D2D-assisted caching on truncated Zipf distribution

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 18, 2018, accepted January 15, 2019, date of publication January 24, 2019, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2894837

D2D-Assisted Caching on Truncated Zipf Distribution

QIANG LI[1], (Member, IEEE), YUANMEI ZHANG[1], ASHISH PANDHARIPANDE[2], (Senior Member, IEEE), XIAOHU GE[1], (Senior Member, IEEE), AND JILIANG ZHANG[3], (Member, IEEE)

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Signify, 5656 AE Eindhoven, The Netherlands
[3]Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S1 4DT, U.K.

Corresponding author: Xiaohu Ge (xhge@mail.hust.edu.cn)

This work was supported in part by the National Key R&D Program of China under Grant 2016YFE0133000, and in part by the EU Horizon2020 under Grant EXICITING-723227.

ABSTRACT In this paper, device-to-device (D2D)-assisted caching is considered to offload traffic from the capacity-stringent backhaul networks to the proximity of users. First, a three-layer hierarchical content provision model is established, where a requested content can be fetched from the local cache directly, from the cache of a proximal device through D2D communications, or from the serving base station through backhaul transmissions. Then, for a general multi-unit-cache equipped at each device, we propose independent content placement in each cache unit and correlated content placement in each cache unit without repetition, based on which the problem of maximizing the edge cache hit ratio is formulated. Instead of optimizing the caching probability for all contents in the library, we propose a parameter-based caching framework based on a truncated Zipf distribution, where only the position of truncation and the Zipf exponent are involved. For jointly determining the optimal values of the two parameters, a genetic algorithm and a two-step search algorithm are designed. The simulation results demonstrate that the correlated content placement outperforms its independent counterpart, and significant performance gains can be achieved by the proposed parameter-based caching framework in comparison with most existing approaches.

INDEX TERMS D2D-assisted caching, multi-unit-cache, content placement, Zipf distribution, cache-hit-ratio.

I. INTRODUCTION

A. BACKGROUND

With the rapid proliferation of smart wireless devices and ubiquitous access to data-hungry services, the global mobile data traffic is experiencing an unprecedented increase. This steep growth in internet traffic is expected to continue at an even higher pace [1], mainly contributed by the newly emerging videos and related multimedia services [2]. To cope with the high data rate and low latency performance requirements in these content-centric applications, caching at the wireless edge, e.g., base stations (BSs) and user devices [3]–[6], has been proposed as a key-enabling technique for alleviating the capacity bottleneck of cellular backhaul networks while enhancing the efficiency of content delivery.

Motivated by the fact that today's electronic devices are typically installed with 64GB/128GB memory, it is beneficial and cost-effective to perform caching at the user devices, e.g., smart phones, tablets, vehicles etc., for fast retrieval of previously viewed contents. On the other hand, exploiting multiple communication technologies including LTE, WiFi-Direct, and mmWave [**?**], device-to-device (D2D) communications are available between proximal devices, with or without the involvement of BSs [8]–[11]. Thus with D2D-assisted caching, the requested content, upon a cache hit, can be directly provisioned through D2D communications [12]–[14]. This facilitates native support to content-centric applications so that more traffic can be offloaded from the backhaul networks [15], [16] to the proximity of end users, which significantly reduces the transmission latency [17], [18].

B. RELATED WORKS

Due to the limited storage available at each user device, the problem of which contents should be cached is of critical

concern to tap the potential D2D communication opportunities. In order to achieve intelligent cache content placement, the popularity distribution of contents is assumed to be available in advance in most existing works [2]–[6], [11]–[14], [17]–[28]. While it has been pointed out that it is generally not optimal to cache just the most popular content for maximizing the cache hit probability, it is suggested that the caching distribution should be skewed towards the most popular content, and at the same time exploit the diversity of contents cached among devices [14], [17]. For content caching in random wireless networks with spatially distributed devices, a commonly adopted strategy is the probabilistic content placement, where each device caches contents following a certain probability distribution [20]–[23].

In order to improve the performance of D2D-assisted caching networks, the optimal way to place contents has been studied from different perspectives, by formulating different problems of optimizing the cache hit probability [19], [20], the density of successful receptions [21], the success probability of content delivery [22], the cache-aided throughput [23], the throughput-outage tradeoff [24], and the network average delay. For the spatially distributed devices that request content following a Zipf distribution, the problem of minimizing the average caching failure probability is formulated in [20], towards which a low-complexity dual-solution searching algorithm is proposed to determine the optimal caching probabilities of individual contents. To exploit the spatial diversity of cached contents, a spatially correlated caching strategy is proposed in [19] to guarantee that the D2D nodes caching the same file are never closer to each other than a predefined exclusion radius, which plays the role of a substitute for caching probability. Randomized caching following a Zipf distribution is considered for D2D networks in the presence of interference and noise in [21]. By employing results from stochastic geometry, an optimization problem is formulated to find the caching distribution that maximizes the density of successful receptions. In [22], probabilistic content placement is studied to control the cache-based channel selection diversity and network interference in a stochastic wireless caching helper network. By using stochastic geometry, the optimal caching probabilities are derived to maximize the average success probability of content delivery. A probabilistic caching scheme is proposed for D2D networks in [23] for optimizing the cache-aided throughput, which measures the density of successfully served requests by cache-enabled devices, by accounting for the reliability of D2D transmissions. In [25], a new architecture for D2D caching with inter-cluster cooperation is proposed, where the users cache popular files and share them with other users either in the local cluster via D2D communication or in the remote clusters using cellular transmission. Performance improvements in terms of both the network average delay and throughput per request are demonstrated. Furthermore, in view of the resource scarcity and interference in wireless transmissions, different mechanisms have been proposed in D2D-assisted caching systems for efficient interference control [26], transmission scheduling and resource management [27], [28].

Apart from the information-theoretic aspects, the social relationship between users has also been taken into account for improving the performance of D2D-assisted caching systems [29], [30]. In [29], in view of the inherent social characteristics that play important roles in the interaction among mobile users, the challenges and possible solutions are discussed for social-aware mobile device caching. In view of the additional energy and storage consumption in caching and provisioning content, a social-aware caching game is proposed to incentivize the selfish devices to cache contents and participate in D2D communications in [30].

