Caching deployment based on energy efficiency in device-to-device cooperative networks

Weiguang Wang, Hui Li, Yang Liu, Wei Cheng and Haoyang Qin

Abstract
The rapid growth of mobile data traffic demand will cause congestion to the future communication network. The cache-enabled device-to-device communication has been proven to effectively enhance the performance of wireless communication networks. This article investigates the caching deployment problem from the energy efficiency in the cache-enabled device-to-device networks. According to the random geometry theory modeling, the closed form expression of energy efficiency is derived, which measures the average number of successful transmitted file bits per unit time and per unit power consumption. And then we establish an optimization problem to maximize energy efficiency. As the formulated optimization problem is a multiple-ratio fractional programming problem that cannot be solved conveniently, we propose a quadratic transformation method to nest in the energy efficiency maximization problem. To tackle this problem, an iterative optimization algorithm is proposed to optimize the caching policy and network energy efficiency. The simulation results demonstrate that the proposed policy can achieve higher energy efficiency and hit probability in the cache-enabled device-to-device network.

Keywords
Wireless caching network, device-to-device communications, energy efficiency, caching policy

Date received: 29 September 2020; accepted: 2 December 2020
Handling Editor: Peio Lopez Iturri

Introduction
Recently, the demand for video delivery services has dramatically increased, which promotes the exponential growth of wireless data traffic.\(^1\)\(^-\)\(^3\) However, conventional solutions like the ultra-dense network with increased base station (BS) deployment density, the millimeter-wave communication using higher frequency spectrum communication, and the multiple input multiple output (MIMO) technology cost too much and have reached their limits.\(^4\) Thus, new paradigms need to be studied to enhance the performance of traditional cellular network architecture.\(^5\) Since device-to-device (D2D) communication can directly communicate with nearby user equipments (UEs) without data forwarding through the BSs, as a result, D2D communication has attracted widespread attention in recent years.\(^6\)\(^-\)\(^8\) From a practical perspective, the storage of UEs is growing rapidly at low cost. Inspired by these facts, the cache-enabled D2D communication was proposed in Golrezaei and colleagues\(^9\)\(^,\)\(^10\) to offload more transmitted data traffic. Incorporating cache-enabled D2D communication into the traditional network brings an amount of benefits such as improving offload gain, decreasing communication delay, and enhancing spectral and energy efficiency (EE).\(^7\) Simultaneously, the

Corresponding author:
Hui Li, School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China.
Email: lh@nwpu.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
The cache-enabled D2D communication has proven to be a promising technology that can effectively offload network traffic and reduce congestion. Recently, cache policy in the cache-enabled D2D communication network often focuses on maximizing the content hit ratio or reducing transmission delay. The main goal of the above caching policy is to offload as much network data traffic as possible. However, they completely ignore the energy cost of data transmission and data storage. Furthermore, an optimal hit rate does not indicate that EE will be optimal. Therefore, it is very meaningful for us to design the cache policy from the perspective of network EE. The design of cache policy based on EE not only guarantees the requirements of green communication but also can maintain the EE of future wireless networks at a reasonable level. Work in this direction, we have carried out an in-depth study on cache deployment in the cache-enabled D2D network. Then, we find that the design of cache policy considering EE is promising, and there are few articles in this area. This means that the design of energy-efficient caching policies in D2D networks is very important, but this problem has not been solved in existing research work.

In this work, we consider the cache-enabled D2D networks in which UEs can obtain desired content through D2D communications and self-cache with different energy costs. Motivated by the fact that a large number of redundant transmissions and high transmission energy cost, we use network EE to establish the model for the problem and propose caching policy design. The design goal is to maximize network EE. As such, the research question about the design of a green content caching policy can be expressed as: how to design a caching strategy to maximize network EE while ensuring data traffic offload? how to turn the non-convex cache policy design problem into a convex optimization problem? In this article, we focus on the solutions for these two problems. As far as we know, this is the first initiative to investigate this research issue.

The main contributions of this article are listed as follows:

- By jointly considering the influence of D2D-cache and self-cache, we derive the hit ratio and EE in closed form based on the stochastic geometry theory modeling. By utilizing the hit ratio and EE formulations, we propose a cache policy design problem based on EE. As we know, the cache deployment optimization of EE based on the cache-enabled D2D network has not been considered explicitly before.
- Through analysis, it is prove that the proposed EE optimization problem is a multiple-ratio fractional programming (FP) problem. We propose a novel method to solve the EE maximization problem of multiple-ratio FP, which converts the concave–convex multiple-ratio FP problem into a sequence of the convex optimization problem. We can easily analyze the performance with low complexity. Then we propose an alternate optimization algorithm to obtain the optimal caching policy design based on network EE.
- We evaluated the proposed design and analyses through simulation. The simulation results of our proposed cache policy are compared with the baseline and the max-hit-rate design using different key parameters. Finally, the feasibility of our proposed caching policy was verified.

