Cache-Enabled Heterogeneous Cellular Networks: Optimal Tier-Level Content Placement

Juan Wen, Kaibin Huang, Sheng Yang, Victor O. K. Li

Cache-Enabled Heterogeneous Cellular Networks: Optimal Tier-Level Content Placement

Juan Wen, Kaibin Huang, Sheng Yang and Victor O. K. Li

***Abstract*—Caching popular contents at *base stations* (BSs) of a *heterogeneous cellular network* (HCN) avoids frequent information passage from content providers to the network edge, thereby reducing latency and alleviating traffic congestion in backhaul links. The potential of caching at the network edge for tackling 5G challenges has motivated the recent studies of optimal content placement in large-scale HCNs. However, due to the complexity of network performance analysis, the existing strategies were designed mostly based on approximation, heuristics and intuition. In general, the optimal strategies for content placement in HCNs remain largely unknown and deriving them forms the theme of this paper. To this end, we adopt the popular random HCN model where K tiers of BSs are modeled as independent *Poisson point processes* (PPPs) distributed in the plane with different densities. Further, the *random caching* scheme is considered where each of a given set of M files with corresponding popularity measures is placed at each BS of a particular tier with a corresponding probability, called *placement probability*. The probabilities are identical for all BSs in the same tier but vary over tiers, giving the name *tier-level content placement*. We consider the network performance metric, *hit probability*, defined as the probability that a file requested by the typical user is delivered successfully to the user. Leveraging existing results on HCN performance, we maximize the hit probability over content placement probabilities, which yields the optimal tier-level placement policies. For the case of uniform received signal-to-interference thresholds for successful transmissions for BSs in different tiers, the policy is in closed-form where the placement probability for a particular file is proportional to the *square-root of the corresponding popularity measure* with an offset depending on BS caching capacities. For the general case of non-uniform SIR thresholds, the optimization problem is non-convex and a sub-optimal placement policy is designed by approximation, which has a similar structure as in the case of uniform SIR thresholds and shown by simulation to be close-to-optimal.**

***Index Terms*—Cache-enabled wireless networks, heterogeneous cellular networks, content delivery, stochastic geometry.**

I. Introduction

The last decade has seen multimedia contents becoming dominant in mobile data traffic [1]. As a result, a vision for 5G wireless systems is to enable high-rate and low-latency content delivery, e.g., ultra-high-definition video streaming [2]. The key challenge for realizing this vision is that transporting large volumes of data from content providers to end users causes severe traffic congestion in backhaul links, resulting in rate loss and high latency [3]. On the other hand, the dramatic advancement of the hard-disk technology makes it feasible to deploy large storage (several to dozens of TB) at the network edge (e.g., base stations (BSs) and dedicated access points) at low cost [4]. In view of these, caching popular contents at the network edge has emerged as a promising solution, where highly skewed content popularity is exploited to alleviate the heavy burden on backhaul networks and reduce latency in content delivery [5]–[7]. Since popular contents vary at a time scale of several days [8], content placement can be performed every day during off-peak hours without causing an extra burden on the system. Compared with caching in wired networks, the broadcast and superposition natures of the wireless medium make the optimal content placement in wireless networks a much more challenging problem and solving the problem has been the main theme in designing cache-enabled wireless systems and networks [9]. Along the same theme, the current work considers caching for next-generation *heterogeneous cellular networks* (HCNs), adopting the classic K-tier HCN model [10], and focuses on studying the optimal policy for placing contents in different BS tiers.

A. Related Work

Extensive research has been conducted on studying the performance gain for joint content-placement and wireless transmissions as well as designing relevant techniques. From the information-theoretic perspective, the capacity scaling laws were derived for a large cache-enabled wireless network with a hierarchical tree structure [11]. In [12], the novel idea of integrating coding into user caching, called *coded caching*, was proposed to improve substantially the efficiency of content delivery over uncoded caching. Specifically, exploiting joint coding of multiple files and the broadcast nature of downlink channels, the content placement at BSs and delivery were jointly optimized to minimize the communication overhead for content delivery. Coded caching in an erasure broadcast channel was then studied in [13] where the optimal capacity region has been derived in some cases. In parallel, extensive research has also been carried out on the more practical uncoded caching where the focus is the design of strategies for content-placement at BSs (or access points) to optimize the network performance in terms of the expected time for file downloading. Since optimal designs are NP-hard in general [14], [15], most research has resorted to sub-optimal techniques with close-to-optimal performance. Specifically, practical algorithms have been designed for caching contents distributively at access points dedicated for content delivery using greedy algorithms [14] and the theory of belief-propagation [15]. Furthermore, joint transmission and caching

J. Wen, K. Huang, and Victor O. K. Li are with the Deptment of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China (Email: jwen@eee.hku.hk, huangkb@eee.hku.hk, vli@eee.hku.hk).

