Joint Optimization in Cached-Enabled Heterogeneous Network for Efficient Industrial IoT

Jiachen Yang, *Member, IEEE*, Chaofan Ma, Bin Jiang, Guiguang Ding, *Member, IEEE*, Gan Zheng, *Senior Member, IEEE*, and Huihui Wang, *Senior Member, IEEE*

***Abstract*—In the era of industrial 4.0, industrial Internet of Things (IIoT) has brought essential changes to human society. For IIoT, communication in network can be defined as the basic condition for further development and integrated information exchange. In this way, cached-enabled heterogeneous industrial network is necessary to be optimized. In this paper, we consider the optimal geographical placement of contents in cache-enabled heterogeneous networks to minimize the total missing probability. And the probability represents that typical user cannot find requested file in the nearby base stations (BSs). In contract to existing works which only concern content placement, we jointly optimize content placement at BSs and activation densities of BSs of different tiers subject to the cache size limits and the constraint on the BSs energy consumption cost. In addition, the user distribution in this work is modeled by a homogeneous Poisson Point Process. We prove that the original optimization problem can be transformed to a convex problem. The convexity of the optimization problem allows us to apply the KKT conditions to derive useful analytical results of the optimal solution. Based on this, we propose a low-complexity near-optimal algorithm to find the approximated content placement probabilities. We further extend the optimization to heterogeneous networks with the user distribution modeled by the modified Cluster Process. Extensive simulation results show the superior performance of joint optimization of content placement and BSs activation densities compared to only optimizing content placement.**

***Index Terms*—Industrial IoT, BS activation, content placement, cache-enabled heterogeneous networks.**

I. Introduction

FOR Industrial Internet of Things (IIoT), how to enhance operational efficiency through real-time process analysis is an important challenge. The real-time communication has diverse requirements for networks [1], [2]. For advanced industry 4.0, heterogeneous networks (HetNets) designed for IIoT have been a key technology to increase the regional spectral efficiency of industrial network, which can meet the growing wireless data demands by deploying different tiers of small cell base stations (BSs) coexisting with macro cells [3].

The major bottleneck of HetNets is the limited capacity of backhaul links, when transferring massive data between the core networks and the BSs [4]. Based on this observations, caching popular contents at edge BSs is a promising approach to reduce the burden on backhaul links and the latency in HetNets [5], [6]. A critical issue in wireless edge caching is the content placement, which has been extensively studied in the literature [7]. In [8], the optimal placement of files to minimize the expected files downloading time was studied. The problem is NP-hard for uncoded case, while it is convex for coded case.

In [9], the optimization of the allocation of storage capacity among files in order to minimize the cache missing probability was studied. A caching strategy that combines caching the most popular contents and achieving the largest content diversity was proposed in [10]. In this regard, Blaszczyszyn *et al.* focused on the optimal content placement to maximize the users' hit probability in homogeneous networks under different coverage scenarios in [11]. In [12], Berksan *et al.* extended this work to a two-tier HetNet and considered the problem of optimally placing contents at the first tier of BSs independent of the other tier. The contents for the second tier were then placed depending on the placement strategy of the first tier. The optimal probabilistic content placement for cache-enabled HetNets in the interference-limited case was derived in [13]. It was shown that the optimal placement probability is linearly proportional to the square root of the content popularity with an offset depending on cache size. Caching policies to maximize the successful delivery probability and area spectral efficiency of cache-enabled HetNets were investigated in [14], and the results show that the optimal caching probability is less skewed to maximize the successful delivery probability but is more skewed to maximize the area spectral efficiency. Joint BS caching and cooperation for maximizing the successful transmission probability was studied in [15] for a HetNet, and both globally and locally optimal solutions were proposed. In [16], a belief propagation based distributed algorithm was proposed to solve the cache placement problem,

Manuscript received July 16, 2019; revised January 14, 2020; accepted February 26, 2020. Date of publication March 16, 2020; date of current version May 7, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61871283 and in part by the Major Civil-Military Integration Project in Tianjin, China, under Grant 18ZXJMTG00170. *(Corresponding author: Bin Jiang.)*

Jiachen Yang, Chaofan Ma, and Bin Jiang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangjiachen@tju.edu.cn; machaofan@tju.edu.cn; jiangbin@tju.edu.cn).

Guiguang Ding is with the School of Software, Tsinghua University, Beijing 100084, China (e-mail: dinggg@tsinghua.edu.cn).

Gan Zheng is with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: g.zheng@lboro.ac.uk).

Huihui Wang is with the Department of Engineering, Jacksonville University, Jacksonville, FL 32211 USA (e-mail: hwang1@ju.edu).

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

Digital Object Identifier 10.1109/JSAC.2020.2980907

where parallel computations are performed by individual BSs based on limited local information and very few messages passed between neighboring BSs, thus no central coordinator is needed.

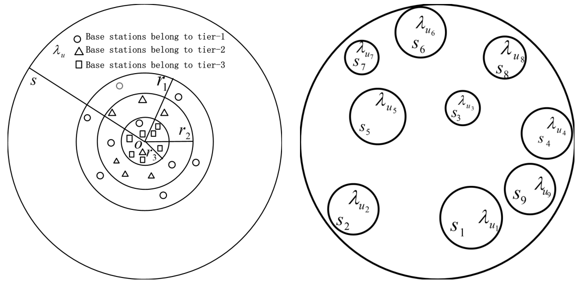
In [17], a typical user is served by the small cell BSs (SBSs) caching the requested file within a circle of certain radius centered at the typical user, which is similar to our work. But [17] assumes the caching capacity of each SBS is one, and thus can not show the impact of the caching capacity. Distributed caching of content in small cells and cooperative transmissions from nearby are combined to improve caching performance in [18]. In [19], Peng *et al.* first considered a single-cell scenario and derived a closed-form expression for maximizing the user success probability which helps reveal the impact of various parameters, then considered a multi-cell scenarios and provided a bisection search algorithm to find the optimal cache size allocation. In [20], it discussed the security issue in industrial networks. In [21], Mehrnaz *et al.* modelled the locations of the users as a uniform Binomial Point Process which is different from the traditional Poisson Point Process and derived a genetic framework to analyze the coverage probability and evaluate the optimal caching probability.

BS sleeping is a key technique to improve the overall network energy efficiency. Studies have shown that BSs are largely under-utilized, i.e., the fraction of time during which traffic load remain below 10% is estimated to be 30% in weekdays and 45% at weekends [22]. Most existing caching strategies in the literature focus on the optimization of content placement only without considering the densities of activated BSs [23]–[25], which do not give practical guideline for the activation of the cache-enabled BSs to serve users with different geographical distributions. Therefore it is essential to jointly optimize the densities of the activated BSs and content placement to match users' demand.

Motivated by this, we jointly optimize content placement and active BSs' densities at each tier of cache-enabled HetNets with considerations on the cache capacities of BSs and total BS energy consumption cost. To accommodate different user distributions, we consider both mutually independent homogeneous PPP based user distribution and the modified Cluster Process (mCP) distributions [26] for the heterogenous user distribution. We adopt the probabilistic content placement strategy, and for the simplicity of mathematical analysis, we suppose that the BSs belonging to the same tier cache the same files with the same probability. Each tier of BSs has different cache capacities and energy consumption costs, so the densities of active BSs at each tier will be optimized to achieve the optimal performance given a total energy consumption cost. We use the total missing probability as the performance metric which has been widely used in the literature [9], [12], and we will see that it is usually not optimal to caching the most popular files to its caching capacity limit.

The main contributions of this paper are summarized as follows:

- We derive the total missing probability in closed form as a function of the content placement probabilities and activated BSs densities at each tier when the user distribution follows a PPP, by extending the analysis in [11].



(a) PPP (b) mCP

Fig. 1. (a) An example of a 3-tier HetNet with users distributed as a homogeneous PPP with density λ_u, and BSs at the k-th tier are modelled as an independent homogeneous PPP with density λ_k and coverage radius r_k; (b) a mCP in which the e-th cluster having the radius s_e and users distributed as a homogeneous PPP with density λ_{u_e}.

- We prove that the joint optimization of the content placement and activated BSs densities can be transformed to a convex optimization problem and reformulate it into a geometric program (GP) problem which can be readily solved.
- By analyzing the Karush-Kuhn-Tucker (KKT) conditions [27], we prove that although different tiers have different content placement probabilities and BS densities. Based on this result, we further propose a low-complexity algorithm that can achieve the near-optimal performance.
- We extend the joint optimization to a more realistic user distribution model characterized by the mCP, and show that GP can be used again to optimally solve the problem.

The rest of this paper is organized as follows. The system model is presented in Section II. We formulate the optimization problem in Section III. Solution and analysis are provided in Section IV. In Section V, we extend the PPP distribution of the users to a mCP distribution. Simulation results and discussions are given in Section VI, followed by conclusions in Section VII.

II. System Model

We consider a heterogeneous circular cellular network comprising of K different types of BSs. At the k-th tier, BSs are deployed with the total density λ_k^{total}, but only part of them are activated. Activated BSs at the k-th tier are modelled as a homogeneous PPP distributed in the plane denoted by Φ_k, with its density λ_k. Suppose that BSs at the same tier has the same caching size denoted by C_k, and the same effective coverage radius denoted by $r_k, \forall k = 1, 2, \ldots, K$. A user can get the requested file cached in an activated BS belonging to the k-th tier if and only if the distance between the user and the BS is less than or equal to r_k [11], [12]. For users in a circular area, there are two kinds of hypothesises on their distributions as illustrated in Fig. 1:

i) the independent homogeneous PPP denoted by Φ_u [25], which has the density λ_u, and radius s denoted in Fig. 1(a).
ii) the mCP with the e-th cluster denoted by Φ_{u_e} [26], which has the density λ_{u_e}, and the radius s_e, illustrated

in Fig. 1(b). Within each cluster, users are distributed as a homogeneous PPP as shown in Fig. 1(a).