C. OUR CONTRIBUTIONS

Motivated by the aforementioned studies, in this paper we consider a D2D-assisted caching system with caching performed at the spatially distributed user devices. To achieve efficient traffic offloading, we focus on the cache-hit-ratio and a three-layer hierarchical content search and provision model is first established. In this model, a requested content can be self-served or served by proximal devices through D2D transmissions upon a cache hit, or served by the BS through backhaul transmissions upon a cache miss at the network edge. The main contributions of this paper are summarized as follows:

- We consider a general multi-unit-cache equipped at each device. By contrast to most existing works where a single-unit-cache is considered at each device [11], [12], [21], [31], multiple contents can be stored at each device in the considered D2D-assisted caching system. Although there are several works also considering multi-unit-cache, see [19], [25], how to place contents in different cache units of a device is not straightforward. As such, we propose two fine-grained cache content placement schemes, namely: the independent content placement in each cache unit and the correlated content placement in each cache unit without repetition.
- Based on an arbitrary caching probability distribution, a general problem of maximizing the edge cache-hit-ratio is formulated, in which a special case of self-offloading [11], [12], which is usually overlooked in the literature [19], [31]–[33], is considered for fetching the requested content from the local cache of the requesting device. To solve this optimization problem, we propose a new parameter-based caching framework based on a truncated Zipf distribution. By contrast to most existing works where the caching probability has to be optimized for each content in the library [20]–[23], in the proposed framework only two parameters of the position of truncation and the Zipf exponent need to be optimized.
- To jointly determine the optimal values of the position of truncation and the Zipf exponent, an asymptotically optimal genetic algorithm (GA) [34], [35] and a simple suboptimal two-step search (TSS) algorithm are designed, respectively. Simulation results

demonstrate that the proposed correlated cache content placement, without placing duplicated contents in different cache units of a device, outperforms its independent counterpart. Furthermore, a very close performance can be achieved by the proposed parameter-based caching framework in comparison to the solution where the caching probabilities of all contents are jointly optimized [20].

The remainder of this paper is organized as follows. Section II describes the D2D-assisted caching system model. Then based on multi-unit-cache, two content placement schemes are proposed and the problem of maximizing the edge cache-hit-ratio is formulated in Section III. A parameter-based caching framework based on a truncated Zipf distribution is proposed in Section IV, where two algorithms are designed for searching the optimal parameters. Simulation results are presented in Section V. Finally, Section VI concludes this paper.

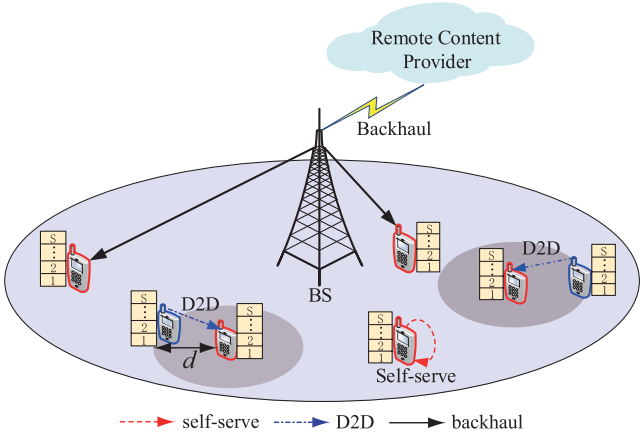


FIGURE 1. An illustration of the D2D-assisted caching system with multi-unit-cache, where a requested content can be possibly served by the requesting device itself, by the proximal devices through D2D transmissions, or by the BS through backhaul transmissions.

II. SYSTEM MODEL

As illustrated in Fig. 1, we consider a D2D-assisted caching system where the locations of user devices are modeled by a homogeneous Poisson point process (PPP) with density λ [20]–[23]. Assuming that the BS provides a common communication channel for control message exchange for content search and D2D pairing [36], D2D communications can be potentially initiated when a proximal device is within a fixed radius d of the requesting device. It is assumed that each device is equipped with a finite storage for caching contents. Then upon a cache hit at the D2D device, the backhaul traffic can be effectively offloaded to the proximity of user devices. To focus on the content placement phase for D2D-assisted caching system, D2D overlaid cellular network is considered where there is no cross-tier interference between D2D and cellular communications [20], [23], and the path loss, fading or interference in the transmission phase are omitted [19], [20].

A. REQUEST GENERATION MODEL

For ease of exposition, we define a library M consisting of M distinct content files, from which the distributed devices make requests. As adopted in most existing works [20]–[23], it is assumed that the requests of contents follow a Zipf distribution. The more frequently a content is requested, the more popular this content is. Arranging the contents in the library in a popularity-descending order, the popularity of the m-th ranked content can thus be expressed as

$$P_r(m) = \frac{m^{-\alpha}}{\sum_{i=1}^{M} i^{-\alpha}}, \text{ where } m \in \{1, 2, \cdots, M\}, \qquad (1)$$

which characterizes the frequency that the m-th ranked content is requested by devices. In other words, for an arbitrary request generated, $P_r(m)$ denotes the probability that the request is for the m-th ranked content. Here $0 < \alpha < 1$ denotes the Zipf exponent that describes the skewness in the request pattern. To be specific, a higher α means that the content requests are more concentrated on the high ranked contents.

To maintain cache consistency, the popularity distribution of the content files (1) is assumed to be relatively time-nonvarying within a certain duration [4]–[6], [11]–[14], [20]–[23]. This assumption is valid in examples including popular news containing short videos that are updated every 2-3 hours, new movies that are posted every week, new music videos that are posted every month. Then the popularity distribution of contents can be learnt periodically or whenever necessary and known to the system by using learning techniques or big data analytics [37], [38].

B. CONTENT SEARCH AND PROVISION MODEL

We define cache hit as the event that the content requested by a user happens to be found in the local cache or in the cache of a proximal device within the effective D2D communication range. As illustrated in Fig. 1, depending on at which place the requested content is hit, a three-layer hierarchical content search and provision model is established, as detailed in the following:

1) Once a request is generated, the requesting device first searches whether the requested content is stored in its local cache. Upon a local cache hit, the requesting device can be self-served by directly fetching content from its cache.
2) Otherwise if the search in the local cache fails, the requesting device enquires if the requested content is cached by proximal devices within the effective D2D range. Upon a cache hit in a proximal device, the content is fetched through D2D communications directly. If there are multiple D2D devices having the requested content, the content is delivered from the nearest one [23]. We define a probability P_s indicating the willingness of a D2D device to share cached content with the requesting device [29], which is irrelevant to the cached contents and depends on factors such as battery status, social relationship, and security.

3) Upon a cache miss that the requested content fails to be found throughout the proximal D2D devices, the BS directs the request up to the cloud to fetch the corresponding content from the remote content provider, which requires backhaul transmissions and incurs additional latency.

For ease of reference, a list of abbreviations and symbols that appear in this paper is presented in Table 1.

TABLE 1. A list of abbreviations and symbols.

D2D	device-to-device
BS	base station
GA	genetic algorithm
TSS	two-step search
PPP	Poisson point process
UC	uniform caching
MPC	maximal-popularity caching
RC	random caching
λ	density of user devices
d	radius of effective D2D communication range
$\mathbf{M}$	library consisting of M distinct content files
$P_r(m)$	requested probability of the m-th ranked content
α	Zipf exponent of the requested content distribution
P_s	willingness probability of D2D sharing
S	number of cache units equipped at each device, $S \geq 1$
$P_c(m)$	caching probability of the m-th ranked content
$H_L(m)$	local cache-hit-ratio upon requesting m-th ranked content
I_s	index of the stored content in the s-th cache unit
H_{local}	average local cache-hit-ratio
H_{device}	average cache-hit-ratio at D2D devices
H_{edge}	average cache-hit-ratio at the network edge
H_{miss}	average cache-miss-ratio
T	position of truncation, $1 \leq T \leq M$
β	Zipf exponent of the cached content distribution, $\beta > 0$

III. CACHE CONTENT PLACEMENT WITH MULTI-UNIT-CACHE

To focus on D2D-assisted caching designs, it is assumed that the content files are of the same unit size [4]–[6], [20]–[23] and each device, equipped with a cache consisting of S cache units where $S \geq 1$, is able to store a content in each cache unit [20], [23].