The other sections of this article are arranged as follows. Section “Related works” explores a related literature review. Section “System model” constructs the system model. Section “Problem formulation and analysis” describes the problem formulation and analysis. Section “Caching policy for EE optimization” describes the caching policy for EE optimization. Section “Simulation and numerical results” demonstrates simulation and numerical results. Finally, section “Conclusion” concludes this article.

Related works

The wireless edge cache technology has been widely used in various network scenarios. Due to the limited cache capacity, it is impractical to cache all content on mobile user devices. Through the research in recent years, scholars have proposed caching policies with different optimization objectives for the cache-enabled D2D network. Chen and Yang optimizes cache policy by user interests and activity status to maximize the offload rate of the D2D network. Malak et al. in the presence of interference and noise, study the cache strategy and transmission in D2D network and derive the closed expression of the optimal cache strategy. In Chen et al., the authors consider the gain of self-caching, and propose a caching strategy based on D2D caching network. These works mainly focus on transmission in single-hop networks, and in Krishnan et al. the authors extends single-hop transmission to multi-hop application retransmission. They mainly studied the impact of content retransmission on the optimal content placement strategy in static and mobile UE scenarios. Dehghan et al. propose an optimized
cache placement policy based on the routing, by reducing the transmission delay for each link. Based on random geometry and optimization theory, Chen et al.\textsuperscript{15} investigate an optimal content placement strategy, which maximizes the cache hit ratio of D2D communication network. In Deng et al.,\textsuperscript{18} the authors construct the content deployment problem in the D2D network scenario, which jointly considers the effects of UEs’ mobile characteristics, cache capacity, and the number of content codes. Chen et al.\textsuperscript{19} jointly optimize the cache policy and user scheduling in D2D cache network to maximize the success rate of data traffic offloading. Giatsoglou et al.\textsuperscript{20} propose a D2D caching strategy for millimeter-wave networks and study its offload gain. In Gregori et al.,\textsuperscript{21} the authors propose a method to optimize content caching and delivery in the same time frame. In Sadeghi et al.,\textsuperscript{22} the authors use local and global Markov processes to model user requests and propose a Q-learning-based file placement algorithm to find the optimal cache policy when the probability of movement is unknown. Chen et al.\textsuperscript{23} consider the remaining battery capacity of mobile users in the cache-enabled D2D. They studied the correlation between the offload gain of the network and the energy cost of each content provider. The author introduces a user centered protocol to reduce the energy consumption of helping users transfer files. Then, the cache strategy is optimized to maximize the offload opportunity and the transmission power of each assistant to maximize the offload probability. Lee and Molisch\textsuperscript{24} study the potential benefits of individual preferences to establish the cache design problem and reduce the average energy consumption of the system through the resulting content placement strategy. In Ji et al.,\textsuperscript{25} the optimal caching policy cannot maximize the offload traffic in the wireless network, and the energy consumption of UEs in the network is very high.

The above research shows that most researchers are optimizing caching policies for different UEs with limited caching capacity to increase the offload gain of the caching network, but they ignore the cost of data transmission. In the actual network, because of the UEs’ limited battery capacity and selfish behavior, some users do not want to be the cache-enabled users because it will waste their energy to transmit the files. Even if the UEs are willing to cache the content and provide the requested content to the requesters, what they really want to achieve is to maximize the EE in the D2D network.\textsuperscript{26} It has been proven in Perabathini et al.\textsuperscript{27} that caching the contents in the cache-enabled D2D UEs can provide significant gains in terms of EE. However, the design of caching policy based on EE analysis in the cache-enabled D2D network is still a research subject to be solved, and need to be further investigated. In this research work, we propose a novel caching policy which considers the EE of the network. This caching policy can maximize EE and keep the network traffic offloading at a reasonable level.

**System model**

This research considers a D2D caching network model, in which UEs are modeled as a homogeneous Poisson point process (HPPP) with density $\lambda_p$, as shown in Figure 1. If the UE caches the required content in the self-cache, the UEs will obtain the required content directly. Otherwise, the UEs will get the required contents from the neighboring devices via D2D links, where the neighboring devices are defined as the UEs with distance less than $d$ from the nth UE. In this scenario, all UEs are equipped with single antenna and the same transmit power $P_t$. The signal is usually affected by the large-scale fading and small-scale fading during transmission. Specifically, for the large-scale fading, we adopt the standard path loss propagation model $d^{-a}$, where $d$ is the link distance and $a$ is the path loss exponent, which satisfies $a>2$. For small-scale fading, the signal will experience the effects of Rayleigh fading and the channel gain follows zero-mean complex Gaussian distribution with unit variance. Each D2D pair communicates through an orthogonal sub-channel with bandwidth $W$. The BS is aware of UEs’ channel state and location and also coordinates D2D communications.\textsuperscript{28}

We considered a limited content library $F = \{1, 2, ..., F\}$, where the content is sorted by popularity from 1 to $F$, and the first file means the most popular file. Due to the limited storage space of the UEs, we set that each UE can cache $S$ files (\(S \ll F\)) and the content have equal unit size. We further set the content popularity to obey the Zipf distribution,\textsuperscript{29} and each UE can independently and identically request the required content from the content library. The probability of requesting the $f$-ranked file is

$$P_f = \frac{f^{-\zeta}}{\sum_{i=1}^F i^{-\zeta}}$$

(1)

