# Joint Optimization Algorithm Based on DCA for Three-tier Caching in Heterogeneous Cellular Networks

**Jun Zhang[1,2] and Qi Zhu[1,2*]**
[1] Jiangsu Key Laboratory of Wireless Communications, Nanjing University of Posts and Telecommunications
Nanjing 210003, Jiangsu - P. R. China
[e-mail: zhuqi@njupt.edu.cn]
[2] Engineering Research Center of Health Service System Based on Ubiquitous Wireless Networks, Nanjing University of Posts and Telecommunications
Nanjing 210003, Jiangsu - P. R. China
[*]Corresponding author: Qi Zhu

*Received October 30, 2020; revised May 4, 2021; accepted June 24, 2021; published July 31, 2021*

## Abstract

In this paper, we derive the expression of the cache hitting probability with random caching policy and propose the joint optimization algorithm based on difference of convex algorithm (DCA) in the three-tier caching heterogeneous cellular network assisted by macro base stations, helpers and users. Under the constraint of the caching capacity of caching devices, we establish the optimization problem to maximize the cache hitting probability of the network. In order to solve this problem, a convex function is introduced to convert the nonconvex problem to a difference of convex (DC) problem and then we utilize DCA to obtain the optimal caching probability of macro base stations, helpers and users for each content respectively. Simulation results show that when the density of caching devices is relatively low, popular contents should be cached to achieve a good performance. However, when the density of caching devices is relatively high, each content ought to be cached evenly. The algorithm proposed in this paper can achieve the higher cache hitting probability with the same density.

***Keywords:*** cache hitting probability, DCA, edge caching, heterogeneous cellular networks, Poisson point process (PPP)

This work is supported by the National Natural Science Foundation of China (61971239, 92067201), Jiangsu Provincial Key Research and Development Program (No. BE2020084-1).

**http://doi.org/10.3837/tiis.2021.07.019** **ISSN : 1976-7277**

## 1. Introduction

According to a forecast by Cisco, the traffic of the demand for data is increasing exponentially in the coming years [1-2], which will cause the congestion of the network. The request for multimedia services tends to be identical, making contents lots of duplicate transmissions in the network. To release the pressure of the network, mobile edge caching is proposed. Edge caching can store some contents in the storage of edge caching devices (e.g. base stations or helpers) in advance, which makes end-users able to get the requested content from the storage of edge caching devices, instead of from the distant core network [3]. Besides, edge caching can also reduce the traffic of backhaul links and the delay of end-users obtaining the requested contents [2,4]. Due to the fact that edge caching has such advantages, it has attracted a lot of attention.

The caching policy has a great effect on the performance of edge caching networks and some research has been done on it. Aimed at maximizing the secure delivery probability, [5] investigates the secure delivery of files with random caching policy, jointly optimizing the caching probability for each content and the redundant rate. [6] analyses the interference in two cells, and contrary to the traditional approach of only caching the most popular files, it finds out that caching some interfering contents which are likely to get from the neighbor cell can increase the achievable rate of users by exploiting interference neutralization. [7] studies the cache size allocation in backhaul limited cellular networks with the most popular caching (MPC) strategy. In the single-cell scenario, with a user success probability threshold, a closed-form expression of the cache size allocation is derived, and then in the multi-cell scenario, the optimal cache size allocation that maximizes the overall user success probability threshold is obtained. [8] investigates the average energy consumption of individual popularity preferences with random caching policy, pointing out that taking individual popularity preferences into consideration can achieve a better performance in the aspect of average energy consumption. However, these references aforementioned do not analyze the cache hitting probability directly.

The cache hitting probability has become an important performance metric in edge caching system, on which different caching policies and architectures of the network have a vital influence. [9] analyses the cache hitting probability of two-tier caching network composed of users and helpers, employing DCA to obtain the optimal caching probability for each content with the goal of maximizing the cache hitting probability of the network. [10] studies the cache hitting probability of two-tier caching network consisting of helpers and edge servers, and proposes a caching probability conversion algorithm to achieve the solution of optimal caching placement that maximizes the cache hitting probability of the network. [11] investigates the cache hitting probability of D2D caching, finding out that when the request of users follows a Zipf distribution, the optimal caching placement for this network also follows a Zipf distribution. However, [9] does not consider the caching ability of base stations, and the D2D caching is not taken into account in [10], and [11] does not take the base station and helper caching into consideration. In [12], the average ergodic rate and outage probability in the downlink is analyzed in the heterogeneous network with the most popular multimedia contents pushed and cached via broadcasting. In [13], authors propose a combined MPC and LCD caching strategy with joint and parallel cooperative transmission in cluster-centric cache-enabled small cell networks and find that there exists an inherent tradeoff between transmission diversity and content diversity. In [14], the revenue maximization problem for content-oriented wireless caching networks with both repair and recommendation considerations is studied.

In this paper, we consider the three-tier caching network consisting of macro base stations, helpers and users, then deriving the expression of the cache hitting probability of the network, and finally propose the joint optimization algorithm based on DCA for three-tier caching in the heterogeneous cellular network to maximize the cache hitting probability of the network so that it can provide a better experience for users.

The main contributions of this paper are presented as follows:

1. We construct the system model of the three-tier caching composed of macro base stations, helpers and users, assuming that the distribution of macro base stations, helpers and users follows a PPP respectively. When a user makes a request for a content, it checks the self-cache, cache of users, cache of helpers and cache of macro base stations in its vicinity in turn. Exploiting the property of PPP, we derive the expression of the cache hitting probability of the network.

2. Given that the limited caching capacity of caching devices, we establish the optimization problem aimed at maximizing the cache hitting probability of the network. In order to solve this problem conveniently, we first convert this problem to a minimization problem. However, the minimization problem is a nonconvex problem, which is hard to solve. Then, we introduce a convex function to convert the above nonconvex problem to the form of the difference of two convex functions. Finally, we utilize DCA to solve the problem so that the optimal caching probability in all caching devices for each content is obtained.

3. Simulation results show that when the density of caching devices is considerably low, popular contents should be stored to achieve a good performance. However, when the density of caching devices is relatively high, each content ought to be stored with the equal probability. The algorithm proposed in this paper can achieve a better performance under the same circumstance.

The remainder of this paper is organized as follows. In Section 2, we present the system model. In Section 3, we give the analysis of the cache hitting probability. In Section 4, we formulate the problem and propose the caching algorithm based on DCA to solve it. The simulation results and the analyses for them are presented in Section 5. In Section 6, we conclude this paper.

## 2. System Model

As illustrated in **Fig. 1**, the system model of this paper is the three-tier cache-enabled heterogeneous network consisting of macro base stations, helpers and users. Each macro base station, helper and user has the limited caching capacity respectively, denoted by $S_M$ contents, $S_H$ contents and $S_{UE}$ contents, and $S_M > S_H > S_{UE}$. The coverage radius of macro base stations, helpers and users is $R_M$, $R_H$ and $R_{UE}$ respectively, and $R_M >> R_H >> R_{UE}$. We suppose that part of users has the ability to cache contents. Let $\alpha$ denote the proportion of users able to cache contents, where $0 < \alpha \le 1$.