For ease of exposition, given a cache unit, we define $P_c(m)$ as the probability of caching the m-th ranked content, where $0 \leq P_c(m) \leq 1$ and $\sum_{m=1}^{M} P_c(m) = 1$. Then, for a device equipped with single-unit-cache where only a single content can be stored [11], [12], [21], [31], i.e., $S = 1$, $P_c(m)$ denotes the local cache-hit-ratio upon a request for the m-th ranked content. In other words, conditioned on a request for the m-th ranked content, the probability of finding the m-th ranked content in the cache of the requesting device locally is given by

$$H_L(m) = P_c(m). \tag{2}$$

However, for a device equipped with multi-unit-cache where multiple contents can be stored, i.e., $S > 1$, the corresponding local cache-hit-ratio upon a request for the m-th ranked content is not straightforward. To provide insights on the caching operation within each cache unit, next we propose independent content placement in each cache unit and correlated content placement in each cache unit without repetition.

A. INDEPENDENT CONTENT PLACEMENT IN EACH CACHE UNIT

For independent content placement, the decision on what content should be stored in each cache unit is made independently from one another. Then upon a request for the m-th ranked content, the local cache-hit-ratio at the requesting device under independent content placement in each cache unit is expressed as

$$H_L(m) = 1 - [1 - P_c(m)]^S, \tag{3}$$

which denotes the probability that there is at least one copy of the m-th ranked content stored in the S cache units of the requesting device. Based on Bernoulli's inequality, $[1 - P_c(m)]^S \geq 1 - S \cdot P_c(m)$. Thus we have

$$\sum_{m=1}^{M} H_L(m) \leq S. \tag{4}$$

B. CORRELATED CONTENT PLACEMENT IN EACH CACHE UNIT WITHOUT REPETITION

It is apparent that for independent cache content placement, there exists a non-zero probability of storing multiple copies of the same content in the cache of a device. To avoid this storage waste, next we propose a correlated content placement where distinct content is stored in each of the S cache units without repetition. To be specific, once a content is stored in a cache unit of a device for the first time, this content will be excluded from being cached in the rest of the cache units. For ease of exposition, we define I_s as the index of the stored content in the s-th cache unit where $I_s \in M = \{1, 2, \cdots, M\}$ and $s \in \{1, \cdots, S\}$. Following an increasing order of s to store content in each cache unit, the probability that the m-th ranked content is exclusively stored in the s-th cache unit, where $s \in \{1, \cdots, S\}$, is expressed as

$$\Pr\{I_1 = m\} = P_c(m), \tag{5a}$$

$$\begin{aligned}\Pr\{I_2 = m\} &= \sum_{I_1 \in M \setminus m} P_c(I_1) \cdot \frac{P_c(m)}{1 - P_c(I_1)} \\ &= \sum_{I_1 \in M \setminus m} \frac{P_c(I_1) \cdot \Pr\{I_1 = m\}}{1 - P_c(I_1)},\end{aligned} \tag{5b}$$

$$\begin{aligned}\Pr\{I_3 = m\} &= \sum_{I_1 \in M \setminus m} P_c(I_1) \cdot \sum_{I_2 \in M \setminus \{m, I_1\}} \frac{P_c(I_2)}{1 - P_c(I_1)} \\ &\quad \cdot \frac{P_c(m)}{1 - P_c(I_1) - P_c(I_2)} \\ &= \sum_{I_2 \in M \setminus \{m, I_1\}} \frac{P_c(I_2) \cdot \Pr\{I_2 = m\}}{1 - P_c(I_1) - P_c(I_2)},\end{aligned} \tag{5c}$$

$$\vdots$$

$$\begin{aligned}\Pr\{I_S = m\} = &\sum_{I_{S-1} \in M \setminus \{m, I_1, \cdots, I_{S-2}\}} \frac{P_c(I_{S-1})}{1 - \sum_{s=1}^{S-1} P_c(I_s)} \\ &\cdot \Pr\{I_{S-1} = m\}.\end{aligned} \tag{5d}$$

From (5b)-(5d), for the already cached contents up to the $(s-1)$-th cache unit, where $s \in \{2, \cdots, S\}$, they will be excluded from the library $\mathbf{M}$ and only the remaining contents are qualified to be cached in the s-th cache unit, following a truncated distribution of $\frac{P_c(I_s)}{1-\sum_{i=1}^{s-1} P_c(I_i)}$ where $I_s \in \{M \setminus \{I_1, \cdots, I_{s-1}\}\}\ \forall\, s \in \{2, \cdots, S\}$.

Then upon a request for the m-th ranked content, the local cache-hit-ratio at the requesting device under the correlated content placement in each cache unit is expressed as

$$H_L(m) = \sum_{s=1}^{S} \Pr\{I_s = m\}, \tag{6}$$

for which we have from (5a)-(5d)

$$\begin{aligned} \sum_{m=1}^{M} H_L(m) &= \sum_{m=1}^{M}\sum_{s=1}^{S} \Pr\{I_s = m\} \\ &= \sum_{s=1}^{S}\sum_{m=1}^{M} \Pr\{I_s = m\} = S. \end{aligned} \tag{7}$$

C. EDGE CACHE-HIT-RATIO AND PROBLEM FORMULATION

As one of the key performance metrics of cache-enabled systems, next we theoretically analyze the edge cache-hit-ratio [3]–[5] achieved by the considered D2D-assisted caching system. Generally, a higher edge cache-hit-ratio translates into more traffic offloaded, which effectively alleviates the capacity bottleneck of backhaul networks while reducing the network energy consumption [32], [39] and transmission latency experienced by users [2], [4], [11], [17], [18], [25]. However, the quantitative modeling of the network energy consumption or transmission latency is non-trivial, which requires more involved parameter and definitions and is beyond the scope of this paper.