![Figure 1. Cache-enabled D2D network model.](image)
where $e$ represents how skewed the popularity distribution is. Note that estimating the popularity distribution of documents is always an open research problem, which is not in the scope of this article (see the literature\textsuperscript{30,31}). Each UE follows the set caching policy $q = [q_1, ..., q_f, ..., q_F]$ to individually cache the content in the content library, where $q_f$ is the probability that the UE caches the $f$th content and $\sum_{f}^{F} q_f = S$ is predefined due to the storage limit. Based on the characteristics of the Poisson process, the UE with the $f$th content in cache obey HPPP with density $q_f \lambda_p$.

When requesters want to obtain content from the content library $F$, there may be two content acquisition methods in the network:

- **Case 1: Self-cache.** If UEs can cache the required content in their own cache, they will get the content directly from their own cache without establishing any communication links.

- **Case 2: D2D cache hit.** If the required content cannot be found in its own cache, it will be obtained from the neighboring cache UEs (i.e. within the D2D radius $d$) by establishing a D2D link. If there are many UEs caching the requested content, the required content will be retrieved from the most recent cached UE.

Notice that UE preference follows the order above. When the required content is not found in the above case, the content will be downloaded from the core network to the nearest BS through the backhaul and then transmitted to the UE, which will generate a large energy cost.

**Problem formulation and analysis**

In this section, we analyze the UEs’ access probabilities, and these conclusions will serve as the basis for the conclusions obtained in the following sections. We formally define the main optimization problem aiming at EE in this article. We strive to find an optimal cache placement policy $q = [q_1, ..., q_f, ..., q_F]$, which can maximize the EE of the cache-enabled D2D communication network.

**Cache hit probability analysis**

We define the cache hit ratio as the probability that UEs can find the requested content in local cache, including the self-cache and the D2D cache hit cases:

1. **Self-cache:** When the requester makes a request message, the probability that the content is found in its own cache space is defined as self-caching, so the probability of self-service should be the same as the probability of caching

$$p_{self} = \sum_{f=1}^{F} p_f q_f$$

2. **D2D cache hit:** If the requester cannot find the required content in the self-caching, they will seek help from nearby cache UEs. The probability that there are $k$ active UEs nearby in the PPP model can be expressed as

$$\Pr(K = k) = \frac{(\pi \lambda_p d^2)^k}{k!} e^{-\pi \lambda_p d^2}$$

The probability that at least one cache UE has cached the required content within the requesters’ communication range and the requester can obtain the required content through D2D communication can be calculated as

$$p_{D2D} = \sum_{f=1}^{F} (1 - q_f) \prod_{k=0}^{k-1} \Pr(K = k) [1 - (1 - q_f)^k]$$

$$= \sum_{f=1}^{F} (1 - q_f) \times \left(1 - e^{-\pi \lambda_p d^2}ight)$$

Finally, through equations (2) and (4), we can obtain the hit ratio of the overall caching system in closed form as

$$\rho = \sum_{f=1}^{F} p_f [q_f + (1 - q_f) \times \left(1 - e^{-\pi \lambda_p d^2}\right)]$$

**The EE analysis for cache-enabled D2D network**

We considered a linear power consumption model in this article. The power consumption of active cache UEs and inactive cache UEs are expressed as $E_{p, act} = (1/\nu)P_t + P_c$ and $E_{p, inact} = P_c$, respectively, where $\nu$ is the power amplifier efficiency, $P_t$ and $P_c$ are the transmitting power and consuming circuit power of the content provider, respectively. Therefore, the total power consumption for obtaining the required files through the D2D link can be calculated as

$$E = \hat{\lambda}_p E_{p, act} + (\lambda_p - \hat{\lambda}_p) E_{p, inact}$$

$$= \sum_{f=1}^{F} \lambda_p f \left(1 - \frac{1}{\nu} p_{D2D} P_t + P_c\right)$$

where $\hat{\lambda}_p$ is the active intensity of users and $\lambda_p f$ is the intensity of users caching file $f$. We define the ratio of the average number of successfully transferred content
bits per unit time to the total power consumption required as the EE of the network, which can be shown as

\[ \eta_{EE} = \frac{\vartheta}{E} \]  

(7)

where \( \vartheta \) is the average number of successfully transmitted content bits per unit time, \( \vartheta = R_t \cdot \rho(q_f, z) \), \( R_t \) is the data transfer rate between the D2D UEs. With the assumption of the channel model, we consider that the average data rate can be derived as \( R_t = W \log_2(P_d d^{-\alpha}/\sigma_0^2) \), where \( \gamma(r) \) is the signal-to-interference-plus-noise ratio (SINR) of the D2D receiver and \( W \) is the bandwidth. The EE can be further formulated as

\[ \eta_{EE} = \frac{R_t \cdot \rho(q_f, d)}{E} \]  

(8)

