mobile device ID that cache specific content, mobile users' content request pattern, content preference of users, and so on.

*2) BSs:* In the network, BSs with limited resources of caching, computing and communication can effectively manage the associated mobile devices including their D2D sharing, and provide services to mobile users via cellular links. Specifically, each BS needs a list for recording the information such as associated mobile device ID, association time with mobile devices, geographic mobility pattern of associated mobile users, content ID, content size, content popularity, content request pattern of associated mobile users, content preference of associated mobile users, the status of D2D sharing, and so on.

*3) Mobile Devices:* Each mobile device also owns limited resources of caching, computing and communication, and can build D2D communications with other mobile devices. Particularly, each mobile device needs a list for recording the information such as physical/social relationship between its owner and other mobile users, encountered mobile device ID, encounter time, content ID, content size, content preference, content popularity and so on.

*4) Communications:* The dynamics of the information lists collected at BSs and mobile devices need to be shared with the MNO core at the costs of a small traffic overhead that can be neglected, aiming to design content caching frameworks with their effective cooperation. Particularly, mobile devices and BSs provide wireless transmissions of content delivery to associated mobile users via D2D links and cellular links, respectively. Besides, BS-BS cooperation and direct downloads from the SPs to BSs via the MNO core provide wired transmissions of content delivery through BS-BS links and backhaul links, respectively.

## III. Hierarchical Edge Caching Framework Design

In this section, we first investigate and decompose the problem of hierarchical edge caching in the D2D-aided mobile network, then analyze the caching cooperation in each level based on the topology, and finally propose the corresponding low-complexity caching schemes.

### *A. Problem Definition and Decomposition*

We investigate the problem of optimizing hierarchical edge caching in the considered D2D aided mobile network.

Particularly, based on mobile users' social behavior and preference, heterogenous cache sizes and the derived system topology, we aim to investigate the maximum capacity of the network infrastructure on offloading the network traffic and reducing system costs while satisfying mobile users' content requests inside the network.

However, as shown in [24] and [40], the content caching problem in the hierarchical caching infrastructure is *NP-hard* even without using coded caching. Besides, popular content is usually *large-scale* in the real-world systems, and the corresponding caching problems are hard to solve and even impossible to get the optimal solutions with centralized control. Thus, it is very important to design efficient and low-complexity caching schemes from the perspective of engineering implementation.

Moreover, there are two practical observations in the real-world systems as follow: 1) from the perspective of the MNO, shorter paths of content delivery generate less traffic in the mobile networks, and thus to minimize the network traffic as well as system costs, it is more favorable to satisfy content requests in the bottom level; 2) from the economic and energy consumption perspectives of mobile users, D2D sharing in the bottom level is much cheaper than establishing cellular communications for accessing content from BSs in the upper level since D2D communications are generally free or have a little of payment in the daily life and can also achieve higher spectral efficiency due to much shorter communication distances and the usage of much higher spectrum frequency (e.g., 2.4GHz/5GHz). Thus, from the perspectives of the MNO and mobile users, it is potential to first explore the D2D sharing in bottom level and then using cellular communications in order to satisfy mobile users' content requests.

Based on the above considerations as well as the derived system topology, we propose an appropriate decomposition of the complex hierarchical edge caching problem that is *large-scale* and *NP-hard*, thereby finding an efficient and practical suboptimal solution. Particularly, we decompose the complex problem into two simpler subproblems (i.e., device caching and BS caching) that focus on the cooperation in different levels. Besides, we propose a low-complexity and globally/locally optimal solution for solving the first subproblem, and solve the second subproblem with low-complexity suboptimal solutions. We firstly explore the cooperation in the bottom level via D2D sharing among mobile users in Sec. III-B, and secondly discuss the cooperation in the upper level based on the content delivery among BSs in Sec. III-C. Finally, we propose a content request routing scheme for both levels in Sec. III-D.

### B. Cooperation in Bottom Level

As shown in Fig. 2, mobile devices in the bottom level are able to disseminate content via their direct D2D links. As similar to [24] and [30], we regard $\lambda_u^f s_f$ as the incurred average traffic load for the requests of content $f$ from user $u$. In particular, for a pair of users $u$ and $v$, if user $u$ caches content $f$ (i.e., $x_u^f = 1$) and user $v$ does not (i.e., $x_v^f = 0$), user $u$ can share content $f$ to user $v$ via D2D links with the probability of $p_{uv}^D$. Then we can obtain the corresponding average supported traffic load as $p_{uv}^D \lambda_v^f s_f$. Thus, whether user $u$ caches content $f$ or not, we can get the average supported traffic load $L_u^f$ via D2D sharing as

$$L_u^f = x_u^f \lambda_u^f s_f + \sum_{v \in \mathcal{U}} p_{uv}^D [x_u^f (1 - x_v^f)] \lambda_v^f s_f, \ \forall u \in \mathcal{U}, \ \forall f \in \mathcal{F}, \tag{10}$$

where the first term and second term denote the corresponding average supported traffic load of the requests for content $f$ from user $u$ and other users, respectively.

In the bottom level, we aim to maximize the total amount of average supported traffic load via D2D sharing under the constraints of cache storage capacity of mobile devices. Thus, the corresponding problem can be formulated as

$$\max_{\{x_u^f\}} \sum_{u \in \mathcal{U}} \sum_{f \in \mathcal{F}} L_u^f \tag{11a}$$

$$s.t. \sum_{f \in \mathcal{F}} x_u^f s_f \le S_u^D, \quad \forall u \in \mathcal{U}, \tag{11b}$$

$$x_u^f \in \{0, 1\}, \quad \forall u \in \mathcal{U}, \ \forall f \in \mathcal{F}. \tag{11c}$$

Here, (11b) denotes the constraints of the cache storage capacity of all mobile devices while (11c) denotes that each content is either cached entirely or not in each mobile device.

Due to the quadratic terms $\{x_u^f(1 - x_v^f)\}$ in the objective in (11a), the problem in (11) is a binary integer linearly constrained quadratic programming (LCQP) problem. Denote $z_{uv}^f = x_u^f(1 - x_v^f) \in \{0, 1\}, \forall u \in \mathcal{U}, \forall v \in \mathcal{U}, \forall f \in \mathcal{F}$. Note that $z_{uu}^f = 0, \forall u \in \mathcal{U}, \forall f \in \mathcal{F}$. However, this binary integer LCQP problem can be equally transformed into a binary integer linear programming (BILP) problem by using *Theorem 1* as follows.

*Theorem 1:* For the problem in (11), the defined $z_{uv}^f = x_u^f(1 - x_v^f)$ is equivalent to two inequalities, i.e., $z_{uv}^f \le x_u^f$ and $z_{uv}^f \le 1 - x_v^f$.

*Proof:* Please see Appendix A. ∎

Thus, the problem in (11) can be equivalent to a BILP problem as

$$\max_{\{x_u^f, z_{uv}^f\}} \sum_{f \in \mathcal{F}} \sum_{u \in \mathcal{U}} x_u^f \lambda_u^f s_f + \sum_{f \in \mathcal{F}} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{U}} z_{uv}^f p_{uv}^D \lambda_v^f s_f \tag{12a}$$

$$s.t. \sum_{f \in \mathcal{F}} x_u^f s_f \le S_u^D, \quad \forall u \in \mathcal{U}, \tag{12b}$$

$$z_{uv}^f \le x_u^f, z_{uv}^f \le 1 - x_v^f, \quad \forall u \in \mathcal{U}, \ \forall v \in \mathcal{U}, \ \forall f \in \mathcal{F}, \tag{12c}$$

$$x_u^f \in \{0, 1\}, \quad z_{uv}^f \in \{0, 1\}, \ \forall u \in \mathcal{U}, \ \forall v \in \mathcal{U}, \ \forall f \in \mathcal{F}. \tag{12d}$$

Considering the above BILP problem in (12) is generally large-scale, we use the exact penalty method (EPM) proposed in [41] to solve it and provide a low-complexity solution. We firstly transform the binary variable constraint in (12d) into some equivalent constraints. Particularly, we define a variable matrix $\boldsymbol{\pi} \in \{-1, 1\}^{U \times (U+1) \times F}$ as: 1) $\pi_{uv}^f = 2z_{uv}^f - 1$ for $\forall u \in \mathcal{U}, \forall v \in \mathcal{U}, \forall f \in \mathcal{F}$ and 2) $\pi_{u,U+1}^f = 2x_u^f - 1$ for $\forall u \in \mathcal{U}, \forall f \in \mathcal{F}$. Then the BILP problem in (12)

is equivalent to

$$\min_{\boldsymbol{\pi}} F(\boldsymbol{\pi}) = -\sum_{f\in\mathcal{F}}\sum_{u\in\mathcal{U}}\left(\pi^f_{u,U+1}\lambda^f_u s_f + \sum_{v\in\mathcal{U}}\pi^f_{uv}p^D_{uv}\lambda^f_v s_f\right) \tag{13a}$$

$$s.t.\ \sum_{f\in\mathcal{F}}\pi^f_{u,U+1}s_f \le 2S^D_u - \sum_{f\in\mathcal{F}}s_f,\quad \forall u\in\mathcal{U}, \tag{13b}$$

$$\pi^f_{uv}\le\pi^f_{u,U+1}, \pi^f_{uv}\le -\pi^f_{v,U+1},\quad \forall u\in\mathcal{U},\ \forall v\in\mathcal{U},\quad \forall f\in\mathcal{F}, \tag{13c}$$

$$\boldsymbol{\pi}\in\{-1,1\}^{U\times(U+1)\times F}. \tag{13d}$$

We define $\Omega \triangleq \left\{\boldsymbol{\pi}\in\mathbb{R}^{U\times(U+1)\times F}\middle|\ (13b),(13c)\right\}$, and provide an extension of the theorem proposed in [41] as follows.

*Theorem 2:* Denote two matrices $\mathbf{a}\in\mathbb{R}^{m\times n\times k}, \mathbf{v}\in\mathbb{R}^{m\times n\times k}$, and define $\Phi \triangleq \left\{(\mathbf{a},\mathbf{v})|\langle\mathbf{a},\mathbf{v}\rangle = mnk, -1\preceq\mathbf{a}\preceq 1, \|\mathbf{v}\|_2^2\le mnk\right\}$. If $(\mathbf{a},\mathbf{v})\in\Phi$, then we have $\mathbf{a}\in\{-1,1\}^{m\times n\times k}, \mathbf{v}\in\{-1,1\}^{m\times n\times k}$, and $\mathbf{a}=\mathbf{v}$.