We will focus on the analysis and optimization for the PPP user distribution in Section III and IV, and then extend the results to the mCP user distribution in Section V.

We assume that there is a finite file library which contains F files with normalized unit size. We denote the f-th most popular file as c_f, and its request popularity as q_f that follows a general distribution. Without loss of generality, we assume the files are sorted according to a descending order of q_f and $\sum_{f=1}^{F} q_f = 1$. We assume the file library is large and cannot be stored at any single BS, i.e., $F \gg C_k, \forall k = 1, \cdots, K$.

We adopt a probabilistic caching model where the files are independently stored with the same probability at all BSs of the same tier [11]. We can then express the content placement probability matrix below:

$$\boldsymbol{P}_{K\times F} = \begin{pmatrix} p_{11} & \cdots & p_{1F} \\ \vdots & \ddots & \vdots \\ p_{K1} & \cdots & p_{KF} \end{pmatrix}, \tag{1}$$

where p_{kj} denotes the probability of the j-th file cached at a k-th tier BS. Obviously we need $\sum_{j=1}^{F} p_{kj} \leq C_k, \forall k$ due to the cache capacity constraint.

III. Problem Formulation

We first consider the users that are distributed according to a homogeneous PPP distribution Φ_u with its density denoted by λ_u, i.e., the first case described in Section II. We choose the total missing probability as the performance metric, which is the probability that a typical user can not find the requested files from caches at activated BSs of all tiers within the coverage area. Then we will derive the total missing probability, and then formulate the problem of jointly optimizing the content placement probabilities and activated BS densities.

We define a point set for the activated BSs of k-th tier with a radius r_k:

$$B_{r_k} = \{\mathbf{v} | \|\mathbf{v}\| \leq r_k\}, \tag{2}$$

where $\mathbf{v}$ denotes the location of the BS and $\|.\|$ denotes the Euclidean norm. For a typical user located at the origin o, it can get the file cached from the k-th tier BS located at $\mathbf{v}$ if $\mathbf{v} \in B_{r_k}$. Then, according to the PPP distribution of the BSs, we can determine the probability that, there are n_k BSs located in B_{r_k} at the k-th tier, as:

$$P\{\Phi_k(B_{r_k}) = n_k\} = \exp(-\lambda_k \pi r_k^2) \frac{(\lambda_k \pi r_k^2)^{n_k}}{n_k!}. \tag{3}$$

When the typical user requires the file f, the probability that the user can not find the file in the BSs belonging to the k-th tier is $(1-p_{kf})^{n_k}$. To derive the average probability that the typical user can not find the file f at the whole tier k, we take the expectation of $(1-p_{kf})^{n_k}$, which leads to

$$\begin{aligned} E_{kf} &= E_{n_k}[(1-p_{kf})^{n_k}] \\ &= \sum_{n_k=0}^{\infty} (1-p_{kf})^{n_k} P\{\Phi_k(B_{r_k}) = n_k\}. \end{aligned} \tag{4}$$

Now we can obtain the average probability that the typical user can not find the file f at all tiers using the assumptions that BSs at different tiers are independently distributed:

$$E_f = \prod_{k=1}^{K} E_{kf}. \tag{5}$$

So we get the total missing probability f_0 by considering all files in the library:

$$\begin{aligned} f_0 &= \sum_{f=1}^{F} q_f E_f \\ &= \sum_{f=1}^{F} q_f \prod_{k=1}^{K} \sum_{n_k=0}^{\infty} (1-p_{kf})^{n_k} P\{\Phi_k(B_{r_k}) = n_k\} \\ &= \sum_{f=1}^{F} q_f \prod_{k=1}^{K} \\ &\quad \times \left[\exp(-\lambda_k \pi r_k^2) \sum_{n_k=0}^{\infty} \frac{\left((1-p_{kf})\lambda_k \pi r_k^2\right)^{n_k}}{n_k!} \right]. \end{aligned} \tag{6}$$

The above total missing probability f_0 has a complicated expression, so we first simplify this formula as follows. According to the Taylor formula:

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \tag{7}$$

we get

$$\sum_{n_k=0}^{\infty} \frac{\left((1-p_{kf})\lambda_k \pi r_k^2\right)^{n_k}}{n_k!} = \exp((1-p_{kf})\lambda_k \pi r_k^2), \tag{8}$$

and after simple substitution, we can derive the following result for the total missing probability f_0:

$$\begin{aligned} f_0 &= \sum_{f=1}^{F} q_f \prod_{k=1}^{K} \exp(-\lambda_k \pi r_k^2 p_{kf}) \\ &= \sum_{f=1}^{F} q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}). \end{aligned} \tag{9}$$

With the above simplified objective function, we can formulate the optimization problem of minimizing the total missing probability subject to the cache capacity and cost constraints as follows:

$$\mathbb{P}_0 : \min_{\{p_{kf}, \lambda_k\}} \; f_0 \tag{10}$$

$$\text{s.t.} \; \sum_{f=1}^{F} p_{kf} \leq C_k, \quad \forall\, 1 \leq k \leq K, \tag{11}$$

$$\begin{aligned} &\sum_{k=1}^{K} t_k \lambda_k \pi r_k^2 \leq \\ &T\lambda_u \pi s^2 - \sum_{k=1}^{K} \beta_k \lambda_k^{total} \pi r_k^2, \end{aligned} \tag{12}$$

$$0 \leq \lambda_k \leq \lambda_k^{total}, \quad 1 \leq k \leq K,$$

$$0 \leq p_{kf} \leq 1, \quad \forall\, 1 \leq k \leq K, \; 1 \leq f \leq F.$$

In $\mathbb{P}_0$, Eq. 11 is due to the cache size limit of BSs at each tier. Eq. 12 reflects the total energy consumption cost of K-tier BSs. We define α_k as the energy consumption cost of an active BS, and β_k as the energy consumption cost of a sleep BS at the k-th tier. So the total energy consumption cost constraint can be described as follows:

$$\sum\nolimits_{k=1}^{K} \alpha_k \lambda_k \pi r_k^2 + \sum\nolimits_{k=1}^{K} \beta_k (\lambda_k^{total} - \lambda_k) \pi r_k^2 \leqslant T \lambda_u \pi s^2, \tag{13}$$

where the total energy consumption cost is in proportion to the total number of users $\lambda_u \pi s^2$ with a cost coefficient T. For mathematical convenience, Eq. 13 can be reformulated to Eq. 12 by defining $t_k \triangleq \alpha_k - \beta_k$, where the right hand side is a constant.

IV. Solutions and Analysis

The problem $\mathbb{P}_0$ is complicated to solve because the content placement probabilities and activated BS densities are coupled together across different tiers. In the following, we will first convert it to a convex optimization problem and then convert the convex optimization problem to geometric programming (GP) problem that can be efficiently solved. We further investigate the analytical properties of the optimal content placement probabilities, and propose a near-optimal low-complexity algorithm.

A. Reformulation to a GP Problem

To reformulate the problem $\mathbb{P}_0$, we need to use new notations for ease of composition. First, we define some new variables:

$$\begin{pmatrix} \lambda_1 \pi r_1^2 p_{11} & \lambda_1 \pi r_1^2 p_{12} & \cdots & \lambda_1 \pi r_1^2 p_{1F} & \lambda_1 \\ \lambda_2 \pi r_2^2 p_{21} & \lambda_2 \pi r_2^2 p_{22} & \cdots & \lambda_2 \pi r_2^2 p_{2F} & \lambda_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_K \pi r_K^2 p_{K1} & \lambda_K \pi r_K^2 p_{K2} & \cdots & \lambda_K \pi r_K^2 p_{KF} & \lambda_K \end{pmatrix}$$
$$\triangleq \begin{pmatrix} z_1 & z_{K+1} & \cdots & z_{(F-1)\times K+1} & z_{F\times K+1} \\ z_2 & z_{K+2} & \cdots & z_{(F-1)\times K+2} & z_{F\times K+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ z_K & z_{2K} & \cdots & z_{F\times K} & z_{(F+1)\times K} \end{pmatrix}. \tag{14}$$

Then we combine all new variables into a $(F+1) \times K$-dimensional vector $\mathbf{z}$:

$$\mathbf{z} = [\ z_1 \quad \cdots \quad z_K \quad \cdots\cdots \quad z_{F\times K+1} \quad \cdots \quad z_{(F+1)\times K}\ ]^T. \tag{15}$$

Second, we introduce some vectors as follows.

- $\mathbf{o}$ is a K-dimensional zero vector.
- $\mathbf{w}$ is a K-dimensional all-one vector.
- $\mathbf{u}_k, 1 \leq k \leq K$, is a K-dimensional vector with the k-th element being 1 and the rest elements being 0.

Now, we construct some $(F+1)K$-dimensional vectors using $\mathbf{o}, \mathbf{w}, \mathbf{u_k}$:

- $\mathbf{a}_f, \forall 1 \leq f \leq F$, with elements indexed from $(f-1)K+1$ to $(f-1)K+K$ being -1, and the rest being 0. For example, $\mathbf{a}_1 = \left[\ -\mathbf{w}^T\ \mathbf{o}^T\ \cdots \mathbf{o}^T\ \mathbf{o}^T \right]^T$.
- $\mathbf{b}_k = \left[\ \mathbf{u}_k^T\ \mathbf{u}_k^T\ \cdots \mathbf{u}_k^T\ -\pi r_k^2 C_k \mathbf{u}_k^T \right]^T, \forall 1 \leq k \leq K$.
- $\mathbf{g} = \frac{1}{T\lambda_u \pi s^2 - \sum_{k=1}^{K} \beta_k \lambda_k^{total} \pi r_k^2}$ $\left[\ \mathbf{o}^T\ \mathbf{o}^T\ \cdots \mathbf{o}^T\ \sum_{k=1}^{K} t_k \pi r_k^2 \mathbf{u}_k^T \right]^T$.
- $\mathbf{c}_k = \left[\ \mathbf{o}^T\ \mathbf{o}^T\ \cdots \mathbf{o}^T\ -\mathbf{u}_k^T \right]^T, \forall 1 \leq k \leq K$.
- $\mathbf{e}_{(f-1)\times K+k}, \forall 1 \leq f \leq F, 1 \leq k \leq K$, with $[(f-1)\times K+k]$-th element being 1, the $[F\times K+k]$-th element being $-\pi r_k^2$, and the rest being 0. For example, $\mathbf{e}_{(F-1)\times K+K} = \left[\ \mathbf{o}^T\ \mathbf{o}^T\ \cdots \mathbf{u}_K^T\ -\pi r_K^2 \mathbf{u}_K^T \right]^T$.
- $\mathbf{d}_{(f-1)\times K+k}, \forall 1 \leq f \leq F, 1 \leq k \leq K$, with the $[(f-1)\times K+k]$-th element being -1, and the rest equal to 0. For example, $\mathbf{d}_{(F-1)\times K+K} = \left[\ \mathbf{o}^T\ \mathbf{o}^T\ \cdots -\mathbf{u}_K^T\ \mathbf{o}^T \right]^T$.