Finally, substituting equations (5) and (6) into equation (8), we can derive the expression of EE as

\[ \eta_{EE} = \frac{R_t \sum_{f=1}^{F} \lambda_p p_f p_d}{\sum_{f=1}^{F} \lambda_p p_f (P_{DD2D} + P_c)} \]  

(9)

### Caching policy for EE optimization

First, we make some alternative transformations to equation (9). The optimal caching policy design problem for maximizing the cache-enabled D2D network EE can be described as the following nonlinear optimization problem

\[
\begin{align*}
\text{maximize} & \quad \eta_{EE}(q) \\
& = \frac{R_t \sum_{f=1}^{F} \lambda_p p_f p_d}{\sum_{f=1}^{F} \lambda_p p_f (P_{DD2D} + P_c)} \\
\text{subject to} & \quad \sum_{f=1}^{F} q_f \leq S, \forall f \\
& \quad 0 \leq q_f \leq 1, \forall f 
\end{align*}
\]  

(10)

The objective function (equation (10a)) is derived from equation (9). Constraint (equation (10b)) is the cache capacity of the cache-enabled UE. Constraint (equation (10c)) is the probability that the file \( f \) is cached by the cache-enabled UE. In order to further solve our optimized problem, we use Theorem 1 as Shen and Yu’s derivation.33

**Theorem 1.** Given \( M \) pairs of nonnegative function \( A_m(x) : \mathbb{R}^d \rightarrow \mathbb{R}^+ \) and denominator functions

\[ B_m(x) : \mathbb{R}^d \rightarrow \mathbb{R}^+ \]  

for \( m = 1, \ldots, M \). Then, the multi-ratio sum optimization problem

\[
\begin{align*}
\text{maximize} & \quad \sum_{m=1}^{M} \frac{A_m(x)}{B_m(x)} \\
\text{subject to} & \quad x \in \mathcal{X}
\end{align*}
\]  

(11)

So, the multi-ratio sum optimization problem can be transformed into an optimization problem in the following equation (12)

\[
\begin{align*}
\text{maximize} & \quad \sum_{m=1}^{M} \left( \frac{2y_m A_m(x)}{B_m(x)} - y_m^2 \right) \\
\text{subject to} & \quad x \in \mathcal{X}
\end{align*}
\]  

(12)

It has been proved in Shen and Yu33 that the optimal value of the objective function through the above transformation is the same as that of the original optimization variable. Where \( y = [y_1, \ldots, y_M] \) is an auxiliary variable.

**Proof.** We can easily obtain that the objective function (12) is concave in \( y \). Thus, we can further get the partial derivative of the objective function (12) with respect to \( y_m \)

\[ y_m^* = \frac{\sqrt{A_m(x)}}{B_m(x)}, \forall m = 1, \ldots, M \]  

(13)

The objective function in equation (11) can be obtained by substituting the value of equation (13) into the objective function in equation (12).

We apply Theorem 1 to the multi-ratio sum term in equation (10). Then, we can equate the original optimization problem to the following optimization problem

\[ \text{PI : maximize } \eta_q(q, y) \]  

\[ \text{subject to } \sum_{f=1}^{F} q_f \leq S, 0 \leq q_f \leq 1, \forall f \in \mathcal{F} \]  

\[ y_f \in \mathbb{R}, \forall f \in \mathcal{F}. \]  

(14)

where \( y \) represents the set \( \{y_1, \ldots, y_F\} \), and the new objective \( \eta_q \) can be written as

\[
\eta_q(q, y) = \sum_{f=1}^{F} 2y_f \sqrt{R_p f [q_f + (1 - q_f) \times \left( 1 - e^{-\alpha_q q_f d^2} \right)]} \\
- \sum_{f=1}^{F} y_f^2 \left[ \frac{P_d}{(1 - q_f) \times \left( 1 - e^{-\alpha_q q_f d^2} \right)} \right]
\]  

(15)

Clearly, \( y_f \) is an auxiliary variable introduced by the quadratic transform. We use alternate optimization
method to optimize $y_f$ and $q_f$, respectively. The optimal $y_f$ for fixed $q_f$ is

$$y_f^* = \frac{1}{P_f} \left[ \frac{P_f}{R_{tpf} q_f + (1 - q_f) \times \left( 1 - e^{-\alpha y_f^* T^2} \right) / C_3} \right], \forall f \in F. \quad (16)$$

**Proposition 1.** The EE-based caching policy optimization problem $P_1$ in the cache-enabled D2D networks is a convex programming problem.

**Proof.** See Appendix A.

When $y_f$ is fixed, each $A_f(q_f)$ is concave with respect to $q_f$, each $B_f(q_f)$ is convex with respect to $q_f$, and the square-root function is concave and increasing, so the quadratic transform is concave in $q_f$ for fixed $y_f$. The optimal $q_f$ can thus be efficiently obtained through standard convex solver in MATLAB. The entire caching strategy optimization method is described in Algorithm 1.