*Proof:* Please see Appendix B. ∎

Based on *Theorem 2*, by introducing a new variable matrix $\mathbf{v}\in\mathbb{R}^{U\times(U+1)\times F}$ and setting $G=U(U+1)F$, the constraint $\boldsymbol{\pi}\in\{-1,1\}^{U\times(U+1)\times F}$ can be equivalent to three constraints $\langle\boldsymbol{\pi},\mathbf{v}\rangle = G, -1\preceq\boldsymbol{\pi}\preceq 1$ and $\|\mathbf{v}\|_2^2\le G$. Thus, the problem in (13) can be equivalently transformed as

$$\min_{\boldsymbol{\pi},\mathbf{v}} F(\boldsymbol{\pi}) \tag{14a}$$

$$s.t.\ -1\preceq\boldsymbol{\pi}\preceq 1, \|\mathbf{v}\|_2^2\le G, \boldsymbol{\pi}\in\Omega, \tag{14b}$$

$$\langle\boldsymbol{\pi},\mathbf{v}\rangle = G, \tag{14c}$$

where $\boldsymbol{\pi}\in\mathbb{R}^{U\times(U+1)\times F}$. As a result, the original discrete integer optimization problem in (13) is equivalently transformed into a continuous optimization problem in (14). Then by introducing the penalty with the EPM for solving the problem in (14), we define a new optimization problem as

$$\min_{\boldsymbol{\pi},\mathbf{v}} \mathcal{L}(\boldsymbol{\pi},\mathbf{v},\rho) = F(\boldsymbol{\pi}) + \rho(G - \langle\boldsymbol{\pi},\mathbf{v}\rangle) \tag{15a}$$

$$s.t.\ -1\preceq\boldsymbol{\pi}\preceq 1, \|\mathbf{v}\|_2^2\le G, \boldsymbol{\pi}\in\Omega, \tag{15b}$$

where $\rho$ is the penalty parameter that is iteratively increased to enforce the constraint in (14c). The problem in (15) is a biconvex optimization problem.

To solve the problem in (14), in each iteration with fixed $\rho$, we solve the problem in (15) by minimizing $\mathcal{L}(\boldsymbol{\pi},\mathbf{v},\rho)$ over $\boldsymbol{\pi}$ and $\mathbf{v}$ alternatingly. We detail the iteration steps of EPM for solving the problem in (14) as shown in Algorithm 1. The parameter $T$ is the number of inner iterations for solving the problem in (15). Note that though the main idea of Algorithm 1 is similar to the proposed EPM in [41], Algorithm 1 is a more complex extension since it focuses on solving a biconvex optimization problem with 3-dimensional matrix variables while the proposed EPM in [41] aims to solve a biconvex optimization problem with vector (that can be regarded as 1-dimensional matrix) variables.

In Algorithm 1, we initialize $\mathbf{v}^{(0)}=\mathbf{0}$ to find a reasonable local minima in the first iteration, as it reduces to LP convex relation for the binary optimization problem in (13). Besides, $\boldsymbol{\pi}$-subproblem in (18) is a LP convex optimization problem, and its optimal solution can be achieved with the interior point method that takes polynomial time (i.e., $O(G^{0.5}\log\frac{1}{\epsilon})$) for converging to an $\epsilon$-accurate solution [42]. For $\mathbf{v}$-subproblem in (19), it is also a convex optimization problem and can be rewritten as

$$\mathbf{v}^{(t+1)} = \operatorname{argmin}_{\mathbf{v}}\ \langle\mathbf{v}, -\boldsymbol{\pi}^{(t+1)}\rangle,\quad s.t.\ \|\mathbf{v}\|_2^2\le G. \tag{16}$$

For the problem in (16), if $\boldsymbol{\pi}^{(t+1)}=\mathbf{0}$, any feasible solution will be an optimal solution; otherwise, the optimal solution will be achieved in the boundary with $\|\mathbf{v}\|_2^2 = G$, and the problem in (16) is equivalent to solving $\min_{\|\mathbf{v}\|_2^2=G}\ \frac{1}{2}\|\mathbf{v}\|_2^2 - \langle\mathbf{v},\boldsymbol{\pi}^{(t+1)}\rangle$. Thus, we can get its optimal solution in a closed form as

$$\mathbf{v}^{(t+1)} = \begin{cases}\sqrt{G}\cdot\boldsymbol{\pi}^{(t+1)}/\|\boldsymbol{\pi}^{(t+1)}\|_2, & \text{if } \boldsymbol{\pi}^{(t+1)}\neq\mathbf{0},\\ \text{any } \mathbf{v} \text{ with } \|\mathbf{v}\|_2^2\le G, & \text{otherwise.}\end{cases} \tag{17}$$

Besides, $F(\boldsymbol{\pi})$ is linear and thus is a $L$-Lipschitz continuous convex function on $-1\preceq\boldsymbol{\pi}\preceq 1$, i.e., $|F(\boldsymbol{\pi}_1) - F(\boldsymbol{\pi}_2)|\le L\parallel\boldsymbol{\pi}_1-\boldsymbol{\pi}_2\parallel_2$, where $-1\preceq\boldsymbol{\pi}_1,\boldsymbol{\pi}_2\preceq 1$ and $L$ is a constant. Thus, the used EPM has the following properties as shown in Theorem 3 and Theorem 4 [41].

*Theorem 3 (Exactness of the Penalty Function):* When $\rho > 2L$, the biconvex optimization problem in (15) has the same local and global minima with the primary problem in (14).

*Proof:* Please see Appendix C. ∎

*Theorem 4 (Convergence Rate and Asymptotic Monotone Property of Algorithm 1): Given the convergence condition $|\langle\boldsymbol{\pi},\boldsymbol{v}\rangle - G|\le\varepsilon$, Algorithm 1 converges to the first-order Karush-Kuhn-Tucker (KKT) point within $\lceil(\ln(L\sqrt{2G}) - \ln(\varepsilon\rho^{(0)}))/\ln\Delta\rceil$ outer iterations.*[3] *Besides, after $\langle\boldsymbol{\pi},\boldsymbol{v}\rangle = G$*

**Algorithm 1** EPM for Solving the Problem in (14)

1: Initialize $t=0$, $\boldsymbol{\pi}^{(0)}=\mathbf{v}^{(0)}=\mathbf{0}$, $\rho^{(0)}>0$, $\Delta>0$, $\varepsilon>0$.
2: **while** not converge **do**
3: Update $\boldsymbol{\pi}^{(t+1)}$ by solving $\boldsymbol{\pi}$-subproblem:

$$\boldsymbol{\pi}^{(t+1)} = \arg\min_{\boldsymbol{\pi}}\mathcal{L}(\boldsymbol{\pi},\mathbf{v}^{(t)},\rho^{(t)}), s.t.\ -1\preceq\boldsymbol{\pi}\preceq 1, \boldsymbol{\pi}\in\Omega. \tag{18}$$

4: Update $\mathbf{v}^{(t+1)}$ by solving $\mathbf{v}$-subproblem:

$$\mathbf{v}^{(t+1)} = \arg\min_{\mathbf{v}}\ \mathcal{L}(\boldsymbol{\pi}^{(t+1)},\mathbf{v},\rho^{(t)}),\ s.t.\ \|\mathbf{v}\|_2^2\le G. \tag{19}$$

5: Check the convergence condition: $|\langle\boldsymbol{\pi}^{(t+1)},\mathbf{v}^{(t+1)}\rangle - G|\le\varepsilon$.
6: Update the penalty in every $T$ iterations (if necessary):

$$\rho^{(t+1)} = \min\{2L, \rho^{(t)}\times\Delta\}. \tag{20}$$

7: Set $t\leftarrow t+1$.
8: **end while**
9: Transform $\boldsymbol{\pi}$ into $\{x^f_u, z^f_{uv}\}$.
10: **Output**: $\{x^f_u, z^f_{uv}\}$.

[3]Each time the parameter $\rho$ is increased, we call it one outer iteration.

*is achieved, the sequence of* $\{F(\boldsymbol{\pi}^{(t)})\}$ *generated by Algorithm 1 is monotonically non-increasing.*

*Proof:* Please see Appendix D. ∎

*Remark 1: Theorem 3 shows that when the penalty parameter* $\rho$ *is larger than some threshold, the biconvex optimization problem in (15) is equivalent to the primary problem in (14), which importantly implies the theoretical convergence of Algorithm 1. Besides, Theorem 4 shows the convergence rate and asymptotic monotone property of Algorithm 1.*

In all, the problem via D2D sharing in the bottom level is firstly formulated as a binary integer LCQP problem in (11). Then based on the derived theorems, we transform it into an equivalent BILP problem in (12) and then into an equivalent biconvex optimization problem in (13). After the equivalent transformations, we further transform the integer variables into equivalent continuous variables, and the biconvex optimization problem becomes an equivalent continuous optimization problem in (14) that is finally solved with the proposed Algorithm 1.

### C. Cooperation in Upper Level

After the cooperation in the bottom level via D2D sharing among mobile users, the average arrival rate of the *unsatisfied* requests for content $f$ from user $u$ via D2D sharing, denoted by $\widetilde{\lambda}_u^f$, can be calculated as

$$\widetilde{\lambda}_u^f = \lambda_u^f \cdot \min\{1 - x_u^f, \sum_{v\in\mathcal{U}} (1 - z_{vu}^f) p_{vu}^D\}, \ \forall u \in \mathcal{U}, \ \forall f \in \mathcal{F}, \tag{21}$$

which has to be served by the associated local BS (say BS $n$) of user $u$ with the conditional probability of $p^B\{u|\text{BS } n\}$. Thus, the average arrival rate of the requests received at BS $n$ for content $f$ from all users, denoted by $\theta_n^f$, is calculated as

$$\theta_n^f = \sum_{u\in\mathcal{U}} [p^B\{u|\text{BS } n\}\widetilde{\lambda}_u^f + (1 - x_u^f) p_{un}^B \lambda_u^f], \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}. \tag{22}$$

In the upper level, the aim of content caching in BSs with their cooperation is to minimize the total amount of average system cost incurred by two factors: 1) downloading the required packets of content from SPs to BSs via backhaul links; 2) packet delivery among BSs via BS-BS links. Denote $c_o$ and $c_{kn}$ as the average cost per unit traffic load for downloading a packet from SPs to BSs and delivering a packet from BS $k$ to BS $n$, respectively. In practice, $c_o$ is usually much greater than $c_{kn}$, i.e., $c_o > c_{kn}, \forall n \in \mathcal{N}, \forall k \in \mathcal{N}$ as the backhaul links connecting BSs to SPs is of many-fold further than the BS-BS links. This makes it cost-effective to fetch packets from the caches in neighboring BSs whenever possible instead of downloading them from SPs. As well, we regard $h_n^f \theta_n^f B_0$ and $\phi_{kn}^f \theta_n^f B_0$ as the corresponding average traffic load for the requests of content $f$ from SPs to BS $n$ and from BS $k$ to BS $n$, respectively. Then this optimization problem is formulated as

$$\min_{\{y_n^f, \phi_{kn}^f, h_n^f\}} \sum_{f\in\mathcal{F}} \sum_{n\in\mathcal{N}} h_n^f c_o \theta_n^f B_0 + \sum_{f\in\mathcal{F}} \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{N}} \phi_{kn}^f c_{kn} \theta_n^f B_0 \tag{23a}$$

$$\text{s.t.} \ \sum_{f\in\mathcal{F}} y_n^f \le \lfloor S_n^B / B_0 \rfloor, \ \ \forall n \in \mathcal{N}, \tag{23b}$$

$$\sum_{f\in\mathcal{F}} \phi_{kn}^f \theta_n^f B_0 \le C_{kn}, \ \ \forall k \ne n \in \mathcal{N}, \tag{23c}$$

$$\phi_{kn}^f \le y_k^f, \ \ \phi_{nn}^f = 0, \ \forall k \in \mathcal{N}, \ \forall n \in \mathcal{N}, \forall f \in \mathcal{F}, \tag{23d}$$

$$y_n^f + \sum_{k\in\mathcal{N}} \phi_{kn}^f + h_n^f \ge K_f, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}, \tag{23e}$$

$$y_n^f \in \{0, 1, \ldots, K_f\}, \ \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}, \tag{23f}$$

$$\phi_{kn}^f \in \{0, 1, \ldots, K_f\}, \ \ \forall k \in \mathcal{N}, \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}, \tag{23g}$$

$$h_n^f \in \{0, 1, \ldots, K_f\}, \ \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}. \tag{23h}$$

Here, (23b) denotes the constraints of the cache storage capacity of all BSs. (23c) denotes the constraints of the capacity of BS-BS links. (23d) denotes the constraints of the packet delivery of content among BSs. (23e) denotes the constraints that content can be successfully decoded from packets by BSs. The problem in (23) is an integer linear programming (ILP) problem, and thus is NP-hard. To solve this problem, we discuss two cases of whether the BS-BS link capacity is sufficiently large or not for supporting all the packet delivery among BSs, i.e., unlimited capacity case and limited capacity case, as follow.