Based on the three-layer hierarchical model established in Section II, for an arbitrary request that is generated following the Zipf distribution (1), the average local cache-hit-ratio that the requested content is found in the local cache can be expressed as

$$\begin{aligned} H_{\text{local}} &= \sum_{m=1}^{M} P_r(m) H_L(m) \\ &= \sum_{m=1}^{M} \frac{m^{-\alpha} H_L(m)}{\sum_{i=1}^{M} i^{-\alpha}}. \end{aligned} \tag{8}$$

If however the search in the local cache fails, the requesting device enquires if the desired content is cached by proximal devices in its effective D2D range. Since λ denotes the density of devices, as a result of independent thinning, the distribution of the devices that happen to cache the m-th ranked content and at the same time willing to share it with other devices follows a homogeneous PPP with density $\lambda' = \lambda P_s H_L(m)$. Thus the probability that the requesting device can be potentially served by n D2D devices is expressed as

$$h(n; d) = \frac{(\pi d^2 \lambda')^n}{n!} e^{-\pi d^2 \lambda'}, \text{ where } n \in \{0, 1, \cdots\}. \tag{9}$$

Thus for an arbitrary request that is generated following the Zipf distribution (1), the average cache-hit-ratio at a proximal D2D device can be expressed as

$$\begin{aligned} H_{\text{device}} &= \sum_{m=1}^{M} P_r(m)\,[1 - H_L(m)]\,[1 - h(0; d)] \\ &= \sum_{m=1}^{M} \frac{m^{-\alpha}\,[1 - H_L(m)]}{\sum_{i=1}^{M} i^{-\alpha}} \\ &\quad \cdot \left[1 - e^{-\pi d^2 \lambda P_s H_L(m)}\right]. \end{aligned} \tag{10}$$

We define $H_{\text{edge}} = H_{\text{local}} + H_{\text{device}}$ as the cache-hit-ratio at the network edge. Together with (8) and (10), H_{edge} can be expressed as

$$\begin{aligned} H_{\text{edge}} &= \sum_{m=1}^{M} P_r(m)\,\{1 - [1 - H_L(m)]\,h(0; d)\} \\ &= 1 - \sum_{m=1}^{M} \frac{m^{-\alpha}\,[1 - H_L(m)]\,e^{-\pi d^2 \lambda P_s H_L(m)}}{\sum_{i=1}^{M} i^{-\alpha}}. \end{aligned} \tag{11}$$

If however the requested content fails to be found throughout the network edge, which occurs with a probability

$$H_{\text{miss}} = 1 - H_{\text{edge}}, \tag{12}$$

then the BS directs the request up to the cloud to fetch the corresponding content. This relies on backhaul transmissions that will incur additional overhead, e.g., energy and latency.

Based on the above analysis, to achieve efficient caching at the distributed devices while exploiting the potential D2D opportunities as much as possible, the problem of maximizing the cache-hit-ratio at the network edge can be expressed as

$$\begin{aligned} &\max_{\{P_c(m)|m\in\{1,\cdots,M\}\}} H_{\text{edge}} \\ &s.t. \sum_{m=1}^{M} P_c(m) = 1, \quad 0 \le P_c(m) \le 1. \end{aligned} \tag{13}$$

From (11), since the second order derivative of H_{edge} with respect to $H_L(m)$ is strictly negative, the problem in (13) corresponds to a non-linear convex optimization problem [20], which can be solved by conventional iterative approaches, e.g., the subgradient method [40]. Thus the solution to (13) corresponds to a joint optimization of all caching probabilities $P_c(m),\ m \in \{1, \cdots, M\}$ [20]–[23], for which the resultant optimal H_{edge} will be demonstrated later in Section V.

Remark 1: Since there are usually a very large number of contents in the library, it would be of overwhelming complexity to analytically calculate the optimal caching probability for each and every content in the library, i.e., $P_c(m)$ for $m \in \{1, \cdots, M\}$, in practical systems. In addition, for optimizing the M caching probabilities jointly, the solution to (13) corresponds to a *non-parametric* caching strategy, from

which limited insight can be obtained. To shed light on the design of efficient D2D-assisted caching so as to effectively enhance the edge cache-hit-ratio, we propose a parameter-based caching framework based on a truncated Zipf distribution, where only two parameters-the position of truncation and the Zipf exponent, need to be optimized.

IV. PROPOSED CACHING FRAMEWORK ON TRUNCATED ZIPF DISTRIBUTION

Due to limited cache storage, only a small fraction of contents in the library can be accommodated at the user devices. To exclude those hardly requested contents from being cached, we define an integer T, for $S \leq T \leq M$, as the position of truncation. That is, only a set of the T most popular contents in the library are qualified to be cached by edge devices. With this truncation, the caching probability of the m-th ranked content in the library can thus be expressed as

$$P_c(m) = \begin{cases} \dfrac{m^{-\beta}}{\sum_{i=1}^{T} i^{-\beta}}, & m \in \{1, \cdots, T\} \\ 0, & m \in \{T+1, \cdots, M\} \end{cases} \tag{14}$$

where $\beta \geq 0$ denotes the Zipf exponent that characterizes the skewness pattern in the cached content distribution among edge devices.

Remark 2: It is worth pointing out that the advantages of the proposed caching framework are two-fold with a joint design on T and β. One, the contents to be cached can be assigned with different caching priorities, which guarantees the availability of the most popular contents while maintaining a reasonable diversity of the contents cached by the edge devices. Two, unpopular contents that are hardly requested can be excluded from being cached, which results in an efficient utilization of the limited cache storage.

From (14), it is observed that when $T = S = 1$, the proposed caching framework corresponds to the most-popular caching (MPC) scheme [20], [33] that caches the single most popular content at each device, where $P_c(1) = 1$. When $T = M$, the proposed caching framework corresponds to the random caching (RC) scheme [21] where the Zipf exponent β needs to be optimized. On the other hand, when $\beta = 0$, the proposed caching framework boils down to the uniform caching (UC) scheme [20] where the qualified contents are uniformly cached by devices with the same probability.

Furthermore, for UC where $\beta = 0$, we have from (3) and (6) $H_L(m) = 1 - (1 - \frac{1}{T})^S$ under independent content placement and $H_L(m) = \frac{S}{T}$ under correlated content placement, respectively. By utilizing Bernoulli's inequality, since $\left(1 - \frac{1}{T}\right)^S \geq 1 - \frac{S}{T}$, then we have $\frac{S}{T} \geq 1 - (1 - \frac{1}{T})^S$. Thus conditioned on a request for an arbitrary content, the correlated content placement always brings a higher local cache-hit-ratio than that of the independent content placement.

A. OPTIMIZATION OF EDGE CACHE-HIT-RATIO

Thus for maximizing the edge cache-hit-ratio achieved within the proposed caching framework, substituting (14) into (11), the optimization problem (13) can be transformed into

$$\max_{T,\beta} \frac{\sum_{m=1}^{T} m^{-\alpha} \left\{1 - [1 - H_L(m)] e^{-\pi d^2 \lambda P_s H_L(m)}\right\}}{\sum_{i=1}^{M} i^{-\alpha}}, \quad s.t. \begin{cases} S \leq T \leq M, \\ \beta \geq 0, \end{cases} \tag{15}$$

where $H_L(m)$ can be obtained by substituting $P_c(m)$ in (14) into (2), (3) and (6). A close look at (15) indicates that the parameters T and β are intertwined in many terms in H_{edge}, which makes it intractable to analytically obtain the jointly optimal position of truncation T^* and Zipf exponent β^*. However some useful insights can be drawn.