S. Yang is with CentraleSupelec, Gif-sur-Yvette Cedex 91192, France (Email: sheng.yang@supelec.fr)

The work was supported by Hong Kong Research Grants Council under the PROCORE-France/Hong Kong Joint Research Scheme with the grant number F-HKU703/15T.

can further improve the network performance [16]–[18]. Sub-optimal solutions were developed to maximize the quality of service for multi-relay networks [16] and two-hop relaying network [17] via decomposing the original problem into several simpler sub-problems. Considering the opportunistic co-operative MIMO, schemes were presented in [18] to leverage multi-time-scale joint optimization of power and cache control to enable real-time video streaming. Recent advancements in wireless caching techniques have been summarized in various journal special issues and survey articles (see e.g., [19]).

It is also crucial to understand the performance gain that caching brings to large-scale wireless networks. Presently, the common approach is to model and design cache-enabled wireless networks using stochastic geometry. The approach leverages the availability of a wide range of existing stochastic geometric network models, ranging from *device-to-device* (D2D) networks to HCNs, and relevant results by adding caching capacities to network nodes [20]–[27]. In the resultant models, BSs and mobiles are typically distributed in the 2-dimensional (2D) plane as Poisson point processes (PPPs). Despite their similarity in the network nodes' spatial distributions, the cache-enabled networks differ from the traditional networks without caching in their functions, with the former aiming at efficient content delivery and the latter at reliable communication. Correspondingly, the performance of a cache-enabled network is typically measured using a metric called *hit probability*, defined as the probability that a file requested by a typical user is not only available in the network but can also be wirelessly delivered to the user [24]. Based on stochastic-geometry network models, the performance of cache-enabled D2D networks [20], [21] and HCNs [22], [23] were analyzed in terms of hit probability as well as average throughput. For small-cell networks, one design challenge is that the cache capacity limitation of BSs affects the availability of contents with low and moderate popularity. A solution was proposed in [26] for multi-cell cooperative transmission/delivery in order to enhance the content availability. Specifically, the proposed content-placement strategy is to partition the cache of each BS into two halves for storing both the most popular files and fractions of other files; then multi-cell cooperation effectively integrates storage spaces at cooperative BSs into a larger cache to increase content availability for improving the network hit probability. Based on approximate performance analysis, the content-placement strategy derived in [26] is heuristic and the optimal one remains unknown.

In the aforementioned work, the content placement at cache-enabled nodes is *deterministic*. An alternative strategy is *probabilistic (content) placement* where a particular file is placed in the cache of a network node (BS or mobile) with a given probability [24], [25], called *placement probability*. The strategy has also been considered in designing large-scale cache-enabled networks [24], [25]. The key characteristic of probabilistic placement is that all files with nonzero placement probabilities are available in a large-scale network with their spatial densities proportional to the probabilities. Given its random nature, the strategy fits the stochastic-geometry models better than the deterministic counterpart as the former allows for tractable analyses for certain networks as demonstrated in this work. The placement probabilities for different content files were optimized to maximize the hit probability for cellular networks in [24] and for D2D networks in [25]. It was found therein that the optimal placement probabilities are highly dependent on, but not identical to, the (content) *popularity measures*, defined as the content-demand distribution over files as they are also functions of network parameters, e.g., wireless-link reliability and cache capacities. To improve content availability, a hybrid scheme combining deterministic and probabilistic content placement was proposed in [27] for HCNs with multicasting where the most popular files are cached at every macro-cell BS and different combinations of other files are randomly cached at pico-cell BSs. Similar to the strategy in [26], the proposed strategy in [27] does not lead to tractable network-performance analysis and was optimized for the approximate hit probability.

B. Motivation, Contributions and Organization

HCNs are expected to be deployed as next-generation wireless networks supporting content delivery besides communication and mobile computing [9]. In view of prior work, the existing strategies for content placement in large-scale HCNs are mostly heuristic and the optimal policies in closed-form remain largely unknown, even though existing results reveal their various properties and dependence on network parameters. This motivates the current work on analyzing the structure of the optimal content-placement policies for HCNs.