*1) Unlimited Capacity Case:* If the BS-BS link capacity is sufficiently large, then (23c) always holds and can be reduced in the ILP problem in (23), and we call it as the *simplified* ILP problem. We first derive some properties of its optimal solutions as follow.

*Theorem 5: For any optimal solution to the simplified ILP problem, denoted by* $(\{(y_n^f)^*\}, \{(\phi_{kn}^f)^*\}, \{(h_n^f)^*\})$*, we can obtain*

$$(h_n^f)^* + (y_n^f)^* + \sum_{k\in\mathcal{N}} (\phi_{kn}^f)^* = K_f, \ \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}, \tag{24}$$

$$(h_n^f)^* = [K_f - \sum_{k\in\mathcal{N}} (y_k^f)^*]^+, \ \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}. \tag{25}$$

*Proof:* Please see Appendix E. ∎

*Remark 2: According to* Theorem 5*, the closed-form expression of* $(h_n^f)^*$ *in (25) is dependent on the value of* $\sum_{k\in\mathcal{N}} (y_k^f)^*$ *and independent on the values of* $\{(\phi_{kn}^f)^*\}$*. Besides, for any* $(n, f)$ *such that* $(h_n^f)^* \ge 0$*, i.e.,* $\sum_{k\in\mathcal{N}} (y_k^f)^* \le K_f$*, we have* $(\phi_{kn}^f)^* = (y_k^f)^*, \forall k \in \mathcal{N} \setminus \{n\}$.

Based on *Theorem 5* and *Remark 2*, the simplified ILP problem can be equivalent to

$$\min_{\{y_n^f, \phi_{kn}^f\}} \sum_{f\in\mathcal{F}} [K_f - \sum_{k\in\mathcal{N}} (y_k^f)]^+ \big( \sum_{n\in\mathcal{N}} c_o \theta_n^f B_0 \big) + \sum_{f\in\mathcal{F}} \sum_{n\in\mathcal{N}} \sum_{k\in\mathcal{N}} \phi_{kn}^f c_{kn} \theta_n^f B_0 \tag{26a}$$

$$\text{s.t.} \ [K_f - \sum_{k\in\mathcal{N}} y_k^f]^+ + y_n^f + \sum_{k\in\mathcal{N}} \phi_{kn}^f = K_f, \ \ \forall n \in \mathcal{N}, \ \forall f \in \mathcal{F}, \tag{26b}$$

$$(23b), \ \ (23d), \ (23f), \text{ and } \ (23g). \tag{26c}$$

**Algorithm 2** Heuristic Method for Solving the Problem in (26) of the Unlimited Capacity Case

1: Initialize $\{y_n^f\} = \mathbf{0}_{N\times F}, \{h_n^f\} = \mathbf{0}_{N\times F}, \{\phi_{kn}^f\} = \mathbf{0}_{N\times N\times F}$.
2: *—Procedure 1. [Decide $\{\sum_{n\in\mathcal{N}} y_n^f\}$ and $\{h_n^f\}$]*
3: Set $S_{\text{tot}} = \sum_{n\in\mathcal{N}} \lfloor S_n^B/B_0 \rfloor$, $\mathbf{N}_{\text{tot}} = \big\{K_f\big(\min\{N, \lceil S_{\text{tot}}/\sum_{f\in\mathcal{F}} K_f\rceil\} - 1\big)\big\}_{F\times 1}$, $\overline{S}_{\text{tot}} = S_{\text{tot}} - \sum_{f\in\mathcal{F}} N_{\text{tot}}(f)$, $\mathbf{A} = \{\sum_{n\in\mathcal{N}} c_o\theta_n^f B_0\}$, $\mathbf{E} = \{A_1, 2A_1, \ldots, K_1A_1, \ldots, A_f, 2A_f, \ldots, K_fA_f, \ldots, A_F, 2A_F, \ldots, K_FA_F\}$, $K_0 = 0$, $i = 1$.
4: Sort $\mathbf{E}$ into $\mathbf{G}$ in a descending order, and label the orginal index vector as $\mathbf{O}$ satisfying $G_j = E_{O_j}, \forall j$.
5: **while** $\overline{S}_{\text{tot}} \neq 0$ & $\sum_{f\in\mathcal{F}} N_{\text{tot}}(f) < N\sum_{f\in\mathcal{F}} K_f$ & $i \leq \sum_{f\in\mathcal{F}} K_f$ **do**
6: Find $f^* = \arg\max_{f\in\mathcal{F}} \big\{f|\sum_{j=0}^{f-1} K_j < O_i\big\}$, and set $N_{\text{tot}}(f^*) \longleftarrow N_{\text{tot}}(f^*)+1$, $\overline{S}_{\text{tot}} \longleftarrow \overline{S}_{\text{tot}} - 1$, $i \longleftarrow i+1$.
7: **end while**
8: Calculate $h_n^f = [K_f - N_{\text{tot}}(f)]^+, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$.
9: *—Procedure 2. [Decide $\{y_n^f\}$]*
10: Set $\overline{S}_n^B = \lfloor S_n^B/B_0 \rfloor, \forall n \in \mathcal{N}$, $\overline{\mathbf{N}}_{\text{tot}} = \mathbf{N}_{\text{tot}}$, $\boldsymbol{\eta} = \big\{(K_f - h_n^f)\theta_n^f \sum_{k\in\mathcal{N}\setminus\{n\}} c_{kn}\big\}_{N\times F}$, $\overline{\boldsymbol{\eta}} = \mathbf{0}_{(NF)\times 1}$, $i = 1$.
11: Reshape $\boldsymbol{\eta}$ into $\overline{\boldsymbol{\eta}}$, where $\eta_n^f = \overline{\eta}_{(f-1)N+n}, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$.
12: Sort $\overline{\boldsymbol{\eta}}$ into $\boldsymbol{\varpi}$ in a descending order, and label the orginal index vector as $\overline{\mathbf{O}}$ satisfying $\varpi_j = \overline{\eta}_{\overline{O}_j}, \forall j$.
13: **while** $\sum_{f\in\mathcal{F}} \overline{N}_{\text{tot}}(f) > 0$ **do**
14: Calculate $n^* = \text{mod}(\overline{O}_i - 1, N) + 1$, $f^* = \frac{\overline{O}_i - n^*}{N} + 1$, and $y_{n^*}^{f^*} = \min\big\{K_{f^*} - h_{n^*}^{f^*}, \overline{N}_{\text{tot}}(f^*), \overline{S}_{n^*}^B\big\}$.
15: Set $\overline{N}_{\text{tot}}(f^*) \longleftarrow \overline{N}_{\text{tot}}(f^*) - y_{n^*}^{f^*}$, $\overline{S}_{n^*}^B \longleftarrow \overline{S}_{n^*}^B - y_{n^*}^{f^*}$, and $i \longleftarrow i+1$.
16: **end while**
17: *—Procedure 3. [Decide $\{\phi_{kn}^f\}$]*
18: Set $c_{nn} = +\infty, \forall n \in \mathcal{N}$, and $\boldsymbol{\Gamma} = \mathbf{0}_{N\times 1}$.
19: **for** $n = 1$ to $N$ **do**
20: Set $\boldsymbol{\Gamma} = \{c_{kn}\}_{k=1}^N$.
21: Sort $\boldsymbol{\Gamma}$ into $\boldsymbol{\Lambda}$ in an ascending order, and label the orginal index vector as $\widehat{\mathbf{O}}$ satisfying $\Lambda_j = \Gamma_{\widehat{O}_j}, \forall j$.
22: **for** $f = 1$ to $F$ **do**
23: **if** $K_f - h_n^f - y_n^f > 0$ **then**
24: Find $j^* = \arg\min_{j\in\mathcal{N}} \big\{j|\sum_{i=1}^{j} y_{\widehat{O}_i}^f \geq K_f - h_n^f - y_n^f\big\}$.
25: **if** $j^* == 1$ **then**
26: Set $\phi_{\widehat{O}_1,n}^f = K_f - h_n^f - y_n^f$.
27: **else**
28: Set $\phi_{\widehat{O}_j,n}^f = y_{\widehat{O}_j}^f, j \in \{1, 2, \ldots, j^*-1\}$, and $\phi_{\widehat{O}_{j^*},n}^f = K_f - h_n^f - y_n^f - \sum_{i=1}^{j^*-1} y_{\widehat{O}_i}^f$.
29: **end if**
30: **end if**
31: **end for**
32: **end for**
33: **Output**: $\{y_n^f, \phi_{kn}^f, h_n^f\}$.

The above large-scale optimization problem in (26) is still NP-hard and even impossible to get its optimal solutions. Thus, from the perspective of practical engineering implementation, we propose a low-complexity greedy heuristic method as shown in Algorithm 2. The proposed Algorithm 2 has three main procedures as:

- **Procedure 1**: Decide the total numbers of packets of each content that will be cached in all the BSs and that will be downloaded from SPs at each BS, i.e., $\{\sum_{n\in\mathcal{N}} y_n^f\}$ and $\{h_n^f\}$, respectively, which is shown in Lines 2-8;
- **Procedure 2**: Decide the number of packets of each content that will be cached in each BS, i.e., $\{y_n^f\}$, which is shown in Lines 9-16;
- **Procedure 3**: Decide the number of packets of each content that will be delivered between each pair of BSs, i.e., $\{\phi_{kn}^f\}$, which is shown in Lines 17-32.