With the above notations, the original objective function f_0 becomes

$$\begin{aligned} f_0 &= \sum\nolimits_{f=1}^{F} q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) \\ &= \sum\nolimits_{f=1}^{F} \exp\left[\mathbf{a}_f^T \mathbf{z} + \ln(q_f)\right], \end{aligned} \tag{16}$$

and we can transform $\mathbb{P}_0$ into the following optimization problem:

$$\begin{aligned} \mathbb{P}_1 : \min_{\mathbf{z}} \quad & \sum\nolimits_{f=1}^{F} \exp\left[\mathbf{a}_f^T \mathbf{z} + \ln(q_f)\right] && (17) \\ \text{s.t.} \quad & \mathbf{b}_k^T \mathbf{z} \leqslant 0, \forall 1 \leqslant k \leqslant K, && (18) \\ & \mathbf{g}^T \mathbf{z} - 1 \leqslant 0, \\ & \mathbf{c}_k^T \mathbf{z} \leqslant 0, \forall 1 \leqslant k \leqslant K, \\ & -(\lambda_k^{total})^{-1} \mathbf{c}_k^T \mathbf{z} - 1 \leqslant 0, \forall 1 \leqslant k \leqslant K, \\ & \mathbf{e}_{(f-1)\times K+k}^T \mathbf{z} \leqslant 0, \forall 1 \leqslant k \leqslant K, \\ & 1 \leqslant f \leqslant F, && (19) \\ & \mathbf{d}_{(f-1)\times K+k}^T \mathbf{z} \leqslant 0, \forall 1 \leqslant k \leqslant K, 1 \leqslant f \leqslant F. \end{aligned}$$

Note that the Eq. 18 comes from the Eq. 11 in problem $\mathbb{P}_0$, i.e.,

$$\begin{aligned} \sum\nolimits_{f=1}^{F} p_{kf} \leqslant C_k \Leftrightarrow \sum\nolimits_{f=1}^{F} \lambda_k \pi r_k^2 p_{kf} - \lambda_k \pi r_k^2 C_k \\ \leqslant 0 \Leftrightarrow \mathbf{b}_k^T \mathbf{z} \leqslant 0 \end{aligned} \tag{20}$$

where we assume that $\lambda_k > 0$ without the loss of generality, since we can remove the tier with density equals to zero. The Eq. 19 is due to the constraint $p_{kf} \leq 1$, i.e.,

$$p_{kf} \leqslant 1 \Leftrightarrow \lambda_k \pi r_k^2 p_{kf} - \lambda_k \pi r_k^2 \leqslant 0 \Leftrightarrow \mathbf{e}_{(f-1)\times K+k}^T \mathbf{z} \leqslant 0. \tag{21}$$

Note that e^x is convex and $\mathbf{a}_f^T \mathbf{z} + \ln(q_f)$ is affine, and all constraints are linear, therefore $\mathbb{P}_1$ is convex.

Notice that the optimization problem $\mathbb{P}_1$ is a special form of the GP optimization, and in the following we will further convert it into a standard GP problem. To do so, we define a new vector variable $\mathbf{y} = [y_1, \cdots, y_{F\times K}]$ below,

$$\begin{pmatrix} \exp(z_1) & \exp(z_{K+1}) & \cdots & \exp(z_{(F-1)\times K+1}) \\ \exp(z_2) & \exp(z_{K+2}) & \cdots & \exp(z_{(F-1)\times K+2}) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(z_K) & \exp(z_{2\times K}) & \cdots & \exp(z_{F\times K}) \end{pmatrix}$$
$$= \begin{pmatrix} y_1 & y_{K+1} & \cdots & y_{(F-1)\times K+1} \\ y_2 & y_{K+1} & \cdots & y_{(F-1)\times K+2} \\ \vdots & \vdots & \ddots & \vdots \\ y_K & y_{2\times K} & \cdots & y_{F\times K} \end{pmatrix} \text{ or } y_i = \exp(z_i). \tag{22}$$

Define variables $h_k = \exp(\lambda_k), \forall 1 \le k \le K$. We can reformulate $\mathbb{P}_1$ into the following optimization problem:

$$\begin{aligned}
\mathbb{P}_2 : \min_{\mathbf{y},\{h_k\}} \quad & \sum\nolimits_{f=1}^{F} q_f \prod\nolimits_{k=1}^{K} \left(y_{(f-1)\times K+k}\right)^{-1} \\
\text{s.t.} \quad & \prod\nolimits_{f=1}^{F} y_{(f-1)\times K+k} \le {h_k}^{\pi r_k^2 C_k}, \forall 1 \le k \le K, \\
& \prod\nolimits_{k=1}^{K} {h_k}^{\pi r_k^2 t_k} \le \\
& \exp(T\lambda_u \pi s^2 - \sum\nolimits_{k=1}^{K} \beta_k \lambda_k^{total} \pi r_k^2), \\
& y_{(f-1)\times K+k} \le {h_k}^{\pi r_k^2}, \quad \forall 1 \le k \le K,\ 1 \le f \le F, \\
& y_{(f-1)\times K+k} \ge 1, \quad \forall 1 \le k \le K,\ 1 \le f \le F, \\
& h_k \ge 1, \quad \forall 1 \le k \le K, \\
& h_k \le e^{\lambda_k^{total}}, \quad \forall 1 \le k \le K. \qquad (23)
\end{aligned}$$

It is easy to see that the optimization problem $\mathbb{P}_2$ is a standard GP problem [28], so it can be readily solved using the CVX toolbox [29].

B. Analytical Properties of the Content Placement Probabilities

In this section, we will investigate the analytical properties in order to gain insights about the optimal content placement probabilities across different tiers and devise a more efficient algorithm that achieve the near-optimal performance.

Notice that the optimization problem $\mathbb{P}_1$ is convex, so we can use the KKT conditions to derive useful analytical results. The Lagrangian function is

$$\begin{aligned}
& L(\mathbf{z}, \boldsymbol{\mu}_1, \mu_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}'_{3k}, \boldsymbol{\mu}_4, \boldsymbol{\mu}_5) \\
& = \sum\nolimits_{f=1}^{F} \exp\left[\mathbf{a}_f^T \mathbf{z} + \ln(q_f)\right] \\
& \quad + \sum\nolimits_{k=1}^{K} \mu_{1k} \mathbf{b}_k^T \mathbf{z} + \mu_2(\mathbf{g}^T \mathbf{z} - 1) + \sum\nolimits_{k=1}^{K} \mu_{3k} \mathbf{c}_k^T \mathbf{z} \\
& \quad + \sum\nolimits_{k=1}^{K} \mu'_{3k}[-(\lambda_k^{total})^{-1} \mathbf{c}_k^T \mathbf{z} - 1] \\
& \quad + \sum\nolimits_{f=1}^{F} \sum\nolimits_{k=1}^{K} \mu_{4[(f-1)\times K+k]} \mathbf{e}_{(f-1)\times K+k}^T \mathbf{z} \\
& \quad + \sum\nolimits_{f=1}^{F} \sum\nolimits_{k=1}^{K} \mu_{5[(f-1)\times K+k]} \mathbf{d}_{(f-1)\times K+k}^T \mathbf{z} \qquad (24)
\end{aligned}$$

where $\boldsymbol{\mu_1}, \mu_2, \boldsymbol{\mu_3}, \boldsymbol{\mu'_3}, \boldsymbol{\mu_4}, \boldsymbol{\mu_5}$ denote non-negative dual variables associated with the constraints.

From the KKT condition $\frac{\partial L}{\partial z_{(f-1)K+k}} = 0$, we can obtain the following result:

$$\begin{aligned}
& q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) \\
& \quad = \exp[\mathbf{a}_f^T z + \ln(q_f)] \\
& \quad = \mu_{1k} + \mu_{4[(f-1)K+k]} - \mu_{5[(f-1)K+k]}, \\
& \qquad \forall k = 1, 2, \ldots, K, f = 1, 2, \ldots, F, \qquad (25)
\end{aligned}$$

which will be used to derive useful insight in Theorem 1 below. We define $\mu_{4kf} \triangleq \mu_{4[(f-1)\times K+k]}$, $\mu_{5kf} \triangleq \mu_{5[(f-1)\times K+k]}$ for easy notation.

Theorem 1: There are the same set of files cached with probability 1 and the same set of files cached with probability 0, at all different tiers. In other words, all tiers store the same files, and there exists a lower threshold L and an upper threshold U of file indices, such that $p_{kf} = 1$ when the file index $f \le L$, and $p_{kf} = 0$ when $f > U, \forall k$.