From (15), for the extreme scenario where either of d and λ approaches infinity, there is an infinite number of devices within the effective D2D communication range, which forms a virtually aggregated cache of infinite space by D2D sharing. In other words, the probability that an arbitrary requested content is hit at the network edge approaches 1, i.e.,

$$\lim_{d \text{ or } \lambda \to \infty} H_{\text{edge}} = \sum_{m=1}^{T} \frac{m^{-\alpha}}{\sum_{i=1}^{M} i^{-\alpha}}. \tag{16}$$

Thus there is no need for truncation, and we have the optimal $T^* = M$ and β becomes irrelevant.

On the other hand, for the extreme scenario where either of d and λ approaches zero, no D2D sharing of the cached contents is permitted between proximal devices. Thus the considered D2D-assisted caching system degrades to a distributed caching system without cooperation, in which either a requested content is self-served, or it has to be fetched by the BS through backhaul transmissions. Then from (8), the optimization problem in (15) becomes

$$\max_{T,\beta} \sum_{m=1}^{T} \frac{m^{-\alpha} H_L(m)}{\sum_{i=1}^{M} i^{-\alpha}}, \quad s.t. \begin{cases} S \leq T \leq M, \\ \beta \geq 0, \end{cases} \tag{17}$$

where $H_L(m)$ can be obtained by substituting $P_c(m)$ in (14) into (2), (3) and (6).

Owing to the complex form of H_{edge} where T and β are intertwined, it is intractable to analytically figure out whether the problem in (15) is a convex optimization problem or not. Thus we present an asymptotically optimal genetic algorithm (GA) and a simple suboptimal two-step search (TSS) algorithm respectively for jointly determining the suitable values of T and β.

B. ALGORITHM DESIGNS

1) A GENETIC ALGORITHM (GA)

Inspired by biological evolution, GA is a useful method to solve complex multi-variable non-linear problems to get the globally optimal solution asymptotically [34], [35].

For ease of description, we define N as the total number of population. Q_j denotes the j-th generation of population, in which $Q_j(i)$ denotes the i-th individual that corresponds to a point in the plane of $\{T, \beta\}$. To search the joint optimal

$\{T^*, \beta^*\}$ iteratively, T_0 and T_e are set as the minimum and the maximum of the position of truncation T, with step size (or precision) T_s. On the other hand, β_0 and β_e are set as the minimum and the maximum of the Zipf exponent β, with precision β_s. To characterize the operations on genes involved in evolution, every individual point on the plane of $\{T, \beta\}$ has two forms of expressions, i.e., real number and binary number, where the conversion from real (or binary) number to binary (or real) number is called encoding (or decoding) of genes. N_e is the number of chosen elite in each generation. $P_{\text{crossover}}$ and P_{mutation} are defined as the probabilities of crossover and mutation, respectively. X denotes the condition for termination, i.e., the iteration terminates when the weighted average relative change in the best fitness function value is less than X. Then the proposed GA is presented in Algorithm 1.

Algorithm 1 Genetic Algorithm (GA)

Require: $N, T_0, T_e, \beta_0, \beta_e, N_e, P_{\text{crossover}}, P_{\text{mutation}}, X$

$j \leftarrow 1$
Step 1: The first generation Q_j is initialized
while X is not satisfied **do**
  Step 2: Fitness of N individuals from Q_j is calculated, based on which elitist strategy is performed.
  Step 3: Genetic operations are conducted to breed N offsprings from Q_j; $j = j + 1$, a new generation of Q_j is bred
end while
Step 4: Termination
return $\{T^*, \beta^*\}$

1) Firstly, N points are randomly selected from the plane of $\{T, \beta\}$ to form the first generation Q_1.
2) Then the fitness function $f(i)$, in terms of the edge cache-hit-ratio H_{edge} as defined in (11), is calculated for each individual $Q_1(i)$ in Q_1, where $i \in \{1, \cdots, N\}$. Elitist strategy is performed where N_e individuals with the highest fitness in Q_1 are directly selected to be parts of the next generation Q_2.
3) A series of genetic operations are conducted to breed the rest of the next generation Q_2.
   a) Crossover: Setting $P(i) = \frac{f(i)}{\sum_{i=1}^{N} f(i)}$ as the probability for selecting point $Q_1(i)$ as a parent, $2(N - N_e)$ parents are selected from the current population Q_1 by using a roulette wheel selection [34], [35]; the selected $2(N - N_e)$ parents are then divided into $(N - N_e)$ pairs randomly, between which scattered crossover on the genes of parent pairs takes place in turn with a probability of $P_{\text{crossover}}$ [34], [35], resulting in the rest $(N - N_e)$ offsprings.
   b) Mutation: With a probability of P_{mutation}, bit-wise mutations take places randomly for each of the N offsprings.
4) The generation Q_j keeps evolving iteratively, until the termination condition X is satisfied. Then the optimal $\{T^*, \beta^*\}$ is returned.

TABLE 2. A list of parameters adopted in the proposed GA and TSS.

Parameters	Definitions	Values
N	population size	100
T_0	the minimum value of T	1
T_e	the maximum value of T	M
T_s	the step size (or precision) of T	1
β_0	the minimum value of β	0
β_e	the maximum value of β	3
β_s	the step size (or precision) of β	0.0001
N_e	the number of chosen elite	10
$P_{\text{crossover}}$	probability of crossover	0.8
P_{mutation}	probability of mutation	0.08
X	termination condition	10^{-6}

2) A TWO-STEP SEARCH (TSS) ALGORITHM

Although asymptotically optimal solutions $\{T^*, \beta^*\}$ can be reached by GA, it comes at a relatively high complexity depending on what values are adopted for the multiple parameters as listed in Table 2. This makes it intractable to analytically characterize the complexity of GA. For fast convergence of the algorithm, we also propose a simple two-step search (TSS) algorithm. By letting $\beta = \alpha$, TSS firstly searches the optimal position of truncation within $[T_0, T_e]$ with precision T_s, referred to as T'. Then with the obtained T', TSS proceeds to search the optimal Zipf exponent β' within $[\beta_0, \beta_e]$ with precision β_s. The obtained $\{T', \beta'\}$ thus constitutes a sub-optimal solution to the problem of (15).

For characterizing the complexity of the two algorithms above, we define $N_1 = \frac{T_e - T_0}{T_s}$ and $N_2 = \frac{\beta_e - \beta_0}{\beta_s}$ as the searching space for T and β, respectively. Thus the corresponding complexity of the TSS algorithm can be readily obtained as $O(N_1 + N_2)$. On the other hand, for the GA, although the corresponding complexity is related to the iteration number and population size, the exact expression of complexity is theoretically intractable as it depends on the specific values of parameters adopted [35], as listed in Table 2.

V. SIMULATION RESULTS

In this section, we present simulation results to demonstrate the edge cache-hit-ratio H_{edge} of the considered D2D-assisted caching system under different caching strategies. Although in reality the library size M is usually very large that contains a gigantic number of contents, here we take $M = 50$ content files in the library for ease of illustration. Similar choices can also be found, e.g., in [22] and [23]. Unless otherwise specified, we let the Zipf exponent of the content popularity distribution $\alpha = 0.7$, the density of the spatially distributed user devices $\lambda = \frac{0.02}{\pi}$ per square meter, the effective D2D range of radius $d = 15$m, and the willingness probability $P_s = 0.8$. For ease of reference, the parameters involved in the GA and TSS algorithms and the values adopted are listed in Table 2.

