Analysis and Optimization of Caching and Multicasting in Large-Scale Cache-Enabled Heterogeneous Wireless Networks

Ying Cui, *MIEEE*, Dongdong Jiang, *StMIEEE*

Abstract

Heterogeneous wireless networks (HetNets) provide a powerful approach to meet the dramatic mobile traffic growth, but also impose a significant challenge on backhaul. Caching and multicasting at macro and pico base stations (BSs) are two promising methods to support massive content delivery and reduce backhaul load in HetNets. In this paper, we jointly consider caching and multicasting in a large-scale cache-enabled HetNet with backhaul constraints. We propose a hybrid caching design consisting of identical caching in the macro-tier and random caching in the pico-tier, and a corresponding multicasting design. By carefully handling different types of interferers and adopting appropriate approximations, we derive tractable expressions for the successful transmission probability in the general region as well as the high signal-to-noise ratio (SNR) and user density region, utilizing tools from stochastic geometry. Then, we consider the successful transmission probability maximization by optimizing the design parameters, which is a very challenging mixed discrete-continuous optimization problem. By using optimization techniques and exploring the structural properties, we obtain a near optimal solution with superior performance and manageable complexity. This solution achieves better performance in the general region than any asymptotically optimal solution, under a mild condition. The analysis and optimization results provide valuable design insights for practical cache-enabled HetNets.

Index Terms

Cache, multicast, backhaul, stochastic geometry, optimization, heterogenous wireless network

Y. Cui and D. Jiang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, China.

I. Introduction

The rapid proliferation of smart mobile devices has triggered an unprecedented growth of the global mobile data traffic. HetNets have been proposed as an effective way to meet the dramatic traffic growth by deploying short range small-BSs together with traditional macro-BSs, to provide better time or frequency reuse [1]. However, this approach imposes a significant challenge of providing expensive high-speed backhaul links for connecting all the small-BSs to the core network [2].

Caching at small-BSs is a promising approach to alleviate the backhaul capacity requirement in HetNets [3]–[5]. Many existing works have focused on optimal cache placement at small-BSs, which is of critical importance in cache-enabled HetNets. For example, in [6] and [7], the authors consider the optimal content placement at small-BSs to minimize the expected downloading time for files in a single macro-cell with multiple small-cells. File requests which cannot be satisfied locally at a small-BS are served by the macro-BS. The optimization problems in [6] and [7] are NP-hard, and low-complexity solutions are proposed. In [8], the authors propose a caching design based on file splitting and MDS encoding in a single macro-cell with multiple small-cells. File requests which cannot be satisfied locally at a small-BS are served by the macro-BS, and backhaul rate analysis and optimization are considered. Note that the focuses of [6]–[8] are on performance optimization of caching design.

In [9]–[11], the authors consider caching the most popular files at each small-BS in large-scale cache-enabled small-cell networks or HetNets, with backhaul constraints. The service rates of uncached files are limited by the backhaul capacity. In [12], the authors propose a partion-based combined caching design in a large-scale cluster-centric small-cell network, without considering backhaul constraints. In [13], the authors consider two caching designs, i.e., caching the most popular files and random caching of a uniform distribution, at small-BSs in a large-scale cache-enabled HetNet, without backhaul constraints. File requests which cannot be satisfied at a small-BS are served by macro-BSs. In [14], the authors consider random caching of a uniform distribution in a large-scale cache-enabled small-cell network, without backhaul constraints, assuming that content requests follow a uniform distribution. Note that the focuses of [9]–[14] are on performance analysis of caching designs.

On the other hand, enabling multicast service at BSs in HetNets is an efficient way to deliver popular contents to multiple requesters simultaneously, by effectively utilizing the broadcast nature of the wireless medium [15]. In [16] and [17], the authors consider a single macro-cell with multiple small-cells with backhaul costs. Specifically, in [16], the optimization of caching and multicasting, which is NP-hard, is considered, and a simplified solution with approximation guarantee is proposed. In [17], the optimization of dynamic multicast scheduling for a given content placement, which is a dynamic programming problem, is considered, and a low-complexity optimal numerical solution is obtained.

The network models considered in [6]–[8], [16], [17] do not capture the stochastic natures of channel fading and geographic locations of BSs and users. The network models considered in [9]–[14] are more realistic and can reflect the stochastic natures of signal and interference. However, the simple identical caching design considered in [9]–[11], [13] does not provide spatial file diversity; the combined caching design in [12] does not reflect the popularity differences of files in each of the three categories; and the random caching design of a uniform distribution in [13], [14] cannot make use of popularity information. Hence, the caching designs in [9]–[14] may not lead to good network performance. On the other hand, [18]–[21] consider analysis and optimization of caching in large-scale cache-enabled single-tier networks. Specifically, [18] considers random caching at BSs, and analyze and optimize the hit probability. Reference [19] considers random caching with contents being stored at each BS in an i.i.d. manner, and analyzes the minimum offloading loss. In [20], the authors study the expected costs of obtaining a complete content under random uncoded caching and coded caching strategies, which are designed only for different pieces of a single content. In [21], the authors consider analysis and optimization of joint caching and multicasting. However, the proposed caching and multicasting designs in [18]–[21] may not be applicable to HetNets with backhaul constraints. In summary, to facilitate designs of practical cache-enabled HetNets for massive content dissemination, further studies are required to understand the following key questions.

- How do physical layer and content-related parameters fundamentally affect performance of cache-enabled HetNets?
- How can caching and multicasting jointly and optimally assist massive content dissemination in cache-enabled HetNets?

In this paper, we consider the analysis and optimization of joint caching and multicasting to improve the efficiency of massive content dissemination in a large-scale cache-enabled HetNet with backhaul constraints. Our main contributions are summarized below.

- First, we propose a hybrid caching design with certain design parameters, consisting of identical caching in the macro-tier and random caching in the pico-tier, which can provide spatial file diversity. We propose a corresponding multicasting design for efficient content dissemination by exploiting broadcast nature of the wireless medium.

- Then, by carefully handling different types of interferers and adopting appropriate approximations, we derive tractable expressions for the successful transmission probability in the general region and the asymptotic region, utilizing tools from stochastic geometry. These expressions reveal the impacts of physical layer and content-related parameters on the successful transmission probability.
- Next, we consider the successful transmission probability maximization by optimizing the design parameters, which is a very challenging mixed discrete-continuous optimization problem. We propose a two-step optimization framework to obtain a near optimal solution with superior performance and manageable complexity. Specifically, we first characterize the structural properties of the asymptotically optimal solutions. Then, based on these properties, we obtain the near optimal solution, which achieves better performance in the general region than any asymptotically optimal solution, under a mild condition.
- Finally, by numerical simulations, we show that the near optimal solution achieves a significant gain in successful transmission probability over some baseline schemes.

II. Network Model

We consider a two-tier HetNet where a macro-cell tier is overlaid with a pico-cell tier, as shown in Fig. 1. The locations of the macro-BSs and the pico-BSs are spatially distributed as two independent homogeneous Poisson point processes (PPPs) Φ_1 and Φ_2 with densities λ_1 and λ_2, respectively, where $\lambda_1 < \lambda_2$. The locations of the users are also distributed as an independent homogeneous PPP Φ_u with density λ_u. We refer to the macro-cell tier and the pico-cell tier as the 1st tier and the 2nd tier, respectively. Consider the downlink scenario. Each BS in the jth tier has one transmit antenna with transmission power P_j ($j = 1, 2$), where $P_1 > P_2$. Each user has one receive antenna. All BSs are operating on the same frequency band of total bandwidth W (Hz). Consider a discrete-time system with time being slotted and study one slot of the network. We consider both large-scale fading and small-scale fading. Due to large-scale fading, a transmitted signal from the jth tier with distance D is attenuated by a factor $\frac{1}{D^{\alpha_j}}$, where $\alpha_j > 2$ is the path loss exponent of the jth tier. For small-scale fading, we assume Rayleigh fading channels [22], [23].

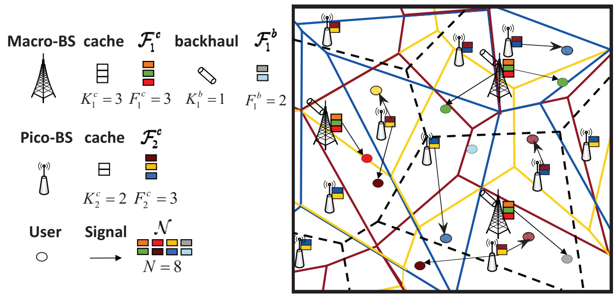


Fig. 1. Network model. The 1st tier corresponds to a Voronoi tessellation (cf. black dashed line segments), determined by the locations of all the macro-BSs. Each file $n \in \mathcal{F}_2^c$ corresponds to a Voronoi tessellation (cf. solid line segments in the same color as the file), determined by the locations of all the pico-BSs storing this file.

Let $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$ denote the set of N files (e.g., data objects or chucks of data objects) in the HetNet. For ease of illustration, assume that all files have the same size.[1] Each file is of certain popularity, which is assumed to be identical among all users. Each user randomly requests one file, which is file $n \in \mathcal{N}$ with probability $a_n \in (0, 1)$, where $\sum_{n \in \mathcal{N}} a_n = 1$. Thus, the file popularity distribution is given by $\mathbf{a} \triangleq (a_n)_{n \in \mathcal{N}}$, which is assumed to be known apriori. In addition, without loss of generality (w.l.o.g.), assume $a_1 > a_2 > \ldots > a_N$.

The HetNet consists of cache-enabled macro-BSs and pico-BSs. Each BS in the jth tier is equipped with a cache of size $K_j^c < N$ to store different files. Assume $K_1^c + K_2^c \leq N$. Each macro-BS is connected to the core network via a wireline backhaul link of transmission capacity $K_1^b < N$ (files/slot), i.e., each macro-BS can retrieve at most K_1^b different files from the core network in each slot.[2] Note that K_1^c, K_2^c and K_1^b reflect the storage and backhaul resources in the cache-enabled HetNet.

III. Joint Caching and Multicasting

We are interested in the case where the storage and backhaul resources are limited, and may not be able to satisfy all file requests. In this section, we propose a joint caching

[1] Files of different sizes can be divided into chunks of the same length. Thus, the results in this paper can be extended to the case of different file sizes.

[2] Note that storing or retrieving more than one copies of the same file at one BS is redundant and will waste storage or backhaul resources.

and multicasting design with certain design parameters, which can provide high spatial file diversity and ensure efficient content dissemination.

A. Hybrid Caching

To provide high spatial file diversity, we propose a *hybrid caching design* consisting of identical caching in the 1st tier and random caching in the 2nd tier, as illustrated in Fig. 1. Let $\mathcal{F}_j^c \subseteq \mathcal{N}$ denote the set of $F_j^c \triangleq |\mathcal{F}_j^c|$ files cached in the jth tier. Specifically, our hybrid caching design satisfies the following requirements: (i) *non-overlapping caching across tiers*: each file is stored in at most one tier; (ii) *identical caching in the 1st tier*: each macro-BS stores the same set $\mathcal{F}_1^c$ of K_1^c (different) files; and (iii) *random caching in the 2nd tier*: each pico-BS randomly stores K_2^c different files out of all files in $\mathcal{F}_2^c$, forming a subset of $\mathcal{F}_2^c$. Thus, we have the following constraint:

$$\mathcal{F}_1^c, \mathcal{F}_2^c \subseteq \mathcal{N},\ \mathcal{F}_1^c \cap \mathcal{F}_2^c = \emptyset,\ F_1^c = K_1^c,\ F_2^c \geq K_2^c. \tag{1}$$

To further illustrate the random caching in the 2nd tier, we first introduce some notations. We say every K_2^c different files in $\mathcal{F}_2^c$ form a combination. Thus, there are $I \triangleq \binom{F_2^c}{K_2^c}$ different combinations in total. Let $\mathcal{I} \triangleq \{1, 2, \cdots, I\}$ denote the set of I combinations. Combination $i \in \mathcal{I}$ can be characterized by an F_2^c-dimensional vector $\mathbf{x}_i \triangleq (x_{i,n})_{n \in \mathcal{F}_2^c}$, where $x_{i,n} = 1$ indicates that file $n \in \mathcal{F}_2^c$ is included in combination i and $x_{i,n} = 0$ otherwise. Note that there are K_2^c 1's in each $\mathbf{x}_i$. Denote $\mathcal{N}_i \triangleq \{n \in \mathcal{F}_2^c : x_{i,n} = 1\} \subseteq \mathcal{F}_2^c$ as the set of K_2^c files contained in combination i. Each pico-BS stores one combination at random, which is combination $i \in \mathcal{I}$ with probability p_i satisfying:[3]

$$0 \leq p_i \leq 1,\ i \in \mathcal{I}, \tag{2}$$

$$\sum_{i \in \mathcal{I}} p_i = 1. \tag{3}$$

Denote $\mathbf{p} \triangleq (p_i)_{i \in \mathcal{I}}$. To facilitate the analysis in later sections, based on $\mathbf{p}$, we also define the probability that file $n \in \mathcal{F}_2^c$ is stored at a pico-BS, i.e.,

$$T_n \triangleq \sum_{i \in \mathcal{I}_n} p_i,\ n \in \mathcal{F}_2^c, \tag{4}$$

[3]In this paper, to understand the natures of joint caching and multicasting in cache-enabled HetNets, we shall first pose the analysis and optimization on the basis of all the file combinations in $\mathcal{I}$ (for the 2nd tier). Then, based on the insights obtained, we shall focus on reducing complexity while maintaining superior performance.

where $\mathcal{I}_n \triangleq \{i \in \mathcal{I} : x_{i,n} = 1\}$ denotes the set of $I_n \triangleq \binom{F_2^c-1}{K_2^c-1}$ combinations containing file $n \in \mathcal{F}_2^c$. Denote $\mathbf{T} \triangleq (T_n)_{n\in\mathcal{F}_2^c}$. Note that $\mathbf{p}$ and $\mathbf{T}$ depend on $\mathcal{F}_2^c$. Thus, in this paper, we use $\mathbf{p}(\mathcal{F}_2^c)$ and $\mathbf{T}(\mathcal{F}_2^c)$ when emphasizing this relation. Therefore, the hybrid caching design in the cache-enabled HetNet is specified by the design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$.

To efficiently utilize backhaul links and ensure high spatial file diversity, we only retrieve files not stored in the cache-enabled HetNet via backhaul links. Let $\mathcal{F}_1^b \subseteq \mathcal{N}$ denote the set of $F_1^b \triangleq |\mathcal{F}_1^b|$ files which can be retrieved by each macro-BS from the core network. Thus, we have the following constraint:

$$\mathcal{F}_1^b = \mathcal{N} \setminus (\mathcal{F}_1^c \cup \mathcal{F}_2^c). \tag{5}$$

Therefore, the file distribution in the cache-enabled HetNet is fully specified by the hybrid caching design $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$.

B. Multicasting

In this part, we propose a multicasting design associated with the hybrid caching design $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$. First, we introduce the user association under the proposed hybrid caching design. In the cache-enabled HetNet, a user accesses to a tier based on its desired file. Specifically, each user requesting file $n \in \mathcal{F}_1^c \cup \mathcal{F}_1^b$ is associated with the nearest macro-BS and is referred to as a macro-user. While, each user requesting file $n \in \mathcal{F}_2^c$ is associated with the nearest pico-BS storing a combination $i \in \mathcal{I}_n$ (containing file n) and is referred to as a pico-user. The associated BS of each user is called its serving BS, and offers the maximum long-term average receive power for its desired file [23]. Note that under the proposed hybrid caching design $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$, the serving BS of a macro-user is its nearest macro-BS, while the serving BS of a pico-user (affected by $\mathbf{p}$) may not be its geographically nearest BS. We refer to this association mechanism as the *content-centric association* in the cached-enabled HetNet, which is different from the traditional *connection-based association* [23] in HetNets.