Specifically, in **Procedure 1**, we denote $N_{\text{tot}}(f) = \{\sum_{n\in\mathcal{N}} y_n^f\}, \forall f \in \mathcal{F}$ as the total number of packets of content $f$ cached in all the BSs, $\mathbf{N}_{\text{tot}} = [N_{\text{tot}}(1), N_{\text{tot}}(2), \ldots, N_{\text{tot}}(F)]$, $S_{\text{tot}} = \sum_{n\in\mathcal{N}} \lfloor S_n^B/B_0 \rfloor$ as the total cache size of all the BSs, and $A_f = \sum_{n\in\mathcal{N}} c_o\theta_n^f B_0$ as the total cost of downloading a packet of content $f$ from SPs to all the BSs. Then **Procedure 1** aims to solve the optimization problem as $\min_{\mathbf{N}_{\text{tot}}} \sum_{f\in\mathcal{F}} [K_f - N_{\text{tot}}(f)]^+ A_f$, $s.t.$, $\sum_{f\in\mathcal{F}} N_{\text{tot}}(f) \leq S_{\text{tot}}, N_{\text{tot}}(f) \in \{0, 1, 2, \ldots, NK_f\}, \forall f \in \mathcal{F}$. In Line 3, we initialize $N_{\text{tot}}(f) = K_f\big(\min\{N, \lceil S_{\text{tot}}/\sum_{f\in\mathcal{F}} K_f\rceil\} - 1\big)$ for the case[4] that $S_{\text{tot}} \geq \sum_{f\in\mathcal{F}} K_f$. Then the following process in Lines 5-7 aims to iteratively update the elements of $\mathbf{N}_{\text{tot}}$ and make them as large as possible to achieve the corresponding optimal solutions. Besides, as shown in Line 8, $\{h_n^f\}$ can be directly achieved based on Theorem 5. In **Procedure 2**, we denote $\eta_n^f = (K_f - h_n^f)\theta_n^f \sum_{k\in\mathcal{N}\setminus\{n\}} c_{kn}, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$. Then **Procedure 2** aims to solve the optimization problem as $\min_{\{y_n^f\}} -\sum_{f\in\mathcal{F}}\sum_{n\in\mathcal{N}} y_n^f\eta_n^f$, $s.t.,(23b),(23f)$, $\sum_{n\in\mathcal{N}} y_n^f = N_{\text{tot}}(f), y_n^f \leq K_f - h_n^f, \forall n \in \mathcal{F}, \forall f \in \mathcal{F}$. The following process in Lines 13-16 iteratively update $\{y_n^f\}$ in a greedy manner. In **Procedure 3**, $\{\phi_{kn}^f\}$ is achieved with given $\{h_n^f, y_n^f\}$ by using a greedy method to minimize the the total amount of average system cost from packet delivery of content among BSs via BS-BS links.

Moreover, the used greedy methods in **Procedure 1** and **Procedure 3** are optimal while the search method in **Procedure 2** is near-optimal. Besides, the whole procedure of Algorithm 2 for solving the unlimited capacity case takes polynomial time, i.e., $O\big(\sum_{f\in\mathcal{F}} K_f \log\big(\sum_{f\in\mathcal{F}} K_f\big) + NF\log(NF) + N^2F\log(N)\big)$, which is mainly bounded by the sorting.

*2) Limited Capacity Case:* In this case, (23c) needs to be considered to solve the ILP problem in (23). Considering its NP-hardness and large scale, we also aim to provide a suboptimal heuristic solution as shown in Algorithm 3. Based on the results in the unlimited capacity case, the proposed

[4]In such a case that may not hold in practice due to the large scale of content and the limits of cache sizes of BSs, using this initialization can effectively reduce the iterations of the proposed method in **Procedure 1**.

**Algorithm 3** Heuristic Method for Solving the Problem in (23) of the Limited Capacity Case

1: **Input:** $(C_{kn})_{N\times N}$ where $C_{nn}=0$.
2: Calculate the results $\{y_n^f, \phi_{kn}^f, h_n^f\}$ in the case of unlimited backhaul capacity.
3: Check the feasibility for whether (23c) holds, and set $\mathcal{N}_1 = \{(k,n)|\sum_{f\in\mathcal{F}}\phi_{kn}^f\theta_n^f B_0 > C_{kn}, k\neq n\in\mathcal{N}\}$.
4: **for** each fixed $(k,n)\in\mathcal{N}_1$ **do**
5: For each specific $f\in\mathcal{F}$ such that $\theta_n^f B_0 > C_{kn}$, set $\phi_{kn}^f = 0$.
6: **while** $\sum_{f\in\mathcal{F}}\phi_{kn}^f\theta_n^f B_0 > C_{kn}$ **do**
7: Find $f^* = \arg\min_{f\in\mathcal{F},\phi_{kn}^f>0}\{\phi_{kn}^f\theta_n^f\}$.
8: Set $\phi_{kn}^{f^*} \leftarrow \phi_{kn}^{f^*} - 1$.
9: **end while**
10: Calculate $h_n^f = K_f - y_n^f - \sum_{k\in\mathcal{N}}(\phi_{kn}^f), \forall f\in\mathcal{F}$.
11: **end for**
12: **Output:** $\{y_n^f, \phi_{kn}^f, h_n^f\}$.

**Algorithm 4** Content Request Routing Strategy

1: After a request for content $f$ is generated by user $u$, satisfy the request if $x_u^f = 1$.
2: If not yet satisfied, user $v$ fetches content $f$ from a nearby user $v$ via D2D links with the probability $p_{vu}^D$ if $x_v^f = 1$.
3: If not yet satisfied via D2D sharing in the bottom level, then user $u$ needs to download content $f$ via cellular links and is associated with BS $n$ with the probability $p_{un}^B$.
4: The local BS $n$ satisfies the corresponding request by the following steps:
5: a) Check the number of the locally cached coded packets of content $f$, i.e., $y_n^f$;
6: b) Fetch $\phi_{kn}^f$ coded packets from BS $k$ if $\phi_{kn}^f > 0$;
7: c) Download $h_n^f$ coded packets over the Internet via the MNO core if $h_n^f > 0$.
8: d) After collecting $K_f$ different coded packets of content $f$, decode the packets for recovering content $f$.
9: e) Deliver content $f$ to user $u$ via cellular links.

Algorithm 3 operates a iterative process in Lines 4-11 for making that the constraint (23c) holds, which takes polynomial time, i.e., $O(|\mathcal{N}_1|F\log(F))$.

In all, to solve the complex optimization problem of BS caching in the upper level, we provide a near-optimal solution with Algorithm 2 in the unlimited capacity case, and provide a suboptimal solution with Algorithm 3 in the limited capacity case.

### D. Content Request Routing

Based on the above decomposed content caching cooperation framework in the D2D-aided mobile network, we can get the strategies of joint content placement and content delivery with the practical considerations on large-scale content distribution. For satisfying a request for content $f$ from user $u$, the details of the proposed content request routing strategy are shown in Algorithm 4.

## IV. Trace-Based Simulation Results

In this section, we evaluate our hierarchical edge caching scheme based on the practical trace from a mobile application *Xender* used for content sharing via D2D communications.

### A. Setup

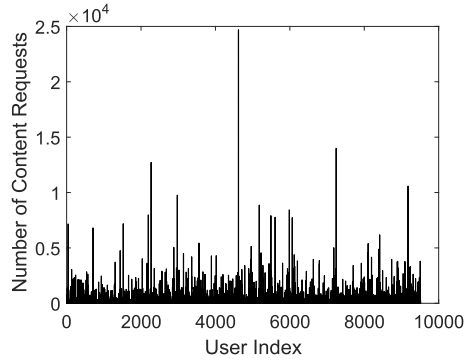
*Xender* is a world-wide popular mobile application for D2D sharing, and mobile users can use it to share different types of content on a large diversity of mobile platforms (e.g., Android, iOS and Windows), instead of using 3G/4G cellular networks [37]. In particular, the D2D links in *Xender* are established mostly via WiFi tethering, while WiFi Direct and Bluetooth are also supported.

We capture *Xender*'s trace for the whole month in February 2016, which consists of selected 9,514 active mobile users, 188,447 content files and 2,107,100 requests. For simulation purpose, we set the number of BSs $N$ as 10, the weighted parameter $\omega_0$ as 0.5, the size of each coded packet $B_o$ as 64 KByte. All mobile users have the same cache size (i.e., $S_u^D \equiv S^D$) and all BSs are also of the same cache size (i.e., $S_n^B \equiv S^B$). Besides, we set the capacity of each BS-BS link as 1 Gbps. The D2D sharing among mobile users is modeled by the practical analysis of *Xender*'s trace while the association of users and BSs is randomly set as modeled in Sec. II-D. Particularly, we employed the same $k$-means ($k=3$) grouping method used in [37] to achieve the parameters $\{\alpha_i\}_{i=1}^3$ in the defined relationship factors. Moreover, the scalability of our proposed framework is not restricted by the aforementioned parameters. We use Python to implement a simulator that constructs the formed edge caching hierarchy.

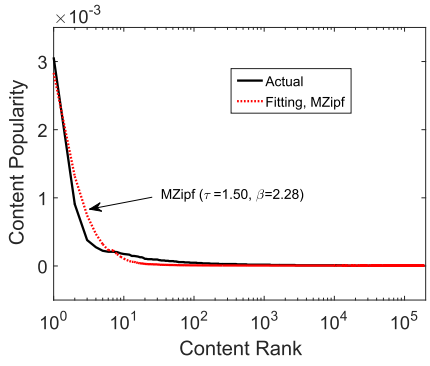
### B. Baseline Schemes and Performance Metrics

In particular, we compare the proposed scheme with four baseline schemes as: 1) *Revised FemtoCaching (FemtoR)*, derived by carefully modifying the FemtoCaching scheme in [6] to address the single-level case of only coded caching in BSs, where D2D sharing is not considered; 2) *Most Popular Caching (MPC)*, derived by caching most popular content in the bottom level and upper level; 3) *Least Recently Used (LRU) Caching*, derived from [43] to address the same case in the paper; 4) *Hierarchical Uncoded Edge Caching (HUEC)*, derived by carefully modifying the proposed hierarchical caching scheme (that is used in IPTV systems) in [24] to address the case where both device caching and BS caching are uncoded.

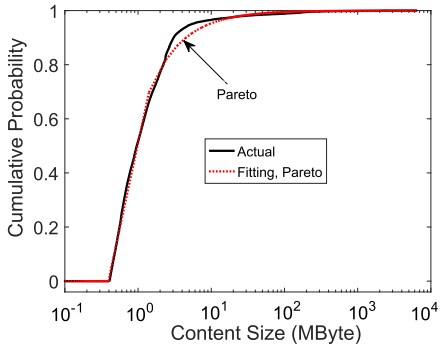
To evaluate the schemes, four performance metrics are used as: 1) *Percentage of traffic offload*, denoting the reduction on the traffic from downloading content via backhaul links; 2) *Percentage of supported requests* via content caching; 3) *Percentage* of reduced costs, denoting the reduction on the costs from delivering content via backhaul links and BS-BS links; 4) *Link utilization*, denoting the utilization ratio of




(a) Content requests of users




(b) Content popularity




(c) Content size

Fig. 5. Distributions of users' content requests, content popularity and content size in the Xender's trace.