We assume that there are $N$ contents in the content library and that all contents have the same size for simplicity. The popularity distribution of the contents is denoted by $\mathrm{q} = \{q_1, \mathrm{L}\ , q_i, \mathrm{L}\ , q_N\}$, where $q_i$ is the request probability for the content $i$. The popularity distribution is characterized by a Zipf distribution with parameter $\gamma$. If the contents are sorted in the descending order of popularity, the popularity of the content $i$ is [15]

$$q_i = \frac{1/i^{\gamma}}{\sum_{j=1}^{N} 1/j^{\gamma}}, \quad i \in \{1, \mathsf{L}, N\} \tag{1}$$

where $\gamma$ represents the skewness of the popularity. As $\gamma$ increases, the requests of users become more focused on popular contents.

When a user makes a request, it first checks whether the requested content is stored in its own cache. If it stores the requested content, the cache is hit. Otherwise, the user checks whether the requested content is stored in the cache of users in its vicinity. If there exists at least one user in its vicinity storing the requested content, the cache is hit. Otherwise, the user checks whether the requested content is stored in the cache of helpers in its vicinity. If there exists at least one helper in its vicinity storing the requested content, the cache is hit. Otherwise, the user checks whether the requested content is stored in the cache of macro base stations in its vicinity. If there exists at least one macro base station in its vicinity storing the requested content, the cache is hit. Otherwise, the user needs to get the requested content through backhaul links.

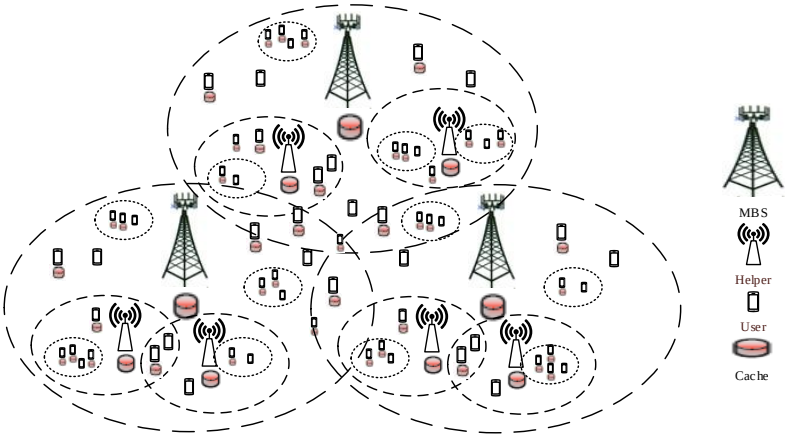


**Fig. 1.** Heterogeneous edge caching model with multiple macro base stations, helpers and users

## 3. Analysis of the Cache Hitting Probability

The cache hitting probability refers to the probability that there is at least one caching device capable of providing the requested content for the user. If any caching device is able to provide the requested content for the user, it is called a cache hit. Otherwise, the user needs to get it via backhaul links, which is seen as a miss. We will analyze the cache hitting probability of the system in the following in that it is an important performance metric of the caching network. Due to the existence of two kinds of users in the network, one is cache-enabled and another is incapable of caching, so we will analyze the cache hitting probability of two kinds of users requesting the content $i$ respectively.

Providing that one device follows a PPP distribution of density $\lambda$, the probability that there exist $n$ devices in the area within the radius $r$ is

$$\Pr(n, r, \lambda) = \frac{(\pi r^2 \lambda)^n}{n!} e^{-\pi r^2 \lambda} \tag{2}$$

We suppose that the positions of macro base stations, helpers and users follow different PPPs of density $\lambda_M$, $\lambda_H$ and $\lambda_{UE}$ respectively, and that $\alpha$ denotes the proportion of cache-enabled users, and that the probability of the macro base station, helper and user caching the content $i$ is $p_i^M$, $p_i^H$ and $p_i^{UE}$ respectively. Therefore, the deployment of macro base stations, helpers and users storing the content $i$ follows different PPPs of density $\lambda_M p_i^M$, $\lambda_H p_i^H$ and $\alpha\lambda_{UE} p_i^{UE}$ respectively.

First, we take analysis of users incapable of caching. Since this kind of users does not have the ability of caching, it can't obtain the requested content without turning to other devices for help. The cases of this kind of users obtaining the requested content are presented as follows:
1) D2D hit: If there exists at least one user storing the requested content within the radius $R_{UE}$ for a specific user, the request will be hit by establishing the D2D communication. According to (2), the probability of the request for the content $i$ is hit via D2D is

$$U(i) = 1 - \Pr(0, R_{UE}, \alpha\lambda_{UE} p_i^{UE}) = 1 - e^{-\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2} \tag{3}$$

2) Helper hit: If D2D does not hit and there exists at least one helper storing the requested content within the radius $R_H$ for a specific user, the request will be hit with the help of helpers. According to (2), the probability of the request for the content $i$ is hit in this way is

$$H(i) = \Pr(0, R_{UE}, \alpha\lambda_{UE} p_i^{UE})[1 - \Pr(0, R_H, \lambda_H p_i^H)] = e^{-\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2}(1 - e^{-\pi\lambda_H p_i^H R_H^2}) \tag{4}$$

3) Macro base station hit: If both D2D and helpers do not hit and there exists at least one macro base station storing the requested content within the radius $R_M$ for a specific user, the request will be hit with the help of macro base stations. According to (2), the probability of the request for the content $i$ is hit in this manner is

$$\begin{aligned} M(i) &= \Pr(0, R_{UE}, \alpha\lambda_{UE} p_i^{UE}) \Pr(0, R_H, \lambda_H p_i^H)[1 - \Pr(0, R_M, \lambda_M p_i^M)] \\ &= e^{-(\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2)}(1 - e^{-\pi\lambda_M p_i^M R_M^2}) \end{aligned} \tag{5}$$

Hence, the cache hitting probability of cache-disabled users for the content $i$ is

$$P_{UN}(i) = U(i) + H(i) + M(i) = 1 - e^{-(\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2 + \pi\lambda_M p_i^M R_M^2)} \tag{6}$$

Compared with cache-disabled users, the cache-enabled users are able to cache, so it has the potential to obtain the requested content from the self-cache without the help of other devices. Due to the fact that, to some extent, the cache-enabled user is equivalent to the cache-disabled user when the cache-enabled user does not cache its requested content, the analysis of other cases of hit for cache-enabled users is similar to that of cache-disabled users. Thus, the cache hitting probability of cache-enabled users for the content $i$ is

$$P_C(i) = p_i^{UE} + (1 - p_i^{UE}) P_{UN}(i) \tag{7}$$

Therefore, the cache hitting probability for the content $i$ is

$$\begin{aligned} P_{i,hit} &= \alpha P_C(i) + (1 - \alpha) P_{UN}(i) \\ &= 1 - (1 - \alpha p_i^{UE}) e^{-(\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2 + \pi\lambda_M p_i^M R_M^2)} \end{aligned} \tag{8}$$

To sum up, the average cache hitting probability for the three-tier cache-enabled heterogeneous network is

$$P_{hit} = \sum_{i=1}^{N} q_i P_{i,hit} \tag{9}$$

where $q_i$ denotes the request probability of users for the content $i$.