Now, we introduce file scheduling in the cache-enabled HetNet. Each BS will serve all the cached files requested by its associated users. Each macro-BS only serves at most K_1^b uncached files requested by its associated users, due to the backhaul constraint for retrieving uncached files. In particular, if the users of a macro-BS request smaller than or equal to K_1^b different uncached files, the macro-BS serves all of them; if the users of a macro-BS request greater than K_1^b different uncached files, the macro-BS will randomly select K_1^b different requested uncached files to serve, out of all the requested uncached files according to the uniform distribution.

We consider multicasting[4] in the cache-enabled HetNet for efficient content dissemination. Suppose a BS schedules to serve requests for k different files. Then, it transmits each of the k files at rate τ (bit/second) and over $\frac{1}{k}$ of total bandwidth W using FDMA. All the users which request one of the k files from this BS try to decode the file from the single multicast transmission of the file at the BS. Note that, by avoiding transmitting the same file multiple times to multiple users, this content-centric transmission (multicast) can improve the efficiency of the utilization of the wireless medium and reduce the load of the wireless links, compared to the traditional connection-based transmission (unicast).

From the above illustration, we can see that the proposed multicasting design is also affected by the proposed hybrid caching design $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$. Therefore, the design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$ affect the performance of the proposed joint caching and multicasting design.

IV. Performance Metric

In this paper, w.l.o.g., we study the performance of the typical user denoted as u_0, which is located at the origin. We assume all BSs are active. Suppose u_0 requests file n. Let j_0 denote the index of the tier to which u_0 belongs, and let $\bar{j}_0$ denote the other tier. Let $\ell_0 \in \Phi_{j_0}$ denote the index of the serving BS of u_0. We denote $D_{j,\ell,0}$ and $h_{j,\ell,0} \overset{d}{\sim} \mathcal{CN}(0,1)$ as the distance and the small-scale channel between BS $\ell \in \Phi_j$ and u_0, respectively. We assume the complex additive white Gaussian noise of power N_0 at u_0. When u_0 requests file n and file n is transmitted by BS ℓ_0, the signal-to-interference plus noise ratio (SINR) of u_0 is given by

$$\mathrm{SINR}_{n,0} = \frac{D_{j_0,\ell_0,0}^{-\alpha_{j_0}} \left|h_{j_0,\ell_0,0}\right|^2}{\sum_{\ell \in \Phi_{j_0} \setminus \ell_0} D_{j_0,\ell,0}^{-\alpha_{j_0}} \left|h_{j_0,\ell,0}\right|^2 + \sum_{\ell \in \Phi_{\bar{j}_0}} D_{\bar{j}_0,\ell,0}^{-\alpha_{\bar{j}_0}} \left|h_{\bar{j}_0,\ell,0}\right|^2 \frac{P_{\bar{j}_0}}{P_{j_0}} + \frac{N_0}{P_{j_0}}}, \ n \in \mathcal{N}. \tag{6}$$

When u_0 requests file $n \in \mathcal{F}_1^c$ ($n \in \mathcal{F}_1^b$), let $K_{1,n,0}^c \in \{1, \cdots, K_1^c\}$ ($\overline{K}_{1,n,0}^c \in \{0, \cdots, K_1^c\}$) and $\overline{K}_{1,n,0}^b \in \{0, \cdots, F_1^b\}$ ($K_{1,n,0}^b \in \{1, \cdots, F_1^b\}$) denote the numbers of different cached and uncached files requested by the users associated with BS $\ell_0 \in \Phi_1$, respectively. When u_0 requests file $n \in \mathcal{F}_2^c$, let $K_{2,n,0}^c \in \{1, \cdots, K_2^c\}$ denote the number of different cached files requested by the users associated with BS $\ell_0 \in \Phi_2$. Note that $K_{1,n,0}^c, \overline{K}_{1,n,0}^b, \overline{K}_{1,n,0}^c, K_{1,n,0}^b, K_{2,n,0}^c$ are discrete random variables, the probability mass functions (p.m.f.s) of which depend on $\mathbf{a}$,

[4]Note that in this paper, the multicast service happens once every slot, and hence no additional delay is introduced.

$$q_1(\mathcal{F}_1^c, \mathcal{F}_2^c) = \sum_{n \in \mathcal{F}_1^c} a_n \Pr\left[\frac{W}{K_{1,n,0}^c + \min\{K_1^b, \overline{K}_{1,n,0}^b\}} \log_2\left(1 + \text{SINR}_{n,0}\right) \geq \tau\right]$$
$$+ \sum_{n \in \mathcal{F}_1^b} a_n \Pr\left[\frac{W}{\overline{K}_{1,n,0}^c + \min\{K_1^b, K_{1,n,0}^b\}} \log_2\left(1 + \text{SINR}_{n,0}\right) \geq \tau, n \text{ is selected}\right] \quad (8)$$
$$q_2(\mathcal{F}_2^c, \mathbf{p}) = \sum_{n \in \mathcal{F}_2^c} a_n \Pr\left[\frac{W}{K_{2,n,0}^c} \log_2\left(1 + \text{SINR}_{n,0}\right) \geq \tau\right] \quad (9)$$

λ_u and the design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$. In addition, if $n \in \mathcal{F}_1^c \cup \mathcal{F}_2^c$, BS ℓ_0 will transmit file n for sure; if $n \in \mathcal{F}_1^b$, for given $K_{1,n,0}^b = k^b \geq 1$, BS ℓ_0 will transmit file n with probability $\frac{\min\{k^b, K_1^b\}}{k^b}$. Given that file n is transmitted, it can be decoded correctly at u_0 if the channel capacity between BS ℓ_0 and u_0 is greater than or equal to τ. Requesters are mostly concerned with whether their desired files can be successfully received. Therefore, in this paper, we consider the successful transmission probability of a file requested by u_0 as the network performance metric. By total probability theorem, the successful transmission probability under the proposed scheme is given by:

$$q(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}) = q_1(\mathcal{F}_1^c, \mathcal{F}_2^c) + q_2(\mathcal{F}_2^c, \mathbf{p}), \quad (7)$$

where $\mathcal{F}_1^b$ is given by (5), and $q_1(\mathcal{F}_1^c, \mathcal{F}_2^c)$ and $q_2(\mathcal{F}_2^c, \mathbf{p})$ are given by (8) and (9), respectively. Note that in (8) and (9), each term multiplied by a_n represents the successful transmission probability of file n.

Later, we shall see that under the proposed caching and multicasting design for content-oriented services in the cache-enabled HetNet, the successful transmission probability is sufficiently different from the traditional rate coverage probability studied for connection-oriented services [23]. In particular, the successful transmission probability considered in this paper not only depends on the physical layer parameters, such as the macro and pico BS densities λ_1 and λ_2, user density λ_u, path loss exponents α_1 and α_2, bandwidth W, backhaul capacity K_1^b and transmit signal-to-noise ratios (SNRs) $\frac{P_1}{N_0}$ and $\frac{P_2}{N_0}$, but also relies on the content-related parameters, such as the popularity distribution $\mathbf{a}$, the cache sizes K_1^c and K_2^c, and the design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}(\mathcal{F}_2^c))$. While, the traditional rate coverage probability only depends on the physical layer parameters. In addition, the successful transmission probability depends on the physical layer parameters in a different way from the traditional rate coverage probability.

For example, the content-centric association leads to different distributions of the locations of serving and interfering BSs; the multicasting transmission results in different file load distributions at each BS [23]; and the cache-enabled architecture makes content availability related to BS densities.

V. Performance Analysis

In this section, we study the successful transmission probability $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ under the proposed caching and multicasting design for given design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}(\mathcal{F}_2^c))$. First, we analyze the successful transmission probability in the general region. Then, we analyze the asymptotic transmission probability in the high SNR and user density region.

A. Performance Analysis in General Region

In this part, we would like to analyze the successful transmission probability in the general region, using tools from stochastic geometry. In general, file loads $K_{1,n,0}^c$, $\overline{K}_{1,n,0}^b$, $\overline{K}_{1,n,0}^c$, $K_{1,n,0}^b$, $K_{2,n,0}^c$ and SINR $\mathrm{SINR}_{n,0}$ are correlated in a complex manner, as BSs with larger association regions have higher file load and lower SINR (due to larger user to BS distance) [24]. For the tractability of the analysis, as in [23] and [24], the dependence is ignored. Therefore, to obtain the successful transmission probability in (7), we analyze the distributions of $K_{1,n,0}^c$, $\overline{K}_{1,n,0}^b$, $\overline{K}_{1,n,0}^c$, $K_{1,n,0}^b$, $K_{2,n,0}^c$ and the distribution of $\mathrm{SINR}_{n,0}$, separately.

First, we calculate the p.m.f.s of $K_{1,n,0}^c$ and $\overline{K}_{1,n,0}^b$ for $n \in \mathcal{F}_1^c$ as well as the p.m.f.s of $\overline{K}_{1,n,0}^c$ and $K_{1,n,0}^b$ for $n \in \mathcal{F}_1^b$. In calculating these p.m.f.s, we need the probability density function (p.d.f.) of the size of the Voronoi cell of ℓ_0 w.r.t. file $m \in \mathcal{F}_1^c \cup \mathcal{F}_1^b \setminus \{n\}$. Note that this p.d.f. is equivalent to the p.d.f. of the size of the Voronoi cell to which a randomly chosen user belongs. Based on a tractable approximated form of this p.d.f. in [25], which is widely used in existing literature [23], [24], we obtain the p.m.f.s of $K_{1,n,0}^c$, $\overline{K}_{1,n,0}^b$, $\overline{K}_{1,n,0}^c$ and $K_{1,n,0}^b$.

Lemma 1 (p.m.f.s of $K_{1,n,0}^c$ $\overline{K}_{1,n,0}^b$, $\overline{K}_{1,n,0}^c$ and $K_{1,n,0}^b$): The p.m.f.s of $K_{1,n,0}^c$ and $\overline{K}_{1,n,0}^b$ for

$n \in \mathcal{F}_1^c$ and the p.m.f.s of $\overline{K}_{1,n,0}^c$ and $K_{1,n,0}^b$ for $n \in \mathcal{F}_1^b$ are given by

$$\Pr\left[K_{1,n,0}^c = k^c\right] = g(\mathcal{F}_{1,-n}^c, k^c - 1), \quad k^c = 1, \cdots, K_1^c, \tag{10}$$

$$\Pr\left[\overline{K}_{1,n,0}^b = k^b\right] = g(\mathcal{F}_1^b, k^b), \quad k^b = 0, \cdots, F_1^b, \tag{11}$$

$$\Pr\left[\overline{K}_{1,n,0}^c = k^c\right] = g(\mathcal{F}_1^c, k^c), \quad k^c = 0, \cdots, K_1^c, \tag{12}$$

$$\Pr\left[K_{1,n,0}^b = k^b\right] = g(\mathcal{F}_{1,-n}^b, k^b - 1), \quad k^b = 1, \cdots, F_1^b, \tag{13}$$

where $g(\mathcal{F}, k) \triangleq \sum_{\mathcal{X} \in \{\mathcal{S} \subseteq \mathcal{F} : |\mathcal{S}| = k\}} \prod_{m \in \mathcal{X}} \left(1 - \left(1 + \frac{a_m \lambda_u}{3.5 \lambda_1}\right)^{-4.5}\right) \prod_{m \in \mathcal{F} \setminus \mathcal{X}} \left(1 + \frac{a_m \lambda_u}{3.5 \lambda_1}\right)^{-4.5}$, $\mathcal{F}_{1,-n}^c \triangleq \mathcal{F}_1^c \setminus \{n\}$ and $\mathcal{F}_{1,-n}^b \triangleq \mathcal{F}_1^b \setminus \{n\}$.

Proof: Please refer to Appendix A. ■

Next, we obtain the p.m.f. of $K_{2,n,0}^c$ for $n \in \mathcal{F}_2^c$. In calculating the p.m.f. of $K_{2,n,0}^c$, we need the p.d.f. of the size of the Voronoi cell of ℓ_0 w.r.t. file $m \in \mathcal{N}_i \setminus \{n\}$ when ℓ_0 contains combination $i \in \mathcal{I}_n$. However, this p.d.f. is very complex and is still unknown. For the tractability of the analysis, as in [21], we approximate this p.d.f. based on a tractable approximated form of the p.d.f. of the size of the Voronoi cell to which a randomly chosen user belongs [25], which is widely used in existing literature [23], [24]. Under this approximation, we obtain the p.m.f. of $K_{2,n,0}^c$.

Lemma 2 (p.m.f. of $K_{2,n,0}^c$): The p.m.f. of $K_{2,n,0}^c$ for $n \in \mathcal{F}_2^c$ is given by

$$\Pr\left[K_{2,n,0}^c = k^c\right] = \sum_{i \in \mathcal{I}_n} \frac{p_i}{T_n} \sum_{\mathcal{X} \in \{\mathcal{S} \subseteq \mathcal{N}_{i,-n} : |\mathcal{S}| = k^c - 1\}} \prod_{m \in \mathcal{X}} \left(1 - \left(1 + \frac{a_m \lambda_u}{3.5 T_m \lambda_2}\right)^{-4.5}\right) \prod_{m \in \mathcal{N}_{i,-n} \setminus \mathcal{X}} \left(1 + \frac{a_m \lambda_u}{3.5 T_m \lambda_2}\right)^{-4.5},$$
$$k^c = 1, \cdots, K_2^c, \tag{14}$$

where $\mathcal{N}_{i,-n} \triangleq \mathcal{N}_i \setminus \{n\}$.

Proof: Please refer to Appendix B. ■

The distributions of the locations of desired transmitters and interferers are more involved than those in the traditional connection-based HetNets. Thus, it is more challenging to analyze the p.d.f. of $\mathrm{SINR}_{n,0}$. When u_0 is a macro-user, as in the traditional connection-based HetNets, there are two types of interferers, namely, i) all the other macro-BSs besides its serving macro-BS, and ii) all the pico-BSs. When u_0 is a pico-user, different from the traditional connection-based HetNets, there are three types of interferers, namely, i) all the other pico-BSs storing the combinations containing the desired file of u_0 besides its serving pico-BS,

ii) all the pico-BSs without the desired file of u_0, and iii) all the macro-BSs. By carefully handling these distributions, we can derive the p.d.f. of $\mathrm{SINR}_{n,0}$, for $n \in \mathcal{F}_1^c \cup \mathcal{F}_1^b$ and $n \in \mathcal{F}_2^c$, respectively.

Then, based on Lemma 1 and Lemma 2 as well as the p.d.f. of $\mathrm{SINR}_{n,0}$, we can derive the successful transmission probability $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$.

Theorem 1 (Performance): The successful transmission probability $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ of u_0 is given by

$$
\begin{aligned}
q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right) = & \sum_{n \in \mathcal{F}_1^c} a_n \sum_{k^c=1}^{K_1^c} \sum_{k^b=0}^{F_1^b} \Pr[K_{1,n,0}^c = k^c] \Pr[\overline{K}_{1,n,0}^b = k^b] f_{1,k^c+\min\{K_1^b,k^b\}} \\
& + \sum_{n \in \mathcal{F}_1^b} a_n \sum_{k^c=0}^{K_1^c} \sum_{k^b=1}^{F_1^b} \Pr[\overline{K}_{1,n,0}^c = k^c] \Pr[K_{1,n,0}^b = k^b] \frac{\min\{K_1^b, k^b\}}{k^b} f_{1,k^c+\min\{K_1^b,k^b\}} \\
& + \sum_{n \in \mathcal{F}_2^c} a_n \sum_{k^c=1}^{K_2^c} \Pr[K_{2,n,0} = k^c] f_{2,k^c}(T_n), \qquad (15)
\end{aligned}
$$

where the p.m.f.s of $K_{1,n,0}^c$ $\overline{K}_{1,n,0}^b$, $\overline{K}_{1,n,0}^c$, $K_{1,n,0}^b$ and $K_{2,n,0}^c$ are given by Lemma 1 and Lemma 2, $f_{1,k}$ and $f_{2,k}(T_n)$ are given by (16) and (17), and T_n is given by (4). Here, $B'\left(x,y,z\right) \triangleq \int_z^1 u^{x-1}\left(1-u\right)^{y-1} \mathrm{d}u$ and $B(x,y) \triangleq \int_0^1 u^{x-1}\left(1-u\right)^{y-1} \mathrm{d}u$ denote the complementary incomplete Beta function and the Beta function, respectively.