**Computational complexity:** Through Algorithm 1, we can easily get the optimal caching policy with computation complexity $O(F^2 \cdot N)$, where $O(\cdot)$ is the big-O notation, $F$ is the number of contents to be calculated per iteration, and $N$ is the number of UEs.

## Simulation and numerical results

In this subsection, we analyze the offloading gain and the EE of the proposed caching policy based on the D2D network by using MATLAB software. In the simulation, unless otherwise stated, the parameters taken in the simulation are given in Table 1. Besides our optimized caching policy (with legend “Proposed EE design”), we also considered the other two caching policies for comparison. (1) the uniform caching policy (i.e. the cache policy obeys uniform distribution, with legend “Uniform-baseline”) as the caching baseline and (2) cache policy for maximizing the cache hit probability (with legend “max-hit-rate design”).

In Figure 2, we compare the proposed EE design, the max-hit-rate design, and the uniform caching probabilities with different Zipf parameters (i.e. an $\varepsilon$ value at 0.6 and 1.2). These two parameters come from the UMass Amherst YouTub experiment and the parameters reported in Lee et al., respectively. The optimal caching probabilities of the files are plotted as a function of the popularity order $f$. We note that in the proposed EE design and the max-hit-rate design, each cache-enabled UE tends to cache the most popular contents with higher probability. When the Zip parameter $\varepsilon = 1.2$, the proposed EE design and the max-hit-rate design are more inclined to cache top-ranking contents than $\varepsilon = 0.6$. With the increase in popularity factor $\varepsilon$, the more concentrated the requested contents, and caching contents with higher popularity can increase the cache throughput. This fact is the same as the fact that only a few contents were repeatedly requested in the same time period in the actual network. Similarly, because UEs have limited storage space, they may not be able to cache the contents that rank lower in popularity.

### Table 1. Simulation parameters.

| Parameters | Value |
|------------|-------|
| The intensity of cache-enabled users $\lambda_p$ | 0.0042 m$^{-2}$ |
| The transmission power of D2D transmitter $P_t$ | 0.25 W |
| D2D bandwidth $W$ | 20 MHz |
| Path loss exponent $\alpha$ | 3.68 |
| Noise power $\sigma^2$ | $-95$ dBm/Hz |
| The number of files $F$ | 20 |
| Each UE cache capacity $2\times$ | 2 files |
| Circuit power consumption $P_c$ | 115.9 mW |
| Power amplifier efficiency $\nu$ | 0.2 |
| Zip parameter $\varepsilon$ | 0.6, 1.2 |
In Figure 3, we compared the EE of different caching policies regarding Zipf parameters. It can be seen from the simulation results that the proposed EE design can show the optimal network EE compared to the other two caching policies. Besides, when the Zip parameter $\varepsilon$ is very small, the network EE of the max-hit-rate design is very close to that of the proposed EE design. As the popularity factor $\varepsilon$ increases, the advantages of the proposed EE design in EE are more significant. The main reason for this phenomenon is that with the increase of Zip parameters, the content of user requests becomes more centralized $\varepsilon$. However, the proposed EE design can cache more popular files, thereby bringing greater benefits to network performance. Since the uniform caching policy is not affected by file popularity, the EE does not change with the popularity and provide the poor performance. The results show that effective inter-user equipment cooperative caching is more important for network EE. However, when the Zipf parameter $\varepsilon = 1.2$, the proposed EE design can increase EE by 18.7% compared to the max-hit-rate design.

In Figure 4, we use the Zip parameters to evaluate the hit rate of the proposed caching policy, that is, the proposed EE design, and compare it with the max-hit-rate design and the uniform caching policy. From the simulation results, we can observe that the hit rate increases as the Zip parameter increases, except that the hit ratio of the uniform caching policy has not changed. This is because the uniform caching policy is evenly distributed. It means that introducing caching into networks does improve the hit rate. In particular, it can be observed that when the Zip parameter $\varepsilon$ is less than 0.1, the cache hit rates of the proposed EE design and the max-hit-rate design are very low. However, as the Zip parameter $\varepsilon$ becomes larger, the cache hit ratio of the proposed EE design increases rapidly. In particular, compared to the max-hit-rate design, the proposed EE design can increase EE by 30.4% when the Zip parameter $\varepsilon = 1.2$. This is mainly due to as the concentration of the request file increases, the proposed EE design can cache more popular files (available in Figure 1) and can bring much greater offload gains.

Figure 5 shows the system EE performance with different distances. In the simulation, we use two sets of Zipf parameters $\varepsilon = 0.6$ and $\varepsilon = 1.2$ to evaluate the network EE of different caching policies. From the simulation results, we can observe that with the increase of
distance, EE of all designs first increases sharply and reaches the maximum value near the cooperation distance $d = 40$ m. However, when the cooperation distance is greater than 40 m, EE will slowly decline. This is intuitive because when the cooperation distance is within the appropriate range, the increase in cooperation distance will increase the probability of users to find the required contents, which will lead to better EE. But when the cooperation distance continues to increase, the energy cost of obtaining the required content through the D2D link will increase significantly, and the total throughput will also decrease. We can also see that the proposed EE design can obtain the optimal network EE and becomes larger with the increased cooperation distance. Again, regardless of the small Zip parameter $\varepsilon = 0.6$ or the large Zip parameter $\varepsilon = 1.2$, the network EE of the proposed EE design is always better than that of the max-hit-rate design, and the uniform caching policy provides very poor network EE. By comparing the abscissa in Figure 5, we can see that the proposed EE design can provide greater cooperation distances. This is because the cache policy proposed by the max-hit-rate design is difficult to balance self-caching and UEs caching.