BS-BS links for packet delivery via the cooperation among BSs. Note that all the percentages of traffic offload, supported requests and reduced costs in the first three performance metrics are normalized by the results of the non-caching scheme.

### C. Distributions

Fig. 5 and Fig. 6 show different distributions in the *Xender*'s trace and random settings. Fig. 5(a) shows different numbers of content requests from different users in the Xender's trace. From Fig. 5(b) and Fig. 5(c), the actual content popularity and size distribution in the Xender's trace can be well fitted by a MZipf distribution with $(\tau, \beta) = (1.50, 2.28)$ and a Pareto distribution, respectively, which agrees with the used models in [34] and [44]. Fig. 6(a) and Fig. 6(b) describe the achieved content factor[5] and relationship factor with $(\alpha_1, \alpha_2, \alpha_3) = (0.9, 0.5, 0.1)$ in the Xender's trace, respectively. Fig. 6(c) shows the achieved D2D sharing probability in the Xender's trace, while Fig. 6(d) shows the association probability between mobile users and BSs in the used random settings. Note that the following results are based on the above practical trace and settings.

### D. Effects of Different Cache Sizes of BSs

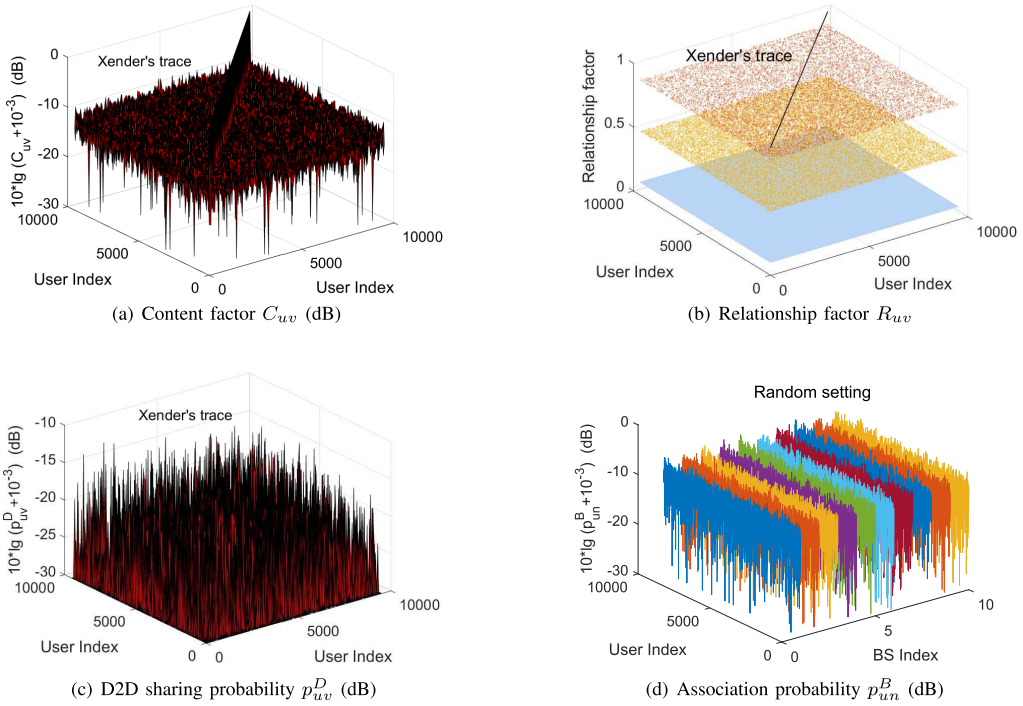
Fig. 7 and Fig. 8 compare the performance of the proposed scheme with the baseline schemes in terms of the four performance metrics versus different cache sizes (percentage to the total content size) of each BS in the entire system and different levels, respectively. Here, the cache size of each mobile user is set as $S^D = 1$ GByte. From Fig. 7, the proposed scheme always outperforms the FemtoR, MPC, LRU and HUEC schemes on all performance metrics for the entire system. Besides, from Fig. 8, as the cache size of each BS increases, the performance of each scheme in the upper level is improved rapidly since BSs can cache more content, while the performance in the bottom level keeps the same due to the fixed cache size of each user.

Moreover, from Fig. 7(a), the proposed scheme can offload the traffic of the entire system by $33.7\%$ to $100\%$, and outperforms the four baseline schemes with at least $38.0\%$, $67.9\%$, $72.4\%$ and $115.0\%$ improvements for the entire system and upper level, respectively. Meanwhile, from Fig. 8(a), in the upper level, the proposed scheme and the HUEC scheme can offload more traffic than the other schemes. From Fig. 7(b), the proposed scheme can support around $28.0\%$, $15.0\%$, $28.8\%$ and $29.0\%$ of total requests more than the above four baseline schemes for the entire system, respectively. Meanwhile, from Fig. 8(b), in the upper level, the percentage of supported requests in the proposed scheme is only slightly less than the MPC scheme since most popular content is cached in the MPC scheme and the number of supported requests is positively correlated with the content popularity. Fig. 7(c) shows that the proposed scheme can reduce the most costs, and achieves up to $4.6\%$, $59.7\%$, $77.8\%$ and $110.8\%$ improvements compared with the four baseline schemes, respectively. From Fig. 7(d), the proposed scheme can achieve the highest and even full link utilization and thus the best cooperation among BSs, and outperforms the four baseline schemes with the improvements of up to $19.2\%$, $197.2\%$, $95.2\%$ and 12 times, respectively.

### E. Effects of Different Cache Sizes of Mobile Devices

Fig. 9 and Fig. 10 evaluate the performance of the considered caching schemes in terms of different performance metrics versus different cache sizes of each mobile user in the entire system and different levels, respectively. In particular, the cache size $S^B$ of each BS is fixed as $5\%$ of the total content size. From Fig. 9, the proposed scheme can also achieve much better performance than the four baseline schemes on

[5]To show the probability diversity, we plot some probabilities (say $x$) in log-domain, e.g., $10\lg(x + 10^{-3})$ dB.



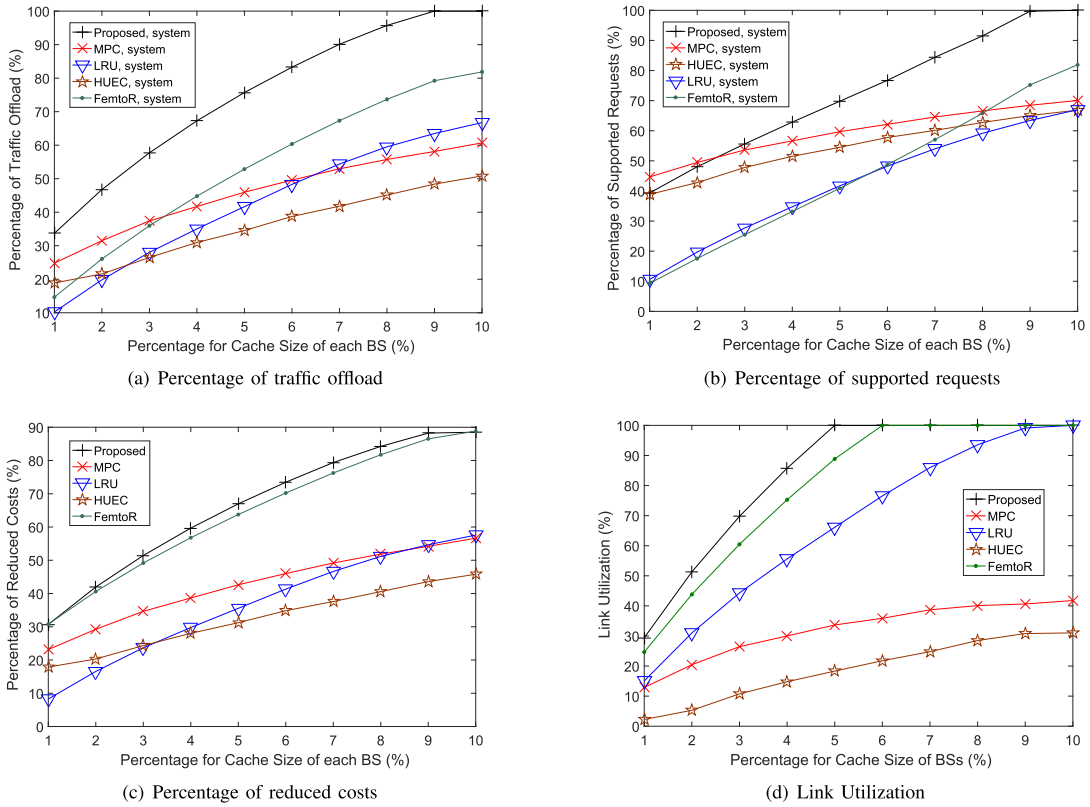

(a) Content factor $C_{uv}$ (dB)

(b) Relationship factor $R_{uv}$

(c) D2D sharing probability $p_{uv}^{D}$ (dB)

(d) Association probability $p_{un}^{B}$ (dB)

Fig. 6. Distributions of content factor, relationship factor, D2D sharing probability and association probability in the Xender's trace and random settings.