## 4. Problem Formulation and Solution

### 4.1 Problem Formulation

The goal of this paper is to find the optimal caching probability for each content that maximizes the cache hitting probability of the network. We assume that the caching devices in the same tier cache the same content with the equal probability. For convenience in the following description, let $\mathbf{P}=[\mathbf{P}_{\mathrm{UE}}\ \mathbf{P}_{\mathrm{H}}\ \mathbf{P}_{\mathrm{M}}]$ denote the caching probability in the user tier, helper tier and macro base station tier, where $\mathbf{P}_{\mathrm{UE}}=[p_1^{UE},p_2^{UE},\cdots,p_i^{UE},\cdots,p_N^{UE}]$, $\mathbf{P}_{\mathrm{H}}=[p_1^{H},p_2^{H},\cdots,p_i^{H},\cdots,p_N^{H}]$ and $\mathbf{P}_{\mathrm{M}}=[p_1^{M},p_2^{M},\cdots,p_i^{M},\cdots,p_N^{M}]$. Under the constraint of caching capacity of caching devices in each tier, the optimization problem for maximizing the cache hitting probability of the network can be formulated as the following problem $\mathbf{P1}$:

$$\mathbf{P1}: \max_{\mathbf{P}} \sum_{i=1}^{N} q_i P_{i,hit}$$
$$s.t. \begin{cases} \sum_{i=1}^{N} p_i^{UE} \le S_{UE} \\ \sum_{i=1}^{N} p_i^{H} \le S_{H} \\ \sum_{i=1}^{N} p_i^{M} \le S_{M} \\ 0 \le p_i^{UE} \le 1, i \in \{1,\cdots,N\} \\ 0 \le p_i^{H} \le 1, i \in \{1,\cdots,N\} \\ 0 \le p_i^{M} \le 1, i \in \{1,\cdots,N\} \end{cases} \tag{10}$$

### 4.2 Solution to the Problem

The optimization problem in this paper aims to find the optimal caching probability for each content in the three-tier caching network maximizing the cache hitting probability of the network. Due to the fact that it is difficult to solve the original problem directly, we convert this original problem to a minimization problem at first. Unfortunately, the transformed minimization problem is a nonconvex problem. Then, we introduce a convex function and exploit it to convert the nonconvex problem above to the difference of two convex functions. Finally, DC programming is adopted to solve the problem so that we can obtain the optimal caching probability for each content.

First, we convert the problem $\mathbf{P1}$ to a minimization problem $\mathbf{P2}$:

$$\mathbf{P2:}\ \min_{\mathbf{P}}\ -\sum_{i=1}^{N} q_i P_{i,hit}$$
$$s.t.\ \begin{cases} \sum_{i=1}^{N} p_i^{UE} \le S_{UE} \\ \sum_{i=1}^{N} p_i^{H} \le S_H \\ \sum_{i=1}^{N} p_i^{M} \le S_M \\ 0 \le p_i^{UE} \le 1, i \in \{1,\mathrm{L}\ ,N\} \\ 0 \le p_i^{H} \le 1, i \in \{1,\mathrm{L}\ ,N\} \\ 0 \le p_i^{M} \le 1, i \in \{1,\mathrm{L}\ ,N\} \end{cases} \tag{11}$$

Let $F(\mathbf{P}) = -\sum_{i=1}^{N} q_i P_{i,hit} = \sum_{i=1}^{N} q_i f_i = \sum_{i=1}^{N} q_i[(1-\alpha p_i^{UE})e^{-(\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2 + \pi\lambda_M p_i^M R_M^2)} - 1]$ denote the objective function of problem **P2**, where $f_i = (1-\alpha p_i^{UE})e^{-(\pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2 + \pi\lambda_M p_i^M R_M^2)} - 1$. Let $F_i$ denote the Hessian matrix of $f_i$ and we can derive $F_i$ as follows:

$$F_i = \begin{bmatrix} \frac{\partial^2 f_i}{\partial (p_i^{UE})^2} & \frac{\partial^2 f_i}{\partial p_i^{UE} \partial p_i^{H}} & \frac{\partial^2 f_i}{\partial p_i^{UE} \partial p_i^{M}} \\ \frac{\partial^2 f_i}{\partial p_i^{H} \partial p_i^{UE}} & \frac{\partial^2 f_i}{\partial (p_i^{H})^2} & \frac{\partial^2 f_i}{\partial p_i^{H} \partial p_i^{M}} \\ \frac{\partial^2 f_i}{\partial p_i^{M} \partial p_i^{UE}} & \frac{\partial^2 f_i}{\partial p_i^{M} \partial p_i^{H}} & \frac{\partial^2 f_i}{\partial (p_i^{M})^2} \end{bmatrix} \tag{12}$$
$$= \begin{bmatrix} \alpha^2 Y[(1-\alpha p_i^{UE})Y+2]e^{-X} & \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & \alpha K[(1-\alpha p_i^{UE})Y+1]e^{-X} \\ \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & Z^2(1-\alpha p_i^{UE})e^{-X} & ZK(1-\alpha p_i^{UE})e^{-X} \\ \alpha K[(1-\alpha p_i^{UE})Y+1]e^{-X} & ZK(1-\alpha p_i^{UE})e^{-X} & K^2(1-\alpha p_i^{UE})e^{-X} \end{bmatrix}$$

where $X = \pi\alpha\lambda_{UE} p_i^{UE} R_{UE}^2 + \pi\lambda_H p_i^H R_H^2 + \pi\lambda_M p_i^M R_M^2 \ge 0$, $Y = \pi\lambda_{UE} R_{UE}^2$, $Z = \pi\lambda_H R_H^2$ and $K = \pi\lambda_M R_M^2$.