Proof: Please refer to Appendix C. ■

From Theorem 1, we can see that in the general region, the physical layer parameters α_1, α_2, W, λ_1, λ_2, λ_u, $\frac{P_1}{N_0}$, $\frac{P_2}{N_0}$, and the design parameters $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$ jointly affect the successful transmission probability $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$. The impacts of the physical layer parameters and the design parameters on $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ are coupled in a complex manner.

B. Performance Analysis in Asymptotic Region

In this part, to obtain design insights, we focus on analyzing the asymptotic successful transmission probability in the high SNR and user density region. Note that in the remaining of the paper, when considering the high SNR region, we assume $P_1 = \beta P$ and $P_2 = P$ for some $\beta > 1$ and $P > 0$, and let $\frac{P}{N_0} \to \infty$. On the other hand, in the high user density region where $\lambda_u \to \infty$, discrete random variables $K_{1,n,0}^c, \overline{K}_{1,n,0}^c \to K_1^c$, $\overline{K}_{1,n,0}^b, K_{1,n,0}^b \to F_1^b$ and $K_{2,n,0}^c \to K_2^c$ in distribution. Define $q_{1,\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right) \triangleq \lim_{\frac{P}{N_0}\to\infty, \lambda_u\to\infty} q_1\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right)$, $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right) \triangleq \lim_{\frac{P}{N_0}\to\infty, \lambda_u\to\infty} q_2\left(\mathcal{F}_2^c, \mathbf{p}\right)$, and $q_{\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right) \triangleq \lim_{\frac{P}{N_0}\to\infty, \lambda_u\to\infty} q(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p})$.

$$f_{1,k} \triangleq 2\pi\lambda_1 \int_0^\infty d \exp\left(-\left(2^{\frac{k\tau}{W}}-1\right) d^{\alpha_1} \frac{N_0}{P_1}\right) \exp\left(-\frac{2\pi\lambda_2}{\alpha_2} d^{\frac{2\alpha_1}{\alpha_2}} \left(\frac{P_2}{P_1}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)$$
$$\times \exp\left(-\frac{2\pi\lambda_1}{\alpha_1} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_1}} B^{'}\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}, 2^{-\frac{k\tau}{W}}\right)\right) \exp\left(-\pi\lambda_1 d^2\right) \mathrm{d}d. \quad (16)$$

$$f_{2,k}(x) \triangleq 2\pi\lambda_2 x \int_0^\infty d \exp\left(-\left(2^{\frac{k\tau}{W}}-1\right) d^{\alpha_2} \frac{N_0}{P_2}\right) \exp\left(-\frac{2\pi\lambda_1}{\alpha_1} d^{\frac{2\alpha_2}{\alpha_1}} \left(\frac{P_1}{P_2}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right)$$
$$\times \exp\left(-\frac{2\pi\lambda_2}{\alpha_2} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} \left(x B^{'}\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right) + (1-x) B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)\right)$$
$$\times \exp\left(-\pi\lambda_2 x d^2\right) \mathrm{d}d. \quad (17)$$

$$f_{1,k,\infty} \triangleq 2\pi\lambda_1 \int_0^\infty d \exp\left(-\pi\lambda_1 d^2\right) \exp\left(-\frac{2\pi\lambda_2}{\alpha_2} d^{\frac{2\alpha_1}{\alpha_2}} \left(\frac{1}{\beta}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)$$
$$\times \exp\left(-\frac{2\pi\lambda_1}{\alpha_1} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_1}} B^{'}\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}, 2^{-\frac{k\tau}{W}}\right)\right) \mathrm{d}d. \quad (18)$$

$$f_{2,k,\infty}(x) \triangleq 2\pi\lambda_2 x \int_0^\infty d \exp\left(-\pi\lambda_2 x d^2\right) \exp\left(-\frac{2\pi\lambda_1}{\alpha_1} d^{\frac{2\alpha_2}{\alpha_1}} \left(\beta\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right)$$
$$\times \exp\left(-\frac{2\pi\lambda_2}{\alpha_2} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} \left(x B^{'}\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right) + (1-x) B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)\right) \mathrm{d}d. \quad (19)$$

Note that when $\lambda_u \to \infty$, q_2 and q become functions of $\mathbf{T}$ instead of $\mathbf{p}$. From Theorem 1, we have the following lemma.

Lemma 3 (Asymptotic Performance): When $\frac{P}{N_0} \to \infty$ and $\lambda_u \to \infty$, we have $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right) = q_{1,\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right) + q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$, where $q_{1,\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right) = f_{1,K_1^c+\min\{K_1^b,F_1^b\},\infty}\left(\sum_{n\in\mathcal{F}_1^b} \frac{\min\{K_1^b,F_1^b\}}{F_1^b} a_n + \sum_{n\in\mathcal{F}_1^c} a_n\right)$ and $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right) = \sum_{n\in\mathcal{F}_2^c} a_n f_{2,K_2^c,\infty}(T_n)$. Here, $f_{1,k,\infty}$ and $f_{2,k,\infty}(T_n)$ are given by (18) and (19), and T_n is given by (4).

Proof: Please refer to Appendix D. ■

Note that $f_{1,K_1^c+\min\{K_1^b,F_1^b\},\infty}$ represents the successful transmission probability for file $n \in \mathcal{F}_1^c \cup \mathcal{F}_1^b$ (given that this file is transmitted), and $f_{2,K_2^c,\infty}(T_n)$ represents the successful transmission probability for file $n \in \mathcal{F}_2^c$, in the asymptotic region. For given $(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T})$, we interpret Lemma 3 below. When $F_1^b \le K_1^b$, the successful transmission probability of file $n_1 \in \mathcal{F}_1^c$ is the same as that of file $n_2 \in \mathcal{F}_1^b$. In other words, when backhaul capacity is sufficient, storing a file at a macro-BS or retrieving the file via the backhaul link makes no difference in successful transmission probability. When $F_1^b > K_1^b$, the successful transmission probability of file $n_1 \in \mathcal{F}_1^c$ is greater than that of file $n_2 \in \mathcal{F}_1^b$. In other words, when backhaul capacity is limited, storing a file at a macro-BS is better than retrieving the file via the backhaul link. Note that $f_{2,k,\infty}(x)$ is an increasing function (Please refer to Appendix E for the proof). Thus, for any $n_1, n_2 \in \mathcal{F}_2^c$ satisfying $T_{n_1} > T_{n_2}$, the successful transmission probability of file $n_1 \in \mathcal{F}_2^c$ is greater than that of file $n_2 \in \mathcal{F}_2^c$. That is, a file of higher probability being cached at a pico-BS has higher successful transmission probability. Later, in Section VII, we shall see that the structure of $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ facilitates the optimization of $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$.

Next, we further study the symmetric case where $\alpha_1 = \alpha_2 \triangleq \alpha$ in the high SNR and user density region. From Lemma 3, we have the following lemma.

Lemma 4 (Asymptotic Performance When $\alpha_1 = \alpha_2$): When $\alpha_1 = \alpha_2 = \alpha$, $\frac{P}{N_0} \to \infty$ and $\lambda_u \to \infty$, we have $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right) = q_{1,\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right) + q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$, where $q_{1,\infty}\left(\mathcal{F}_1^c, \mathcal{F}_2^c\right) = \frac{1}{\omega_{K_1^c + \min\{K_1^b, F_1^b\}}}\left(\sum_{n\in\mathcal{F}_1^c} a_n + \frac{\min\{K_1^b, F_1^b\}}{F_1^b}\sum_{n\in\mathcal{F}_1^b} a_n\right)$ and $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right) = \sum_{n\in\mathcal{F}_2^c}\frac{a_n T_n}{\theta_{2,K_2^c} + \theta_{1,K_2^c} T_n}$. Here, T_n is given by (4), and ω_k, $\theta_{1,k}$ and $\theta_{2,k}$ are given by

$$\omega_k = \frac{2}{\alpha}\left(2^{\frac{k\tau}{W}} - 1\right)^{\frac{2}{\alpha}} B'\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}, 2^{\frac{-k\tau}{W}}\right) + \frac{2\lambda_2}{\alpha\lambda_1}\left(\frac{1}{\beta}\left(2^{\frac{k\tau}{W}} - 1\right)\right)^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) + 1, \tag{20}$$

$$\theta_{1,k} = \frac{2}{\alpha}\left(2^{\frac{k\tau}{W}} - 1\right)^{\frac{2}{\alpha}} B'\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}, 2^{\frac{-k\tau}{W}}\right) - \frac{2}{\alpha}\left(2^{\frac{k\tau}{W}} - 1\right)^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) + 1, \tag{21}$$

$$\theta_{2,k} = \frac{2}{\alpha}\left(2^{\frac{k\tau}{W}} - 1\right)^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right) + \frac{2\lambda_1}{\alpha\lambda_2}\left(\beta\left(2^{\frac{k\tau}{W}} - 1\right)\right)^{\frac{2}{\alpha}} B\left(\frac{2}{\alpha}, 1 - \frac{2}{\alpha}\right). \tag{22}$$

Proof: Please refer to Appendix D. ■

From Lemma 4, we can see that in the high SNR and user density region, when $\alpha_1 = \alpha_2 = \alpha$, the impact of the physical layer parameters α, β and W, captured by ω_k, $\theta_{1,K}$ and $\theta_{2,K}$, and the impact of the design parameters $\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ on the successful transmission probability $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ can be easily separated. Later, in Section VII, we shall see that this separation greatly facilitates the optimization of $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$.

Fig. 2 plots the successful transmission probability versus the transmit SNR $\frac{P}{N_0}$ and the user density λ_u. Fig. 2 verifies Theorem 1 and Lemma 3 (Lemma 4), and demonstrates the accuracy of the approximations adopted. Fig. 2 also indicates that $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ provides a simple and good approximation for $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ in the high transmit SNR region (e.g., $\frac{P}{N_0} \geq 100$ dB) and the high user density region (e.g., $\lambda_u \geq 3 \times 10^{-5}$).

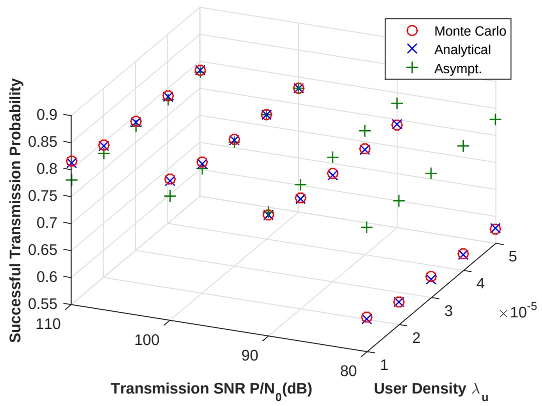


Fig. 2. Successful transmission probability versus transmit SNR $\frac{P}{N_0}$ and user density λ_u. $N = 10$, $K_1^c = 3$, $K_2^c = 2$, $K_1^b = 1$, $\mathcal{F}_1^c = \{1, 2, 3\}$, $\mathcal{F}_1^b = \{7, 8, 9, 10\}$, $\mathcal{F}_2^c = \{4, 5, 6\}$, $\mathbf{p} = (0.7, 0.2, 0.1)$, $\mathcal{N}_1 = \{4, 5\}$, $\mathcal{N}_2 = \{4, 6\}$, $\mathcal{N}_3 = \{5, 6\}$, $\lambda_1 = 5 \times 10^{-7}$, $\lambda_2 = 3 \times 10^{-6}$, $P_1 = 10^{1.5}P$, $P_2 = P$, $\alpha_1 = \alpha_2 = 4$, $W = 20 \times 10^6$, $\tau = 2 \times 10^4$, and $a_n = \frac{n^{-\gamma}}{\sum_{n \in \mathcal{N}} n^{-\gamma}}$ with $\gamma = 1$. In this paper, to simulate the large-scale HetNet, we use a 2-dimensional square of area 15000^2, which is sufficiently large in our case. Note that if the simulation window size is not large enough, the observed interference would be smaller than the true interference due to the edge effect, resulting in larger successful transmission probability than the true value. In addition, the Monte Carlo results are obtained by averaging over 10^5 random realizations.

VI. Optimization Problem Formulation

In this section, we formulate the optimal caching and multicasting design problem to maximize the successful transmission probability $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$, which is a mixed discrete-continuous optimization problem. To facilitate the solution of this challenging optimization problem in the next section, we also formulate the asymptotically optimal caching and multicasting design problem to maximize the asymptotic successful transmission probability $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ in the high SNR and user density region.

A. Optimization Problem

The caching and multicasting design fundamentally affects the successful transmission probability via the design parameters $\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$. We would like to maximize $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ by carefully optimizing $\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$.

Problem 1 (Performance Optimization):

$$q^* \triangleq \max_{\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}} \quad q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$$
$$s.t. \quad (1), (2), (3),$$

where $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ is given by (15).

Note that Problem 1 is a mixed discrete-continuous optimization problem with two main challenges. One is the choice of the sets of files $\mathcal{F}_1^c$ and $\mathcal{F}_2^c$ (discrete variables) stored in the two tiers, and the other is the choice of the caching distribution $\mathbf{p}(\mathcal{F}_2^c)$ (continuous variables) of random caching for the 2nd tier. We thus propose an equivalent alternative formulation of Problem 1 which naturally subdivides Problem 1 according to these two aspects.

Problem 2 (Equivalent Optimization):

$$q^* = \max_{\mathcal{F}_1^c, \mathcal{F}_2^c} \quad q_1(\mathcal{F}_1^c, \mathcal{F}_2^c) + q_2^*\left(\mathcal{F}_2^c\right) \tag{23}$$
$$s.t. \quad (1).$$

$$q_2^*(\mathcal{F}_2^c) \triangleq \max_{\mathbf{p}} \quad q_2\left(\mathcal{F}_2^c, \mathbf{p}\right) \tag{24}$$
$$s.t. \quad (2), (3).$$

For given $\mathcal{F}_2^c$, the optimization problem in (24) is in general a non-convex optimization problem with a large number of optimization variables (i.e., $I = \binom{F_2^c}{K_2^c}$ optimization variables), and it is difficult to obtain the global optimal solution and calculate $q_2^*\left(\mathcal{F}_2^c\right)$. Even given $q_2^*\left(\mathcal{F}_2^c\right)$, the optimization problem in (23) is a discrete optimization problem over a very large constraint set, and is NP-complete in general. Therefore, Problem 2 is still very challenging.

B. Asymptotic Optimization Problem

To facilitate the solution of the challenging mixed discrete-continuous optimization problem, we also formulate the optimization of the asymptotic successful transmission probability $q_\infty\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{T}\right)$ given in Lemma 3, i.e., which has a much simpler form than $q\left(\mathcal{F}_1^c, \mathcal{F}_2^c, \mathbf{p}\right)$ given in Theorem 1. Equivalently, we can consider the asymptotic version of Problem 2 in the high SNR and user density region.

Problem 3 (Asymptotic Optimization):

$$q_\infty^* = \max_{\mathcal{F}_1^c, \mathcal{F}_2^c} \quad q_{1,\infty}(\mathcal{F}_1^c, \mathcal{F}_2^c) + q_{2,\infty}^*\left(\mathcal{F}_2^c\right) \tag{25}$$
$$s.t. \quad (1).$$

The optimal solution to the optimization in (25) is written as $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*})$ and $q^*_{2,\infty}(\mathcal{F}_2^c)$ is given by

$$q^*_{2,\infty}(\mathcal{F}_2^c) \triangleq \max_{\mathbf{p}} \quad q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{p}\right) \tag{26}$$
$$s.t. \quad (2), (3), (4),$$

where the optimal solution to the optimization in (26) is written as $\mathbf{p}^*(\mathcal{F}_2^c)$. The optimal solution to Problem 3 is given by $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{p}^*(\mathcal{F}_2^{c*}))$, which is the asymptotic optimal solution to Problem 2 (Problem 1).