(a) Percentage of traffic offload

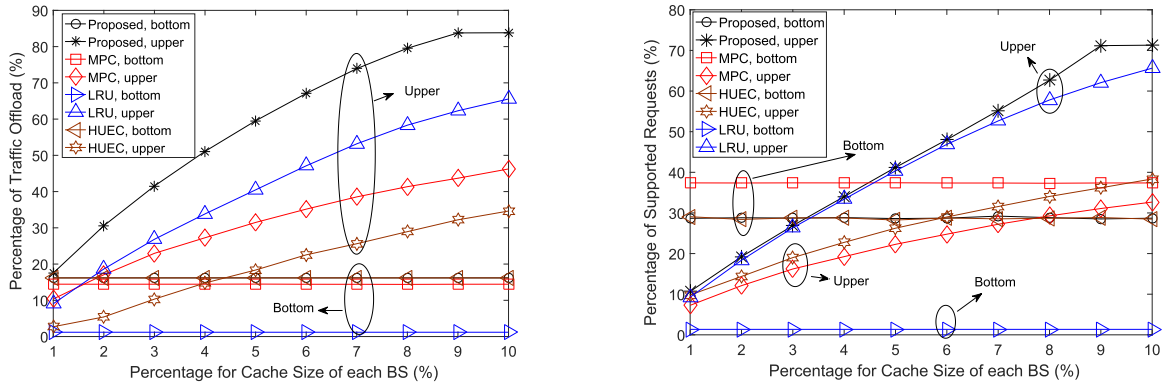
(b) Percentage of supported requests

(c) Percentage of reduced costs

(d) Link Utilization

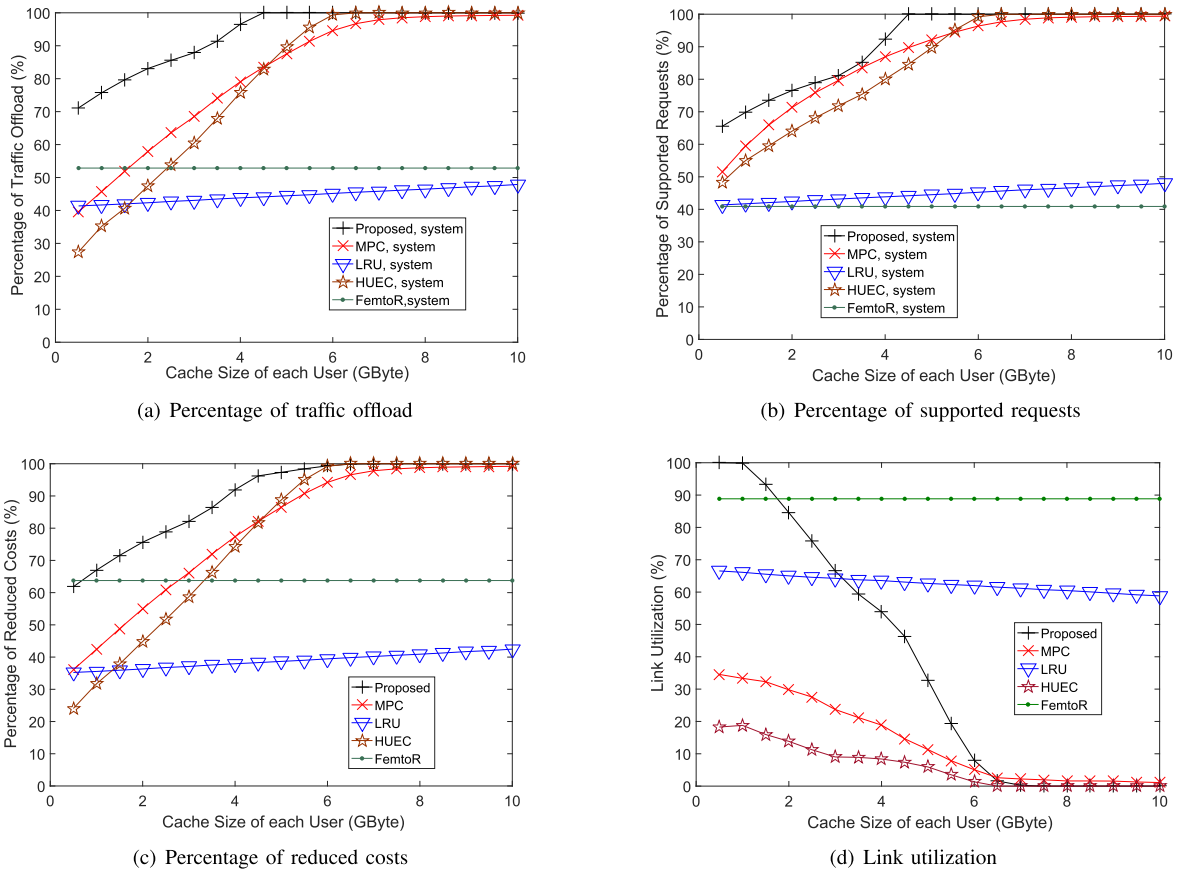
Fig. 7. The percentage of traffic offload, the percentage of supported requests, the percentage of reduced costs and the link utilization versus different cache sizes (percentage to the total content size) of each BS in the entire system, where $S^D = 1$ GByte.

all the first three performance metrics for the entire system. Besides, as the cache size of each mobile user increases, the performance of the FemtoR scheme keeps the same since it only considers the BS caching, and the performance of the other schemes is improved rapidly in the bottom level since mobile users can cache more content while their performance in the considered four metrics degrades in the upper level. Specifically, from Fig. 10, when the cache size of each user is




(a) Percentage of traffic offload (b) Percentage of supported requests

Fig. 8. The percentage of traffic offload and the percentage of supported requests versus different cache sizes (percentage to the total content size) of each BS in the bottom level and upper level, where $S^D = 1$ GByte.



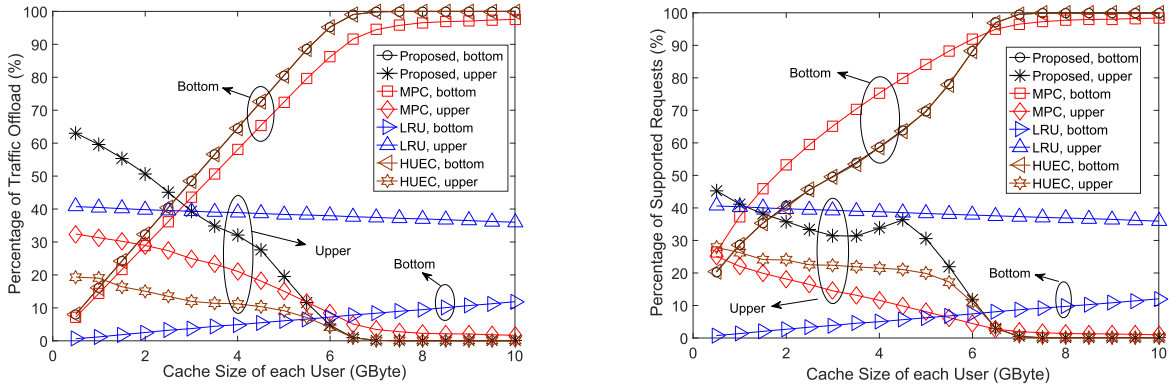

(a) Percentage of traffic offload (b) Percentage of supported requests

(c) Percentage of reduced costs (d) Link utilization

Fig. 9. The percentage of traffic offload, the percentage of supported requests, the percentage of reduced costs and the link utilization versus different cache sizes of each mobile user in the entire system, where the cache size $S^B$ of each BS is set as 5% of the total content size.

sufficiently large, D2D sharing in the bottom level can satisfy all the content requests and BS caching in the upper level is not necessary.

Moreover, Fig. 9(a) shows that the proposed scheme can offload the traffic of the entire system by 71.1% to 100%, and outperforms the FemtoR, MPC, LRU and HUEC schemes with up to 89.2%, 79.6%, 96.3% and 151.0% improvements for the entire system and bottom level, respectively. Meanwhile, Fig. 10(a) shows that the proposed scheme and the HUEC scheme can also offload more traffic than the other schemes in the bottom level. From Fig. 9(b), the proposed scheme can support up to 59.1%, 14.0%, 55.0% and 17.2% of total requests more than the above four baseline schemes for the entire system, respectively. Meanwhile, Fig. 10(b) shows that the proposed scheme only supports slightly less requests than the MPC scheme in the bottom level. From Fig. 9(c), the proposed scheme can reduce the costs by at least 63.8%, and obtains up to 56.8%, 75.5%, 88.4% and 157.0% improvements compared




(a) Percentage of traffic offload

(b) Percentage of supported requests

Fig. 10. The percentage of traffic offload and the percentage of supported requests versus different cache sizes of each mobile user in the bottom level and upper level, where the cache size $S^B$ of each BS is set as $5\%$ of the total content size.

with the four baseline schemes, respectively. Fig. 9(d) shows that the FemtoR scheme achieves the same link utilization since it does not consider D2D sharing, while the link utilization achieved by the other schemes decreases to even zero with the increase of the cache size of each mobile user since D2D sharing satisfies increasing content requests in the bottom level. Particularly, the link utilization achieved by the proposed scheme is mostly larger than those achieve by both the MPC and HUEC schemes.

## V. Conclusion

In this paper, we have proposed a collaborative hierarchical edge caching framework in D2D aided mobile networks. Specifically, based on the analysis of social behavior and preference of mobile users, heterogenous cache sizes and the derived system topology, we have optimized the maximum capacity of the network infrastructure on offloading the network traffic, reducing system costs and supporting mobile users' content requests inside the mobile network, and proposed the corresponding low-complexity solutions from the perspective of engineering implementation. Trace-based simulation results have shown that the proposed hierarchical edge caching framework has excellent performance and outperforms the considered four baseline schemes in terms of offloading network traffic, satisfying content requests and reducing system costs.

## Appendix

### A. Proof of Theorem 1

We first prove $z_{uv}^f = \min\{x_u^f, 1 - x_v^f\}$ by discussions as: 1) if $(x_u^f, x_v^f) \in \{(0,0),(0,1),(1,1)\}$, then we have $z_{uv}^f = \min\{x_u^f, 1 - x_v^f\} = 0$; 2) otherwise, i.e., $(x_u^f, x_v^f) = (1,0)$, then we have $z_{uv}^f = \min\{x_u^f, 1 - x_v^f\} = 1$.

Considering that the problem in (11) is a maximization problem and all the coefficients $\{p_{uv}^D \lambda_v^f s_f\}$ associated with the quadratic terms $\{x_u^f(1-x_v^f)\}$ are non-negative, the proved $z_{uv}^f = \min\{x_u^f, 1 - x_v^f\}$ can further be equivalent to two inequalities, i.e., $z_{uv}^f \le x_u^f$ and $z_{uv}^f \le 1 - x_v^f$.

### B. Proof of Theorem 2

First, we prove that $\mathbf{a} \in \{-1,1\}^{m\times n\times k}$. By using the definition of $\Phi$ and the Cauchy-Schwarz Inequality, we have $mnk = \langle \mathbf{a}, \mathbf{v}\rangle \le \|\mathbf{a}\|_2 \cdot \|\mathbf{v}\|_2 \le \|\mathbf{a}\|_2 \cdot \sqrt{mnk}$. Thus, we can obtain $\|\mathbf{a}\|_2^2 \ge mnk$. Considering $-1 \preceq \mathbf{a} \preceq 1$, we have $\|\mathbf{a}\|_2^2 \le mnk$. Combining the above two aspects, we have $\|\mathbf{a}\|_2^2 = mnk$. Then considering $-1 \preceq \mathbf{a} \preceq 1$ again, we have $\mathbf{a} \in \{-1,1\}^{m\times n\times k}$.

Second, we prove that $\mathbf{v} \in \{-1,1\}^{m\times n\times k}$. We have

$$mnk = \langle \mathbf{a}, \mathbf{v}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} a_{ijl}v_{ijl} \le \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} |a_{ijl}||v_{ijl}|$$
$$\le \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{l=1}^{k} |v_{ijl}| \le \sqrt{mnk}\|\mathbf{v}\|_2. \quad (27)$$

Thus, we can obtain $\|\mathbf{v}\|_2^2 \ge mnk$. Combining that $\|\mathbf{v}\|_2^2 \le mnk$, we have $\|\mathbf{v}\|_2^2 = mnk$. With the Squeeze Theorem, all the equalities in (27) hold automatically. Using the equality condition for the Cauchy-Schwarz Inequality, we can obtain $\mathbf{v} \in \{-1,1\}^{m\times n\times k}$.

At last, considering $\mathbf{a} \in \{-1,1\}^{m\times n\times k}$, $\mathbf{v} \in \{-1,1\}^{m\times n\times k}$, and $\langle \mathbf{a}, \mathbf{v}\rangle = mnk$, we can obtain $\mathbf{a} = \mathbf{v}$.