Proof : Because we have assumed an ordered file popularity probabilities $q_1 \ge q_2 \ge \ldots \ge q_F$, it must hold that $p_{k1} \ge p_{k2} \ge \ldots \ge p_{kF}, \forall k$. For the k-th tier of the HetNet, we define a lower threshold L_k and an upper threshold U_k of file indices as follows:

$$p_{kf} \begin{cases} = 1, & 0 < f \le L_k \Rightarrow \mu_{5kf} = 0; \\ \in (0,1), & L_k < f \le U_k \Rightarrow \mu_{4kf} = \mu_{5kf} = 0; \\ = 0, & U_k < f \le F \Rightarrow \mu_{4kf} = 0. \end{cases} \qquad (26)$$

So the KKT condition, Eq. 25, can be simplified in the following cases:

- when $0 < f \le L_k$, we have
$$q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) = \mu_{1k} + \mu_{4kf}. \qquad (27)$$
- when $L_k < f \le U_k$, the condition becomes
$$q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) = \mu_{1k}. \qquad (28)$$
- when $U_k < f \le F$, the condition is
$$q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) = \mu_{1k} - \mu_{5kf}. \qquad (29)$$

Note that in Eq. 28, the left hand side is a function of the file index f while the right hand side is a function of the file index k, so we conclude that when $L_k < f \le U_k$,

$$q_f \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) = \mu_{1k} = A = \text{const.} \qquad (30)$$

When $f = L_k$, $p_{kL_k} = 1 = p_{k(L_k-n)}, \forall 0 < n < L_k$

$$\begin{aligned}
& q_{L_k} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kL_k}) \\
& \quad = \mu_{1k} + \mu_{4kL_k} \\
& \quad < q_{L_k-n} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{k(L_k-n)}). \qquad (31)
\end{aligned}$$

If $\mu_{4kL_k} = 0$ is also satisfied in Eq. 31, then we have

$$\begin{aligned}
& q_{L_k} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{kL_k}) = A \\
& \qquad < q_{L_k-n} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{k(L_k-n)}), \qquad (32)
\end{aligned}$$

where $0 < n < L_k$. Otherwise, $q_{L_k} \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kL_k}) > A$.

Depending on whether $\mu_{4kL_k} = 0$ or not, next we define ${L_k}'$ as

$${L_k}' = \begin{cases} L_k - 1, & \mu_{4kL_k} = 0; \\ L_k, & \mu_{4kL_k} \ne 0. \end{cases} \qquad (33)$$

Similarly, we define ${U_k}'$ as

$${U_k}' = \begin{cases} U_k + 1, & \mu_{5k(U_k+1)} = 0; \\ U_k, & \mu_{5k(U_k+1)} \ne 0. \end{cases} \qquad (34)$$

Now we can conclude that $\forall k$,

$$\begin{aligned}
& q_{{L_k}'} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{k{L_k}'}) \ne A \\
& \qquad \ne q_{{U_k}'+1} \exp(-\sum\nolimits_{k=1}^{K} \lambda_k \pi r_k^2 p_{k({U_k}'+1)}). \qquad (35)
\end{aligned}$$

Next we proceed to prove that L_k is the same for all K tiers by contradiction. For any two different tiers k_1, k_2, if ${L_{k_1}}' \neq {L_{k_2}}'$, without loss of generality, we suppose ${L_{k_1}}' > {L_{k_2}}'$. Then for the k_1-th tier, according to Eq. 35, we have

$$q_{{L_{k_1}}'} \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{k{L_{k_1}}'}) \neq A. \tag{36}$$

According to Eq. 30, when ${L_{k_2}}' < f \leq {U_{k_2}}'$,

$$q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) = A. \tag{37}$$

Because ${L_{k_1}}' > {L_{k_2}}'$, we have

$$q_{{L_{k_1}}'} \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{k{L_{k_1}}'}) = A, \tag{38}$$

which clearly contradicts the result in Eq. 36. The only way to eliminate this contradiction is to rely on the result that ${L_{k_1}}' = {L_{k_2}}'$ must hold. So for all K tiers, the lower bound file indices satisfy ${L_k}' = L', \forall k$. Likewise, we can show that the upper bound of the indices should also be the same across tiers, i.e., ${U_k}' = U', \forall k$, then we get $L_k = L, U_k = U, \forall k$. This completes the proof. ■

From Theorem 1, we can get some useful analytical results about the objective function (total missing probability). We define $B = \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2)$, and then

$$\begin{aligned} f_0 &= \sum_{f=1}^{F} q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) \\ &= \sum_{f=1}^{L} q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2) \\ &\quad + (U-L)A + \sum_{f=U+1}^{F} q_f \\ &= \sum_{f=1}^{L} q_f B + (U-L)A + \sum_{f=U+1}^{F} q_f. \end{aligned} \tag{39}$$

It is observed that the total missing probability in Eq. 39 can be decomposed into three parts. The first L files' contribution to the total missing probability is the sum of their popularity probabilities multiplied by a decay factor B. The second part is composed of files with indices between L and U. Each file in this part has a contribution equal to the same constant A, so their total contribution is $(U-L)A$. The third part are made of those files that are not cached, so their contribution is the sum of their popularity probabilities.

C. A Near-Optimal Low-Complexity Algorithm

Motivated by the results in Theorem 1, in this subsection, we design a near-optimal low-complexity solution to find the content placement probabilities. The BS densities $\{\lambda_k\}$ will still be solved using GP similar to $\mathbb{P}_0$ with reduced number of variables. In the following we focus on deriving the solution to the content placement probabilities with given BS densities.

For convenience, we define

$$Q_f \triangleq q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}), \tag{40}$$

which can be seen as the contribution to the total missing probability of the f-th file. Now we can derive the relation between the lower threshold L (if $L > 0$) and the upper threshold U.

First notice that

$$Q_L = q_L \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2) \approx Q_{L+1} = A, \tag{41}$$

where the first equality is because $p_{kf} = 1, \forall f \leq L, \forall k$ and the last equality comes from Eq. 30. The approximation is because of the intuition that the two adjacent L-th and $(L+1)$-th files have similar contributions to the total missing probability.

Next we approximate A as

$$A = q_U \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kU}) \approx q_U, \tag{42}$$

where the equality is due to Eq. 30 and the approximation is due to the fact that $p_{kU} \approx 0$.

Combining Eq. 41 and 42, we can get

$$q_L \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2) \approx q_U, \tag{43}$$

which is very useful because it characterizes the relation between L and U and also helps design the near-optimal algorithm. Once L is known, the closest U can be easily estimated via Eq. 43. The optimal L can be found by exhausted search from 1 to $\min\{C_k\} - 1$, and $L = 0$ will be treated as a special case because only U needs to be searched.

When L and U are known, the remaining task is only to calculate $p_{kf}, \forall L < f \leq U, \forall k$, because $p_{kf} = 1$ for $1 \leq f \leq L$ and $p_{kf} = 0$ for $f > U$. Again we will use Eq. 30 and Eq. 42 to find an approximated solution.

Combining Eq. 30 and 42, we can have the following approximation for p_{kf} when $L < f \leq U$:

$$q_f \exp(-\sum_{k=1}^{K} \lambda_k \pi r_k^2 p_{kf}) \approx q_U. \tag{44}$$

For a given file f, based on Eq. 44 alone, we can not solve K unknown quantities p_{kf} for $1 \leq k \leq K$. In order to find an approximated solution, we will need an additional assumption on the relation between these K unknowns.

According to Theorem 1, all tiers in the HetNets share the same lower threshold L and upper threshold U. $C_k - L$ is the remaining cache capacity for the rest files that are not cached with probability 1. To decrease the algorithm complexity, a simple yet intuitive choice is to impose linear dependence between content placement probability of each tier, i.e.,

$$p_{kf} = \frac{C_k - L}{C_K - L} p_{Kf}, \quad 1 \leq k \leq K. \tag{45}$$

Numerical simulation will show this heuristic solution will provide good approximation to the optimal content placement popularity.

With Eq. 44 and 45, we can easily solve the approximated p_{kf}. If some quantities of p_{kf} are greater than 1 which are invalid, they will be reduced to 1. We observe that if the content library size F is much greater than the caching capacity of the base stations, we normally get valid solution $p_{kf} < 1$.

Algorithm 1 The Proposed Near-Optimal Low-Complexity Algorithm

```
1: Initialise the densities $\{\lambda_k\}$ of each tier, a threshold $\epsilon$, iteration index $i=1$ and $f_0(0)=0, f_0(1)=1$.
2: while $|f_0(i)-f_0(i-1)| \geq \epsilon$ do
3:   for $c=1$ to $\min\{C_k\}-1$ do
4:     Find the closest $U_c$ that satisfies (43).
5:     Given $L=c$ and $U=U_c$, calculate the caching probability matrix $\boldsymbol{P_{K\times F}}^{(c)}$ using (44) and (45).
6:     Calculate the total missing probability $f_0^{(c)}$ using $\boldsymbol{P_{K\times F}}^{(c)}$ and given $\{\lambda_k\}$.
7:   end for
8:   Find the optimal caching probability matrix $\boldsymbol{P_{K\times F}}(i_{opt})$, where $i_{opt}=\arg_c \min f_0^{(c)}$, and update $f_0(i)=f_0^{(i_{opt})}$.
9:   Optimize the activated BS densities $\{\lambda_k\}$ given the caching probability matrix $\boldsymbol{P_{K\times F}}(i_{opt})$ using the GP approach similar to $\mathbb{P}_2$.
10:  Set $i=i+1$.
11: end while
```

Based on the above analysis and derivation, we can summarize the proposed near-optimal algorithm as follows.

It can be seen that because $F \gg K$, the complexity of solving only K variables $\{\lambda_k\}$ in the above Step 9 is much lower than the optimal solution that requires solving $\mathbb{P}_0$ with both variables $\{\lambda_k\}$ and $\{p_{kf}\}$. The comparison results of both performance and complexity between the optimal solution and the near-optimal solution will be presented in Section VI.