Based on Lemma 2 in [21], we know that the optimization in (26) is equivalent to the following optimization for given $\mathcal{F}_2^c$

$$q^*_{2,\infty}(\mathcal{F}_2^c) \triangleq \max_{\mathbf{T}} \quad q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right) \tag{27}$$
$$s.t. \qquad 0 \le T_n \le 1, \;\; n \in \mathcal{F}_2^c, \tag{28}$$
$$\sum_{n \in \mathcal{F}_2^c} T_n = K_2^c, \tag{29}$$

where the optimal solution is written as $\mathbf{T}^*(\mathcal{F}_2^c)$. In addition, any $\mathbf{p}^*(\mathcal{F}_2^c)$ in convex polyhedron $\mathcal{P}^*(\mathcal{F}_2^c) \triangleq \{\mathbf{p}^*(\mathcal{F}_2^c) : (2), (3), (30)\}$ is an optimal solution to the optimization in (26), where (30) is given by:

$$\sum_{i \in \mathcal{I}_n} p_i^*(\mathcal{F}_2^c) = T_n^*(\mathcal{F}_2^c), \;\; n \in \mathcal{F}_2^c. \tag{30}$$

The vertices of the convex polyhedron $\mathcal{P}^*(\mathcal{F}_2^c)$ can be obtained based on the simplex method, and any $\mathbf{p}^*(\mathcal{F}_2^c) \in \mathcal{P}^*(\mathcal{F}_2^c)$ can be constructed from all the vertices using convex combination. Thus, when optimizing the asymptotic performance for given $\mathcal{F}_2^c$, we can focus on the optimization in (27) instead of the optimization in (26).

VII. Near Optimal Solution

In this section, we propose a two-step optimization framework to obtain a near optimal solution with manageable complexity and superior performance in the general region. We first characterize the structural properties of the asymptotically optimal solutions. Then, based on these properties, we obtain a near optimal solution in the general region.

A. Asymptotically Optimal Solution

In this part, we study the continuous part and the discrete part of the asymptotic optimization in Problem 3, respectively, to obtain design insights into the solution in the general region.

$$f'_{2,k,\infty}(x) \triangleq 2\pi\lambda_2 x \int_0^\infty d \exp\left(-\frac{2\pi\lambda_2}{\alpha_2} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} \left(xB'\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right) + (1-x)\, B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)\right)$$
$$\times \exp\left(-\pi\lambda_2 x d^2\right) \exp\left(-\frac{2\pi\lambda_1}{\alpha_1} d^{\frac{2\alpha_2}{\alpha_1}} \left(\frac{P_1}{P_2}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right)$$
$$\times \left(-\pi\lambda_2 d^2 - \frac{2\pi\lambda_2}{\alpha_2} d^2 \left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} \left(B'\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right) - B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)\right) \mathrm{d}d$$
$$+ \frac{f_{2,k,\infty}(x)}{x}. \tag{32}$$

---

1) Continuous Optimization: As the structure of $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$ is very complex, it is difficult to obtain the closed-form optimal solution $\mathbf{T}^*(\mathcal{F}_2^c)$ to the optimization in (27). By exploring the structural properties of $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$, we know that files of higher popularity get more storage resources.

Lemma 5 (Structural Property of Optimization in (27)*):* Given any $\mathcal{F}_2^c \subseteq \mathcal{N}$ satisfying $F_2^c \geq K_2^c$ and $n_1, n_2 \in \mathcal{F}_2^c$, if $n_1 < n_2$, then $T^*_{n_1}(\mathcal{F}_2^c) \geq T^*_{n_2}(\mathcal{F}_2^c)$.

Proof: Please refer to Appendix E. ■

Now, we focus on obtaining a numerical solution to the optimization in (27). For given $\mathcal{F}_2^c \subseteq \mathcal{N}$ satisfying $F_2^c \geq K_2^c$, the optimization in (27) is a continuous optimization of a differentiable function $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$ over a convex set. In general, it is difficult to show the convexity of $f_{2,k,\infty}(x)$ in (19). A stationary point to the optimization in (27) can be obtained using standard gradient projection methods [26, pp. 223]. Here, we consider the diminishing stepsize [26, pp. 227] satisfying

$$\epsilon(t) \to 0 \text{ as } t \to \infty, \ \sum_{t=1}^{\infty} \epsilon(t) = \infty, \sum_{t=1}^{\infty} \epsilon(t)^2 < \infty, \tag{31}$$

and propose Algorithm 1. In Step 2 of Algorithm 1, $\frac{\partial q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}(t)\right)}{\partial T_n(t)} = a_n f'_{2,k,\infty}(x)$, where $f'_{2,k,\infty}(x)$ is given by (32). Step 3 is the projection of $\bar{T}_n(t+1)$ onto the set of the variables satisfying the constraints in (28) and (29). It is shown in [26, pp. 229] that $\mathbf{T}(t)$ in Algorithm 1 converges to a stationary point of the optimization in (27) as $t \to \infty$.

On the other hand, as illustrated in the discussion of Lemma 3, $f_{2,k,\infty}(x)$ is actually a cumulative density function (c.d.f.), and is concave in most of the cases we are interested in. If $f_{2,k,\infty}(x)$ in (17) is concave w.r.t. x, the differentiable function $q_{2,\infty}\left(\mathcal{F}_2^c, \mathbf{T}\right)$ is concave w.r.t. $\mathbf{T}$, and hence, the optimization in (27) is a convex problem. Then, $\mathbf{T}(t)$ in Algorithm 1

Algorithm 1 Asymptotically Optimal Solution

1: Initialize $t = 1$ and $T_n(1) = \frac{1}{F_2^c}$ for all $n \in \mathcal{F}_2^c$.

2: For all $n \in \mathcal{F}_2^c$, compute $\bar{T}_n(t+1)$ according to $\bar{T}_n(t+1) = T_n(t) + \epsilon(t)\frac{\partial q_{2,\infty}(\mathcal{F}_2^c, \mathbf{T}(t))}{\partial T_n(t)}$, where $\{\epsilon(t)\}$ satisfies (31).

3: For all $n \in \mathcal{F}_2^c$, compute $T_n(t+1)$ according to $T_n(t+1) = \min\left\{\left[\bar{T}_n(t+1) - \nu^*\right]^+, 1\right\}$, where ν^* satisfies $\sum_{n\in\mathcal{F}_2^c} \min\left\{\left[\bar{T}_n(t+1) - \nu^*\right]^+, 1\right\} = K_2^c$.

4: Set $t = t + 1$ and go to Step 2.

converges to the optimal solution $\mathbf{T}^*(\mathcal{F}_2^c)$ to the optimization in (27) as $t \to \infty$. In other words, under a mild condition (i.e., $f_{2,k,\infty}(x)$ is convex), we can obtain the optimal solution $\mathbf{T}^*(\mathcal{F}_2^c)$ to the optimization in (27) using Algorithm 1.

Next, we consider the symmetric case, i.e., $\alpha_1 = \alpha_2 = \alpha$, in the high SNR and user density region. In this case, we can easily verify that $q_{2,\infty}(\mathcal{F}_2^c, \mathbf{T}) = \sum_{n\in\mathcal{F}_2^c} \frac{a_n T_n}{\theta_{2,K_2^c} + \theta_{1,K_2^c} T_n}$ (given in Lemma 4) is convex and Slater's condition is satisfied, implying that strong duality holds. Using KKT conditions, we can obtain the closed-form solution to the optimization in (27) in this case.

Lemma 6 (Asymptotically Optimal Solution when $\alpha_1 = \alpha_2$): For given $\mathcal{F}_2^c$, when $\alpha_1 = \alpha_2 = \alpha$, $\frac{P}{N_0} \to \infty$ and $\lambda_u \to \infty$, the optimal solution to the optimization in (27) is given by

$$T_n^*(\mathcal{F}_2^c) = \min\left\{\left[\frac{1}{\theta_{1,K_2^c}}\sqrt{\frac{a_n \theta_{2,K_2^c}}{\nu^*}} - \frac{\theta_{2,K_2^c}}{\theta_{1,K_2^c}}\right]^+, 1\right\}, \ n \in \mathcal{F}_2^c, \tag{33}$$

where $[x]^+ \triangleq \max\{x, 0\}$ and ν^* satisfies

$$\sum_{n\in\mathcal{F}_2^c} \min\left\{\left[\frac{1}{\theta_{1,K_2^c}}\sqrt{\frac{a_n \theta_{2,K_2^c}}{\nu^*}} - \frac{\theta_{2,K_2^c}}{\theta_{1,K_2^c}}\right]^+, 1\right\} = K_2^c. \tag{34}$$

Here, $\theta_{1,k}$ and $\theta_{2,k}$ are given by (21) and (22), respectively.

Proof: Please refer to Appendix F. ■

As the water-level in the traditional water-filling power control, the root ν^* to the equation in (34) can be easily solved. Thus, by Lemma 6, we can efficiently compute $\mathbf{T}^*(\mathcal{F}_2^c)$ when $\alpha_1 = \alpha_2$.

Lemma 6 can be interpreted as follows. As illustrated in Fig. 3, $\mathbf{T}^*(\mathcal{F}_2^c)$ given by Lemma 6 has a reverse water-filling structure. The file popularity distribution $\{a_n : n \in \mathcal{F}_2^c\}$ and the physical layer parameters (captured in θ_{1,K_2^c} and θ_{2,K_2^c}) jointly affect ν^*. Given ν^*, the physical layer parameters (captured in θ_{1,K_2^c} and θ_{2,K_2^c}) affect the caching probabilities of all the files in

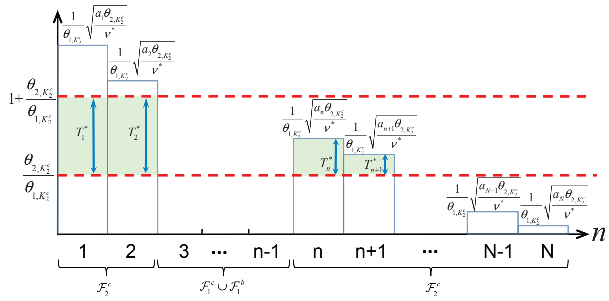


Fig. 3. Illustration of the optimality structure in Lemma 6. In this example, $\mathcal{F}_1^c \cup \mathcal{F}_1^b = \{3, 4 \cdots, n-1\}$, and $\mathcal{F}_2^c = \{1, 2, n, n+1, \cdots, N\}$ for some $n \in \mathcal{N}$.

the same way, while the popularity of file $n \in \mathcal{F}_2^c$ (i.e., a_n) only affects the caching probability of file n (i.e., T_n^*). From Lemma 6, we know that for any $n_1, n_2 \in \mathcal{F}_2^c$ such that $n_1 < n_2$, we have $T_{n_1}^* > T_{n_2}^*$, as $a_{n_1} > a_{n_2}$. In other words, files in $\mathcal{F}_2^c$ of higher popularity get more storage resources in the 2nd tier. In addition, there may exist $\bar{n} \in \mathcal{F}_2^c$ such that $T_n^* > 0$ for all $n \in \mathcal{F}_2^c$ and $n < \bar{n}$, and $T_n^* = 0$ for all $n \in \mathcal{F}_2^c$ and $n \geq \bar{n}$. In other words, some files in $\mathcal{F}_2^c$ of lower popularity may not be stored in the 2nd tier. For a popularity distribution with a heavy tail, more different files in $\mathcal{F}_2^c$ can be stored in the 2nd tier.

2) Discrete Optimization: Given $q_{2,\infty}^*(\mathcal{F}_2^c) = q_{2,\infty}(\mathcal{F}_2^c, \mathbf{T}^*(\mathcal{F}_2^c))$, the optimization in (25) is a discrete optimization. It can be shown that the number of possible choices for $(\mathcal{F}_1^c, \mathcal{F}_2^c)$ satisfying (1) is given by $\sum_{F_2^c=K_2^c}^{N-K_1^c} \binom{N}{F_2^c}\binom{N-F_2^c}{K_1^c} = \Theta(N^N)$. Thus, a brute-force solution to the discrete optimization in (25) is not acceptable. Now, we explore the structural properties of the discrete optimization in (25) to facilitate the design of low-complexity asymptotically optimal solutions.

Theorem 2 (Structural Properties of Optimization in (25)): There exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*})$ to the optimization in (25) satisfying the following conditions: (i) $F_2^{c*} \in \{F_{2,lb}^{c*}, F_{2,lb}^{c*}+1, \cdots, N-K_1^c\}$, where $F_{2,lb}^{c*} \triangleq \max\{K_2^c, N-K_1^c-K_1^b\}$; and (ii) there exists $n_1^c \in \{1, 2, \cdots, F_2^{c*}+1\}$, such that $\mathcal{F}_1^{c*} = \{n_1^c, n_1^c+1, \cdots, n_1^c+K_1^c-1\}$ and $\mathcal{F}_2^{c*} = \mathcal{N} \setminus (\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*})$, where $\mathcal{F}_1^{b*} = \{n_1^c+K_1^c, n_1^c+K_1^c+1, \cdots, n_1^c+K_1^c+N-(K_1^c+F_2^{c*})-1\}$.

Proof: Please refer to Appendix G. ■

Theorem 2 can be interpreted as follows. Property (ii) indicates that there is an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*})$ to the optimization in (25) satisfying that the files in $\mathcal{F}_1^{c*}$, $\mathcal{F}_1^{b*}$ and $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ are consecutive, and the files in $\mathcal{F}_1^{c*}$ are more popular than those in $\mathcal{F}_1^{b*}$. This can

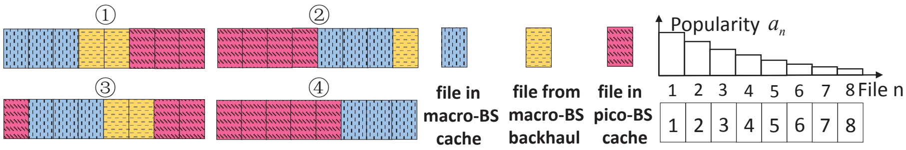


Fig. 4. Illustration of the structural properties in Theorem 2 and Lemma 7. $K_1^c = 3$, $K_1^b = 2$ and $K_2^c = 2$.

be easily understood from Fig. 4. It can be shown that the number of possible choices for $(\mathcal{F}_1^c, \mathcal{F}_2^c)$ satisfying the properties in Theorem 2 is given by $\sum_{F_2^c = F_{2,lb}^{c*}}^{N-K_1^c} \sum_{n_1^c=1}^{F_2^c+1} 1 = \Theta(N^2)$, which is much smaller than the number of possible choices just satisfying (1), i.e., $\Theta(N^N)$. By restricting to the choices for $(\mathcal{F}_1^c, \mathcal{F}_2^c)$ satisfying the properties in Theorem 2, we can greatly reduce the complexity for solving the optimization in (25) without losing any optimality.

In some special cases, we can obtain extra properties other than those in Theorem 2.

Lemma 7 (Structural Properties of Optimization in (25) *in Special Cases):* (i) If $f_{1,K_1^c+K_1^b,\infty} > f_{2,K_2^c,\infty}(1)$, then n_1^c in Theorem 2 satisfies $n_1^c = 1$; (ii) If $f_{1,K_1^c,\infty} < f_{2,K_2^c,\infty}\left(\frac{K_2^c}{N-K_1^c}\right)$, then n_1^c in Theorem 2 satisfies $n_1^c \geq 2$.