In Figure 6, we show the hit ratio with different caching policies versus Zipf distribution parameter $\varepsilon$ and cooperation distance. We can see from the figure that the hit rate curves for all designs increase sharply before the cooperation distance $d = 60$ m, and then increase slightly until all files are hit. As the cooperation distance between UEs increases, all cache policies can reach the maximum hit ratio. Obviously, the increase in the cooperation distance also increases the probability that the UE will find the desired content within the communication range. When Zipf parameter is very small, the proposed EE design increases rapidly and can provide near-optimal performance. However, when Zipf parameter $\varepsilon = 1.2$, the proposed EE design is obviously better than the other two caching policies. Similarly, no matter what Zipf parameters, the uniform caching policy performance is always the worst. This means that when the requested files are concentrated, it is more beneficial to increase the chance of requesters to find the desired file by caching contents with Zipf parameter, which justifies our assumption in the previous analysis.

In Figure 7, we use Zipf parameters $\varepsilon = 0.6$ and $\varepsilon = 1.2$ to evaluate the network EE of different caching policies. From the figure, we can see that before the density of active UEs reaches $\lambda_p = 0.24$, the EE of all designs increases sharply. As the density of active UEs increases, the uniform caching policy begins to deviate from the proposed EE DESIGN and the max-hit-rate design. However, when the density of active UEs approaches the maximum value, the EE of all designs increases slowly and becomes almost stationary. Although a larger number of active UEs improve the probability of D2D-serve, the increase in the density of active UEs means the condition of co-channel interference is more stringent. Nonetheless, the network EE of the caching policy designed by the proposed EE design is significantly better than the other caching policies. More interestingly, when the density of active UEs is greater than a certain value (about $\lambda_p = 0.4$), the network EE of the uniform caching policy no longer increases, and it is far lower than the other two policies. This also confirms that in the cache-enabled D2D network considering the actual popular distribution, the cooperative cache communication between UEs is beneficial to improve the EE of the network.
In Figure 8, we examine the impact of user density on hit ratio. As expected, the hit ratio of all caching policies increases rapidly with UE density. This is because the dense deployment of active cache UEs in the real network can increase the network storage space and the increase in the number of cache UEs also increases the probability of establishing D2D communication. We noticed that when the density of active UEs increases significantly, the max-hit-rate design and the uniform caching policy starts to diverge from the best proposed EE design. We also can observe that when the Zipf parameter is $e = 0.6$, as the active UE density increases, the hit rate of the proposed EE design is slightly lower than the max-hit-rate design first and then significantly better than the max-hit-rate design. When the Zipf parameters are $e = 0.6$ and $e = 1.2$, the data hit rate designed by the max-hit-rate design does not change significantly. However, in the case of the above two key parameters, the proposed EE design can provide good performance when the Zipf parameter $e = 1.2$. This is because a larger $e$, such as $e = 1.2$ usually means a higher concentration of the content being requested compared to $e = 0.6$. Thus, the proposed EE design can provide better offload performance. The uniform caching policy performs poorly because it does not judiciously design the cache policy.

**Conclusion**

The cache-enabled D2D communication is a potential technology to solve future network congestion, which can maximize the overall network EE and offload gain. In this article, we propose a novel D2D caching policy for the content-related EE in the D2D communication network. Based on random geometry theory modeling, we deduce the theoretical expression of the hit rate and define the EE. And then we formulate a cache deployment problem as maximizes the overall EE of the D2D communication network. By applying the quadratic transform, we transform the multiple-ratio FP problem into a solvable convex programming problem, then we solve the problem by using the iterative optimization algorithm. Furthermore, we compare the proposed caching policies with other caching policies by utilizing some key parameters, such as the cooperation distance between users, the user density, and the Zipf distribution. We evaluate our analysis results by simulation and numerical and the results show that the performance of the proposed caching policy outperforms previously proposed schemes.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (Nos. 61571364 and 61401360) and the Seed Foundation of Innovation and Creation for Graduate Students in Northwestern Polytechnical University (No. CX2020149).

**ORCID iDs**

Weiguang Wang https://orcid.org/0000-0002-7989-4937  
Wei Cheng https://orcid.org/0000-0002-0874-9927

**Data accessibility statement**

The data used to support the findings of this study are available from the corresponding author upon request (Email: Laughing_wwg@163.com). The lemma or theorem cited in this article can be publicly obtained according to the reference.