**Proposition 1**: $F(\mathbf{P})$ is nonconvex of $\mathbf{P}$.

**Proof**: Due to the fact that any order leading principal minor of a given matrix which is greater than zero is equivalent to the matrix being positive definite, when one order leading principal minor of a given matrix is not greater than zero, this matrix is not positive definite.
1) The 1st order leading principal minor of $F_i$ is

$$D_{F1} = \alpha^2 Y[(1-\alpha p_i^{UE})Y+2]e^{-X} > 0 \tag{13}$$

2) The 2nd order leading principal minor of $F_i$ is

$$D_{F2}=\begin{vmatrix}\alpha^2 Y[(1-\alpha p_i^{UE})Y+2]e^{-X} & \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X}\\ \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & Z^2(1-\alpha p_i^{UE})e^{-X}\end{vmatrix}=-\alpha^2 Z^2 e^{-2X}<0 \quad (14)$$

According to (14), the 2nd order leading principal minor of $F_i$ is less than zero, so the matrix $F_i$ is not positive definite.

Hence, $f_i$ is nonconvex of $\mathbf{P}$. Since $F(\mathbf{P})$ is the linear combination of $f_i$, $F(\mathbf{P})$ is also nonconvex of $\mathbf{P}$. So, **Proposition 1** is proved.

Then, we introduce a new function $H(\mathbf{P})=\sum_{i=1}^{N} q_i h_i$, where $h_i=(\alpha\pi\lambda_H R_H^2+\alpha\pi\lambda_M R_M^2)(p_i^{UE^2}+p_i^{H^2}+p_i^{M^2})$. Let $H_i$ denote the Hessian matrix of $h_i$ and $H_i$ is derived as follows:

$$H_i=\begin{bmatrix}\frac{\partial^2 h_i}{\partial(p_i^{UE})^2} & \frac{\partial^2 h_i}{\partial p_i^{UE}\partial p_i^{H}} & \frac{\partial^2 h_i}{\partial p_i^{UE}\partial p_i^{M}}\\ \frac{\partial^2 h_i}{\partial p_i^{H}\partial p_i^{UE}} & \frac{\partial^2 h_i}{\partial(p_i^{H})^2} & \frac{\partial^2 h_i}{\partial p_i^{H}\partial p_i^{M}}\\ \frac{\partial^2 h_i}{\partial p_i^{M}\partial p_i^{UE}} & \frac{\partial^2 h_i}{\partial p_i^{M}\partial p_i^{H}} & \frac{\partial^2 h_i}{\partial(p_i^{M})^2}\end{bmatrix} \quad (15)$$

$$=\begin{bmatrix}2\alpha\pi(\lambda_H R_H^2+\lambda_M R_M^2) & 0 & 0\\ 0 & 2\alpha\pi(\lambda_H R_H^2+\lambda_M R_M^2) & 0\\ 0 & 0 & 2\alpha\pi(\lambda_H R_H^2+\lambda_M R_M^2)\end{bmatrix}$$

Thus, $H_i$ is a positive definite matrix and $h_i$ is convex of $\mathbf{P}$. Since $H(\mathbf{P})$ is the linear combination of $h_i$ and the combination coefficient $q_i$ is greater than zero, $H(\mathbf{P})$ is convex of $\mathbf{P}$.

Afterwards, let $G(\mathbf{P})=F(\mathbf{P})+H(\mathbf{P})=\sum_{i=1}^{N} q_i g_i$, where $g_i=(1-\alpha p_i^{UE})e^{-(\pi\alpha\lambda_{UE}p_i^{UE}R_{UE}^2+\pi\lambda_H p_i^H R_H^2+\pi\lambda_M p_i^M R_M^2)}$ -1+$(\alpha\pi\lambda_H R_H^2+\alpha\pi\lambda_M R_M^2)(p_i^{UE^2}+p_i^{H^2}+p_i^{M^2})$. Let $G_i$ denote the Hessian matrix of $g_i$ and we can derive $G_i$ as follows:

$$G_i=\begin{bmatrix}\frac{\partial^2 g_i}{\partial(p_i^{UE})^2} & \frac{\partial^2 g_i}{\partial p_i^{UE}\partial p_i^{H}} & \frac{\partial^2 g_i}{\partial p_i^{UE}\partial p_i^{M}}\\ \frac{\partial^2 g_i}{\partial p_i^{H}\partial p_i^{UE}} & \frac{\partial^2 g_i}{\partial(p_i^{H})^2} & \frac{\partial^2 g_i}{\partial p_i^{H}\partial p_i^{M}}\\ \frac{\partial^2 g_i}{\partial p_i^{M}\partial p_i^{UE}} & \frac{\partial^2 g_i}{\partial p_i^{M}\partial p_i^{H}} & \frac{\partial^2 g_i}{\partial(p_i^{M})^2}\end{bmatrix} \quad (16)$$

$$=\begin{bmatrix}\alpha^2 Y[(1-\alpha p_i^{UE})Y+2]e^{-X}+2\alpha(Z+K) & \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & \alpha K[(1-\alpha p_i^{UE})Y+1]e^{-X}\\ \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & Z^2(1-\alpha p_i^{UE})e^{-X}+2\alpha(Z+K) & ZK(1-\alpha p_i^{UE})e^{-X}\\ \alpha K[(1-\alpha p_i^{UE})Y+1]e^{-X} & ZK(1-\alpha p_i^{UE})e^{-X} & K^2(1-\alpha p_i^{UE})e^{-X}+2\alpha(Z+K)\end{bmatrix}$$

where $X=\pi\alpha\lambda_{UE}p_i^{UE}R_{UE}^2+\pi\lambda_H p_i^H R_H^2+\pi\lambda_M p_i^M R_M^2\geq 0$, $Y=\pi\lambda_{UE}R_{UE}^2$, $Z=\pi\lambda_H R_H^2$ and $K=\pi\lambda_M R_M^2$.

**Proposition 2**: $G(\mathbf{P})$ is convex of $\mathbf{P}$.

**Proof**: According to the necessary and sufficient condition of positive definite matrix, when any order leading principal minor of a given matrix is greater than zero, this matrix is positive definite.

1) The 1st order leading principal minor of $G_i$ is

$$D_{G1} = \alpha^2 Y[(1-\alpha p_i^{UE})Y + 2]e^{-X} + 2\alpha(Z+K) > 0 \tag{17}$$

2) The 2nd order leading principal minor of $G_i$ is

$$\begin{aligned} D_{G2} &= \begin{vmatrix} \alpha^2 Y[(1-\alpha p_i^{UE})Y+2]e^{-X} + 2\alpha(Z+K) & \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} \\ \alpha Z[(1-\alpha p_i^{UE})Y+1]e^{-X} & Z^2(1-\alpha p_i^{UE})e^{-X} + 2\alpha(Z+K) \end{vmatrix} \\ &= 4\alpha^2(Z+K)^2 - \alpha^2 Z^2 e^{-2X} + 2\alpha(Z+K)[2\alpha^2 Y + (1-\alpha p_i^{UE})(Z^2+\alpha^2 Y^2)]e^{-X} \end{aligned} \tag{18}$$