Proof: Please refer to Appendix H. ■

Lemma 7 can be interpreted as follows. Property (i) implies that the most popular files are served by the 1st tier (cf. Case 1 in Fig. 4), if $f_{1,K_1^c+K_1^b,\infty} > f_{2,K_2^c,\infty}(1)$. This condition holds when $\frac{P_1}{P_2}$ and $\frac{\lambda_1}{\lambda_2}$ are above some thresholds, respectively. In this case, macro-BSs intend to multicast the most popular files, as they can offer relatively higher receive power, and hence higher successful transmission probability for the most popular files. Note that when the condition in (i) holds, by Theorem 2 and Lemma 7, we can directly determine $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*})$. Property (ii) implies that the most popular file, i.e., file 1, is not served by the 1st tier (cf. Cases 2-4 in Fig. 4), if $f_{1,K_1^c,\infty} < f_{2,K_2^c,\infty}\left(\frac{K_2^c}{N-K_1^c}\right)$. This condition holds when $\frac{P_1}{P_2}$ and $\frac{\lambda_1}{\lambda_2}$ are below some thresholds, respectively. In this case, pico-BSs intend to multicast the most popular file, as they can offer relatively higher receive power, and hence higher successful transmission probability for the most popular file. When the condition in (ii) holds, we can use Lemma 7 together with Theorem 2 to reduce the set of possible choices for $(\mathcal{F}_1^c, \mathcal{F}_2^c)$, to further reduce the complexity for solving the optimization in (25) without losing any optimality.

$$q_2\left(\mathcal{F}_2^c, \mathbf{p}^*\right)$$
$$= \sum_{i \in \mathcal{I}} \left(\sum_{n \in \mathcal{N}_i} \frac{a_n}{T_n^*} \sum_{k^c=1}^{K_2^c} f_{2,k^c}(T_n^*) \sum_{\mathcal{X} \in \{\mathcal{S} \subseteq \mathcal{N}_{i,-n} : |\mathcal{S}| = k^c - 1\}} \prod_{m \in \mathcal{X}} \left(1 - \left(1 + \frac{a_m \lambda_u}{3.5 T_m^* \lambda_2}\right)^{-4.5}\right) \prod_{m \in \mathcal{N}_{i,-n} \setminus \mathcal{X}} \left(1 + \frac{a_m \lambda_u}{3.5 T_m^* \lambda_2}\right)^{-4.5} \right)$$
$$\triangleq q_2\left(\mathcal{F}_2^c, \mathbf{p}^*, \mathbf{T}^*\right) \tag{35}$$

B. Near Optimal Solution in General Region

First, we consider the near optimal solution for the continuous part (for given $\mathcal{F}_2^c$). As illustrated in Section VII-A, based on $\mathbf{T}^*(\mathcal{F}_2^c)$ obtained using Algorithm 1 or Lemma 6 (when $\alpha_1 = \alpha_2$), we can determine $\mathcal{P}^*(\mathcal{F}_2^c)$. As illustrated in Section VI-B, any $\mathbf{p}^* \in \mathcal{P}^*(\mathcal{F}_2^c)$ is an optimal solution to the optimization in (26). In other words, for given $\mathcal{F}_2^c$, we have a set of asymptotically optimal solutions in the high SNR and user density region. Substituting $\mathbf{p}^*$ satisfying (30) into $q_2\left(\mathcal{F}_2^c, \mathbf{p}\right)$ in Theorem 1, we have $q_2\left(\mathcal{F}_2^c, \mathbf{p}^*, \mathbf{T}^*\right)$ given in (35). For given $\mathcal{F}_2^c$ (and $\mathbf{T}^*(\mathcal{F}_2^c)$), we would like to obtain the best asymptotically optimal solution which maximizes the successful transmission probability $q_2\left(\mathcal{F}_2^c, \mathbf{p}^*, \mathbf{T}^*\right)$ in the general region among all the asymptotically optimal solutions in $\mathcal{P}^*(\mathcal{F}_2^c)$.

Problem 4 (Optimization of $\mathbf{p}^*$ under Given $\mathcal{F}_2^c$ (and $\mathbf{T}^*$)):

$$q_2^\dagger(\mathcal{F}_2^c) \triangleq \max_{\mathbf{p}^*} \quad q_2(\mathcal{F}_2^c, \mathbf{p}^*, \mathbf{T}^*)$$
$$s.t. \quad (2), (3), (30).$$

The optimal solution is denoted as $\mathbf{p}^\dagger(\mathcal{F}_2^c)$.

Problem 4 is a linear programming problem. To reduce the complexity for solving Problem 4, we first derive some caching probabilities which are zero based on the relationship between $\mathbf{p}$ and $\mathbf{T}$. In particular, for all $i \in \mathcal{I}_n$ and $n \in \{n \in \mathcal{F}_2^c : T_n^* = 0\}$, we have $p_i^\dagger(\mathcal{F}_c^c) = 0$; for all $i \notin \mathcal{I}_n$ and $n \in \{n \in \mathcal{F}_2^c : T_n^* = 1\}$, we have $p_i^\dagger(\mathcal{F}_c^c) = 0$. Thus, we have

$$p_i^\dagger(\mathcal{F}_c^c) = 0, \ i \in \mathcal{I}', \tag{36}$$

where $\mathcal{I}' \triangleq \cup_{n \in \{n \in \mathcal{F}_2^c : T_n^* = 0\}} \mathcal{I}_n \cup \left(\mathcal{I} \setminus \cup_{n \in \{n \in \mathcal{F}_2^c : T_n^* = 1\}} \mathcal{I}_n\right)$. Then, we can compute the remaining caching probabilities for the combinations in $\mathcal{I} \setminus \mathcal{I}'$ using the simplex method (refer to Step 6 of Algorithm 2 for details). Therefore, using the above approach, for given $\mathcal{F}_2^c$, we can obtain the best asymptotically optimal solution $\mathbf{p}^\dagger(\mathcal{F}_2^c)$ in $\mathcal{P}^*(\mathcal{F}_2^c)$ to the optimization in (24).

Next, we consider the near optimal solution for the discrete part. Specifically, after obtaining $q_2^\dagger(\mathcal{F}_2^c)$ using the above approach for the continuous part, we consider the optimization of $q_1(\mathcal{F}_1^c, \mathcal{F}_2^c) + q_2^\dagger(\mathcal{F}_2^c)$ over the set of $(\mathcal{F}_1^c, \mathcal{F}_2^c)$ satisfying Theorem 2 (and Lemma 7). Let $(\mathcal{F}_1^{c\dagger}, \mathcal{F}_2^{c\dagger})$ and $q^\dagger$ denote the optimal solution and the optimal value.

Finally, combining the above discrete part and continuous part, we can obtain the near optimal solution $(\mathcal{F}_1^{c\dagger}, \mathcal{F}_2^{c\dagger}, \mathbf{p}^\dagger(\mathcal{F}_2^{c\dagger}))$ to Problem 1 (Problem 2), as summarized in Algorithm 2. We can show that in the general region, under a mild condition (i.e., $f_{2,k,\infty}(x)$ is convex), the near optimal solution $(\mathcal{F}_1^{c\dagger}, \mathcal{F}_2^{c\dagger}, \mathbf{p}^\dagger(\mathcal{F}_2^{c\dagger}))$ obtained by Algorithm 2 achieves the successful transmission probability $q^\dagger = q(\mathcal{F}_1^{c\dagger}, \mathcal{F}_2^{c\dagger}, \mathbf{p}^\dagger(\mathcal{F}_2^{c\dagger}))$ greater than or equal to that of any optimal solution to Problem 3, i.e., any asymptotically optimal solution to Problem 1 (Problem 2).

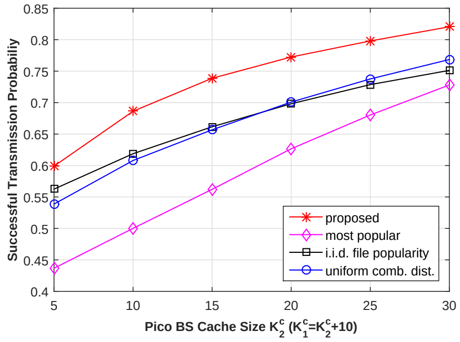
Lemma 8: We have $q(\mathcal{F}_1^{c\dagger}, \mathcal{F}_2^{c\dagger}, \mathbf{p}^\dagger(\mathcal{F}_2^{c\dagger})) \geq q(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{p}^*(\mathcal{F}_2^{c*}))$, for all $\mathbf{p}^*(\mathcal{F}_2^{c*}) \in \mathcal{P}^*(\mathcal{F}_2^{c*})$, where $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{p}^*(\mathcal{F}_2^{c*}))$ is an optimal solution to Problem 3.

Algorithm 2 Near Optimal Solution

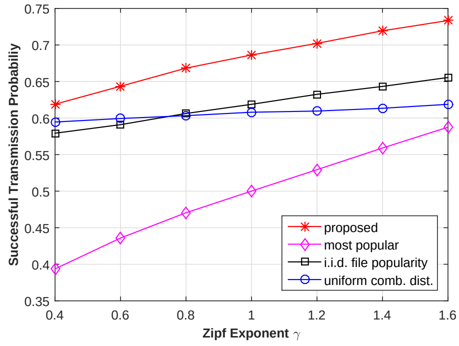
```
Initialize q† = 0.
for F_2^c = max{K_2^c, N − K_1^c − K_1^b} : N − K_1^c do
    for n_1^c = 1 : F_2^c + 1 do
        Choose F_1^c and F_2^c according to Theorem 2 (and Lemma 7).
        Obtain the optimal solution T*(F_2^c) to the optimization in (27) using Algorithm 1 or Lemma 6 (when
        α_1 = α_2).
        Determine I' and choose p_i†(F_2^c) = 0 for all i ∈ I' according to (36). Then, obtain {p_i† : i ∈ I \ I'}
        and q_2†(F_2^c) by solving Problem 4 (under the constraint in (36)) using the simplex method.
        Compute q_1(F_1^c, F_2^c) + q_2†(F_2^c) ≜ q_∞. If q_∞† < q_∞, set q_∞† = q_∞ and (F_1^c†, F_2^c†, p†(F_2^c†)) =
        (F_1^c, F_2^c, p†(F_2^c)).
    end for
end for
```

VIII. Numerical Results

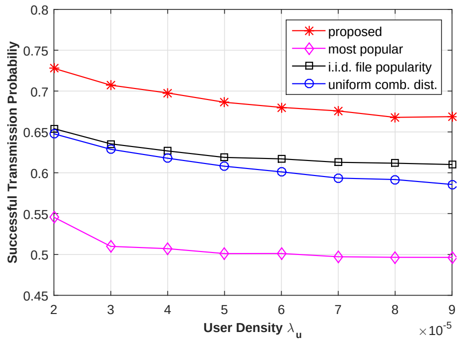
In this section, we compare the proposed near optimal design given by Algorithm 2 with three schemes. Baseline 1 (most popular) refers to the design in which each macro-BS selects the most $K_1^c + K_1^b$ popular files to store and fetch, and each pico-BS selects the most K_2^c popular files to store [9]–[11]. Baseline 2 (i.i.d. file popularity) refers to the design in which each macro-BS selects $K_1^c + K_1^b$ files to store and fetch, and each pico-BS selects K_2^c files



(a) Cache size at $\gamma = 1$ and $\lambda_u = 5 \times 10^{-5}$.



(b) Zipf exponent at $K_2^c = 10$ and $\lambda_u = 5 \times 10^{-5}$.



(c) User density at $K_2^c = 10$ and $\gamma = 1$.

Fig. 5. Successful transmission probability versus cache size K_1^c and K_2^c, Zipf exponent γ and user density λ_u. $K_1^c = K_2^c + 10$, $K_1^b = 15$, $\lambda_1 = 5 \times 10^{-7}$, $\lambda_2 = 3 \times 10^{-6}$, $\frac{P_1}{P_2} = 16$dB, $\alpha_1 = \alpha_2 = 4$, $W = 20 \times 10^6$, $\tau = 2 \times 10^4$ and $N = 100$.

to store, in an i.i.d. manner with file n being selected with probability a_n [19]. Note that under this scheme, each (macro or pico) BS may cache multiple copies of one file, leading to storage waste, and each macro-BS may fetch multiple copies of one file, leading to backhaul waste. Baseline 3 (uniform comb. dist.) refers to the design in which each macro-BS randomly selects a combination of $K_1^c + K_1^b$ different files to store and fetch, and each pico-BS randomly selects a combination of K_2^c different files to store, according to the uniform distribution [13], [14]. Under the three baseline schemes, each user requesting file n is associated with the BS which stores file n and offers the maximum long-term average receive power at this user. In addition, the three baseline schemes also adopt the same multicasting scheme as in our design. In the simulation, we assume the popularity follows Zipf distribution, i.e., $a_n = \frac{n^{-\gamma}}{\sum_{n \in \mathcal{N}} n^{-\gamma}}$, where γ is the Zipf exponent.

Fig. 5 illustrates the successful transmission probability versus different parameters. From Fig. 5, we can observe that the proposed design outperforms all the three baseline schemes. In addition, the proposed design, Baseline 2 and Baseline 3 have much better performance than Baseline 1, as they provide file diversity to improve the network performance, when the storage and backhaul resources are limited and the cache-enabled HetNet with backhaul constraints may not be able to satisfy all file requests.

Specifically, Fig. 5 (a) illustrates the successful transmission probability versus the cache sizes K_1^c and K_2^c. We can see that the performance of all the schemes increases with K_1^c and K_2^c. This is because as K_1^c and K_2^c increase, each BS can store more files, and the probability that a randomly requested file is cached at a nearby BS increases. Fig. 5 (b) illustrates the successful transmission probability versus the Zipf exponent γ. We can observe that the performance of the proposed design, Baseline 1 and Baseline 2 increases with the Zipf exponent γ faster than Baseline 3. This is because when γ increases, the tail of popularity distribution becomes small, and hence, the average network file load decreases. The performance increase of Baseline 3 with γ only comes from the decrease of the average network file load. While, under the proposed design, Baseline 1 and Baseline 2, the probability that a randomly requested file is cached at a nearby BS increases with γ. Thus, the performance increases of the proposed design, Baseline 1 and Baseline 2 with γ are due to the decrease of the average network file load and the increase of the chance of a requested file being cached at a nearby BS. Fig. 5 (c) illustrates the successful transmission probability versus the user density λ_u. We can see that the performance of all the schemes decreases with λ_u. This is because the probability of a cached file being requested by at least one user increases, as λ_u increases.

IX. Conclusion

In this paper, we considered the analysis and optimization of caching and multicasting in a large-scale cache-enabled HetNet with backhaul constraints. We proposed a hybrid caching design and a corresponding multicasting design to provide high spatial file diversity and ensure efficient content dissemination. Utilizing tools from stochastic geometry, we analyzed the successful transmission probability in the general region and the asymptotic region. Then, we formulated a mixed discrete-continuous optimization problem to maximize the successful transmission probability by optimizing the design parameters. By exploring the structural properties, we obtained a near optimal solution with superior performance and manageable complexity, based on a two-step optimization framework. The analysis and optimization results offered valuable design insights for practical cache-enabled HetNets.