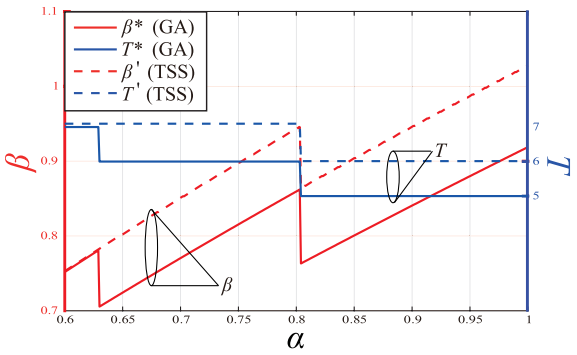


FIGURE 2. The solution $\{T^*, \beta^*\}$ and $\{T', \beta'\}$ with respect to α.

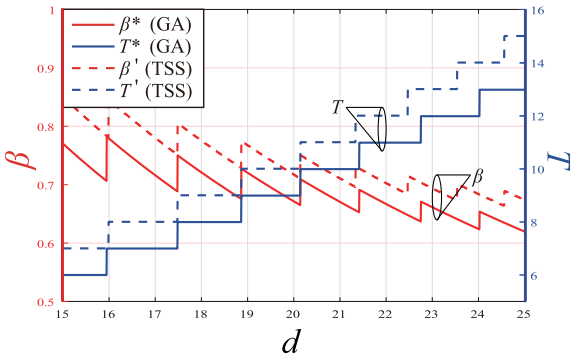


FIGURE 3. The solution $\{T^*, \beta^*\}$ and $\{T', \beta'\}$ with respect to d.

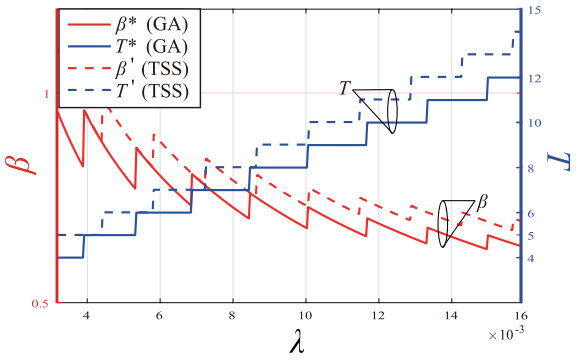


FIGURE 4. The solution $\{T^*, \beta^*\}$ and $\{T', \beta'\}$ with respect to λ.

In Fig. 2-Fig. 4, we demonstrate the obtained $\{T, \beta\}$ under GA and TSS when $S = 1$, where the impact of different system parameters is illustrated. Overall, it is observed that the obtained $\{T, \beta\}$ under GA and TSS follow the same trend, although there exists a gap between them. Fig. 2 demonstrates $\{T, \beta\}$ with respect to varying values of α. It is observed that with an increase of α, the obtained β and T behave in an opposite manner in general. This is reasonable as a greater α means that the requests are more concentrated on the high-ranked contents, thus a smaller set of T contents should be truncated with a higher β to adapt to the skewness in the request pattern. On the other hand, it is worth noting that although β exhibits an overall upward trend with an increase in α, it is not the case at the instants when T is decreased by a unit value. This is reasonable as each time T is decreased by a unit value, with α being unchanged, the diversity of the cached contents needs to be maintained by a more evenly distribution of the cached contents among edge devices, i.e., a smaller β.

Fig. 3 demonstrates $\{T, \beta\}$ with respect to varying values of d. Similarly, it is observed that T and β behave oppositely with an increase of d. This is because with increasing d, a larger virtual cache can be formed and shared by the proximal devices through D2D communications, which has the potential to accommodate more distinct contents at the network edge, thus resulting in a greater T. On the other hand, with increasing d, a lower β is required to match the greater T, which enhances the richness of the cached contents such that D2D sharing of the cached contents can be facilitated between proximal devices. Similarly, although β exhibits an overall downward trend with an increase in d, each time T is increased by a unit value, with α being unchanged, a larger β is needed for maintaining the skewness in the cache content distribution.

Fig. 4 demonstrates $\{T, \beta\}$ with respect to varying values of λ. With an increase of λ, since more devices are located within the effective D2D range, similar phenomena can be observed as in Fig. 3.

Next, by adopting the optimal $\{T^*, \beta^*\}$ searched by GA, the edge cache-hit-ratio H_{edge} is demonstrated with respect to varying values of S in Fig. 5, under independent and correlated content placement respectively. It is observed that with an increase of S, since more storage is available for caching contents, overall a higher H_{edge} is achieved. In addition, it is observed that the correlated content placement, which avoids caching the same content for multiple times at a device, always achieves a better performance than that of the independent content placement, and this performance gap becomes larger with the increase of S. Fig. 5(a) demonstrates H_{edge} under different values of α, in which a greater α means that the requests are more concentrated on the high-ranked contents, thus resulting in a higher H_{edge}. Similarly, with an increase of d and λ, since more devices are located within the effective D2D range, cooperative sharing of the cached contents through D2D communications can be facilitated, which results in a higher H_{edge}, as illustrated in Fig. 5(b) and Fig. 5(c), respectively.

Next, we adopt the correlated content placement in the proposed caching framework and compare its performance with the existing caching strategies, as listed in the following:

- UC: In uniform caching (UC) strategy, with $\beta = 0$ and $T = M$, each content in the library will be cached with equal caching probability [20].
- MPC: In maximal-popularity caching (MPC) strategy, each device caches the most popular content until the storage capacity is reached [33].
- Truncated-UC: In truncated-UC strategy, with $\beta = 0$, the optimal position of truncation T is determined for optimizing H_{edge}.
- RC: In random caching (RC) strategy, with $T = M$, the optimal Zipf exponent β is determined for optimizing H_{edge} [21].