Due to

$$\begin{aligned} 4\alpha^2(Z+K)^2 - \alpha^2 Z^2 e^{-2X} &= [2\alpha(Z+K)]^2 - (\alpha Z e^{-X})^2 \\ &= [2\alpha(Z+K) + \alpha Z e^{-X}][2\alpha(Z+K) - \alpha Z e^{-X}] \\ &= [2\alpha(Z+K) + \alpha Z e^{-X}][\alpha Z(2-e^{-X}) + 2\alpha K] > 0 \end{aligned} \tag{19}$$

and

$$2\alpha(Z+K)[2\alpha^2 Y + (1-\alpha p_i^{UE})(Z^2+\alpha^2 Y^2)]e^{-X} > 0 \tag{20}$$

therefore

$$D_{G2} > 0 \tag{21}$$

3) The 3rd order leading principal minor of $G_i$ is

$$\begin{aligned} D_{G3} = |G_i| &= 8\alpha^3(Z+K)^3 - 2\alpha^3(Z+K)(Z^2+K^2)e^{-2X} \\ &+ 4\alpha^2(Z+K)^2[2\alpha^2 Y + (1-\alpha p_i^{UE})(\alpha^2 Y^2 + K^2 + Z^2)]e^{-X} \end{aligned} \tag{22}$$

Due to

$$0 < (Z+K)(Z^2+K^2) < (Z+K)^3, 0 < e^{-2X} \le 1 \Rightarrow (Z+K)(Z^2+K^2)e^{-2X} < (Z+K)^3 \tag{23}$$

$$8\alpha^3(Z+K)^3 - 2\alpha^3(Z+K)(Z^2+K^2)e^{-2X} > 6\alpha^3(Z+K)^3 > 0 \tag{24}$$

and

$$4\alpha^2(Z+K)^2[2\alpha^2 Y + (1-\alpha p_i^{UE})(\alpha^2 Y^2 + K^2 + Z^2)]e^{-X} > 0 \tag{25}$$

hence

$$D_{G3} > 0 \tag{26}$$

According to (17), (21) and (26), any order leading principal minor of $G_i$ is greater than zero, so $G_i$ is a positive definite matrix and $g_i$ is convex of $\mathbf{P}$. $G(\mathbf{P})$ is also convex of $\mathbf{P}$ in that $G(\mathbf{P})$ is the linear combination of $g_i$. So, **Proposition 2** is proved.

Thus, $F(\mathbf{P})$ can be denoted as a difference of the two convex functions $G(\mathbf{P})$ and $H(\mathbf{P})$:

$$F(\mathbf{P}) = G(\mathbf{P}) - H(\mathbf{P}) \tag{27}$$

Due to the fact that $\dfrac{\partial H(\mathbf{P})}{\partial \mathbf{P}}$ is continuous and the constraint of problem **P2** is a convex set, we can adopt the DC programming to solve this problem. DCA can solve the DC programming well, which is usually able to obtain the global optimal solution of a nonconvex function in the form of the difference of two convex functions and has a quick convergence. Therefore,

we utilize DCA to solve this problem, as described in **Algorithm 1**.

**Algorithm 1**. Joint optimization algorithm based on DCA

1: Initialize: $\mathbf{P}_0^{UE}=\frac{S_{UE}}{N},\mathbf{P}_0^{H}=\frac{S_H}{N},\mathbf{P}_0^{M}=\frac{S_M}{N}$;

2: Solve the convex optimization problem:

$$\min_{\mathbf{P}} G(\mathbf{P})-H(\mathbf{P}_k)-(\mathbf{P}-\mathbf{P}_k)\frac{\partial H(\mathbf{P}_k)}{\partial \mathbf{P}}$$

$$s.t.\begin{cases}\sum_{i=1}^{N} p_i^{UE}\le S_{UE}\\ \sum_{i=1}^{N} p_i^{H}\le S_H\\ \sum_{i=1}^{N} p_i^{M}\le S_M\\ 0\le p_i^{UE}\le 1, i\in\{1,\text{L}\ ,N\}\\ 0\le p_i^{H}\le 1, i\in\{1,\text{L}\ ,N\}\\ 0\le p_i^{M}\le 1, i\in\{1,\text{L}\ ,N\}\end{cases};$$

3: The solution of step 2 is $\mathbf{P}_{k+1}$;

4: If $\left|F(\mathbf{P}_k)-F(\mathbf{P}_{k+1})\right|\le\delta$ or $\left|\mathbf{P}_k-\mathbf{P}_{k+1}\right|\le\delta$ with $\delta$ set as 0.00001 in this paper, $\mathbf{P}_k$ is the optimal solution of $F(\mathbf{P})$; otherwise, return to step2;

5: Return: the result is: $-F(\mathbf{P}_k)$, and the solution is $\mathbf{P}_k$.

## 5. Simulations

In this section, we exploit MATLAB to provide some numerical results of the proposed algorithm with the system model illustrated in **Fig. 1**. We suppose that the communication range of D2D, helpers and macro base stations is 15 meters, 100 meters and 500 meters respectively and that the size of content library is 30, with other specific simulation parameters listed in **Table 1**. In order to evaluate the algorithm proposed in this paper, we compare the performance of the proposed algorithm with that of the MPC policy, the equal probability random caching (EPRC) policy and the algorithm in [9] of which caching system is two-tier caching composed of users and helpers with base station incapable of caching. We will discuss the impact of the density of users $\lambda_{UE}$, the density of helpers $\lambda_H$, the density of macro base stations $\lambda_M$, the skewness of the popularity $\gamma$, the size of content library $N$ and the proportion of cache-enabled users $\alpha$ on the cache hitting probability respectively.

**Table 1.** Parameter settings

| Parameter | Value |
|---|---|
| Caching capacity of macro base stations: $S_M$ (contents) | 10 |
| Caching capacity of helpers: $S_H$ (contents) | 5 |
| Caching capacity of users: $S_{UE}$ (contents) | 2 |
| Proportion of cache-enabled users: $\alpha$ | 0.5 |
| Skewness of popularity: $\gamma$ | 1 |
| Density of macro base stations: $\lambda_M$ (/m$^2$) | $1/\pi 500^2$ |
| Density of helpers: $\lambda_H$ (/m$^2$) | $25/\pi 500^2$ |
| Density of users: $\lambda_{UE}$ (/m$^2$) | $5000/\pi 500^2$ |

**Fig. 2** shows that the cache hitting probability increases with user density $\lambda_{UE}$. It can be seen from this figure that the performance of the proposed algorithm outperforms that of other three caching placements at any density $\lambda_{UE}$. As $\lambda_{UE}$ increases, the gap between the proposed algorithm and the algorithm in [9] is gradually narrowing. When $\lambda_{UE}$ is considerable large, the algorithm in [9] approximates to the proposed algorithm. That is because the requested contents of users are mainly obtained from cache-enabled users via D2D and seldom obtained from macro base stations when $\lambda_{UE}$ is relatively large. Whether the macro base stations are capable of caching is becoming an unimportant factor in this instance, so the performance gap of the two algorithms is gradually diminishing.