To this end, the cache-enabled HCN is modeled by adopting the classic K-tier HCN model for the spatial distributions of BSs and mobiles [10]. To be specific, the locations of different tiers of BSs and mobiles are modeled as independent homogeneous PPPs with non-uniform densities. Besides density, each tier is characterized by a set of additional parameters including BS transmission power, finite cache capacity and minimum received *signal-to-interference* (SIR) threshold required for successful content delivery. Note that the use of SIR is based on the implicit assumption that the network is interference limited. A user is associated with the nearest BS where the requested file is available. It is assumed that there exists a content database comprising M files characterized by corresponding popularity measures. Each user generates a random request for a particular file based on the discrete popularity distribution. In the paper, we propose a tractable approach of probabilistic *tier-level content placement (TLCP)* for the HCN where the placement probabilities are identical for all BSs belonging to the same tier but are different across tiers. The goal of the current work is to analyze the structure of the optimal policies for TLCP given the network-performance metric of hit probability. The main contributions are summarized as follows.

1) **Hit Probability Analysis**. By extending the results on outage probability for HCNs in [10], the hit probability for cache-enabled HCNs are derived in closed form. The results reveal that the metric is determined not only by the physical-layer related parameters, including BS density, transmission power, and path-loss exponent, but also the

Table I: Summary of Notations

Symbol	Meaning
K	Total number of tiers in a HCN
M	Total number of files in a datbase
$\mathcal{F}_m$	The m-th file
Φ_k	Point process of BSs in the k-th tier
Φ_{mk}, Φ^c_{mk}	Point process of BSs in the k-th tier *with, without* file $\mathcal{F}_m$
λ_k, P_k	Density and transmission power of BSs in the k-th tier
β_k	SIR threshold of BSs in the k-th tier
h	Rayleigh fading gain with unit mean
α	Path-loss exponent
C_k	Cache capacity of BSs in the k-th tier
q_m	Popularity measure for file $\mathcal{F}_m$
p_{mk}	Placement probability for file $\mathcal{F}_m$ in the k-th tier BSs

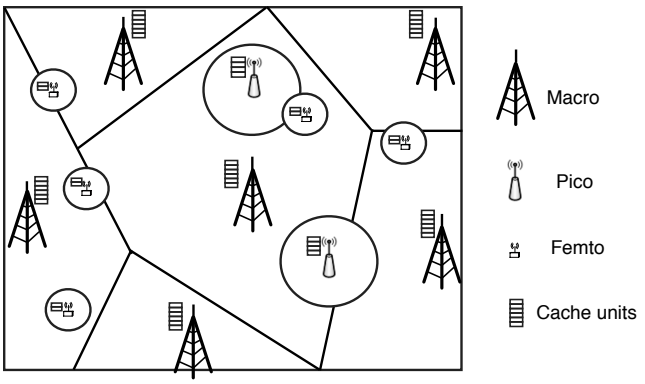


Figure 1: A cache-enabled heterogenous cellular network.

content-related parameters, including content-popularity measures and placement probabilities. With uniform SIR thresholds for all tiers, the hit probability is observed to be a monotone increasing function of the placement probability and converges to a constant independent of BS density and transmission power as the placement probabilities approach 1.

2) **Optimal Content Placement for Multi-Tier HCNs**. For a multi-tier HCN, the placement probabilities form an $M \times K$ matrix whose rows and columns correspond to the M files and the K tiers, respectively. First, consider a multi-tier HCN with uniform SIR thresholds for all tiers. Building on the results derived for single-tier HCNs, a weighted sum (over tiers) of the placement probabilities for a particular file has the structure that it is *proportional to the square root of the popularity measure with a fixed offset*. Using this result, we derive the expressions for individual placement probabilities and reveal a useful structure allowing for a simple sequential computation of the probabilities. An algorithm is proposed to realize the aforementioned procedure. Next, consider the general case of a multi-tier HCN with non-uniform SIR thresholds for different tiers. In this case, finding the optimal content placement is non-convex and it is thus difficult to derive the optimal policy in closed-form. However, a sub-optimal algorithm can be designed leveraging the insights from the optimal policy structures for the previous cases. Our numerical results show that the performance of the proposed scheme is close-to-optimal.

The remainder of the paper is organized as follows. The network model and metric are described in Section II. The hit probability and optimal content placement for cache-enabled HCNs are analyzed in Sections III and IV, respectively. Numerical results are provided in Section V followed by the conclusion in Section VI.