Appendix A: Proof of Lemma 1

When typical user u_0 requests file $n \in \mathcal{F}_1^c \cup \mathcal{F}_1^b$, let random variable $Y_{m,n} \in \{0,1\}$ denote whether file $m \in \mathcal{F}_1^c \cup \mathcal{F}_1^b \setminus \{n\}$ is requested by the users associated with serving macro-BS ℓ_0. Specifically, when u_0 requests file $n \in \mathcal{F}_1^c$, we have $K_{1,n,0}^c = 1 + \sum_{m \in \mathcal{F}_1^c \setminus \{n\}} Y_{m,n}$ and $\overline{K}_{1,n,0}^b = \sum_{m \in \mathcal{F}_1^b} Y_{m,n}$. When u_0 requests file $n \in \mathcal{F}_1^b$, we have $\overline{K}_{1,n,0}^c = \sum_{m \in \mathcal{F}_1^c} Y_{m,n}$ and $K_{1,n,0}^b = 1 + \sum_{m \in \mathcal{F}_1^b \setminus \{n\}} Y_{m,n}$. Thus, we have

$$\Pr\left[K_{1,n,0}^c = k^c\right] = \sum_{\mathcal{X} \in g(\mathcal{F}_{1,-n}^c, k^c-1)} \prod_{m \in \mathcal{X}} (1 - \Pr[Y_{m,n}=0]) \prod_{m \in \mathcal{F}_{1,-n}^c \setminus \mathcal{X}} \Pr[Y_{m,n}=0], \quad k^c = 1 \cdots K_1^c,$$

$$\Pr\left[\overline{K}_{1,n,0}^b = k^b\right] = \sum_{\mathcal{X} \in g(\mathcal{F}_1^b, k^b)} \prod_{m \in \mathcal{X}} (1 - \Pr[Y_{m,n}=0]) \prod_{m \in \mathcal{F}_1^b \setminus \mathcal{X}} \Pr[Y_{m,n}=0], \quad k^b = 0 \cdots F_1^b,$$

$$\Pr\left[\overline{K}_{1,n,0}^c = k^c\right] = \sum_{\mathcal{X} \in g(\mathcal{F}_1^c, k^c)} \prod_{m \in \mathcal{X}} (1 - \Pr[Y_{m,n}=0]) \prod_{m \in \mathcal{F}_1^c \setminus \mathcal{X}} \Pr[Y_{m,n}=0], \quad k^c = 0 \cdots K_1^c,$$

$$\Pr\left[K_{1,n,0}^b = k^b\right] = \sum_{\mathcal{X} \in g(\mathcal{F}_{1,-n}^b, k^b-1)} \prod_{m \in \mathcal{X}} (1 - \Pr[Y_{m,n}=0]) \prod_{m \in \mathcal{F}_{1,-n}^b \setminus \mathcal{X}} \Pr[Y_{m,n}=0], \quad k^b = 1 \cdots F_1^b.$$

To prove (10), (11), (12) and (13), it remains to calculate $\Pr[Y_{m,n}=0]$. The p.m.f. of $Y_{m,n}$ depends on the p.d.f. of the size of the Voronoi cell of macro-BS ℓ_0, i.e., the p.d.f. of the size of the Voronoi cell to which a randomly chosen user belongs [25]. Thus, we can calculate the p.m.f. of $Y_{m,n}$ using Lemma 3 of [25] as follows

$$\Pr[Y_{m,n}=0] = \left(1 + 3.5^{-1} \frac{a_m \lambda_u}{\lambda_1}\right)^{-4.5}, \quad m \in \mathcal{F}_1^c \cup \mathcal{F}_1^b \setminus \{n\}. \tag{37}$$

Therefore, we complete the proof.

Appendix B: Proof of Lemma 2

When typical user u_0 requests file $n \in \mathcal{F}_2^c$, let random variable $Y_{m,n,i} \in \{0,1\}$ denote whether file $m \in \mathcal{N}_i \setminus \{n\}$ is requested by the users associated with serving pico-BS ℓ_0 when pico-BS ℓ_0 contains combination $i \in \mathcal{I}_n$. When u_0 requests file $n \in \mathcal{F}_2^c$ and serving pico-BS ℓ_0 contains combination $i \in \mathcal{I}_n$, we have $K_{2,n,0}^c = 1 + \sum_{m \in \mathcal{N}_{i,-n}} Y_{m,n,i}$. Thus, we have

$$\begin{aligned}&\Pr\left[K_{2,n,0}^c = k^c | \text{pico-BS } \ell_0 \text{ contains combination } i \in \mathcal{I}_n\right]\\ =& \sum_{\mathcal{X} \in g(\mathcal{N}_{i,-n}, k^c-1)} \prod_{m \in \mathcal{X}} (1 - \Pr[Y_{m,n,i}=0]) \prod_{m \in \mathcal{N}_{i,-n} \setminus \mathcal{X}} \Pr[Y_{m,n,i}=0], \quad k^c = 1, \cdots, K_2^c.\end{aligned} \tag{38}$$

The probability that pico-BS ℓ_0 contains combination $i \in \mathcal{I}_n$ is $\frac{p_i}{T_n}$. Thus, by the law of total probability, we have

$$\Pr\left[K^c_{2,n,0} = k^c\right] = \sum_{i\in\mathcal{I}_n} \frac{p_i}{T_n} \Pr\left[K^c_{2,n,0} = k^c | \text{pico-BS } \ell_0 \text{ contains combination } i \in \mathcal{I}_n\right], \quad k^c = 1, \cdots, K^c_2.$$

Thus, to prove (14), it remains to calculate $\Pr[Y_{m,n,i} = 0]$. The p.m.f. of $Y_{m,n,i}$ depends on the p.d.f. of the size of the Voronoi cell of pico-BS ℓ_0 w.r.t. file $m \in \mathcal{N}_{i,-n}$ when pico-BS ℓ_0 contains combination $i \in \mathcal{I}_n$, which is unknown. We approximate this p.d.f. based on the known result of the p.d.f. of the size of the Voronoi cell to which a randomly chosen user belongs [25]. Under this approximation, we can calculate the p.m.f. of $Y_{m,n,i}$ using Lemma 3 of [25] as follows

$$\Pr[Y_{m,n,i} = 0] = \left(1 + 3.5^{-1} \frac{a_m \lambda_u}{T_m \lambda_2}\right)^{-4.5}, \quad m \in \mathcal{N}_{i,-n},\ i \in \mathcal{I}_n. \tag{39}$$

Appendix C: Proof of Theorem 1

Based on (7), to prove Theorem 1, we calculate $q_1\left(\mathcal{F}^c_1, \mathcal{F}^c_2\right)$ and $q_2\left(\mathcal{F}^c_2, \mathbf{p}\right)$, respectively.

Calculation of $q_1\left(\mathcal{F}^c_1, \mathcal{F}^c_2\right)$

When u_0 is a macro-user, as in the traditional connection-based HetNets, there are two types of interferers, namely, i) all the other macro-BSs besides its serving macro-BS, and ii) all the pico-BSs. Thus, we rewrite the SINR expression in (6) as follows:

$$\text{SINR}_{n,0} = \frac{D^{-\alpha_1}_{1,\ell_0,0} \left|h_{1,\ell_0,0}\right|^2}{\sum_{\ell\in\Phi_1\setminus\{\ell_0\}} D^{-\alpha_1}_{1,\ell,0} \left|h_{1,\ell,0}\right|^2 + \sum_{\ell\in\Phi_2} D^{-\alpha_2}_{2,\ell,0} \left|h_{2,\ell,0}\right|^2 \frac{P_2}{P_1} + \frac{N_0}{P_1}} = \frac{D^{-\alpha_1}_{1,\ell_0,0} \left|h_{1,\ell_0,0}\right|^2}{I_1 + I_2 \frac{P_2}{P_1} + \frac{N_0}{P_1}}, \ n \in \mathcal{F}^c_1 \cup \mathcal{F}^b_1, \tag{40}$$

where $I_1 \triangleq \sum_{\ell\in\Phi_1\setminus\{\ell_0\}} D^{-\alpha_1}_{1,\ell,0} \left|h_{1,\ell,0}\right|^2$ and $I_2 \triangleq \sum_{\ell\in\Phi_2} D^{-\alpha_2}_{2,\ell,0} \left|h_{2,\ell,0}\right|^2$.

Next, we calculate the conditional successful transmission probability of file $n \in \mathcal{F}^c_1 \cup \mathcal{F}^b_1$ requested by u_0 conditioned on $D_{1,\ell_0,0} = d$ when the file load is k, i.e.,

$$\begin{aligned}
q_{k,n,D_{1,\ell_0,0}}(d) &\triangleq \Pr\left[\frac{W}{k}\log_2\left(1 + \text{SINR}_{n,0}\right) \geq \tau \Big| D_{1,\ell_0,0} = d\right] \\
&\overset{(a)}{=} \mathrm{E}_{I_1,I_2}\left[\Pr\left[\left|h_{1,\ell_0,0}\right|^2 \geq \left(2^{\frac{k\tau}{W}} - 1\right) D^{\alpha_1}_{1,\ell_0,0} \left(I_1 + I_2 \frac{P_2}{P_1} + \frac{N_0}{P_1}\right) \Big| D_{1,\ell_0,0} = d\right]\right] \\
&\overset{(b)}{=} \mathrm{E}_{I_1,I_2}\left[\exp\left(-\left(2^{\frac{k\tau}{W}} - 1\right) d^{\alpha_1} \left(I_1 + I_2 \frac{P_2}{P_1} + \frac{N_0}{P_1}\right)\right)\right] \\
&\overset{(c)}{=} \underbrace{\mathrm{E}_{I_1}\left[\exp\left(-\left(2^{\frac{k\tau}{W}} - 1\right) d^{\alpha_1} I_1\right)\right]}_{\triangleq \mathcal{L}_{I_1}(s,d)|_{s=\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_1}}} \underbrace{\mathrm{E}_{I_2}\left[\exp\left(-\left(2^{\frac{k\tau}{W}} - 1\right) d^{\alpha_1} I_2 \frac{P_2}{P_1}\right)\right]}_{\triangleq \mathcal{L}_{I_2}(s,d)|_{s=\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_1}}} \\
&\quad \times \exp\left(-\left(2^{\frac{k\tau}{W}} - 1\right) d^{\alpha_1} \frac{N_0}{P_1}\right),
\end{aligned} \tag{41}$$

where (a) is obtained based on (40), (b) is obtained by noting that $|h_{1,\ell_0,0}|^2 \overset{d}{\sim} \exp(1)$, and (c) is due to the independence of the Rayleigh fading channels and the independence of the PPPs. To calculate $q_{k,n,D_{1,\ell_0,0}}(d)$ according to (41), we first calculate $\mathcal{L}_{I_1}(s,d)$ and $\mathcal{L}_{I_2}(s,d)$, respectively. The expression of $\mathcal{L}_{I_1}(s,d)$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{I_1}(s,d) &= \mathrm{E}\left[\exp\left(-s\sum_{\ell\in\Phi_1\backslash\{\ell_0\}} D_{1,\ell,0}^{-\alpha_1}|h_{1,\ell,0}|^2\right)\right] = \mathrm{E}\left[\prod_{\ell\in\Phi_1\backslash\{\ell_0\}}\exp\left(-sD_{1,\ell,0}^{-\alpha_1}|h_{1,\ell,0}|^2\right)\right]\\ &\overset{(d)}{=}\exp\left(-2\pi\lambda_1\int_d^\infty\left(1-\frac{1}{1+sr^{-\alpha_1}}\right)r\mathrm{d}r\right)\\ &\overset{(e)}{=}\exp\left(-\frac{2\pi}{\alpha_1}\lambda_1 s^{\frac{2}{\alpha_1}}B'\left(\frac{2}{\alpha_1},1-\frac{2}{\alpha_1},\frac{1}{1+sd^{-\alpha_1}}\right)\right),\end{aligned} \tag{42}$$

where (d) is obtained by utilizing the probability generating functional of PPP [27, Page 235], and (e) is obtained by first replacing $s^{-\frac{1}{\alpha_1}}r$ with t, and then replacing $\frac{1}{1+t^{-\alpha_1}}$ with w. Similarly, the expression of $\mathcal{L}_{I_2}(s,d)$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{I_2}(s,d) =&\mathrm{E}\left[\exp\left(-s\sum_{\ell\in\Phi_2} D_{2,\ell,0}^{-\alpha_2}|h_{2,\ell,0}|^2\frac{P_2}{P_1}\right)\right] = \mathrm{E}\left[\prod_{\ell\in\Phi_2}\exp\left(-sD_{2,\ell,0}^{-\alpha_2}|h_{2,\ell,0}|^2\frac{P_2}{P_1}\right)\right]\\ =&\exp\left(-2\pi\lambda_2\int_0^\infty\left(1-\frac{1}{1+\frac{P_2}{P_1}sr^{-\alpha_2}}\right)r\mathrm{d}r\right)\\ =&\exp\left(-\frac{2\pi}{\alpha_2}\lambda_2\left(\frac{P_2}{P_1}s\right)^{\frac{2}{\alpha_2}}B\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2}\right)\right).\end{aligned} \tag{43}$$

Substituting (42) and (43) into (41), we obtain $q_{k,n,D_{1,\ell_0,0}}(d)$ as follows:

$$\begin{aligned}q_{k,n,D_{1,\ell_0,0}}(d) =&\exp\left(-\frac{2\pi}{\alpha_1}\lambda_1 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_1}}B'\left(\frac{2}{\alpha_1},1-\frac{2}{\alpha_1},2^{-\frac{k\tau}{W}}\right)\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_1}\frac{N_0}{P_1}\right)\\ &\times\exp\left(-\frac{2\pi}{\alpha_2}\lambda_2 d^{\frac{2\alpha_1}{\alpha_2}}\left(\frac{P_2}{P_1}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_2}}B\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2}\right)\right).\end{aligned} \tag{44}$$

Now, we calculate $q_1(\mathcal{F}_1^c,\mathcal{F}_2^c)$ by first removing the condition of $q_{k,n,D_{1,\ell_0,0}}(d)$ on $D_{1,\ell_0,0}=d$. Note that we have the p.d.f. of $D_{1,\ell_0,0}$ as $f_{D_{1,\ell_0,0}}(d)=2\pi\lambda_1 d\exp\left(-\pi\lambda_1 d^2\right)$. Thus, we have:

$$\begin{aligned}&\int_0^\infty q_{k,n,D_{1,\ell_0,0}}(d)f_{D_{1,\ell_0,0}}(d)\mathrm{d}d\\ =&2\pi\lambda_1\int_0^\infty d\exp\left(-\frac{2\pi}{\alpha_1}\lambda_1 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_1}}B'\left(\frac{2}{\alpha_1},1-\frac{2}{\alpha_1},2^{-\frac{k\tau}{W}}\right)\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_1}\frac{N_0}{P_1}\right)\\ &\times\exp\left(-\frac{2\pi}{\alpha_2}\lambda_2 d^{\frac{2\alpha_1}{\alpha_2}}\left(\frac{P_2}{P_1}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_2}}B\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2}\right)\right)\exp\left(-\pi\lambda_1 d^2\right)\mathrm{d}d.\end{aligned} \tag{45}$$

Therefore, by (8) and by letting $k=k^c+k^b$ in (45), we have

$$\begin{aligned}q_1(\mathcal{F}_1^c,\mathcal{F}_2^c) =& \sum_{n\in\mathcal{F}_1^c}a_n\sum_{k^c=1}^{K_1^c}\sum_{k^b=0}^{F_1^b}\Pr[K_{1,n,0}^c=k^c]\Pr[\overline{K}_{1,n,0}^b=k^b]\int_0^\infty q_{k,n,D_{1,\ell_0,0}}(d)f_{D_{1,\ell_0,0}}(d)\mathrm{d}d\\ &+\sum_{n\in\mathcal{F}_1^b}a_n\sum_{k^c=0}^{K_1^c}\sum_{k^b=1}^{F_1^b}\Pr[\overline{K}_{1,n,0}^c=k^c]\Pr[K_{1,n,0}^b=k^b]\frac{\min\{K_1^b,k^b\}}{k^b}\int_0^\infty q_{k,n,D_{1,\ell_0,0}}(d)f_{D_{1,\ell_0,0}}(d)\mathrm{d}d.\end{aligned} \tag{46}$$

Calculation of $q_2\left(\mathcal{F}_2^c, \mathbf{p}\right)$

When u_0 is a pico-user, different from the traditional connection-based HetNets, there are three types of interferers, namely, i) all the other pico-BSs storing the combinations containing the desired file of u_0 besides its serving pico-BS, ii) all the pico-BSs without the desired file of u_0, and iii) all the macro-BSs. Thus, we rewrite the SINR expression in (6) as follows:

$$\begin{aligned}\mathrm{SINR}_{n,0} =& \frac{D_{2,\ell_0,0}^{-\alpha_2}\left|h_{2,\ell_0,0}\right|^2}{\sum_{\ell\in\Phi_{2,n}\backslash\{\ell_0\}} D_{2,\ell,0}^{-\alpha_2}\left|h_{2,\ell,0}\right|^2 + \sum_{\ell\in\Phi_{2,-n}} D_{2,\ell,0}^{-\alpha_2}\left|h_{2,\ell,0}\right|^2 + \sum_{\ell\in\Phi_1} D_{1,\ell,0}^{-\alpha_1}\left|h_{1,\ell,0}\right|^2 \frac{P_1}{P_2} + \frac{N_0}{P_2}} \\ =& \frac{D_{2,\ell_0,0}^{-\alpha_2}\left|h_{2,\ell_0,0}\right|^2}{I_{2,n} + I_{2,-n} + I_1\frac{P_1}{P_2} + \frac{N_0}{P_2}},\ n\in\mathcal{N}, \end{aligned} \tag{47}$$

where $\Phi_{2,n}$ is the point process generated by pico-BSs containing file combination $i \in \mathcal{I}_n$, $\Phi_{2,-n}$ is the point process generated by pico-BSs containing file combination $i \notin \mathcal{I}_n$, $I_{2,n} \triangleq \sum_{\ell\in\Phi_{2,n}\backslash\{\ell_0\}} D_{2,\ell,0}^{-\alpha_2}\left|h_{2,\ell,0}\right|^2$, $I_{2,-n} \triangleq \sum_{\ell\in\Phi_{2,-n}} D_{2,\ell,0}^{-\alpha_2}\left|h_{2,\ell,0}\right|^2$ and $I_1 \triangleq \sum_{\ell\in\Phi_1} D_{1,\ell,0}^{-\alpha_1}\left|h_{1,\ell,0}\right|^2$. Due to the random caching policy and independent thinning [27, Page 230], we obtain that $\Phi_{2,n}$ is a homogeneous PPP with density $\lambda_2 T_n$ and $\Phi_{2,-n}$ is a homogeneous PPP with density $\lambda_2\left(1-T_n\right)$.