V. Extension to the mCP Model of User Distribution

In this section, we extend the previous study to the user distribution modelled by the mCP. The mCP shown in Fig. 1(b) can be viewed as the assemble of E different PPP clusters in Fig. 1(a). Suppose there are E different circular clusters in a 2D-plane where the e-th cluster is centered at $\boldsymbol{w_e}$ and has the radius s_e. We assumes that in each cluster users are distributed according to the PPP with the density denoted as λ_{u_e}, where $e=1,2,\ldots,E$. In each cluster there exists a K-tier heterogeneous network distributed as the K mutually independent homogeneous PPP, with the density matrix $\boldsymbol{\lambda_{BS}}$ defined as follows:

$$\boldsymbol{\lambda_{BS}} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1E} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2E} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{K1} & \lambda_{K2} & \cdots & \lambda_{KE} \end{pmatrix}, \tag{46}$$

where λ_{ke} denotes the density of active BSs at the k-tier of the e-th cluster, and the associated coverage radius is r_k. λ_{ke}^{total} denotes the total density of BSs including the active BSs and sleeping BSs at the k-tier of the e-th cluster. We define the content placement probability matrix in the e-th cluster as follows:

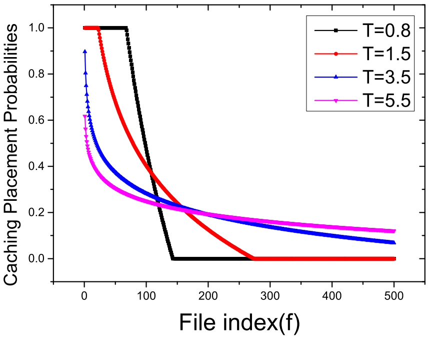


Fig. 2. Content placement probabilities of a single-tier network with different cost coefficients T.

$$\boldsymbol{P^{(e)}} = \begin{pmatrix} p_{11}^{(e)} & p_{12}^{(e)} & \cdots & p_{1F}^{(e)} \\ p_{21}^{(e)} & p_{22}^{(e)} & \cdots & p_{2F}^{(e)} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1}^{(e)} & p_{K2}^{(e)} & \cdots & p_{KF}^{(e)} \end{pmatrix}. \tag{47}$$

We further assume that the distance between any two adjacent clusters is large enough or orthogonal resources are used, so that any adjacent clusters will not interfere each other. Similar to the single-cluster case, we aim to minimize the overall missing probability subject to the cache capacity and energy consumption cost constraints. We choose the objective function to be a weighted sum of individual clusters' missing probabilities and the weight is in proportion to the number of users in each cluster. Mathematically, the optimization problem for the mCP user distribution model is formulated as

$$\begin{aligned} \min_{\{p_{kf}^{(e)},\lambda_{ke}\}} & \sum_{e=1}^{E} \lambda_{u_e}\pi s_e^2 \frac{\sum\limits_{f=1}^{F} q_f \exp\left(-\sum\limits_{k=1}^{K}\lambda_{ke}\pi r_k^2 p_{kf}^{(e)}\right)}{\sum\limits_{e=1}^{E}\lambda_{u_e}\pi s_e^2} \\ \text{s.t.} & \sum\nolimits_{f=1}^{F} p_{kf}^{(e)} \leq C_k, \quad \forall k=1,2,\ldots,K,\ e=1,2,\ldots,E, \\ & \sum_{e=1}^{E}\sum_{k=1}^{K} t_k\lambda_{ke}\pi r_k^2 \leq \\ & T(\sum_{e=1}^{E}\lambda_{u_e}\pi s_e^2) - \sum\nolimits_{k=1}^{K}\sum\nolimits_{e=1}^{E}\beta_k\lambda_{ke}^{total}\pi r_k^2, \\ & 0 \leq \lambda_{ke} \leq \lambda_{ke}^{total}, \quad \forall k=1,2,\ldots,K,\ e=1,2,\ldots,E, \\ & 0 \leq p_{kf}^{(e)} \leq 1, \quad \forall k=1,2,\ldots,K, \\ & \qquad f=1,2,\ldots,F,\ e=1,2,\ldots,E, \end{aligned} \tag{48}$$

where C_k is the cache capacity of a BS at the k-th tier of the l-th cluster, t_k is the energy consumption cost of activating a BS at the k-th tier of the e-th cluster, and the total energy consumption cost is in proportion to the total number of users in all clusters with a cost coefficient T.

It is not difficult to see that the Eq. 48 and the Eq. 10 have the similar structure and the Eq. 48 can also be transformed into a GP. For convenience, we give the final reformulated GP problem below:

$$
\begin{aligned}
&\mathbb{P}_3: \\
&\min_{\{y^{(e)}_{(f-1)\times K+k}, h_{ke}\}} \quad \sum_{e=1}^{E} \lambda_{u_e}\pi s_e^2 \\
&\qquad\qquad \frac{\sum_{f=1}^{F} q_f \left(\prod_{k=1}^{K} y^{(e)}_{(f-1)\times K+k}\right)^{-1}}{\sum_{e=1}^{E} \lambda_{u_e}\pi s_e^2} \\
&\text{s.t. } \prod_{k=1}^{K} y^{(e)}_{(f-1)\times K+k} \leq (h_{ke})^{\pi r_{ke}^2 C_k}, \quad \forall e = 1, \cdots, E, \\
&\quad \prod_{e=1}^{E}\prod_{k=1}^{K} (h_{ke})^{t_k \pi r_k^2} \\
&\quad \leq \exp\left[T\left(\sum_{e=1}^{E} \lambda_{u_e}\pi s_e^2\right) \right.\\
&\qquad\qquad \left. -\sum_{k=1}^{K}\sum_{e=1}^{E} \beta_k \lambda_{ke}^{total}\pi r_k^2\right], \\
&\quad h_{ke} \geq 1, \\
&\quad h_{ke} \leq exp(\lambda_{ke}^{total}), \\
&\quad y^{(e)}_{(f-1)\times K+k} \leq (h_{ke})^{\pi r_k^2}, \\
&\quad y^{(e)}_{(f-1)\times K+k} \geq 1.
\end{aligned}
\tag{49}
$$

Once the optimal $\{y^{(e)}_{(f-1)\times K+k}, h_{ke}\}$ is found, the original variables $\{\lambda_{ke}, p^{(e)}_{kf}\}$ can be solved via the transformations $h_{ke} = \exp(\lambda_{ke})$ and $y^{(e)}_{(f-1)\times K+k} = \exp(\lambda_{ke}\pi r_k^2 p^{(e)}_{kf}), \forall e, k, f$.

VI. Performance Evaluation

In this section, computer simulations are carried out to assess the performance of the proposed algorithms. The proposed joint content placement and BS activation scheme for K-tier HetNets will be compared with the fixed BS deployment and homogeneous networks. We assume the request popularity q_f for the f-th most popular file follows the Zipf distribution, i.e,

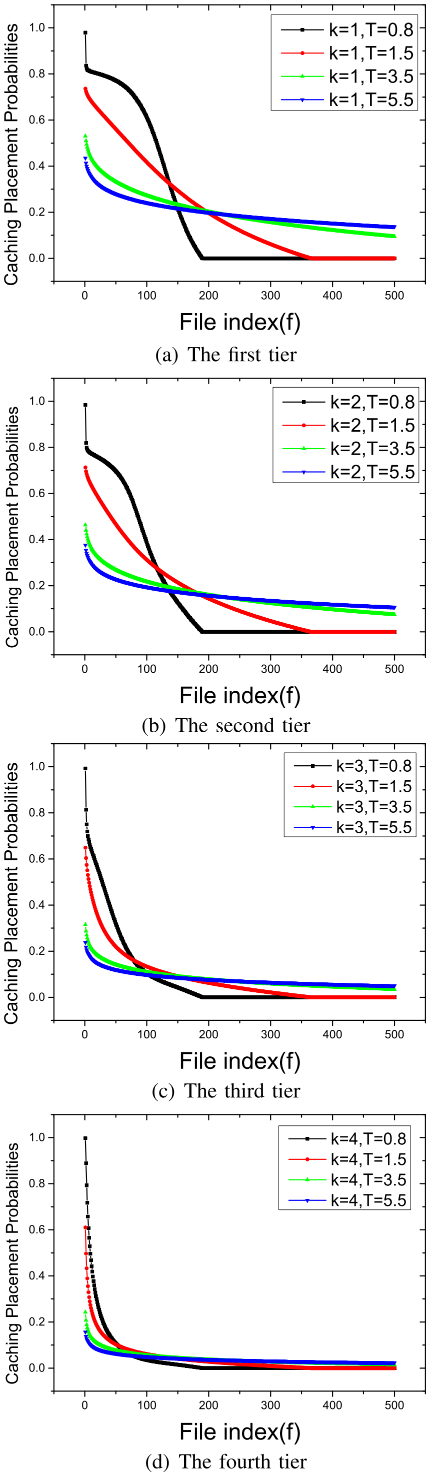
$$q_f = \frac{f^{-\gamma}}{\sum_{f=1}^{F} f^{-\gamma}}, \tag{50}$$

where γ denotes the Zipf parameter.

We consider a library with $F = 500$ files, cost coefficient $\lambda_u \pi s^2 = 25$ and Zipf parameter $\gamma = 1$ unless otherwise specified. Generally, we normalize the effective radius r of tier-1 in HetNets, which represents $0.5km$, and normalize the sleeping BS energy consumption cost β of tier-1 in HetNets, which represents $10W$. So, the units of some parameters are as follows: $r : 0.5km$; $s : km$; $\lambda_u : user/km^2$; $t, \beta : 10W$. For example, there are $\boldsymbol{r} = (1, 0.8, 0.4, 0.2)$, $\boldsymbol{\beta} = (1, 0.8, 0.4, 0.2)$, $\boldsymbol{t} = (10, 8, 4, 2)$, which means $\boldsymbol{r} = (0.5km, 0.4km, 0.2km, 0.1km)$, $\boldsymbol{\beta} = (10W, 8W, 4W, 2W)$, $\boldsymbol{t} = (100W, 80W, 40W, 20W)$.