II. Network Model and Metric

In this section, we describe the mathematical model for the cache-enabled HCN illustrated in Fig. 1 and define its performance metric. The symbols used therein and their meanings are tabulated in Table I.

A. Network Topology

The spatial distributions of BSs are modeled using the classic K-tier stochastic-geometry model for the HCN described as follows [10]. The network comprises K tiers of BSs modeled as K independent homogeneous PPPs distributed in the plane. The k-th tier is denoted by Φ_k with the BS density and transmission power represented by λ_k and P_k, respectively. Assuming an interference-limited network, the transmission by a BS to an associated user is successful if the received SIR exceeds a given threshold, denoted by β_k, identical for all links in the k-th tier.

We consider a particular frequency-flat channel, corresponding to a single frequency sub-channel of a broadband system. Single antennas are deployed at all BSs and users. Furthermore, the BSs are assumed to transmit continuously in the unicast mode. The users are assumed to be Poisson distributed. As a result, based on Slyvnyak's theorem [28], it is sufficient to consider in the network-performance analysis a typical user located at the origin, which yields the expected experience for all users. The channel is modeled in such a way that the signal power received at the user from a k-th tier BS located at $X_k \in \mathbb{R}^2$ is given by $P_k h_{X_k} ||X_k||^{-\alpha}$, where the random variable $h_{X_k} \sim \exp(1)$ models the Rayleigh fading and $\alpha > 2$ is the path-loss exponent [1]. Based on the channel model [2], the interference power measured at the typical user, denoted by I_0, can be written as

$$I_0 = \sum\nolimits_{k=1}^{K} \sum\nolimits_{X \in \Phi_k \setminus X_k} P_k h_X ||X||^{-\alpha}, \qquad (1)$$

where the fading coefficients $\{h_X\}$ are assumed to be independent and identically distributed (i.i.d.).

[1] In practice, the path-loss exponent may vary over the tiers. The corresponding conditional hit probability does not have a closed-form expression as in Lemma 3, resulting in an intractable optimization problem. The current solution for the simpler case of uniform path-loss exponent can provide useful insights into designing practical content placement schemes for the said general case.

[2] The effect of shadowing on network performance is omitted in the current model for simplicity but can be captured by modifying the model following the method in [29], namely appropriately scaling the transmission power of BSs in each tier. However, the corresponding modifications of the analysis and algorithmic design are straightforward without changing the key results and insights.

B. Probabilistic Content Placement

In this paper, we consider a content (such as video file) database containing M files with normalized size equal to 1 following the literature [24], [26], [27].[3] As illustrated in Fig. 1, the BSs from different tiers are assumed to have different cache capacities which are denoted by C_k for the k-th tier with $k = 1, 2, \cdots, K$. We make the practical assumption that not all BSs have sufficient capacities for storing the whole database, i.e., $C_k \leq M, \forall k$. We adopt a probabilistic content placement scheme similar to the one in [24] to randomly select files for caching at different tiers under their cache capacity constraints:

$$\sum_{m=1}^{M} p_{mk} \leq C_k, \forall k. \tag{2}$$

Specifically, the m-th file, denoted by $\mathcal{F}_m$, is cached at a tier-k BS with a fixed probability p_{mk} called a *placement probability*. The placement probabilities, $(p_{1k}, p_{2k}, \cdots, p_{Mk})$, are identical for all BSs in the same tier k, $k = 1, \ldots, K$. They specify the *tier-level content placement* (TLCP). Grouping the placement probabilities yields the following *placement probability matrix*:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & & \vdots \\ p_{M1} & p_{M2} & \cdots & p_{MK} \end{bmatrix}. \tag{3}$$

The rows and columns of $\mathbf{P}$ correspond to different files and different tiers, respectively. Given the placement probabilities in $\mathbf{P}$ and under the cache-capacity constraints in (2), there exist specific strategies of randomly placing contents at individual BSs such that File $\mathcal{F}_m$ is available at a tier-k BS with a probability exactly equal to p_{mk} [24]. One of such strategies is illustrated in [24, Fig. 1]. Given the existence of random-placement strategies for achieving the content availability specified by $\mathbf{P}$, this paper focuses on optimizing $\mathbf{P}$ for maximizing the hit probability.