**References**

1. Cisco visual networking index: global mobile data traffic forecast 2017-2022, www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-forecast-qua.pdf (accessed 18 February 2019).
2. Romero E, Blesa J, Rozas A, et al. Energy efficiency strategy in D2D cognitive networks using channel selection based on game theory and collaboration. *Int J Distrib Sens Netw* 2016; 12(8): 1–13.
3. Liu PL, Li Y, Cheng W, et al. Energy-efficient power allocation for millimeter wave beamspace MIMO-NOMA systems. *IEEE Access* 2019; 7: 114582–114592.
4. Golrezaei N, Dimakis AG and Molisch AF. Device-to-device collaboration through distributed storage.
2012 IEEE global communications conference (GLOBECOM), Anaheim, CA, 3–7 December 2012. New York: IEEE.

5. Yuan P, Cai Y, Liu Y, et al. ProRec: a unified content caching and replacement framework for mobile edge computing. *Wireless Netw* 2020; 26(4): 2929–2941.

6. Zhou H, Wang H, Li X, et al. A survey on mobile data offloading technologies. *IEEE Access* 2018; 6: 5101–5111.

7. Asadi A, Wang Q and Mancuso V. A survey on device-to-device communication in cellular networks. *IEEE Commun Surv Tuts* 2014; 16: 1801–1819.

8. Chen B, Yang C and Wang G. High-throughput opportunistic cooperative device-to-device communications with caching. *IEEE Trans Veh Technol* 2017; 66: 7527–7539.

9. Golrezaei N, Mansourifard P, Molisch A, et al. Base-station assisted device-to-device communications for high-throughput wireless video networks. *IEEE Trans Wireless Commun* 2014; 13: 3665–3676.

10. Golrezaei N, Molisch AF, Dimakis AG, et al. Femtocaching and device-to-device collaboration: a new architectural perspective for wireless video distribution. *IEEE Commun Mag* 2013; 51: 142–1495.

11. You XH, Wang CX, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. *Sci China Inform Sci* 2020; 64: 110301.

12. Wang X, Chen M, Taleb T, et al. Cache in the air: exploiting content caching and delivery techniques for 5G systems. *IEEE Commun Mag* 2014; 52: 131–139.

13. Chen BQ and Yang C. Caching policy for cache-enabled D2D communications by learning user preference. *IEEE Trans Commun* 2018; 66: 6586–6601.

14. Malal D, Al-Shalash M and Andrews JG. Optimizing content caching to maximize the density of successful receptions in device-to-device networking. *IEEE Trans Commun* 2016; 64(10): 4365–4380.

15. Chen Z, Pappas N and Kontouris M. Probabilistic caching in wireless D2D networks: cache hit optimal versus throughput optimal. *IEEE Commun Lett* 2017, 21(3): 584–587.

16. Krishnan S, Afshang M and Dhillon HS. Effect of retransmissions on optimal caching in cache-enabled small cell networks. *IEEE Trans Veh Technol* 2017; 66: 11383–11387.

17. Dehghan M, Seetharam A, Jiang B, et al. On the complexity of optimal routing and content caching in heterogeneous networks. In: *2015 IEEE conference on computer communications (INFOCOM)*, Kowloon, Hong Kong, 26 April–1 May 2015, pp.936–944. New York: IEEE.

18. Deng T, Ahani G, Fan P, et al. Cost-optimal caching for D2D networks with user mobility: modeling, analysis, and computational approaches. *IEEE Trans Wireless Commun* 2018; 17: 3082–3094.

19. Chen B, Yang C and Xiong Z. Optimal caching and scheduling for cache-enabled D2D communications. *IEEE Commun Lett* 2017; 21: 1155–1158.

20. Giatsoglou N, Ntontin K, Kartsakli E, et al. D2D-aware device caching in mmWave-cellular networks. *IEEE J Sel Areas Commun* 2017; 35: 2025–2037.

21. Gregori M, Gómez-Vilardebé J, Matamoros J, et al. Wireless content caching for small cell and D2D networks. *IEEE J Select Area Commun* 2016; 35(5): 1222–1234.

22. Sadeghi A, Sheikholeslami F and Giannakis GB. Optimal and scalable caching for 5G using reinforcement learning of space-time popularities. *IEEE J Select Topic Signal Pr* 2018; 12: 180–190.

23. Chen B, Yang C and Molisch AF. Cache-enabled device-to-device communications: offloading gain and energy cost. *IEEE Trans Wireless Commun* 2017; 16(7): 4519–4536.

24. Lee MC and Molisch AF. Individual preference aware caching policy design in wireless D2D networks. *IEEE Trans Wireless Commun* 2020; 19(8): 5589–5604.

25. Ji M, Caire G and Molisch A. Wireless device-to-device caching networks: basic principles and system performance. *IEEE J Sel Areas Commun* 2016; 34: 176–189.

26. Chen BQ, Yang C and Xiong Z. Optimal caching and scheduling for cache-enabled D2D communications. *IEEE Commun Lett* 2017; 21: 1155–1158.

27. Perabathini B, Ba E, Kontouris M, et al. Caching at the edge: a green perspective for 5G networks. In: *2015 IEEE international conference on communication workshop (ICCW)*, London, 8–12 June 2015, pp.2830–2835. New York: IEEE.