A. The Impact of System Parameters on Content Placement Probabilities and Densities of HetsNets

We first study the impact of system parameters on the total missing probability, including the cost coefficient T, Zipf distribution's parameter γ and caching size C_k. For comparison purposes, we show the results for both single-tier homogeneous and multi-tier HetNets.



(a) The first tier

(b) The second tier

(c) The third tier

(d) The fourth tier

Fig. 3. Content placement probabilities of a 4-tier networks with different cost coefficients T.

1) Cost Coefficient T: We evaluate the content placement probability of a single-tier network with different cost coefficients T or equivalently the BS density λ in Fig. 2. Some other parameters are set as follows: $r = 1$, $C = 100$, $t = 10$, $\beta = 1$, $\lambda^{total} = 4$. Because the BSs' caching capacity is

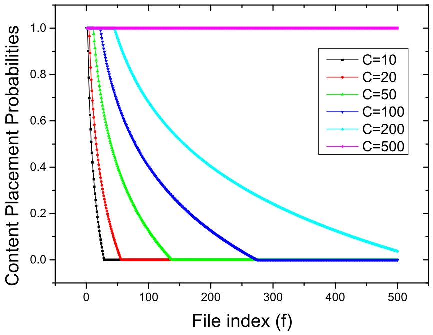


Fig. 4. Content placement probabilities vs. caching capacity of a single-tier network.

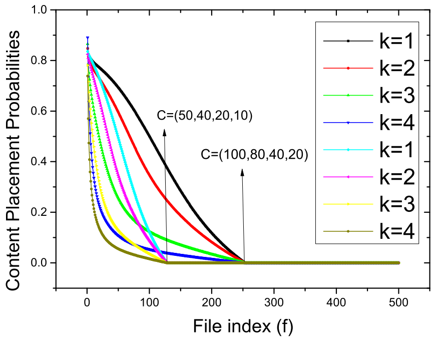


Fig. 5. Content placement probabilities vs different caching capacity of a 4-tier network.

limited, Fig. 2 shows that the BSs tend to store some most popular files with probability approximately 1 while avoiding caching the files less popular as T or the BS density decreases. When the cost coefficient T increases, the trend goes to the opposite, i.e., the BSs tends to increase the file diversity to achieve better performance.

The results are different in a 4-tier HetNet shown in Fig. 3. The parameters are set as follows: $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, $\mathbf{C} = (100, 80, 40, 20)$, $\mathbf{t} = (10, 8, 4, 2)$, $\boldsymbol{\beta} = (1, 0.8, 0.4, 0.2)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (0.1, 0.5, 1, 2)$. In Fig. 3, there are no files cached by the BSs with probability 1 because of the cooperation of the HetNets. But, the first tier (Fig. 3(a)) and the second tier (Fig. 3(b)) have similar caching probability characteristics as that in Fig. 2. However, Fig. 3(c) and 3(d) show a different characteristic: BSs at the 3rd and 4th tiers do not cache the most popular files with the same high probability when the cost coefficient T is small, and this is because of the constraint of small caching capacities at the 3rd and 4th tiers and there are not enough cache capacities at these tiers.

2) Cache Capacity $\{C_k\}$: In Fig. 4, we show the impact of the cache capacity on the content placement probability for a single-tier network. Some other parameters are set as follows: $r = 1$, $t = 10$, $\beta = 1$, $\lambda^{total} = 4$, $T = 1.5$. As expected, as C increases, more files are stored with probability 1. Intuitively, when $C = F = 500$, the content placement probability of each file is equal to 1. In Fig. 5, we investigate how the cache capacities influence the content placement probability of each file for a 4-tier HetNet. Some other parameters are set as follows: $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, and $\mathbf{t} = (10, 8, 4, 2), \boldsymbol{\beta} = (1, 0.8, 0.4, 0.2), \boldsymbol{\lambda}^{\boldsymbol{total}} = (0.1, 0.5, 1, 2)$, $T = 1.05$. The results show that the cache capacity has more direct influence on the upper threshold of the content placement probabilities in HetNets, but usually the lower threshold is still equal to 0 as the cache capacity increases. While for single-tier networks, cache capacity will affect both the lower and the upper thresholds of the content placement probabilities.

3) The Impact of the Zipf Parameters γ: The impact of the Zipf parameter γ on the content placement probabilities in the single-tier network is shown in Fig. 6 with other system parameters set as follows: $r = 1, C = 100, t = 10$, $\beta = 1, \lambda^{total} = 4, T = 1.50$. It is observed that γ has a similar impact on the content placement probabilities as the cost coefficient T. When γ increases, the BSs tend to cache more most popular files with the probability approximately 1 and ignore the less popular files. Similar results are produced in Fig. 7(a) and 7(b) for a 4-tier HetNet. There does exist some differences between the single-tier network and HetNet. When γ increases, the single-tier network tends to cache the most popular files with probability 1, while in HetNets, they usually cache the most popular files with a probability less than 1. In Fig. 7(c) and 7(d), we assume very limited cache size allocated to the 3rd and 4th tiers, so we can see a quite different characteristic: the caching probability decreases rapidly as the file index increases. This is the result of limited cache capacities at the 3rd and 4th tiers.

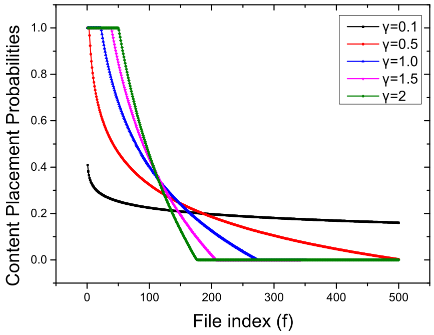
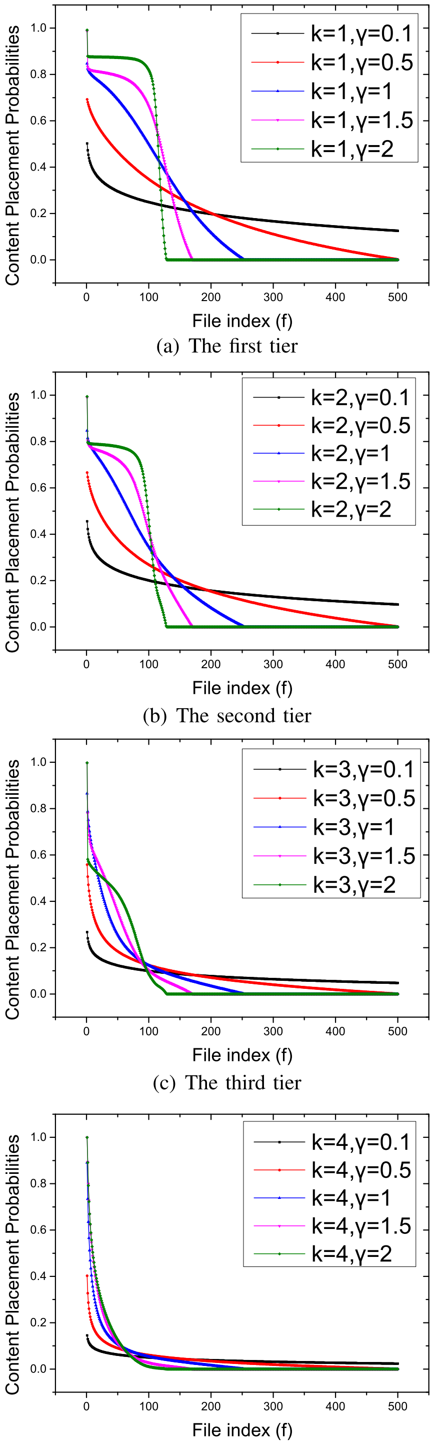


Fig. 6. Content placement probabilities vs. Zipf parameters γ in a single-tier network.

4) Simulations of Densities of HetNets: The impact of the Zipf parameter γ and energy consumption cost coefficient T on the densities of 4-tier HetNets is shown in Fig. 8. In Fig. 8(a), some other parameters are set as follows: $\gamma = 0.8$, $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, $\mathbf{C} = (50, 40, 30, 25)$, $\mathbf{t} = (5, 4, 3, 2.5)$, $\boldsymbol{\beta} = (0.5, 0.4, 0.3, 0.25)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (0.5, 1, 5, 25)$. In Fig. 8(b), some other parameters are set as follows: $T = 1.2$, $\mathbf{r} = (0.5, 0.45, 0.4, 0.35)$, $\mathbf{C} = (50, 45, 40, 35)$, $\mathbf{t} = (5, 4.5, 4, 3.5)$, $\boldsymbol{\beta} = (0.5, 0.45, 0.4, 0.35)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (2, 3, 5, 20)$. Fig. 8(a) shows that densities of HetNets increase linearly with the cost coefficient T, which coincides with Eq. 12. Fig. 8(b) shows the change of optimal densities of HetNets as γ increases. When $\gamma \leq 1.1$, optimal densities of HetNets remain relatively static. HetNets will gradually converge one single tier as the content popularity distribution becomes more skewed.