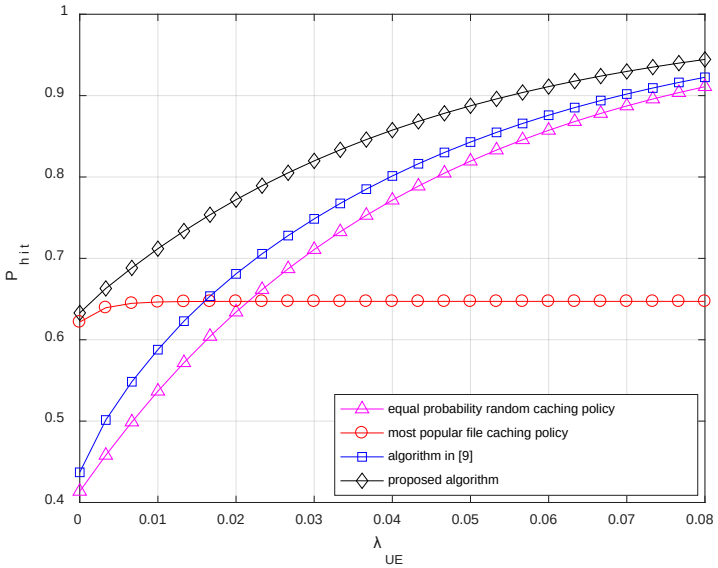


**Fig. 2.** The impact of $\lambda_{UE}$ on the cache hitting probability, where $\lambda_M = 1/\pi 500^2$, $\lambda_H = 25/\pi 500^2$.

**Fig. 3** shows the impact of helper density $\lambda_H$ on the cache hitting probability. It can be found from this figure that the performance of the proposed algorithm outperforms that of other three caching placements at any density $\lambda_H$. With the increase of $\lambda_H$, the gap between the proposed algorithm and the algorithm in [9] is gradually becoming smaller. When $\lambda_H$ is considerable large, the algorithm in [9] approximates to the proposed algorithm. That is because the requested contents of users are mainly obtained from helpers and seldom obtained from macro base stations when $\lambda_H$ is relatively large. Hence, whether the macro base stations are capable of caching is not a notable factor in this situation, so the performance gap of the two algorithms is gradually narrowing.

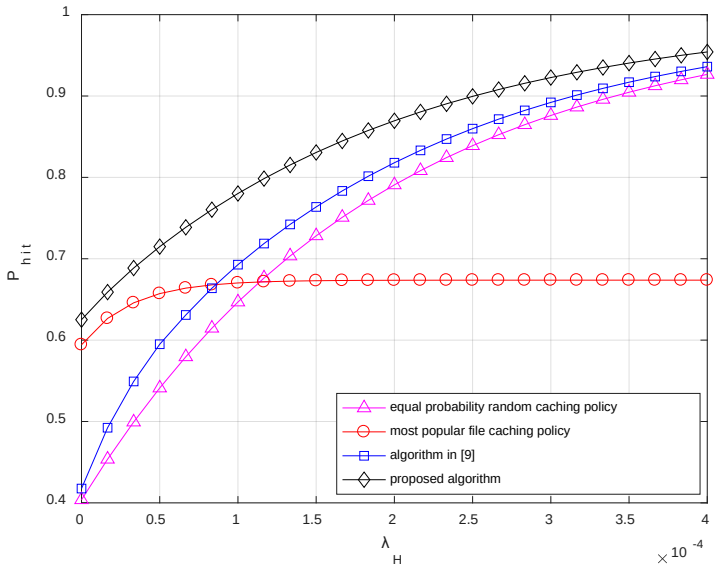


**Fig. 3.** The impact of $\lambda_H$ on the cache hitting probability, where $\lambda_M = 1/\pi 500^2$, $\lambda_{UE} = 5000/\pi 500^2$.

**Fig. 4** shows the impact of macro base station density $\lambda_M$ on the cache hitting probability. It can be seen from this figure that the performance of the proposed algorithm outperforms that of other three caching placements at any density $\lambda_M$. When $\lambda_M = 0$, the performance of the proposed algorithm is equal to that of the algorithm in [9], because both of them are the optimal caching placement for the two-tier caching system consisting of users and helpers without macro base stations joining in caching contents in this instance. However, with the increase of $\lambda_M$, the performance of the proposed algorithm improves distinctly while the performance of the algorithm in [9] is not affected by $\lambda_M$. That is because the caching ability of macro base stations is not taken into account in [9] all the time.

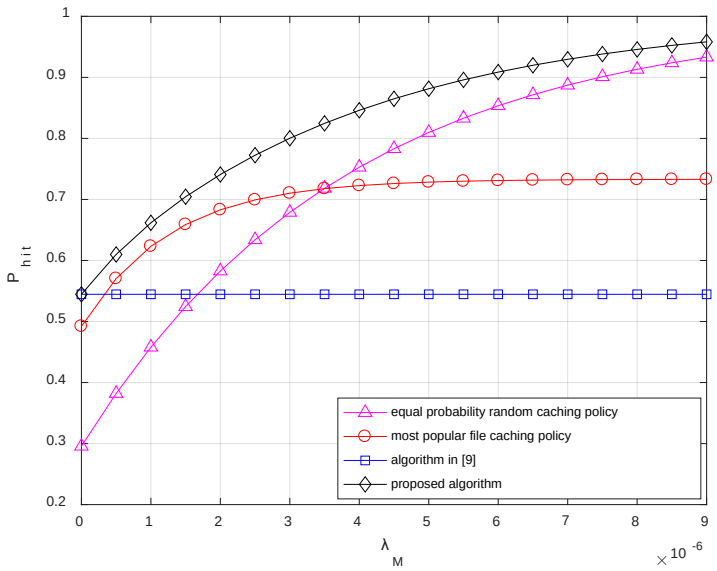


**Fig. 4.** The impact of $\lambda_M$ on the cache hitting probability, where $\lambda_H = 25/\pi 500^2$, $\lambda_{UE} = 5000/\pi 500^2$.

Moreover, from **Fig. 2**, **Fig. 3** and **Fig. 4**, we can see that when all of $\lambda_{UE}$, $\lambda_H$ and $\lambda_M$ are relatively small, the performance of the EPRC policy is the worst one. With the increase of $\lambda_{UE}$, $\lambda_H$ or $\lambda_M$, the performance of the EPRC scheme becomes better and gradually outperforms the MPC policy. That is because when the density of caching devices is relatively low in the network, popular contents need to be stored by users, helpers and macro base stations to tackle the large request probabilities. Thus, the MPC scheme performs well in this situation; When there are many caching devices in the network, the requests for the most popular contents can be hit easily, and the remaining storage of caching devices can be utilized to cache other less popular contents. Therefore, the EPRC scheme performs well under this circumstance. Besides, we can also find that the performance of the MPC policy no longer improves when the density of caching devices becomes relatively high in that the MPC policy only caches these popular contents all the time, regardless of other less popular contents, leading to many redundant contents in the storage of caching devices, and the contents in the low ranking do not have the chance to be stored, hence unable to be hit. Thus, the corresponding performance does not improve any more. When $\lambda_{UE}$, $\lambda_H$ or $\lambda_M$ is relatively large, the performance of the EPRC scheme approximates to the optimal caching placement. That is because the EPRC scheme gives an overall consideration in all contents and most of the requests of users can be hit in this instance.