Next, we calculate the conditional successful transmission probability of file $n \in \mathcal{F}_2^c$ requested by u_0 conditioned on $D_{2,\ell_0,0} = d$ when the file load is k, denoted as

$$q_{k,n,D_{2,\ell_0,0}}\left(\mathbf{p}, d\right) \triangleq \Pr\left[\frac{W}{k}\log_2\left(1+\mathrm{SINR}_{n,0}\right) \geq \tau \middle| D_{2,\ell_0,0} = d\right].$$

Similar to (41) and based on (47), we have:

$$\begin{aligned} & q_{k,n,D_{2,\ell_0,0}}\left(\mathbf{p}, d\right) \\ =& \mathrm{E}_{I_{2,n},I_{2,-n},I_1}\left[\Pr\left[\left|h_{2,\ell_0,0}\right|^2 \geq \left(2^{\frac{k\tau}{W}}-1\right) D_{2,\ell_0,0}^{\alpha_2}\left(I_{2,n}+I_{2,-n}+I_1\frac{P_1}{P_2}+\frac{N_0}{P_2}\right)\middle| D_{2,\ell_0,0}=d\right]\right] \\ =& \underbrace{\mathrm{E}_{I_{2,n}}\left[\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}I_{2,n}\right)\right]}_{\triangleq \mathcal{L}_{I_{2,n}}(s,d)\big|_{s=\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}}} \underbrace{\mathrm{E}_{I_{2,-n}}\left[\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}I_{2,-n}\right)\right]}_{\triangleq \mathcal{L}_{I_{2,-n}}(s,d)\big|_{s=\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}}} \\ & \times \underbrace{\mathrm{E}_{I_1}\left[\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}I_1\frac{P_1}{P_2}\right)\right]}_{\triangleq \mathcal{L}_{I_1}(s,d)\big|_{s=\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}}} \exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}\frac{N_0}{P_2}\right). \end{aligned} \tag{48}$$

To calculate $q_{k,n,D_{2,\ell_0,0}}\left(\mathbf{p}, d\right)$ according to (48), we first calculate $\mathcal{L}_{I_{2,n}}(s,d)$, $\mathcal{L}_{I_{2,-n}}(s,d)$ and $\mathcal{L}_{I_1}(s,d)$, respectively. Similar to (42) and (43), we have:

$$\mathcal{L}_{I_{2,n}}(s,d) = \exp\left(-\frac{2\pi}{\alpha_2}T_n\lambda_2 s^{\frac{2}{\alpha_2}} B'\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, \frac{1}{1+sd^{-\alpha_2}}\right)\right), \tag{49}$$

$$\mathcal{L}_{I_{2,-n}}(s,d) = \exp\left(-\frac{2\pi}{\alpha_2}(1-T_n)\lambda_2 s^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right), \tag{50}$$

$$\mathcal{L}_{I_1}(s,d) = \exp\left(-\frac{2\pi}{\alpha_1}\lambda_1\left(\frac{P_1}{P_2}s\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right) . \tag{51}$$

Substituting (49), (50) and (51) into (48), we obtain $q_{k,n,D_{2,\ell_0,0}}(\mathbf{p},d)$ as follows:

$$\begin{aligned} q_{k,n,D_{2,\ell_0,0}}(\mathbf{p},d) =& \exp\left(-\frac{2\pi}{\alpha_2}T_n\lambda_2 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} B'\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right)\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}\frac{N_0}{P_2}\right) \\ &\times \exp\left(-\frac{2\pi}{\alpha_2}(1-T_n)\lambda_2 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right) \\ &\times \exp\left(-\frac{2\pi}{\alpha_1}\lambda_1 d^{\frac{2\alpha_2}{\alpha_1}}\left(\frac{P_1}{P_2}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right). \end{aligned} \tag{52}$$

Now, we calculate $q_2(\mathcal{F}_2^c,\mathbf{p})$ by first removing the condition of $q_{k,n,D_{2,\ell_0,0}}(\mathbf{p},d)$ on $D_{2,\ell_0,0} = d$. Note that we have the p.d.f. of $D_{2,\ell_0,0}$ as $f_{D_{2,\ell_0,0}}(d) = 2\pi T_n\lambda_2 d\exp\left(-\pi T_n\lambda_2 d^2\right)$, as pico-BSs storing file n form a homogeneous PPP with density $T_n\lambda_2$. Thus, we have:

$$\begin{aligned} &\int_0^\infty q_{k,n,D_{2,\ell_0,0}}(\mathbf{p},d) f_{D_{2,\ell_0,0}}(d)\mathrm{d}d \\ =& 2\pi T_n\lambda_2\int_0^\infty d\exp\left(-\frac{2\pi}{\alpha_2}T_n\lambda_2 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} B'\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}, 2^{-\frac{k\tau}{W}}\right)\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha_2}\frac{N_0}{P_2}\right) \\ &\times \exp\left(-\frac{2\pi}{\alpha_2}(1-T_n)\lambda_2 d^2\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)\right)\exp\left(-\pi T_n\lambda_2 d^2\right) \\ &\times \exp\left(-\frac{2\pi}{\alpha_1}\lambda_1 d^{\frac{2\alpha_2}{\alpha_1}}\left(\frac{P_1}{P_2}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)\right)\mathrm{d}d. \end{aligned} \tag{53}$$

Therefore, by (9) and by letting $k = k^c$ in (53), we have

$$q_2(\mathcal{F}_2^c,\mathbf{p}) = \sum_{n\in\mathcal{F}_2^c} a_n \sum_{k=1}^{K_2^c}\Pr[K_{2,n,0}^c = k^c]\int_0^\infty q_{k,n,D_{2,\ell_0,0}}(\mathbf{p},d) f_{D_{2,\ell_0,0}}(d)\mathrm{d}d. \tag{54}$$

Appendix D: Proof of Lemma 3 and Lemma 4

Proof of Lemma 3

When $\frac{P}{N}\to\infty$, $\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha}\frac{N_0}{P_1}\right)\to 1$ and $\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^{\alpha}\frac{N_0}{P_2}\right)\to 1$. When $\lambda_u\to\infty$, discrete random variables $K_{1,n,0}^c, \overline{K}_{1,n,0}^c \to K_1^c$, $\overline{K}_{1,n,0}^b, K_{1,n,0}^b \to F_1^b$ and $K_{2,n,0}^c \to K_2^c$ in distribution. Thus, when $P_1 = \beta P$, $P_2 = P$, $\frac{P}{N_0}\to\infty$, and $\lambda_u\to\infty$, we can show $f_{1,K_1^c+\min\{K_1^b,F_1^b\}} \to f_{1,K_1^c+\min\{K_1^b,F_1^b\},\infty}$ and $f_{2,K_2^c}(x)\to f_{2,K_2^c,\infty}(x)$. Thus, we can prove Lemma 3.

Proof of Lemma 4

When $P_1 = \beta P$, $P_2 = P$, $\frac{P}{N_0} \to \infty$, $\lambda_u \to \infty$, and $\alpha_1 = \alpha_2 = \alpha$ we have:

$$f_{1,K_1^c+\min\{K_1^b,F_1^b\},\infty} = 2\pi\lambda_1 \int_0^\infty d \exp\left(-\pi\lambda_1 \omega_{K_1^c+\min\{K_1^b,F_1^b\}} d^2\right) \mathrm{d}d = \frac{1}{\omega_{K_1^c+\min\{K_1^b,F_1^b\}}}, \tag{55}$$

$$f_{2,K_2^c,\infty}(x) = 2\pi\lambda_2 \int_0^\infty d \exp\left(-\pi\lambda_2 d^2 \left(\theta_{2,K_2^c} + x\theta_{1,K_2^c}\right)\right) \mathrm{d}d = \frac{x}{\theta_{2,K_2^c} + \theta_{1,K_2^c} x}, \tag{56}$$

where ω_k, $\theta_{1,k}$ and $\theta_{2,k}$ are given by (20), (21) and (22). Noting that $\int_0^\infty d \exp\left(-cd^2\right) \mathrm{d}d = \frac{1}{2c}$ (c is a constant), we can solve integrals in (18) and (19). Thus, by Lemma 3, we can prove Lemma 4.

APPENDIX E: PROOF OF LEMMA 5

To prove Lemma 5, we first have the following lemma.

Lemma 9 (Monotonicity of $f_{2,k,\infty}(x)$): $f_{2,k,\infty}(x)$ is an increasing function of x.

Proof: By replacing $\exp\left(-\pi x \lambda_2 d^2\right)$ with y in (19), we have:

$$f_{2,k,\infty}(x) = \int_0^1 y^{-\frac{2}{\alpha_2}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}}\left(B\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2}\right)-B'\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2},2^{-\frac{k\tau}{W}}\right)\right)} y^{\frac{2}{\alpha_2 x}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_2}} B\left(\frac{2}{\alpha_2},1-\frac{2}{\alpha_2}\right)}$$
$$\times\, y^{\left(\frac{-\ln y}{\pi\lambda_2}\right)^{\frac{\alpha_2}{\alpha_1}-1} \frac{2}{\alpha_1}\left(\frac{1}{x}\right)^{\frac{\alpha_2}{\alpha_1}} \frac{\lambda_1}{\lambda_2}\left(\frac{P_1}{P_2}\right)^{\frac{2}{\alpha_1}}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha_1}} B\left(\frac{2}{\alpha_1},1-\frac{2}{\alpha_1}\right)} \mathrm{d}y. \tag{57}$$

When $y \in (0,1)$ and $a \in (0,\infty)$, y^a is a decreasing function of a. Because $B\left(\frac{2}{\alpha_2}, 1-\frac{2}{\alpha_2}\right)$, $B\left(\frac{2}{\alpha_1}, 1-\frac{2}{\alpha_1}\right)$ and $2^{\frac{k\tau}{W}}-1 > 0$, and $\frac{1}{x}$ and $\left(\frac{1}{x}\right)^{\frac{\alpha_2}{\alpha_1}}$ are decreasing functions of x. The integrand is an increasing function of x for all $y \in (0,1)$. Therefore, we can show that $f_{2,k,\infty}(x)$ is an increasing function of x. ■

Now, we prove Lemma 5. Let $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ denote an optimal solution to Probelm 3. Consider $n_1, n_2 \in \mathcal{F}_2^{c*}$ satisfying $a_{n_1} > a_{n_2}$. Suppose $T_{n_1}^* < T_{n_2}^*$. Based on Lemma 9, we have $f_{2,K_2^c,\infty}(T_{n_1}^*) < f_{2,K_2^c,\infty}(T_{n_2}^*)$. Now, we construct a feasible solution $(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}')$ to Problem 3 by choosing $\mathcal{F}_1^{c'} = \mathcal{F}_1^{c*}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*}$, $T'_{n_1} = T^*_{n_2}, T'_{n_2} = T^*_{n_1}$, and $T'_n = T^*_n$ for all $n \in \mathcal{F}_2^c \setminus \{n_1, n_2\}$. Thus, by Lemma 3 and the optimality of $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q_\infty\left(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*\right) = \left(a_{n_1} - a_{n_2}\right)\left(f_{2,K_2^c,\infty}(T^*_{n_2}) - f_{2,K_2^c,\infty}(T^*_{n_1})\right) \le 0. \tag{58}$$

Since $a_{n_1} > a_{n_2}$, by (58), we have $f_{2,K_2^c,\infty}(T^*_{n_2}) - f_{2,K_2^c,\infty}(T^*_{n_1}) \le 0$, which contradicts the assumption. Therefore, by contradiction, we can prove Lemma 5.

APPENDIX F: PROOF OF LEMMA 6

For given $\mathcal{F}_2^c$, when $\alpha_1 = \alpha_2 = \alpha$, $\frac{P}{N_0} \to \infty$, and $\lambda_u \to \infty$, the Lagrangian of the optimization in (27) is given by

$$L(\mathbf{T}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \nu) = \sum_{n \in \mathcal{F}_2^c} \frac{a_n T_n}{\theta_{2,K_2^c} + \theta_{1,K_2^c} T_n} + \sum_{n \in \mathcal{F}_2^c} \lambda_n T_n + \sum_{n \in \mathcal{F}_2^c} \eta_n (1 - T_n) + \nu \left(K_2^c - \sum_{n \in \mathcal{F}_2^c} T_n \right),$$

where λ_n and $\eta_n \geq 0$ are the Lagrange multipliers associated with (28), ν is the Lagrange multiplier associated with (29), $\boldsymbol{\lambda} \triangleq (\lambda_n)_{n \in \mathcal{F}_2^c}$, and $\boldsymbol{\eta} \triangleq (\eta_n)_{n \in \mathcal{F}_2^c}$. Thus, we have

$$\frac{\partial L}{\partial T_n}(\mathbf{T}, \boldsymbol{\lambda}, \boldsymbol{\eta}, \nu) = \frac{a_n \theta_{2,K_2^c}}{(\theta_{2,K_2^c} + \theta_{1,K_2^c} T_n)^2} + \lambda_n - \eta_n - \nu.$$

Since strong duality holds, primal optimal $\mathbf{T}^*$ and dual optimal $\boldsymbol{\lambda}^*$, $\boldsymbol{\eta}^*$, ν^* satisfy KKT conditions, i.e., (i) primal constraints: (28), (29), (ii) dual constraints $\lambda_n \geq 0$ and $\eta_n \geq 0$ for all $n \in \mathcal{F}_2^c$, (iii) complementary slackness $\lambda_n T_n = 0$ and $\eta_n(1 - T_n) = 0$ for all $n \in \mathcal{F}_2^c$, and (iv) $\frac{a_n \theta_{2,K_2^c}}{(\theta_{2,K_2^c} + \theta_{1,K_2^c} T_n)^2} + \lambda_n - \eta_n - \nu = 0$ for all $n \in \mathcal{F}_2^c$. By (ii), (iii), and (iv), when $T_n = 0$, we have $\lambda_n \geq 0$, $\eta_n = 0$, and $\nu \geq \frac{a_n}{\theta_{2,K_2^c}}$; when $0 < T_n < 1$, we have $\lambda_n = 0$, $\eta_n = 0$, $T_n = \frac{1}{\theta_{1,K_2^c}} \sqrt{\frac{a_n \theta_{2,K_2^c}}{\nu}} - \frac{\theta_{2,K_2^c}}{\theta_{1,K_2^c}}$, and $\frac{a_n \theta_{2,K_2^c}}{(\theta_{2,K_2^c} + \theta_{1,K_2^c})^2} < \nu < \frac{a_n}{\theta_{2,K_2^c}}$; when $T_n = 1$, we have $\lambda_n = 0$, $\eta_n \geq 0$, and $\nu \leq \frac{a_n \theta_{2,K_2^c}}{(\theta_{2,K_2^c} + \theta_{1,K_2^c})^2}$. Therefore, we have $T_n^* = \min\left\{ \left[\frac{1}{\theta_{1,K_2^c}} \sqrt{\frac{a_n \theta_{2,K_2^c}}{\nu^*}} - \frac{\theta_{2,K_2^c}}{\theta_{1,K_2^c}} \right]^+, 1 \right\}$. Combining (29), we can prove Lemma 6.