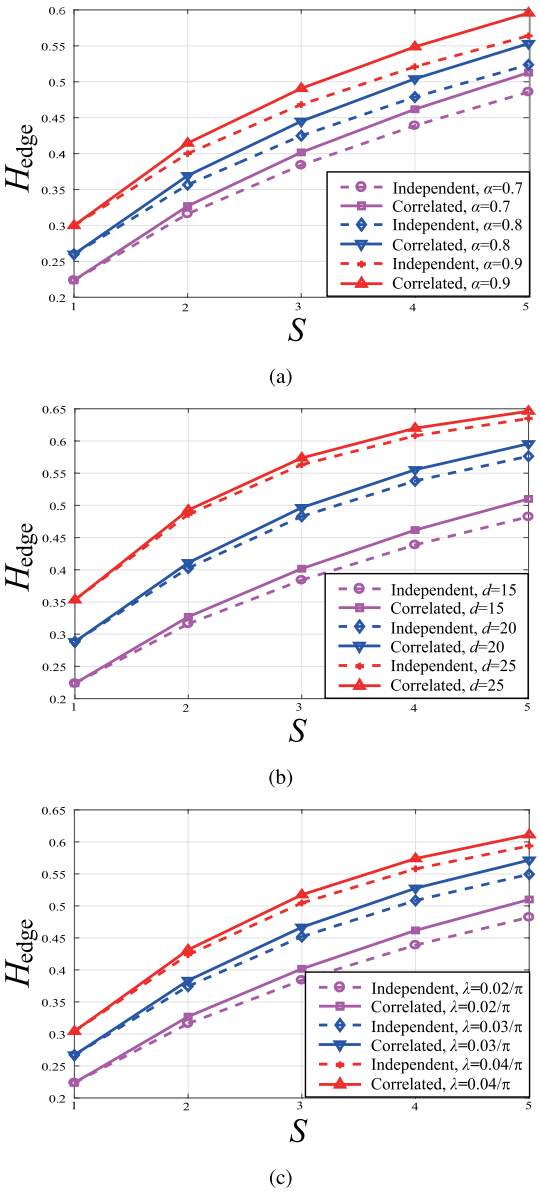


FIGURE 5. Edge cache-hit-ratio H_{edge} with respect to S. (a) H_{edge} under different values of α. (b) H_{edge} under different values of d. (c) H_{edge} under different values of λ.

- Subgradient method: The caching probabilities for all contents in the library are jointly optimized by using the subgradient method [40] iteratively.

In Fig. 6-Fig. 8, the edge cache-hit-ratio H_{edge} is demonstrated for UC, MPC, Truncated-UC, RC, subgradient method and the proposed caching framework when $S = 3$, where the impact of system parameters α, d, λ is evaluated. From Fig. 6, it is observed that the performance of all strategies except UC is improved with an increase of α. This is reasonable as a greater α means that the requests are more concentrated on the high-ranked contents, thus H_{edge} is increased with other parameters being the same. Whereas for UC where the contents are uniformly cached by user devices, H_{edge} is irrelevant to the value of α. On the other hand, with a joint optimization of T and β, it is observed

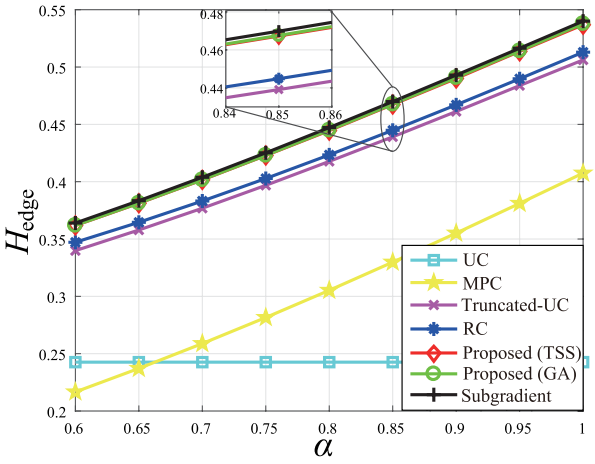


FIGURE 6. H_{edge} with respect to α under different caching strategies.

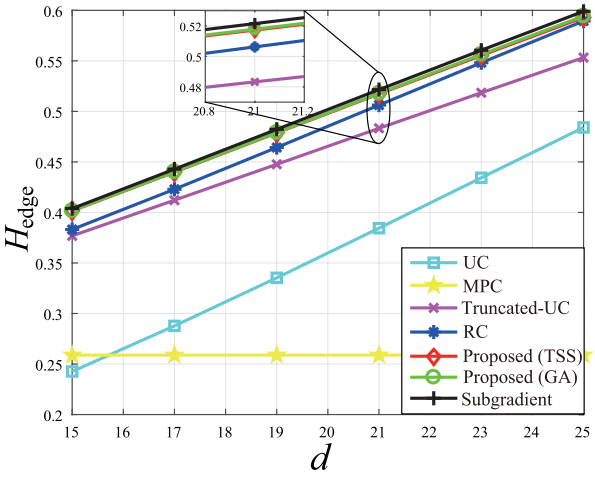


FIGURE 7. H_{edge} with respect to d under different caching strategies.

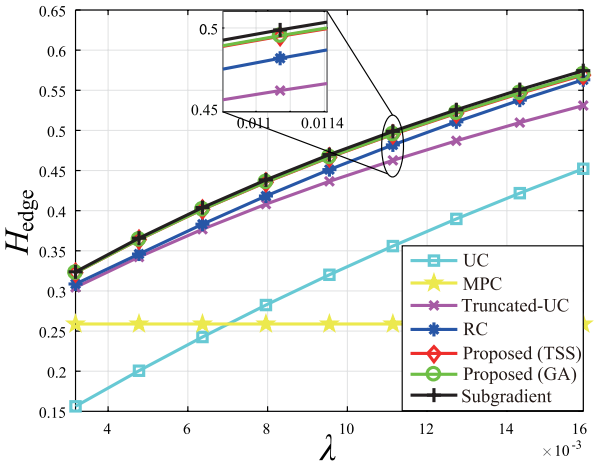


FIGURE 8. H_{edge} with respect to λ under different caching strategies.

that a very close performance is achieved by the proposed caching framework to the solution to (13), where all caching probabilities $P_c(m)$ are jointly optimized by using the subgradient method. Furthermore, while a very close

performance to GA is achieved by TSS but with a relatively low complexity, both of them outperform the strategies of UC, MPC, Truncated-UC, and RC.

Similar phenomena can be observed in Fig. 7 (and Fig. 8) where H_{edge} is demonstrated with respect to d (and λ), except that the performance of UC is improved whereas the performance of MPC becomes irrelevant to the increasing d and λ. This is reasonable as when more devices are located within the effective D2D communication range with the increase of d (or λ), the richness of the cached contents is enhanced by UC, which facilitates cooperative sharing of the cached contents through D2D communications, thus enhancing the edge cache-hit-ratio. By contrast, since the S most popular contents are cached across all devices in MPC, H_{edge} remains constant with an increase in d (or λ).

VI. CONCLUSIONS AND DISCUSSIONS

In this paper we considered a D2D-assisted caching system consisting of spatially distributed cache-enabled devices. For increasing the traffic offloaded to the edge of the network as much as possible, a hierarchical three-layer content provision model was established, in which a special case of self-offloading is considered by fetching the requested content from the local cache of the requesting device. To characterize the caching operations at devices equipped with multi-unit-cache, we proposed independent content placement and correlated content placement without repetition respectively, based on which a general problem of optimizing the edge cache-hit-ratio is formulated. To solve this problem while providing insights on efficient caching designs, we proposed a novel parameter-based caching framework based on the truncated Zipf distribution. Rather than attempting to optimize the caching probability for each content in the library, only the position of truncation and Zipf exponent need to be optimized. Simulation results demonstrate that the correlated content placement outperforms the independent content placement. Furthermore, while the proposed caching framework outperforms most existing D2D-assisted caching strategies, it also achieves a very close performance to the case where the caching probabilities for all contents are jointly optimized.