In **Fig. 5**, we demonstrate the relationship between the caching probability and the user density $\lambda_{UE}$, which is figured out by the proposed algorithm, where $N=4$, $S_{UE}=1$, $S_H=2$, $S_M=3$. It can be seen from this figure that with the increase of $\lambda_{UE}$, the optimal caching placement changes from the MPC policy to the EPRC policy, which is consistent with our analysis above.

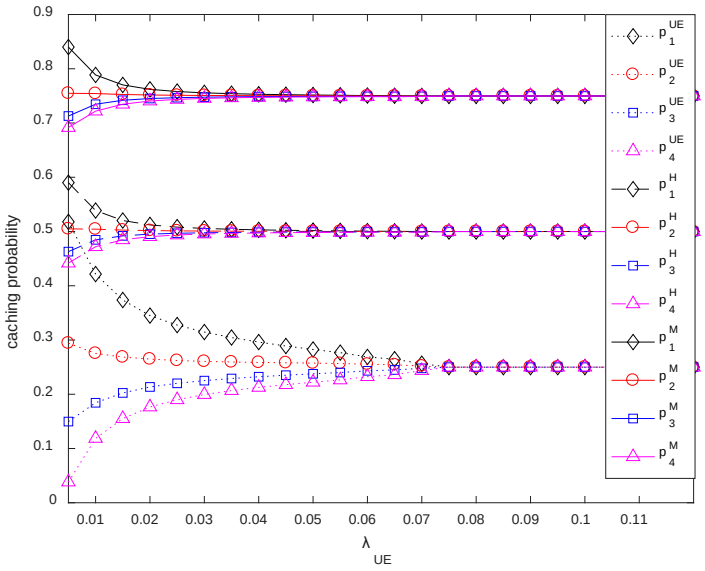


**Fig. 5.** The caching placement of the proposed algorithm, where $N=4$, $S_{UE}=1$, $S_H=2$, $S_M=3$.

**Fig. 6** shows the impact of the skewness of the popularity $\gamma$ on the cache hitting probability. It can be seen from this figure that when $\gamma=0$, the performance of the EPRC scheme is equal to that of the proposed algorithm, because each content has an equal request probability and caching each content with the same probability can achieve a good performance in this situation. When $\gamma$ increases, the requests of users become more focused on the popular contents and the network will store these popular contents with large probabilities. With the increase of $\gamma$, the performance of the MPC scheme improves rapidly while the performance of the EPRC scheme remains unchanged in that it caches each content with the equal probability all the time, regardless of the users' preferences.

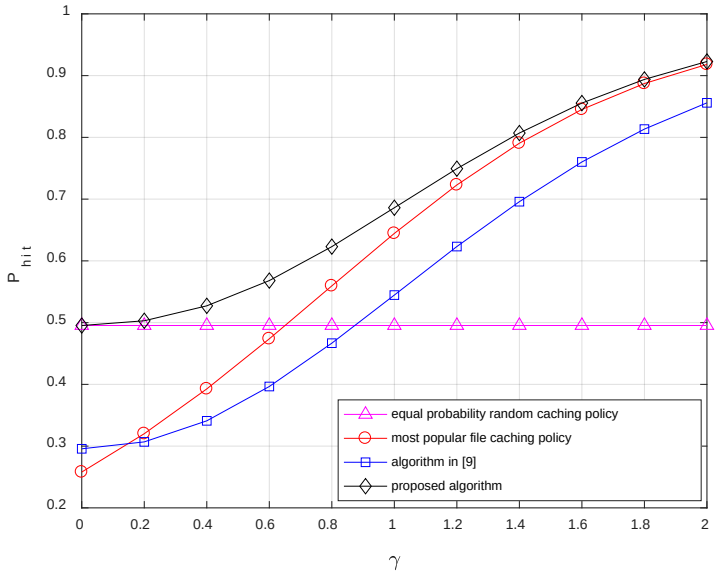


**Fig. 6.** The impact of $\gamma$ on the cache hitting probability.

**Fig. 7** shows the impact of the size of content library $N$ on the cache hitting probability. It can be found from this figure that the cache hitting probability decreases with $N$. We can also find that with the increasing of $N$, the proposed algorithm has a more distinct advantage over other three caching placements, indicating that the proposed algorithm can adjust the caching probability for each content well in the multi-contents situation by a joint optimization in order to keep a good performance.

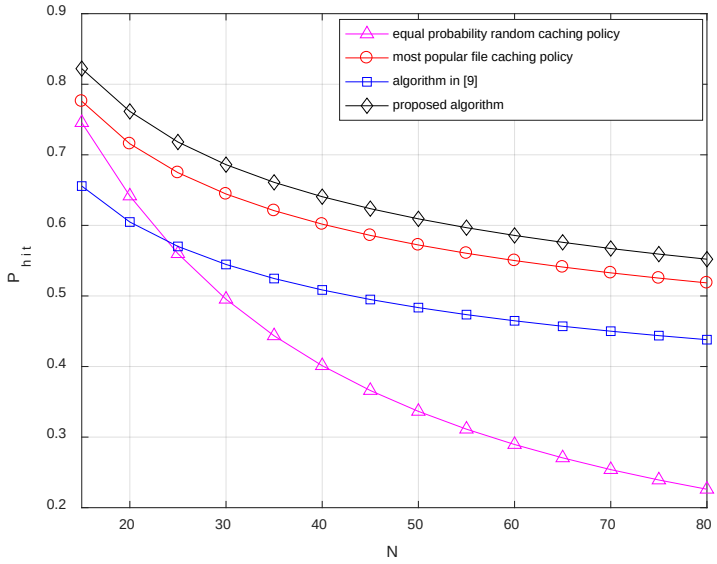


**Fig. 7.** The impact of $N$ on the cache hitting probability.

**Fig. 8** shows the impact of the proportion of cache-enabled users $\alpha$ on the cache hitting probability. It can be seen from this figure that the performance of the proposed algorithm outperforms that of other three caching placements at any proportion $\alpha$. We can also find that the performance of the MPC policy no longer improves when $\alpha$ becomes relatively large in that the MPC policy only caches the popular contents all the time and the caching capacity of users is very limited, leading to many redundant contents in the storage of cache-enabled users, and the contents in the low ranking do not have the chance to be cached, therefore unable to be hit. Even if all the users are capable of caching, the network can only satisfy very few requests. Hence, the corresponding performance does not improve any more.