APPENDIX G: PROOF OF THEOREM 2

Proof of Property (i) of Theorem 2

By constraints (1) and (5), we have $K_2^c \leq F_2^{c*}$, $F_2^{c*} = N - K_1^c - F_1^{b*}$ and $0 \leq F_1^{b*}$. To prove property (i) of Theorem 2, it remains to prove $F_1^{b*} \leq K_1^b$. Suppose there exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Probelm 3 satisfying $F_1^{b*} > K_1^b$, Then we have:

$$q_\infty^* = f_{1,K_1^c + K_1^b,\infty} \left(\sum_{n \in \mathcal{F}_1^{c*}} a_n + K_1^b \sum_{n \in \mathcal{F}_1^{b*}} \frac{a_n}{F_1^{b*}} \right) + \sum_{n \in \mathcal{F}_2^{c*}} a_n f_{2,K_2^c,\infty}(T_n^*). \tag{59}$$

Now, we construct a feasible solution $(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}')$ to Problem 3, where $\mathcal{F}_1^{b'}$ consists of the most K_1^b popular files of $\mathcal{F}_1^{b*}$, $\mathcal{F}_1^{c'} = \mathcal{F}_1^{c*}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup (\mathcal{F}_1^{b*} \setminus \mathcal{F}_1^{b'})$, $T_n' = T_n^*$ for all $n \in \mathcal{F}_2^{c*}$ and $T_n' = 0$ for all $n \in \mathcal{F}_1^{b*} \setminus \mathcal{F}_1^{b'}$. By Lemma 3, we have:

$$q_\infty(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}') - q_\infty^* = f_{1,K_1^c + K_1^b,\infty} \left(\frac{1}{K_1^b} \sum_{n \in \mathcal{F}_1^{b'}} a_n - \frac{1}{F_1^b} \sum_{n \in \mathcal{F}_1^{b*}} a_n \right) K_1^b > 0. \tag{60}$$

Thus, $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ is not an optimal solution, which contradicts the assumption. Therefore, by contradiction, we can prove $F_1^{b*} \leq K_1^b$ for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Probelm 3. Since $F_2^{c*} \geq K_2^c$, $F_2^{c*} = N - K_1^c - F_1^{b*}$ and $0 \leq F_1^{b*} \leq K_1^b$, we have $\max\{K_2^c, N - K_1^c - K_1^b\} \leq F_2^{c*} \leq N - K_1^c$. Therefore, We can prove property (i) of Theorem 2.

Proof of Property (ii) of Theorem 2

First, we prove that there exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T})$ to Problem 3, such that files in $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ are consecutive. By Lemma 3 and $F_1^{b*} \leq K_1^b$ shown in the proof of Property(i) , we have:

$$q_\infty^* = f_{1,K_1^c+F_1^{b*},\infty} \sum_{n\in\mathcal{F}_1^{c*}\cup\mathcal{F}_1^{b*}} a_n + \sum_{n\in\mathcal{F}_2^{c*}} a_n f_{2,K_2^c,\infty}(T_n^*). \tag{61}$$

Let $n_1(n_2)$ denote the most (least) popular file in $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$. Suppose for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Probelm 3, files in $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ are not consecutive, i.e., there exists $n_3 \in \mathcal{F}_2^{c*}$ satisfying $n_1 < n_3 < n_2$. Now, we can construct a feasible solution $(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}')$ to Probelm 3 where files in $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'}$ are consecutive as follows.

- If $f_{1,K_1^c+F_1^{b*},\infty} < f_{2,K_2^c,\infty}(T_{n_3}^*)$, choose $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'} = \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*} \cup \{n_3\} \setminus \{n_1\}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup \{n_1\} \setminus \{n_3\}$, $T'_{n_1} = T_{n_3}^*$ and $T'_n = T_n^*$ for all $n \in \mathcal{F}_2^{c*} \setminus \{n_3\}$. By Lemma 3, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q_\infty^* = (a_{n_1} - a_{n_3})\left(f_{2,K_2^c,\infty}(T_{n_3}^*) - f_{1,K_1^c+F_1^{b*},\infty}\right) > 0, \tag{62}$$

where the inequality is due to $n_1 < n_3$ and $f_{1,K_1^c+F_1^{b*},\infty} < f_{2,K_2^c,\infty}(T_{n_3}^*)$.

- If $f_{1,K_1^c+F_1^{b*},\infty} > f_{2,K_2^c,\infty}(T_{n_3}^*)$, choose $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'} = \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*} \cup \{n_3\} \setminus \{n_2\}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup \{n_2\} \setminus \{n_3\}$, $T'_{n_2} = T_{n_3}^*$ and $T'_n = T_n^*$ for all $n \in \mathcal{F}_2^{c*} \setminus \{n_3\}$. By Lemma 3, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q_\infty^* = (a_{n_2} - a_{n_3})\left(f_{2,K_2^c,\infty}(T_{n_3}^*) - f_{1,K_1^c+F_1^{b*},\infty}\right) > 0, \tag{63}$$

where the inequality is due to $n_3 < n_2$ and $f_{2,K_2^c,\infty}(T_{n_3}^*) < f_{1,K_1^c+F_1^{b*},\infty}$.

- If $f_{1,K_1^c+F_1^{b*},\infty} = f_{2,K_2^c,\infty}(T_{n_3}^*)$, choose $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'} = \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*} \cup \{n_3\} \setminus \{n_1\}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup \{n_1\} \setminus \{n_3\}$, $T'_{n_1} = T_{n_3}^*$ and $T'_n = T_n^*$ for all $n \in \mathcal{F}_2^{c*} \setminus \{n_3\}$. By Lemma 3, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q_\infty^* = (a_{n_1} - a_{n_3})\left(f_{2,K_2^c,\infty}(T_{n_3}^*) - f_{1,K_1^c+F_1^{b*},\infty}\right) = 0, \tag{64}$$

By (62), (63) and (64), we know that if $f_{1,K_1^c+F_1^{b*},\infty} < f_{2,K_2^c,\infty}(T_{n_3}^*)$ or $f_{1,K_1^c+F_1^{b*},\infty} > f_{2,K_2^c,\infty}(T_{n_3}^*)$, $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ is not an optimal solution, which contradicts the assumption; and if $f_{1,K_1^c+F_1^{b*},\infty} = f_{2,K_2^c,\infty}(T_{n_3}^*)$, we can always construct an optimal solution $\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right)$, satisfying that files in $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'}$ are consecutive. Thus, we can prove that there exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T})$ to Problem 3, such that files in $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ are consecutive.

In addition, by (61), we know that whether file $n \in \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ belongs to $\mathcal{F}_1^{c*}$ or $\mathcal{F}_1^{b*}$ makes no difference in the optimal successful transmission probability. Therefore, we can prove the property (ii) of Theorem 2.

Appendix H: Proof of Lemma 7

Proof of Property (i) *of Lemma 7*

We prove that if $f_{1,K_1^c+K_1^b,\infty} > f_{2,K_2^c,\infty}(1)$, the most popular file $n=1$ belongs to $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3. Suppose that there exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3, such that the most popular file $n=1$ belongs to $\mathcal{F}_2^{c*}$. Let n_2 denote a file in $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$. Now, we can construct a feasible solution $(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}')$ to Probelm 3, where $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'} = \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*} \cup \{1\} \setminus \{n_2\}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup \{n_2\} \setminus \{1\}$, $T'_{n_2} = T^*_1$ and $T'_n = T^*_n$ for all $n \in \mathcal{F}_2^{c*} \setminus \{1\}$. By Lemma 3, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q^*_\infty = (a_1 - a_{n_2})\left(f_{1,K_1^c+\min\{K_1^b, F_1^{b*}\},\infty} - f_{2,K_2^c,\infty}(T_1^*)\right). \tag{65}$$

Since $a_1 > a_{n_2}$ and $f_{1,K_1^c+\min\{K_1^b,F_1^{b*}\},\infty} \geq f_{1,K_1^c+K_1^b,\infty} > f_{2,K_2^c,\infty}(1) \geq f_{2,K_2^c,\infty}(T_1^*)$, we have $q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q^*_\infty > 0$. Thus, $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ is not an optimal solution to Problem 3, which contradicts the assumption. By contradiction, we prove that if $f_{1,K_1^c+K_1^b,\infty} > f_{2,K_2^c,\infty}(1)$, the most popular file $n=1$ belongs to $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$ for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3, and hence n_1^c in Theorem 2 (ii) satisfies $n_1^c = 1$.

Proof of Property (ii) *of Lemma 7*

We prove that if $f_{1,K_1^c,\infty} < f_{2,K_2^c,\infty}(\frac{K_2^c}{N-K_1^c})$, the most popular file $n=1$ belongs to $\mathcal{F}_2^{c*}$ for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3. Suppose that there exists an optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3, such that file $n=1$ belongs to $\mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*}$. Let n_2 denote the most popular file in $\mathcal{F}_2^{c*}$. Based on Lemma 5, we have $T^*_{n_2} \geq T^*_n$ for any $n \in \mathcal{F}_2^{c*} \setminus \{n_2\}$, and hence $T^*_{n_2} \geq \frac{K_2^c}{N-K_1^c}$. Now, we can construct a feasible solution $(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}')$ to Probelm 3, where $\mathcal{F}_1^{c'} \cup \mathcal{F}_1^{b'} = \mathcal{F}_1^{c*} \cup \mathcal{F}_1^{b*} \cup \{n_2\} \setminus \{1\}$, $\mathcal{F}_2^{c'} = \mathcal{F}_2^{c*} \cup \{1\} \setminus \{n_2\}$, $T'_1 = T^*_{n_2}$ and $T'_n = T^*_n$ for all $n \in \mathcal{F}_2^{c*} \setminus \{n_2\}$. By Lemma 3, we have:

$$q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q^*_\infty = (a_1 - a_{n_2})\left(f_{2,K_2^c,\infty}(T^*_{n_2}) - f_{1,K_1^c+\min\{K_1^b+\mathcal{F}_1^{b*}\},\infty}\right). \tag{66}$$

Since $a_1 > a_{n_2}$ and $f_{1,K_1^c+\min\{K_1^b+\mathcal{F}_1^{b*}\},\infty} \leq f_{1,K_1^c,\infty} < f_{2,K_2^c,\infty}(\frac{K_2^c}{N-K_1^c}) \leq f_{2,K_2^c,\infty}(T^*_{n_2})$, we have $q_\infty\left(\mathcal{F}_1^{c'}, \mathcal{F}_2^{c'}, \mathbf{T}'\right) - q^*_\infty > 0$. Thus, $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ is not an optimal solution to Problem 3, which contradicts the assumption. Therefore, we prove that if $f_{1,K_1^c,\infty} < f_{2,K_2^c,\infty}(\frac{K_2^c}{N-K_1^c})$, the most popular file $n=1$ belongs to $\mathcal{F}_2^{c*}$ for any optimal solution $(\mathcal{F}_1^{c*}, \mathcal{F}_2^{c*}, \mathbf{T}^*)$ to Problem 3, and hence n_1^c in Theorem 2 (ii) satisfies $n_1^c \geq 2$.

References

[1] J. Hoadley and P. Maveddat, "Enabling small cell deployment with hetnet," *IEEE Wireless Communications*, vol. 19, no. 2, pp. 4–5, April 2012.

[2] J. G. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.

[3] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5g systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 131–139, February 2014.

[4] H. Sarkissian, "The business case for caching in 4g lte networks," *Wireless 20—20*, vol. 20120, 2012.

[5] A. Liu and V. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan 2015.

[6] K. Shanmugam, N. Golrezaei, A. Dimakis, A. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *Information Theory, IEEE Transactions on*, vol. 59, no. 12, pp. 8402–8413, Dec 2013.

[7] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3553–3568, Oct 2015.

[8] V. Bioglio, F. Gabry, and I. Land, "Optimizing MDS codes for caching at the edge," in *2015 IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[9] E. Bastuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, 2015.

[10] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *IEEE Int. Conf. on Commun. (ICC)*, Kuala Lumpur, Malaysia, June 2016. [Online]. Available: http://arxiv.org/abs/1602.08255

[11] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," *Wireless Communications, IEEE Transactions on*, vol. 15, no. 1, pp. 131–145, Jan 2016.

[12] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks," *CoRR*, vol. abs/1601.00321, 2016. [Online]. Available: http://arxiv.org/abs/1601.00321

[13] S. T. ul Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-aho, "Caching in wireless small cell networks: A storage-bandwidth tradeoff," *IEEE Communications Letters*, vol. PP, no. 99, pp. 1–1, 2016.

[14] S. Tamoor-ul-Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-aho, "Modeling and analysis of content caching in wireless small cell networks," *CoRR*, vol. abs/1507.00182, 2015. [Online]. Available: http://arxiv.org/abs/1507.00182

[15] D. Lecompte and F. Gabin, "Evolved multimedia broadcast/multicast service (eMBMS) in LTE-advanced: overview and Rel-11 enhancements," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 68–74, 2012.

[16] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *IEEE WCNC*, April 2014, pp. 2300–2305.

[17] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *submitted to Wireless Communications, IEEE Transactions on*, vol. abs/1509.06611, 2015. [Online]. Available: http://arxiv.org/abs/1509.06611

[18] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *IEEE Int. Conf. on Commun. (ICC)*, London, United Kingdom, June 2015, pp. 1–6.

[19] B. N. Bharath and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," *CoRR*, vol. abs/1504.03632, 2015. [Online]. Available: http://arxiv.org/abs/1504.03632

[20] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *CoRR*, vol. abs/1309.0604, 2013. [Online]. Available: http://arxiv.org/abs/1309.0604

[21] Y. Cui, D. Jiang, and Y. Wu, “Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks,” *CoRR*, vol. abs/1512.06176, 2015. [Online]. Available: http://arxiv.org/abs/1512.06176

[22] J. Andrews, F. Baccelli, and R. Ganti, “A tractable approach to coverage and rate in cellular networks,” *Communications, IEEE Transactions on*, vol. 59, no. 11, pp. 3122–3134, November 2011.

[23] S. Singh, H. S. Dhillon, and J. G. Andrews, “Offloading in heterogeneous networks: modeling, analysis, and design insights,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, March 2013.

[24] S. Singh and J. Andrews, “Joint resource partitioning and offloading in heterogeneous cellular networks,” *Wireless Communications, IEEE Transactions on*, vol. 13, no. 2, pp. 888–901, Feb 2014.

[25] S. M. Yu and S.-L. Kim, “Downlink capacity and base station density in cellular networks,” in *Modeling Optimization in Mobile, Ad Hoc Wireless Networks (WiOpt), 2013 11th International Symposium on*, May 2013, pp. 119–124.

[26] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 1999.

[27] M. Haenggi and R. K. Ganti, “Interference in large wireless networks,” *Foundations and Trends in Networking*, vol. 3, no. 2, pp. 127–248, 2009.