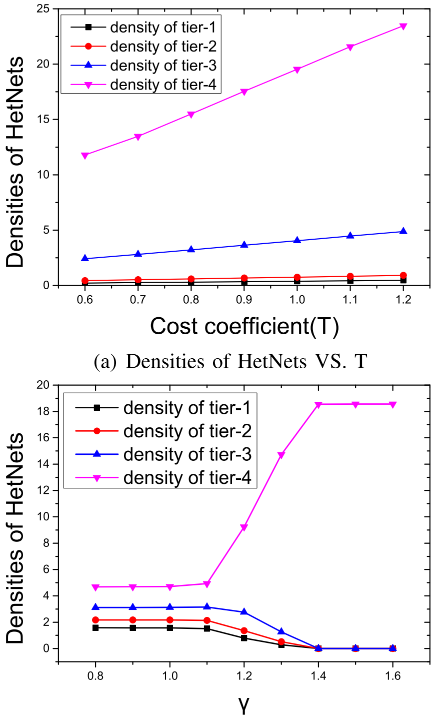
(a) The first tier

(b) The second tier

(c) The third tier

(d) The fourth tier

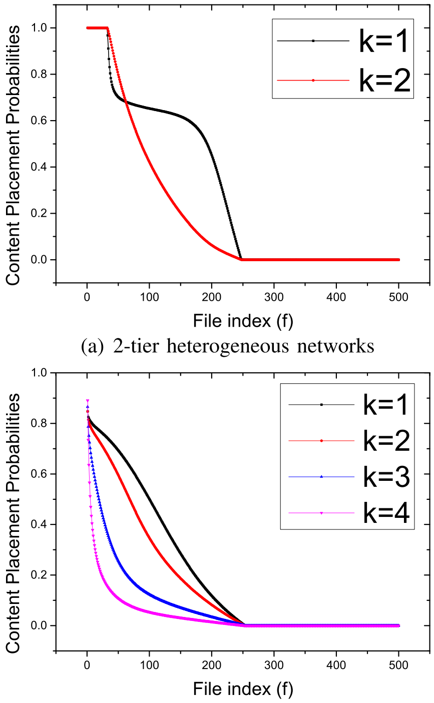
Fig. 7. Content placement probabilities vs. Zipf parameters γ in a 4-tier network.



(a) Densities of HetNets VS. T

(b) Densities of HetNets VS. γ

Fig. 8. Densities of 4-tier HetNets.



(a) 2-tier heterogeneous networks

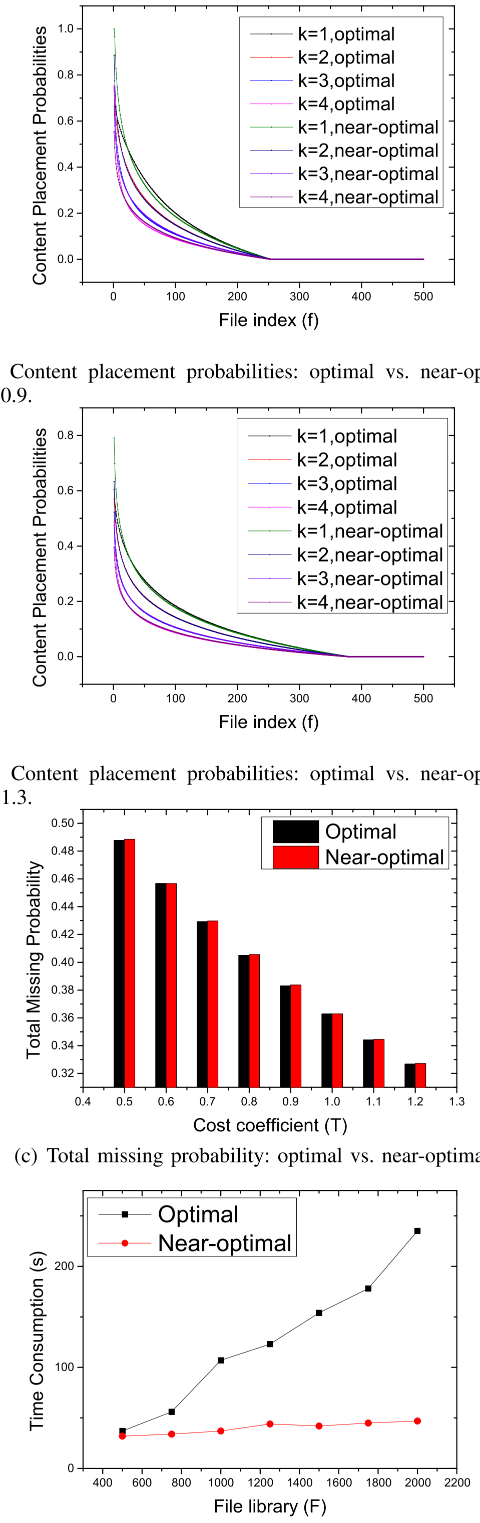
(b) 4-tier heterogeneous networks

Fig. 9. Verification of the thresholds in Theorem 2 for different heterogeneous networks.

B. Multi-Tier Heterogeneous Networks

We continue to present simulation results specifically for the HetNets.

1) Simulation Verification of Theorem 2: Next we validate the results in Theorem 2 for a 2-tier and 4-tier HetNets in Fig. 9(a) and 9(b), respectively. In Fig. 9(a), some other parameters are set as follows: $\mathbf{r} = (1, 0.8)$, $\mathbf{C} = (110, 50)$, $\mathbf{t} = (10, 5)$, $\boldsymbol{\beta} = (1, 0.5)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (0.5, 2)$, $T = 1.14$. In Fig. 9(b), some other parameters are set as follows: $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, $\mathbf{C} = (100, 80, 40, 20)$, $\mathbf{t} = (10, 8, 4, 2)$, $\boldsymbol{\beta} = (1, 0.8, 0.4, 0.2)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (0.1, 0.5, 1, 2)$, $T = 1.05$. As we can see, different tiers have the same lower and upper thresholds although they have different network



(a) Content placement probabilities: optimal vs. near-optimal, T=0.9.

(b) Content placement probabilities: optimal vs. near-optimal, T=1.3.

(c) Total missing probability: optimal vs. near-optimal.

(d) Time consumption: optimal vs. near-optimal.

Fig. 10. Comparisons between the optimal and the near-optimal algorithms.

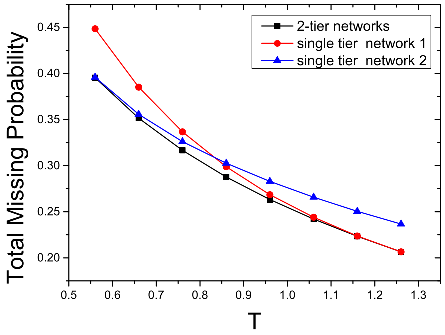


Fig. 11. Comparison of the Total missing probability between the 2-tier networks and single tier network.

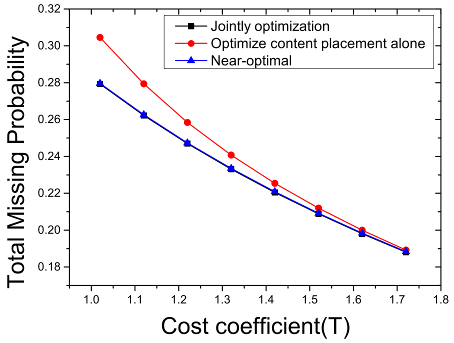
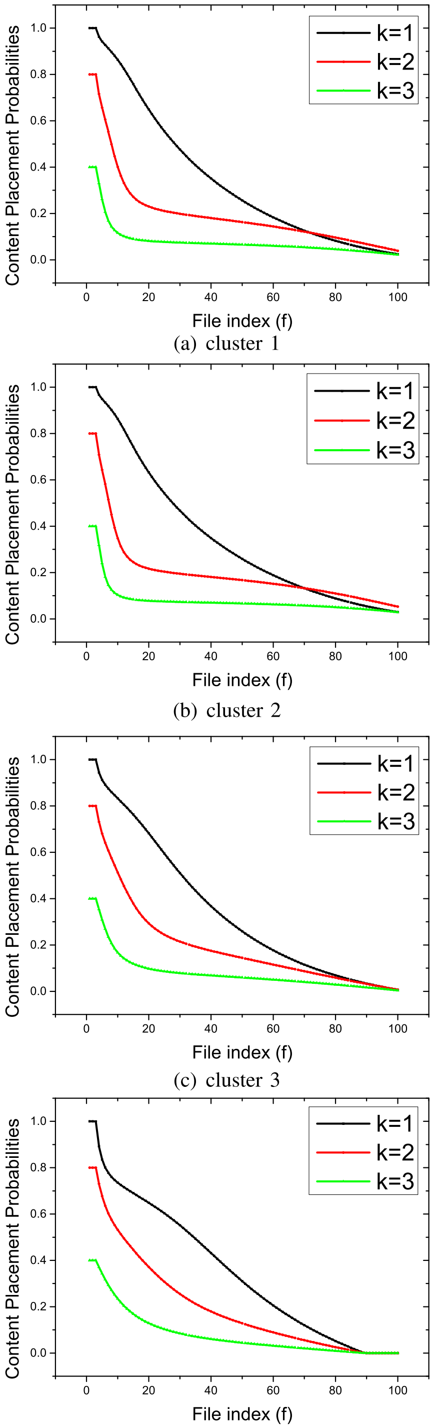


Fig. 12. Comparison of joint optimization of caching placement and BSs deployment, optimization of the caching placement only, and near-optimal low-complexity algorithm.

parameters. In Fig. 9(b), the lower threshold L is equal to 0 for the 4-tier network. We can also find that when $K \geq 3$, BSs usually do not cache the files with probability 1 even for the most popular ones. This is because the cooperation of different tiers makes it unnecessary to cache the files with probability 1 so that more cache space can be saved for the less popular files to increase caching diversity.

2) Performance Evaluation of the Near-Optimal Low-Complexity Algorithm: Next, we will present comparisons between the optimal and the near-optimal algorithms in Fig. 10 in a 4-tier HetNet. Relevant parameters are set as follows: $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, $\mathbf{C} = (50, 40, 30, 25)$, $\mathbf{t} = (5, 4, 3, 2.5)$, $\boldsymbol{\beta} = (0.5, 0.4, 0.3, 0.25)$, $\boldsymbol{\lambda}^{\boldsymbol{total}} = (0.5, 1, 5, 25)$, $\gamma = 0.8$. In Fig. 10(a) and 10(b), we show the content placement probabilities of the optimal and the near-optimal solutions with different cost coefficients T. It is observed that except for the 4-th tier, the content placement probabilities of the optimal solution and the near-optimal solutions match each other. The gap for the 4-th tier is mainly because of our assumption in Eq. 45. In Fig. 10(c), we can see that the resulting total missing probability of the near-optimal algorithm is almost identical to the optimal solution. In Fig. 10(d), we compare the running time of the optimal and the near-optimal algorithms. For a content library that contains $F = 500$ files, two algorithms show the similar running time. When F increases, the gap of running time between two algorithms increases very rapidly. We found that the optimal algorithm can not handle a content library with $F > 3000$ using CVX, while the near-optimal algorithm can deal with a content library containing $F = 50000$ files. These results verify the performance of the near-optimal algorithm and demonstrate its advantage of reduced complexity.