In this paper, a static network scenario is considered where each user requests content following the same popularity distribution known *a priori*. In the practical implementation, however, users are usually of different preferences and issues like varying densities of user devices may arise due to mobility, which results in varying content popularity distributions with time and space. In addition, on the basis of cache-hit-ratio, the savings in network energy consumption and transmission latency are worth further investigation, which require more involved parameters and sophisticated modeling. These issues will be delegated to future work.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021," Cisco, San Jose, CA, USA, White Paper 1454457600805266, Mar. 2017.

[2] X. L. Ge Pan, Q. Li, G. Mao, and S. Tu, "Multipath cooperative communications networks for augmented and virtual reality transmission," *IEEE Trans. Multimedia*, vol. 19, no. 10, pp. 2345–2358, Oct. 2017.

[3] 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.

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

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

[6] Q. Li, W. Shi, X. Ge, and Z. Niu, "Cooperative edge caching in software-defined hyper-cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2596–2605, Nov. 2017.

[7] A. Gupta and E. R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, Jul. 2015.

[8] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 86–92, May 2014.

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

[10] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmi, "Device-to-device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510–1520, Dec. 2014.

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

[12] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

[13] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[14] 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. 176–189, Jan. 2016.

[15] X. Ge, S. Tu, T. Han, and Q. Li, "Energy efficiency of small cell backhaul networks based on Gauss-Markov mobile models," *IET Netw.*, vol. 4, no. 2, pp. 158–167, 2015.

[16] Y. Zhong, X. Ge, H. H. Yang, T. Han, and Q. Li, "Traffic matching in 5G ultra-dense networks," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 100–105, Aug. 2018.

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

[18] 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, 2017.

[19] D. Malak, M. Al-Shalash, and J. G. Andrews, "Spatially correlated content caching for device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 56–70, Jan. 2018.

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

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

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

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

[24] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, Oct. 2015.

[25] R. Amer, M. M. Butt, M. Bennis, and N. Marchetti, "Inter-cluster cooperation for wireless D2D caching networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6108–6121, Sep. 2018.

[26] B. Chen, C. Yang, and G. Wang, "High throughput opportunistic cooperative device-to-device communications with caching," *IEEE Trans. Veh. Tech.*, vol. 66, no. 8, pp. 7527–7539, Aug. 2017.
[27] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.
[28] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1155–1158, May 2017.
[29] Y. Wu *et al.*, "Challenges of mobile social device caching," *IEEE Access*, vol. 4, pp. 8938–8947, 2016.
[30] K. Zhu, W. Zhi, L. Zhang, X. Chen, and X. Fu, "Social-aware incentivized caching for D2D communications," *IEEE Access*, vol. 4, pp. 7585–7593, 2016.
[31] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2014, pp. 863–868.
[32] 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.
[33] 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.
[34] E. S. H. Hou, N. Ansari, and H. Ren, "A genetic algorithm for multiprocessor scheduling," *IEEE Trans. Parallel Distrib. Syst.*, vol. 5, no. 2, pp. 113–120, Feb. 1994.
[35] P. Guo, X. Wang, and Y. Han, "The enhanced genetic algorithms for the optimization design," in *Proc. 3rd Int. Conf. Biomed. Eng. Inform.*, Yantai, China, 2010, pp. 2990–2994.
[36] 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.
[37] E. Zeydan *et al.*, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.
[38] J. Song, M. Sheng, T. Q. S. Quek, C. Xu, and X. Wang, "Learning-based content caching and sharing for wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4309–4324, Oct. 2017.
[39] J. Li, B. Liu, and H. Wu, "Energy-efficient in-network caching for content-centric networking," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 797–800, Apr. 2013.
[40] S. Boyd and A. Mutapcic, "Subgradient methods," Stanford Univ., Stanford, CA, USA, Appl. Notes EE364b, Apr. 2008.

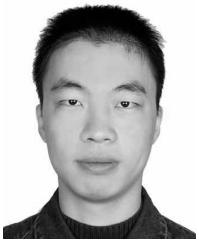
QIANG LI (M'16) received the B.Eng. degree in communication engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2007, and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2011, where he was a Research Fellow, from 2011 to 2013. Since 2013, he has been an Associate Professor with the Huazhong University of Science and Technology, Wuhan, China. He was a Visiting Scholar with The University of Sheffield, Sheffield, U.K., in 2015. His current research interests include future broadband wireless networks, cooperative communication, software-defined networks, device-to-device communication, and edge caching.

YUANMEI ZHANG received the B.Eng. degree from the Wuhan University of Technology, China, in 2017. She is currently pursuing the master's degree with the Huazhong University of Science and Technology, Wuhan, China. Her current research interests include software-defined networks, device-to-device communication, and content caching.

ASHISH PANDHARIPANDE (S'97–AM'98–M'03–SM'08) received the M.S. degrees in electrical and computer engineering and mathematics and the Ph.D. degree in electrical and computer engineering from The University of Iowa, Iowa City, in 2000, 2001, and 2002, respectively. Since 2002, he has been a Postdoctoral Researcher with the University of Florida, a Senior Researcher with the Samsung Advanced Institute of Technology, and a Senior Scientist with Philips Research. He has held visiting positions at AT&T Laboratories, NJ, USA, and the Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, India. He is currently a Lead R&D Engineer with Signify, Eindhoven, The Netherlands.

His research interests include the intersection of sensing, networking, and controls, and system applications in domains, such as smart lighting systems, energy monitoring and control, and cognitive spectrum sharing, and a member of the International Advisory Board, *Lighting Research & Technology*, an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE SENSORS JOURNAL, and an Editor of *EURASIP Journal on Wireless Communications and Networking*.

XIAOHU GE (M'09–SM'11) received the Ph.D. degree in communication and information engineering from the Huazhong University of Science and Technology (HUST), China, in 2003, where he is currently a Full Professor with the School of Electronic Information and Communications. He was a Researcher with Ajou University, South Korea, and the Politecnico di Torino, Italy, from 2004 to 2005. He has been with HUST, since 2005. He was a Visiting Researcher with Heriot-Watt University, Edinburgh, U.K., in 2010. He is also an Adjunct Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. He has published about 190 papers in refereed journals and conference proceedings. He holds about 15 patents in China. His research interests include mobile communication, traffic modeling in wireless networks, green communications, and interference modeling in wireless communication. He received the Best Paper Awards from the IEEE GLOBECOM 2010. He served as the General Chair for the 2015 IEEE International Conference on Green Computing and Communications. He serves as an Associate Editor for the IEEE WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.



JILIANG ZHANG (M'15) received the B.E., M.E., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2007, 2009, and 2014, respectively. He was a Postdoctoral Fellow with the Shenzhen Graduate School, Harbin Institute of Technology, from 2014 to 2016, an Associate Professor with the School of Information Science and Engineering, Lanzhou University, in 2017, and a Researcher with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. He is currently a Marie Curie Research Fellow with the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K. His research interests include wireless systems, MIMO channel measurement and modeling, single radio-frequency MIMO systems, ultra-dense networks, relay systems, device-to-device networks, and wireless ranging systems.