Based on the above analysis, the whole proof is complete.

### C. Proof of Theorem 3

We provide a lemma as below, which is an extension of the conclusion in [41] and can be proved by using a simple parallel-to-serial conversion on the elements of matrices and then using the same method in [41].

*Lemma 1: Consider the following optimization problem:*

$$(\boldsymbol{\pi}_\rho^*, \boldsymbol{\nu}_\rho^*) = \min_{-1\preceq\boldsymbol{\pi}\preceq 1, \|\boldsymbol{\nu}\|_2^2 \le G, \boldsymbol{\pi}\in\Omega} \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\nu}, \rho). \quad (28)$$

*$F(\boldsymbol{\pi})$ is a L-Lipschitz continuous convex function on $-1 \preceq \boldsymbol{\pi} \preceq 1$. When $\rho > 2L$, $\langle \boldsymbol{\pi}_\rho^*, \boldsymbol{\nu}_\rho^*\rangle = G$ will be achieved for any local optimal solution to the problem in (28).*

To prove Theorem 3, we denote $(\boldsymbol{\pi}^*, \mathbf{v}^*)$ as any global optimal solution to the original problem in (14), and $(\boldsymbol{\pi}_\rho^*, \mathbf{v}_\rho^*)$

as any global optimal solution to the dual problem in (15) for some $\rho > 2L$.

First, we prove that $(\boldsymbol{\pi}^*, \mathbf{v}^*)$ is also a global optimal solution to the problem in (15). For any feasible solution $(\boldsymbol{\pi}, \mathbf{v})$ satisfying $-1 \preceq \boldsymbol{\pi} \preceq 1, \|\mathbf{v}\|_2^2 \leq G, \boldsymbol{\pi} \in \Omega$, we can derive

$$\begin{aligned}\mathcal{L}(\boldsymbol{\pi}, \mathbf{v}, \rho) &\geq \min_{-1\preceq\boldsymbol{\pi}\preceq 1, \|\mathbf{v}\|_2^2\leq G, \boldsymbol{\pi}\in\Omega} F(\boldsymbol{\pi}) + \rho(G - \langle \boldsymbol{\pi}, \mathbf{v}\rangle) \\ &= \min_{-1\preceq\boldsymbol{\pi}\preceq 1, \|\mathbf{v}\|_2^2\leq G, \boldsymbol{\pi}\in\Omega} F(\boldsymbol{\pi}), \quad s.t.\ \langle \boldsymbol{\pi}, \mathbf{v}\rangle = G \\ &= F(\boldsymbol{\pi}^*) + \rho(G - \langle \boldsymbol{\pi}^*, \mathbf{v}^*\rangle) \\ &= \mathcal{L}(\boldsymbol{\pi}^*, \mathbf{v}^*, \rho),\end{aligned}$$

where the first equality holds since the constraint $\langle \boldsymbol{\pi}, \mathbf{v}\rangle = G$ is satisfied at the local optimal solution when $\rho > 2L$ (see Lemma 1). Thus, $(\boldsymbol{\pi}^*, \mathbf{v}^*)$ is also a global optimal solution to the problem in (15).

Second, we prove that $(\boldsymbol{\pi}_\rho^*, \mathbf{v}_\rho^*)$ is also a global optimal solution to the problem in (14). For any feasible solution $(\boldsymbol{\pi}, \mathbf{v})$ satisfying $-1 \preceq \boldsymbol{\pi} \preceq 1, \|\mathbf{v}\|_2^2 \leq G, \langle \boldsymbol{\pi}, \mathbf{v}\rangle = G, \boldsymbol{\pi} \in \Omega$, we can derive

$$\begin{aligned}F(\boldsymbol{\pi}_\rho^*) - F(\boldsymbol{\pi}) &= F(\boldsymbol{\pi}_\rho^*) + \rho(G - \langle \boldsymbol{\pi}_\rho^*, \mathbf{v}_\rho^*\rangle) - F(\boldsymbol{\pi}) \\ &\quad - \rho(M - \langle \boldsymbol{\pi}, \mathbf{v}\rangle) \\ &= \mathcal{L}(\boldsymbol{\pi}_\rho^*, \mathbf{v}_\rho^*, \rho) - \mathcal{L}(\boldsymbol{\pi}, \mathbf{v}, \rho) \\ &\leq 0.\end{aligned}$$

Thus, $(\boldsymbol{\pi}_\rho^*, \mathbf{v}_\rho^*)$ is also a global optimal solution to the problem in (14).

Finally, we can conclude that when $\rho > 2L$, the biconvex optimization problem in (15) has the same local and global minima with the primary problem in (14).

### *D. Proof of Theorem 4*

Denote $s$ and $t$ as the numbers of the outer iteration and inner iteration in Algorithm 1, respectively.

First, we prove the convergence rate of Algorithm 1. We assume that Algorithm 1 takes $s$ outer iterations to converge, and denote $F'(\boldsymbol{\pi})$ as the subgradient of $F(\boldsymbol{\pi})$. Based on the $\boldsymbol{\pi}$-subproblem in (18), if $\boldsymbol{\pi}^*$ solves (18), then we can get the following mixed variational inequality condition as

$$\langle \boldsymbol{\pi} - \boldsymbol{\pi}^*, F'(\boldsymbol{\pi}^*)\rangle + \rho(G - \sqrt{G}\|\boldsymbol{\pi}\|_2) - \rho(G - \sqrt{G}\|\boldsymbol{\pi}^*\|_2) \geq 0, \quad \text{for } \forall \boldsymbol{\pi} \in [-1, 1]^{U\times(U+1)\times F} \cap \Omega.$$

Letting $\boldsymbol{\pi}$ be any feasible solution such that $\boldsymbol{\pi} \in \{-1, 1\}^{U\times(U+1)\times F} \cap \Omega$, we can get

$$\begin{aligned}G - \sqrt{G}\|\boldsymbol{\pi}^*\|_2 &\leq G - \sqrt{G}\|\boldsymbol{\pi}\|_2 + \frac{1}{\rho}\langle \boldsymbol{\pi} - \boldsymbol{\pi}^*, F'(\boldsymbol{\pi}^*)\rangle \\ &\leq \frac{1}{\rho}\|\boldsymbol{\pi} - \boldsymbol{\pi}^*\|_2 \|F'(\boldsymbol{\pi}^*)\|_2 \leq L\sqrt{2G}/\rho, \quad (29)\end{aligned}$$

where the second ineuality is achieved with the Cauchy-Schwarz Inequality, and the third inequality is achieved due to the conclusion $\|\mathbf{x} - \mathbf{y}\|_2 \leq \sqrt{2G}, \forall -1 \preceq \mathbf{x}, \mathbf{y} \preceq 1$ and the Lipschitz continuity of $F(\boldsymbol{\pi})$ that $\|F'(\boldsymbol{\pi}^*)\|_2 \leq L$. From (29), we can observe that when $\rho^{(s)} \geq L\sqrt{2G}/\varepsilon$, Algorithm 1 can achieve the accuracy with at least $G - \sqrt{G}\|\boldsymbol{\pi}\|_2 \leq \varepsilon$. Considering that $\rho^{(s)} = \Delta^s \rho^{(0)}$, we have $\Delta^s \geq \frac{L\sqrt{2G}}{\varepsilon\rho^{(0)}}$, and thus $s \geq \ln(L\sqrt{2G}) - \ln(\varepsilon\rho^{(0)}))/\ln\Delta$.

Second, we prove the asymptotic monotone property of Algorithm 1. We can get the following inequalities as

$$\begin{aligned}&F(\boldsymbol{\pi}^{(t+1)}) - F(\boldsymbol{\pi}^{(t)}) \\ &\quad \leq \rho(G - \langle \boldsymbol{\pi}^{(t)}, \mathbf{v}^{(t)}\rangle) - \rho(G - \langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t)}\rangle) \\ &\quad = \rho(\langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t)}\rangle - \langle \boldsymbol{\pi}^{(t)}, \mathbf{v}^{(t)}\rangle) \\ &\quad \leq \rho(\langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t+1)}\rangle - \langle \boldsymbol{\pi}^{(t)}, \mathbf{v}^{(t)}\rangle) = 0,\end{aligned}$$

where the first inequality uses the conclusion that $F(\boldsymbol{\pi}^{(t+1)}) + \rho(G - \langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t)}\rangle) \leq F(\boldsymbol{\pi}^{(t)}) + \rho(G - \langle \boldsymbol{\pi}^{(t)}, \mathbf{v}^{(t)}\rangle)$ holds since $\boldsymbol{\pi}^{(t+1)}$ is the optimal solution to the $\boldsymbol{\pi}$-subproblem in (18); the second inequality uses the conclusion that $-\langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t+1)}\rangle \leq -\langle \boldsymbol{\pi}^{(t+1)}, \mathbf{v}^{(t)}\rangle$ holds since $\mathbf{v}^{(t+1)}$ is the optimal solution to the $\mathbf{v}$-subproblem in (19); the last equality holds due to $\langle \boldsymbol{\pi}, \mathbf{v}\rangle = G$. Note that the equality $\langle \boldsymbol{\pi}, \mathbf{v}\rangle = G$ as well as the feasible set $-1 \preceq \boldsymbol{\pi} \preceq 1, \|\mathbf{v}\|_2^2 \leq G$ also implies that $\boldsymbol{\pi} \in \{-1, 1\}^{U\times(U+1)\times F}$.

Based on the above analysis, the whole proof is complete.

### *E. Proof of Theorem 5*

Clearly, (24) holds since delivering or downloading one more packet leads to a larger value of the objective in (23a). Then we prove (25) holds as follow:

1) Based on (23a) and (23d), we have

$$(h_n^f)^* = [K_f - (y_n^f)^* - \sum_{k\in\mathcal{N}} (\phi_{kn}^f)^*]^+, \quad \forall n \in \mathcal{N},\ \forall f \in \mathcal{F}. \tag{30}$$

Then based on (23d) and (30), we can further obtain $(h_n^f)^* \geq [K_f - \sum_{k\in\mathcal{N}} (y_k^f)^*]^+ \geq 0, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$.

2) We prove that $(h_n^f)^* \leq [K_f - \sum_{k\in\mathcal{N}} (y_k^f)^*]^+, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$ holds. If there exists any $j \in \mathcal{F}$ such that $(h_n^j)^* > [K_j - \sum_{k\in\mathcal{N}} (y_k^j)^*]^+ \geq 0$, then based on (30), we have $(h_n^j)^* = K_j - (y_n^j)^* - \sum_{k\in\mathcal{N}} (\phi_{kn}^j)^* > K_j - \sum_{k\in\mathcal{N}} (y_k^j)^*$, i.e., $\sum_{k\in\mathcal{N}\setminus\{n\}} [(y_k^j)^* - (\phi_{kn}^j)^*] > 0$. Thus, combing (23d), we obtain that there exists $i \in \mathcal{N}$ such that $(y_i^j)^* > (\phi_{in}^j)^*$. However, we can generate another feasible solution by setting $\phi_{in}^j = (\phi_{in}^j)^* + 1$ and $h_n^j = (h_n^j)^* - 1$, which can make the objective value in (23a) smaller than that with the optimal solution. Clearly, this contradicts with the assumption on the optimal solution. Thus, we have $(h_n^f)^* \leq [K_f - \sum_{k\in\mathcal{N}} (y_k^f)^*]^+, \forall n \in \mathcal{N}, \forall f \in \mathcal{F}$.