3) A 2-Tier HetNet vs a Single-Tier Network: Here we compare the total missing probability of the optimal HetNet



(a) cluster 1

(b) cluster 2

(c) cluster 3

(d) cluster 4

Fig. 13. Content placement probabilities in different clusters in the mCP model.

with the optimal single-tier network under the same conditions in Fig. 11. The system parameters are set as follows: the 2-tier network, $\mathbf{r} = (1, 0.4)$, $\mathbf{C} = (100, 40)$, $\mathbf{t} = (8, 4.1)$, $\boldsymbol{\beta} = (0.8, 4.1)$, $\boldsymbol{\lambda^{total}} = (1.20, 5.37)$; and the single-tier network 1, $r = 1$, $C = 100$, $t = 8$, $\beta = 0.8$, $\lambda^{total} = 1.64$; the single-tier network 2, $r = 0.4$, $C = 40$, $t = 4.1$, $\beta = 0.41$, $\lambda^{total} = 20$. For fairness of comparison, the coefficient T related to the BS deployment cost is set to be the same for the 2-tier HetNet and each single-tier network. When we increase T, the density of different tiers will increase correspondingly. In Fig. 11, we can see that for a moderate T, the 2-tier HetNet always leads to a much lower total missing probability, and is superior to both single-tier networks. However, as T decreases or increases, the performance of the HetNet will gradually converge to that of one of the two single-tier networks. This is expected because heterogeneity can provide performance gain over homogeneous networks.

4) Joint Optimization of Content Placement and BS Deployment vs Optimization of the Content Placement Only: In Fig. 12, we demonstrate the performance gain of joint optimization of content placement and BS deployment for a 4-tier HetNet. We compare it with the optimization of the content placement alone given the same BS density across different tiers under the same cost coefficient T. Some other parameters are set as follows: $\mathbf{r} = (1, 0.8, 0.4, 0.2)$, $\mathbf{C} = (100, 80, 40, 20)$, $\mathbf{t} = (10, 8, 4, 2)$, $\boldsymbol{\beta} = (1, 0.8, 0.4, 0.2)$, $\boldsymbol{\lambda^{total}} = (0.5, 1, 5, 50)$. We can see that when T is small, which means that the BS deployment density is small, joint optimization will offer a significantly better performance. When T increases which means that we densify the BSs, the gap between the two schemes gradually narrows. This is because when we place enough BSs on a plane, optimization of the densities of different tiers is less important to the system performance. We can also see that the near-optimal solution is almost identical to the optimal one.

C. mCP Model for User Distribution

Finally, we evaluation the performance for a 4-cluster 3-tier mCP model in Fig. 13. In our model, different clusters have different content placement probabilities and BSs densities. The model parameters are set as follows: $\mathbf{s} = (10, 9.6, 8.4, 7)$, $\mathbf{r} = (1, 0.8, 0.4)$, $\mathbf{C} = (35, 25, 20)$, $\mathbf{t} = (7.5, 5.5, 4.5)$, $\boldsymbol{\lambda_u} = (25, 28, 32, 40)$, $\boldsymbol{\beta} = (0.75, 0.55, 0.45)$, $\boldsymbol{\lambda^{total}} = (1, 1, 1, 1; 1, 1, 1, 1; 5, 5, 5, 5)$, $F = 100, T = 0.004$. In each cluster, we can still find the lower threshold. The trend of the content placement probabilities in the mCP model is similar to that in the PPP model.

VII. Conclusion

For IIoT, the application of low delay is particularly extensive. In industry 4.0, it needs high reliability of network to ensure the safety and efficiency of production process. Some special situations in IIoT will require the improvement of communications [30], [31]. In this work, we have jointly optimized the content placement and the BS activation to minimize the total missing probability in heterogeneous networks. We showed that the original optimization problem can be converted to a convex optimization problem and can be efficiently solved using the GP approach. We further derived an analytical result that all tiers should cache the same set of files with the probability 1, and also the same set files with the probability 0. Based on this result, we devised an efficient algorithm that can achieve near-optimal performance with much reduced complexity. We also extended the optimization to deal with a more realistic heterogeneous network in which the user distribution is modeled by the mCP. Numerical simulations have verified our derivation and analysis, and demonstrated significant performance improvement thanks to the joint optimization of content placement probabilities and BS densities.

References

[1] K.-K.-R. Choo, S. Gritzalis, and J. H. Park, "Cryptographic solutions for industrial Internet-of-things: Research challenges and opportunities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3567–3569, Aug. 2018.

[2] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in SDN-based industrial Internet of Things with edge computing," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1351–1360, Jun. 2018.

[3] 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.

[4] N. Wang, E. Hossain, and V. K. Bhargava, "Joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier HetNets with large-scale antenna arrays," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3251–3268, May 2016.

[5] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[6] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. 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.

[7] 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.

[8] 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.

[9] K. Avrachenkov, X. Bai, and J. Goseling, "Optimization of caching devices with geometric constraints," 2016, *arXiv:1602.03635*. [Online]. Available: http://arxiv.org/abs/1602.03635

[10] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3401–3415, May 2017.

[11] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 3358–3363.

[12] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[13] J. Wen, K. Huang, S. Yang, and V. O. K. Li, "Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5939–5952, Sep. 2017.

[14] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, Jun. 2017.

[15] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," 2017, *arXiv:1701.05761*. [Online]. Available: http://arxiv.org/abs/1701.05761

[16] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[17] S. H. Chae, J. Y. Ryu, T. Q. S. Quek, and W. Choi, "Cooperative transmission via caching helpers," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[18] W. C. Ao and K. Psounis, "Distributed caching and small cell cooperation for fast content delivery," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Hangzhou, China, 2015, pp. 127–136.

[19] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[20] B. Jiang, J. Yang, G. Ding, and H. Wang, "Cyber-physical security design in multimedia data cache resource allocation for industrial networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 12, pp. 6472–6480, Dec. 2019.

[21] M. Afshang and H. S. Dhillon, "Optimal geographic caching in finite wireless networks," in *Proc. IEEE 17th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–5.

[22] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[23] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[24] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[25] S. Krishnan and H. S. Dhillon, "Distributed caching in device-to-device networks: A stochastic geometry perspective," in *Proc. 49th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2015, pp. 1280–1284.

[26] C. Saha, M. Afshang, and H. S. Dhillon, "Poisson cluster process: Bridging the gap between PPP and 3GPP HetNet models," 2017, *arXiv:1702.05706*. [Online]. Available: http://arxiv.org/abs/1702.05706

[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[28] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi, "A tutorial on geometric programming," *Optim. Eng.*, vol. 8, no. 1, pp. 67–127, May 2007.

[29] M. Grant and S. Boyd. (Mar. 2014). *CVX: MATLAB Software for Disciplined Convex Programming*. version 2.1. [Online]. Available: http://cvxr.com/cvx

[30] B. Jiang, J. Yang, H. Xu, H. Song, and G. Zheng, "Multimedia data throughput maximization in Internet-of-Things system based on optimization of cache-enabled UAV," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3525–3532, Apr. 2019.

[31] M. Aazam, K. A. Harras, and S. Zeadally, "Fog computing for 5G tactile industrial Internet of Things: QoE-aware resource allocation model," *IEEE Trans. Ind. Informat.*, vol. 15, no. 5, pp. 3085–3092, May 2019.

Jiachen Yang (Member, IEEE) received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2005 and 2009, respectively. He was a Visiting Scholar with the Department of Computer Science, School of Science, Loughborough University, U.K. He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include multimedia signal processing, mathematical optimization, and deep learning.

Chaofan Ma received the B.S. degree from the School of Materials Science and Engineering, Tianjin University, Tianjin, China, in 2016. He is currently pursuing the M.S. degree with the School of Electrical and Information Engineering, Tianjin University. His research interests lie in wireless edge cache and mathematical optimization for networks.

Bin Jiang received the B.S. and M.S. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree. He is also a Visiting Scholar with the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, where he is a member of the Security and Optimization for Networked Globe Laboratory. His research interests mainly lie in the Internet of Things and information security.

Guiguang Ding (Member, IEEE) received the Ph.D. degree in electronic engineering from Xidian University, Xi’an, China, in 2014. He was a Post-Doctoral Research Fellow with the Department of Automation, Tsinghua University, Beijing, China. He is currently an Associate Professor with the School of Software, Tsinghua University. He has authored 80 articles in major journals and conferences. His current research centers on the area of multimedia information retrieval, computer vision, and machine learning.

Gan Zheng (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in electronic and information engineering from Tianjin University, Tianjin, China, in 2002 and 2004, respectively, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong, in 2008. He is currently a Reader of signal processing for wireless communications with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, U.K. His research interests include machine learning for communications, UAV communications, mobile edge caching, full-duplex radio, and wireless power transfer. He was a recipient of the 2013 IEEE Signal Processing Letters Best Paper Award, the 2015 GLOBECOM Best Paper Award, and the 2018 IEEE Technical Committee on Green Communications and Computing Best Paper Award. He currently serves as an Associate Editor for the IEEE COMMUNICATIONS LETTERS.

Huihui Wang (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from The University of Virginia, Charlottesville, VA, USA, in 2013. In 2011, she was an Engineering Intern with Qualcomm, Inc. She is currently with the Department of Engineering, Jacksonville University, Jacksonville, FL, USA. In 2013, she joined the Department of Engineering, Jacksonville University, where she is also an Assistant Professor and the Founding Chair with the Department of Engineering. She has authored over 30 articles and holds one U.S. patent. Her research interests include cyber-physical systems, the Internet of Things, healthcare, and medical engineering.