28. Chen YY, Wang L, Ma RQ, et al. Cooperative full duplex content sensing and delivery improves the offloading probability of D2D caching. *IEEE Access* 2019; 7: 29076–29084.

29. Breslau L, Cao P, Fan L, et al. Web caching and Zipf-like distributions: evidence and implications. In: *IEEE INFOCOM ’99 conference on computer communications. Proceedings of 18th annual joint conference of the IEEE computer and communications societies. The Future is Now (Cat. No.99CH36320)*, New York, 21–25 March 1999, pp.126–134. New York: IEEE.

30. VU TX, Chatzinotas S, Ottersten B, et al. Full-duplex enabled mobile edge caching: from distributed to cooperative caching. *IEEE Trans Wireless Commun* 2020; 19: 1141–1151.

31. Liu Y, Wang Y, Sun R, et al. Energy efficient downlink resource allocation for D2D-assisted cellular networks with mobile edge caching. *IEEE Access* 2019; 19: 2053–2067.

32. Heath RW, Wu T, Kwon YH, et al. Multiuser MIMO in distributed antenna systems with out-of-cell interference. *IEEE Trans Signal Pr* 2011; 59(10): 4885–4899.

33. Shen K and Yu W. Fractional programming for communication systems-Part II: uplink scheduling via matching. *IEEE T Signal Pr* 2018; 66(10): 2631–2644.

34. Boyd S and Vandenberghe L. *Convex optimization*. Cambridge: Cambridge University Press, 2004.

35. Lee MC, Ji M, Molisch AF, et al. Performance of caching-based D2D video distribution with measured popularity distributions. In: *2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, 9–13 December 2019. New York: IEEE.
Appendix 1

For ease of notation, we set \( A_f(q_f) = R_p q_f [q_f + (1 - q_f) \times (1 - e^{-\pi \lambda_f q_f d^2})] \) and \( B_f(q_f) = \frac{1}{2} P_c/r_c + P_c/r_c (1 - q_f) \times (1 - e^{-\pi \lambda_f q_f d^2}) \). Taking the first-order derivation of \( A_f(q_f) \) and \( B_f(q_f) \) with respect to \( q_f \), respectively, we can obtain that

\[
\frac{\partial A_f(q_f)}{\partial q_f} = R_p q_f - R_p q_f \left( 1 - e^{-\pi \lambda_f q_f d^2} \right) + R_p q_f \pi \lambda_f d^2 \left( 1 - q_f \right) e^{-\pi \lambda_f q_f d^2} \tag{17}
\]

\[
\frac{\partial B_f(q_f)}{\partial q_f} = P_c R_p [ \left( 1 - e^{-\pi \lambda_f q_f d^2} \right) - \pi \lambda_f d^2 \left( 1 - q_f \right) e^{-\pi \lambda_f q_f d^2} ] \quad \frac{1}{\left( 1 - q_f \right)^2} \times \left( 1 - e^{-\pi \lambda_f q_f d^2} \right)^2 \tag{18}
\]

Furthermore, the second-order derivative of \( A_f(q_f) \) and \( B_f(q_f) \) with respect to \( q_f \) is obtained as

\[
\frac{\partial^2 A_f(q_f)}{\partial q_f^2} = -2R_p q_f \pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} - R_p q_f \left( \pi \lambda_f d^2 \right)^2 \left( 1 - q_f \right) e^{-\pi \lambda_f q_f d^2} \tag{19}
\]

\[
\frac{\partial^2 B_f(q_f)}{\partial q_f^2} = P_c R_p [ \frac{2\left( 1 - e^{-\pi \lambda_f q_f d^2} \right) - \pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} \right]}{(1 - q_f)^2} + \frac{\left( -\pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} \right)}{(1 - q_f)^2} \times \left( 1 - e^{-\pi \lambda_f q_f d^2} \right)^2 \right] \right] \tag{20}
\]

\[
\frac{2\left( 1 - e^{-\pi \lambda_f q_f d^2} \right) - \pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} \right]}{(1 - q_f)^2} - \frac{\pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} \left( 1 - q_f \right)}{(1 - q_f)^2} \left( 1 - e^{-\pi \lambda_f q_f d^2} \right)^3 \right] \]

\[
\frac{\pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2} \left( 1 - q_f \right)}{(1 - q_f)^2} \left( 1 - e^{-\pi \lambda_f q_f d^2} \right)^3 \right] \]

The inequality in Equation (20) comes from scaling a term in the numerator. Denote \( g(q_f) = 2(1 - e^{-\pi \lambda_f q_f d^2})q_f - 2\pi \lambda_f d^2 e^{-\pi \lambda_f q_f d^2}(1 - q_f) \). By taking the first-order derivative of \( g(q_f) \), we obtain \( g'(q_f) \) due to \( 0 \leq q_f \leq 1 \)

\[
g'(q_f) = 2\left( 1 - e^{-\pi \lambda_f q_f d^2} \right) \tag{21}
\]