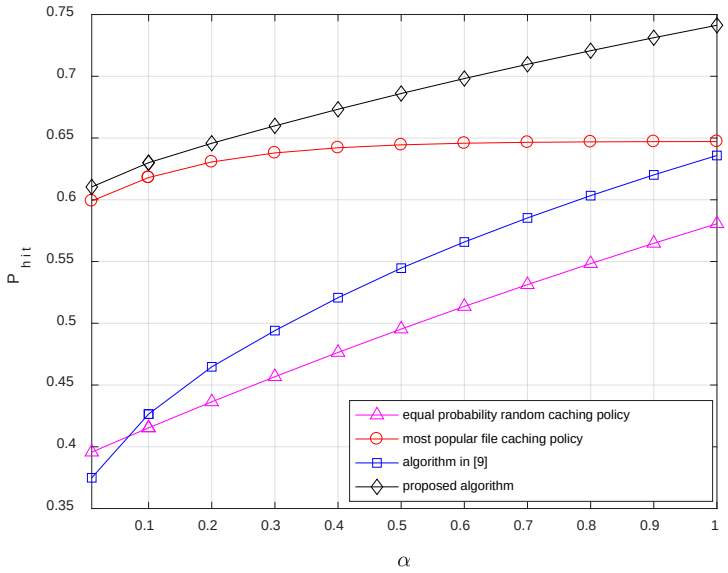


**Fig. 8.** The impact of $\alpha$ on the cache hitting probability.

## 6. Conclusion

In this paper, we derive the expression of the cache hitting probability with random caching policy and the optimization problem is established to obtain the optimal caching placement that maximizes the cache hitting probability of the three-tier caching heterogeneous cellular network assisted by macro base stations, helpers and D2D. Specially, we convert the caching placement problem to a DC problem and employ DCA to solve it. Simulation results show that when the density of caching devices is considerably low, popular contents should be stored for this network. However, when the density of caching devices is considerably high, each content ought to be cached uniformly. The joint optimization algorithm based on DCA proposed in this paper can keep a balance between the MPC policy and the EPRC policy in order to achieve a good performance in terms of the cache hitting probability.

## References

[1] Cisco, “Cisco visual networking index: forecast and trends, 2017-2022,” Cisco Systems Inc., San Jose, CA, USA, 2018.

[2] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. of 2015 IEEE International Conference on Communications*, pp. 3358-3363, Jun. 8-12, 2015. Article (CrossRef Link)
[3] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao and X. Shen, "Cooperative Edge Caching in User-Centric Clustered Mobile Networks," *IEEE Transactions on Mobile Computing*, vol. 17, no. 8, pp. 1791-1805, Aug. 2018. Article (CrossRef Link)
[4] K. Poularakis, G. Iosifidis, A. Argyriou and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *Proc. of 2014 IEEE Conference on Computer Communications*, pp. 1078-1086, Apr. 27-May 2, 2014. Article (CrossRef Link)
[5] S. Zhang, W. Sun, J. Liu and K. Nei, "Physical Layer Security in Large Scale Probabilistic Caching: Analysis and Optimization," *IEEE Communications Letters*, vol. 23, no. 9, pp. 1484-1487, Sept. 2019. Article (CrossRef Link)
[6] J. Song, H. Song and W. Choi, "Which One Is Better to Cache: Requested Contents or Interfering Contents?," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 861-864, Jun. 2019. Article (CrossRef Link)
[7] X. Peng, J. Zhang, S. H. Song and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. of 2016 IEEE International Conference on Communications*, pp. 1-6, May 22-27, 2016. Article (CrossRef Link)
[8] M. Lee and A. F. Molisch, "Individual Preference Aware Caching Policy Design for Energy Efficient Wireless D2D Communications," in *Proc. of 2017 IEEE Global Communications Conference*, pp. 1-7, Dec. 4-8, 2017. Article (CrossRef Link)
[9] J. Rao, H. Feng, C. Yang, Z. Chen and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. of 2016 IEEE International Conference on Communications*, pp. 1-6, May 22-27, 2016. Article (CrossRef Link)
[10] S. Zhang and J. Liu, "Optimal Probabilistic Caching in Heterogeneous IoT Networks," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3404-3414, Apr. 2020. Article (CrossRef Link)
[11] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *Proc. of 2014 IEEE Globecom Workshops*, pp. 863-868, Dec. 8-12, 2014. Article (CrossRef Link)
[12] C. Yang, Y. Yao, Z. Chen and B. Xia, "Analysis on Cache-Enabled Wireless Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 1, pp. 131-145, Jan. 2016. Article (CrossRef Link)
[13] Z. Chen, J. Lee, T. Q. S. Quek and M. Kountouris, "Cooperative Caching and Transmission Design in Cluster-Centric Small Cell Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3401-3415, May 2017. Article (CrossRef Link)
[14] Y. Fu, Q. Yu, T. Q. S. Quek and W. Wen, "Revenue Maximization for Content-Oriented Wireless Caching Networks (CWCNs) With Repair and Recommendation Considerations," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 284-298, Jan. 2021. Article (CrossRef Link)
[15] K. Wang, Z. Chen and H. Liu, "Push-Based Wireless Converged Networks for Massive Multimedia Content Delivery," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2894-2905, May 2014. Article (CrossRef Link)
[16] An Le Thi Hoai, and P. D. Tao, "The DC (Difference of Convex Functions) Programming and DCA Revisited with DC Models of Real World Nonconvex Optimization Problems," *Annals of Operations Research*, vol. 133, no. 1-4, pp. 23-46, 2005. Article (CrossRef Link)
[17] Thi Hoai An Le, V. N. Huynh, and T. P. Dinh, "DC Programming and DCA for General DC Programs," in *Proc. of the 2nd International Conference on Computer Science, Applied Mathematics and Applications*, pp. 15-35, May 8-9, 2014. Article (CrossRef Link)
[18] S. Boyd, and L. Vandenberghe, *Convex Optimization*, New York, NY, USA: Cambridge University Press, 2004. Article (CrossRef Link)

**Jun Zhang** received the bachelor degree in Telecommunication Engineering from Nantong University, Nantong, China, in 2019. He is currently working toward the master degree in Communication and Information System of Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China. His research interests include heterogeneous networks, wireless caching networks and edge computing.

**Qi Zhu** received the bachelor and master degree in Radio Engineering from Nanjing University of Posts and Telecommunications (NUPT), China, in 1986 and 1989, respectively. She is now a full-time professor in the School of Telecommunication and Information Engineering of NUPT. Her research interests focus on technology of next generation communication, broadband wireless access, OFDM, channel and source coding, and dynamic allocation of radio resources.