Based on the above analysis, the conclusion in (25) holds.

## REFERENCES

[1] 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 Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[2] K. Samdanis, T. Taleb, and S. Schmid, "Traffic offload enhancements for eUTRAN," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 3, pp. 884–896, 3rd Quart., 2012.

[3] H. Zhou, H. Wang, X. Li, and V. Leung, "A survey on mobile data offloading technologies," *IEEE Access.*, vol. 6, no. 1, pp. 5101–5111, Jan. 2018.

[4] K. Kanai *et al.*, "Context-aware proactive content caching for mobile video utilizing transportation systems and evaluation through field experiments," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2102–2114, Aug. 2016.

[5] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.
[6] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.
[7] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 49–55, Apr. 2014.
[8] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
[9] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *Proc. MobiHoc*, Jun. 2015, pp. 127–136.
[10] J.-P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 377–388, Jan. 2016.
[11] E. Baştuğ, M. Kountouris, M. Bennis, and M. Debbah, "On the delay of geographical caching methods in two-tiered heterogeneous networks," in *Proc. IEEE SPAWC*, Jul. 2016, pp. 1–5.
[12] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *Proc. IEEE GLOBECOM*, Dec. 2015, pp. 1–6.
[13] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.
[14] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[15] A. Liu and V. K. N. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.
[16] X. Wang, Y. Zhang, V. C. M. Leung, N. Guizani, and T. Jiang, "D2D big data: Content deliveries over wireless device-to-device sharing in large-scale mobile networks," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 32–38, Feb. 2018.
[17] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–188, Jan. 2016.
[18] B. Chen and C. Yang, "Caching policy optimization for D2D communications by learning user preference," in *Proc. IEEE VTC-Spring*, Jun. 2017, pp. 1–7.
[19] W. Zhi, K. Zhu, Y. Zhang, and L. Zhang, "Hierarchically social-aware incentivized caching for D2D communications," in *Proc. IEEE ICPADS*, Dec. 2016, pp. 316–323.
[20] L. Wang, H. Wu, Y. Ding, W. Chen, H. V. Poor, "Hypergraph-based wireless distributed storage optimization for cellular D2D underlays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2650–2666, Oct. 2016.
[21] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, "Caching based socially-aware D2D communications in wireless content delivery networks: A hypergraph framework," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 74–81, Aug. 2016.
[22] C. Yi, S. Huang, and J. Cai, "An incentive mechanism integrating joint power, channel and link management for social-aware D2D content sharing and proactive caching," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 789–802, Apr. 2018.
[23] H. Che, Y. Tung, and Z. Wang, "Hierarchical Web caching systems: Modeling, design and experimental results," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–1314, Sep. 2012.
[24] J. Dai, Z. Hu, B. Li, J. Liu, and B. Li, "Collaborative hierarchical caching with dynamic request routing for massive content distribution," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 2444–2452.
[25] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[26] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 131–145, Jan. 2016.
[27] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1382–1393, May 2017.
[28] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.
[29] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *IEEE Trans. Commun.*, to be published.
[30] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, "Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6936, Oct. 2017.
[31] Y. Wang, J. Wu, and M. Xiao, "Hierarchical cooperative caching in mobile opportunistic social networks," in *Proc. IEEE GLOBECOM*, Dec. 2014, pp. 1–6.
[32] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. IEEE ICC*, May 2016, pp. 1–6.
[33] W. Wang, R. Lan, J. Gu, A. Huang, H. Shan, and Z. Zhang, "Edge caching at base stations with device-to-device offloading," *IEEE Access*, vol. 5, pp. 6399–6410, Mar. 2017.
[34] M. Hefeeda and O. Saleh, "Traffic modeling and proportional partial caching for peer-to-peer systems," *IEEE/ACM Trans. Netw.*, vol. 16, no. 6, pp. 1447–1460, Dec. 2008.
[35] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. ACM KDD*, Aug. 2011, pp. 1100–1108.
[36] Y. Vaizman, B. McFee, and G. Lanckriet, "Codebook-based audio feature representation for music information retrieval," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1483–1493, Oct. 2014.
[37] X. Wang, H. Wang, K. Li, S. Yang, and T. Jiang, "Serendipity of sharing: Large-scale measurement and analytics for device-to-device (D2D) content sharing in mobile social networks," in *Proc. IEEE SECON*, Jun. 2017, pp. 1–9.
[38] C. Wu, X. Chen, Y. Zhou, N. Li, X. Fu, and Y. Zhang, "Spice: Socially-driven learning-based mobile media prefetching," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.
[39] L. Wang, H. Wu, and Z. Han, "Wireless distributed storage in socially enabled D2D communications," *IEEE Access*, vol. 4, pp. 1971–1984, Mar. 2016.
[40] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM J. Comput.*, vol. 38, no. 4, pp. 1411–1429, 2008.
[41] G. Yuan and B. Ghanem, "An exact penalty method for binary optimization based on MPEC formulation," in *Proc. AAAI*, Feb. 2017, pp. 2867–2875.
[42] S. Boyd and L. Vandenbergh. *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[43] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 836–845, Apr. 2016.
[44] A. B. Downey, "The structural cause of file size distributions," in *Proc. IEEE MASCOTS*, Aug. 2001, pp. 361–370.

**Xiuhua Li** (S'12) received the B.S. degree from the Honors School, Harbin Institute of Technology, Harbin, China, in 2011, and the M.S. degree from the School of Electronics and Information Engineering, Harbin Institute of Technology, in 2013. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. His current research interests are resource allocation, optimization design, distributed antenna systems, cooperative edge caching, and traffic offloading in mobile content-centric networks.

**Xiaofei Wang** (S'06–M'13–SM'18) received the B.S. degree from the Department of Computer Science and Technology, Huazhong University of Science and Technology, in 2005, and the M.S. and Ph.D. degrees from the School of Computer Science and Engineering, Seoul National University, in 2008 and 2013, respectively.

He is currently a Professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, China. He was a Post-Doctoral Fellow with The University of British Columbia from 2014 to 2016. Focusing on the research of social-aware cloud computing, cooperative cell caching, and mobile traffic offloading, he has authored over 80 technical papers in the IEEE JSAC, the IEEE TWC, the IEEE WIRELESS COMMUNICATIONS, the IEEE COMMUNICATIONS, the IEEE TMM, the IEEE INFOCOM, and the IEEE SECON. He was a recipient of the National Thousand Talents Plan (Youth) of China. He received the "Scholarship for Excellent Foreign Students in IT Field" by NIPA of South Korea from 2008 to 2011, the "Global Outstanding Chinese Ph.D. Student Award" by the Ministry of Education of China in 2012, and the Peiyang Scholarship from Tianjin University. In 2017, he received the "Fred W. Ellersick Prize" from the IEEE Communication Society. He served as a leading GE for several top magazines, transactions, and journals.

**Peng-Jun Wan** (M'10–SM'13–F'16) received the B.S. degree in applied mathematics from Tsinghua University, Beijing, China, in 1990, the M.S. degree in applied mathematics from the Chinese Academy of Sciences, Beijing, in 1993, and the Ph.D. degree in computer and information science from the University of Minnesota, Minneapolis, MN, USA, in 1997.

He is currently a Professor of computer science with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA. His research interests include wireless networks and algorithm design. He served as the Technical Program Chair for the IEEE INFOCOM 2016 and the ACM MOBIHOC 2008 among others, and an Associate Editor for the IEEE/ACM TRANSACTIONS ON NETWORKING and the *Journal of Computer and System Sciences*.

**Zhu Han** (S'01–M'04–SM'09–F'14) received the B.S. degree in electronic engineering from Tsinghua University in 1997 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, MD, USA, in 1999 and 2003, respectively.

From 2000 to 2002, he was a Research and Development Engineer with JDSU, Germantown, MD, USA. From 2003 to 2006, he was a Research Associate with the University of Maryland. From 2006 to 2008, he was an Assistant Professor with Boise State University, Boise, ID, USA. He is currently a John and Rebecca Moores Professor with the Electrical and Computer Engineering Department and the Computer Science Department, University of Houston, Houston, TX, USA. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. He was a recipient of the NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the *Journal on Advances in Signal Processing* in 2015, and the IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016. He received several best paper awards in IEEE conferences. He is the IEEE Communications Society Distinguished Lecturer. He is a 1% highly cited Researcher 2017 according to Web of Science.

**Victor C. M. Leung** (S'75–M'89–SM'97–F'03) received the B.A.Sc. degree (Hons.) and the Ph.D. degree in electrical engineering from The University of British Columbia (UBC) in 1977 and 1982, respectively. He received the APEBC Gold Medal as the head of the 1977 graduating class in the Faculty of Applied Science. He was a recipient of the Canadian Natural Sciences and Engineering Research Council Postgraduate Scholarship.

From 1981 to 1987, he was a Senior Member of Technical Staff and a Satellite System Specialist with MPR Teltech Ltd., Canada. In 1988, he joined the Department of Electronics, Chinese University of Hong Kong, as a Lecturer. In 1989, he joined UBC as a Faculty Member, where he is currently a Professor and TELUS Mobility Research Chair in Advanced Telecommunications Engineering with the Department of Electrical and Computer Engineering. He has co-authored over 1100 journal conference papers and 40 book chapters and co-edited 14 book titles. Several of his papers had been selected for best paper awards. His research interests are in the broad areas of wireless networks and mobile systems.

Dr. Leung has co-authored papers that received the 2017 IEEE ComSoc Fred W. Ellersick Prize, the 2017 IEEE Systems Journal Best Paper Award, and the 2018 IEEE ComSoc CSIM Best Journal Paper Award. He was a recipient of the IEEE Vancouver Section Centennial Award, the 2011 UBC Killam Research Prize, the 2017 Canadian Award for Telecommunications Research, and the 2018 IEEE ComSoc TGCC Distinguished Technical Achievement Recognition Award. He is a Registered Professional Engineer with the Province of British Columbia, Canada. He is a fellow of the Royal Society of Canada, the Engineering Institute of Canada, and the Canadian Academy of Engineering. He was a Distinguished Lecturer of the IEEE Communications Society. He serves on the editorial boards of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, the IEEE TRANSACTIONS ON CLOUD COMPUTING, the IEEE ACCESS, the *Computer Communications*, and several other journals. He served on the editorial boards of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS-Wireless Communications Series and Series on Green Communications and Networking, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE WIRELESS COMMUNICATIONS LETTERS, and the *Journal of Communications and Networks*. He guest-edited many journal special issues and provided leadership to the organizing committees and technical program committees of numerous conferences and workshops.