The files in the content database differ in popularity, measured by a corresponding set of values $\{q_m\}$ with $q_m \in [0, 1]$ for all m and $\sum_{m=1}^{M} q_m = 1$ [24]–[27]. This set is a probability mass function such that the typical user requests file $\mathcal{F}_m$ with probability q_m. Without loss of generality, it is assumed that the files are ordered in decreasing popularity, i.e., $q_1 > q_2 > \cdots > q_M$.

C. Content-Centric Cell Association

Content-centric cell association accounts for both the factor of link reliability and the factor of content availability. We adopt a common scheme that associates a user with the BS that maximizes the received signal power among those having the requested file (see e.g., [27], [30]).[4] It is important to note that due to limited BS storage, the database cached at BSs is only the popular subset of all contents. Thus, it is possible that a file requested by a user is unavailable at the network edge, which has to be retrieved from a data center across the backhaul network. In such cases, the classic cell association rule is applied to connect the user to the nearest BS. These cases occur infrequently and furthermore are outside the current scope of content placement at the network edge. Thus, they are omitted in our analysis for simplicity following the common approach in the literature (see e.g., [30]). For ease of exposition, we partition the HCN into $M \times K$ effective tiers, called the *content-centric tiers*, according to the file availability within each tier. The (m, k)-th content-centric tier refers to the process of tier-k BSs with file $\mathcal{F}_m$, denoted by Φ_{mk}, while the remaining tier-k BSs are denoted by Φ^c_{mk} with $\Phi_{mk} \cup \Phi^c_{mk} = \Phi_k$. Due to the probabilistic content placement scheme, Φ_{mk} and Φ^c_{mk} are independent PPPs with densities $p_{mk}\lambda_k$ and $(1 - p_{mk})\lambda_k$, respectively. A user is said to be associated with the (m, k)-th content-centric tier if the user requests $\mathcal{F}_m$ and is served by a tier-k BS. Then, conditioned on the typical user requesting file $\mathcal{F}_m$, the serving BS X_k is given by

$$\text{(Cell Association)} \quad X_k = \arg \max_{X \in \bigcup_k \Phi_{mk}} P_X ||X||^{-\alpha}, \tag{4}$$

where P_X denotes BS X's transmission power. In addition, conditioned on the typical user requesting file $\mathcal{F}_m$, the interference power I_0 in (1) can be written in terms of the content-centric tiers as:

$$I_0(\mathcal{F}_m) = \sum\nolimits_{k=1}^{K} \sum\nolimits_{X \in \Phi_{mk} \setminus X_k} P_k h_X ||X||^{-\alpha} + \sum\nolimits_{k=1}^{K} \sum\nolimits_{X \in \Phi^c_{mk}} P_k h_X ||X||^{-\alpha}. \tag{5}$$

D. Network Performance Metric

The network performance is measured by the *hit probability* defined as the probability that a file the typical user requested is not only cached at a BS but also successfully delivered by the BS over the wireless channel (see e.g., [24]). By definition, the hit probability quantifies the reduced fraction of backhaul load. In addition, it can also indicate the reduction of mean latency in the backhaul network (see Appendix A for details). Therefore, we use the hit probability as the main network performance metric in this paper. For the purpose of analysis, let $\mathcal{P}$ denote the (unconditional) hit probability, $\mathcal{P}_m$ denote the conditional hit probability given that the typical user requests file $\mathcal{F}_m$, and q_m denote the content popularity for file $\mathcal{F}_m$. Then

$$\mathcal{P} = \sum_{m=1}^{M} q_m \mathcal{P}_m. \tag{6}$$

[3] In practice, a large content file to be cached is usually divided in to units of equal sizes because they have different popularity. For instance, the beginning 1-minute of a YouTube video is much more popular than the remainder. Thus, to be precise, the equal-size files considered in this paper correspond to content units in practice.

[4] *Coordinated multiple access point* (CoMP) defined in the LTE standard can be applied to improve the network performance via associating each user with multiple BSs. Adopting the technology in the current network model does not lead to tractable analysis. However, it is possible to develop practical content-delivery schemes for HetNets with CoMP by integrating the current optimal TL content placement and the design of cooperative content delivery in [26].

Furthermore, define the *association probability* indexed by (m,k), denoted by A_{mk}, as the probability that the typical user is associated with the (m,k)-th content-centric tier. The hit probability conditional on this event is represented by $\mathcal{P}_{mk}$. It follows that

$$\mathcal{P}_m = \sum_{k=1}^{K} A_{mk}\mathcal{P}_{mk}. \tag{7}$$

III. Analysis of Hit Probability

In this section, the hit probability for the cache-enabled HCN is calculated. To this end, the association probabilities and the probability density function (PDF) of the serving distances are derived in the following two lemmas, via directly modifying Lemmas 3 and 5 in [31] enabled by the interpretation of the HCN as one comprising $M \times K$ content-centric tiers (see Section II-B).

Lemma 1 (Association Probabilities)**.** The association probability that the typical user belongs to the (m,k)-th effective tier is given as

$$A_{mk} = \frac{p_{mk}\lambda_k P_k^{\delta}}{\sum_{j=1}^{K} p_{mj}\lambda_j P_j^{\delta}}, \tag{8}$$

where the constant $\delta = \frac{2}{\alpha}$.

Proof: See Appendix B. □

The result in Lemma 1 shows that the typical user requesting a particular file is more likely to be associated with one of those tiers having not only larger placement probability but also denser BS or higher BS transmission power, aligned with intuition. In addition, it is shown that if δ is small, the placement probability and BS density have more dominant effects on determining the association probability than transmission power, since P_j^{δ} converges to one for all j as $\delta \to 0$.

Lemma 2 (Statistical Serving Distances)**.** The PDF of the serving distance between the typical user and the associated BS in the (m,k)-th effective tier is given as

$$f_R(r) = \frac{2\pi p_{mk}\lambda_k}{A_{mk}} r \exp\left(-\pi \sum_{j=1}^{K} p_{mj}\lambda_j \left(\frac{P_j}{P_k}\right)^{\delta} r^2\right), \tag{9}$$

where A_{mk} is given in (8).

Next, we are ready to derive the hit probabilities using Lemmas 1 and 2. For ease of notation, we define the following two functions $Q(\beta_k)$ and $V(\beta_k)$, which are related to the interference coming from the BSs *with* and *without* the file $\mathcal{F}_m$, respectively:

$$Q(\beta_k) = \frac{\delta\beta_k}{1-\delta}\, {}_2F_1[1, 1-\delta; 2-\delta; -\beta_k], \tag{10}$$

$$V(\beta_k) = \beta_k^{\delta}\, \delta\pi \csc(\delta\pi), \tag{11}$$

where ${}_2F_1[\cdot]$ denotes the Gauss hypergeometric function and $\csc(.)$ is the cosecant-trigonometry function. To further simplify the expression of hit probability, we define the following function:

$$W(\beta_k) = 1 + Q(\beta_k) - V(\beta_k). \tag{12}$$

Then the conditional hit probability can be written as shown in the following lemma.

Lemma 3 (Conditional Hit Probability)**.** In the cache-enabled HCN, the conditional hit probability for the typical user requesting file $\mathcal{F}_m$ is given as

$$\mathcal{P}_m = \sum_{k=1}^{K} \frac{p_{mk}\lambda_k P_k^{\delta}}{W(\beta_k)\sum\limits_{i=1}^{K} p_{mi}\lambda_i P_i^{\delta} + V(\beta_k)\sum\limits_{i=1}^{K}\lambda_i P_i^{\delta}}, \tag{13}$$

where the functions $V(\cdot)$ and $W(\cdot)$ are defined in (11) and (12), respectively.

Proof: See Appendix C. □

Using Lemma 3 and the definition of hit probability in (6), we obtain the first main result of this paper.

Theorem 1 (Hit Probability)**.** The hit probability for the cache-enabled HCNs is given as

$$\mathcal{P} = \sum_{m=1}^{M} q_m \sum_{k=1}^{K} \frac{p_{mk}\lambda_k P_k^{\delta}}{W(\beta_k)\sum\limits_{i=1}^{K} p_{mi}\lambda_i P_i^{\delta} + V(\beta_k)\sum\limits_{i=1}^{K}\lambda_i P_i^{\delta}},$$

where functions $V(\cdot)$ and $W(\cdot)$ are given in (11) and (12), respectively.

Theorem 1 shows that the hit probability is determined by two sets of network parameters: one set is related to the physical layer including the BS density $\{\lambda_k\}$, transmit power $\{P_k\}$, and path-loss parameter δ; the other set contains content-related parameters including the popularity measures $\{q_m\}$ and placement probabilities $\{p_{mk}\}$.

From Theorem 1, we can directly obtain hit probabilities for two special cases, namely, the single-tier HCNs and the multi-tier HCNs with uniform SIR thresholds, as shown in the following two corollaries.

Corollary 1 (Hit Probability for Single-Tier HCNs)**.** Given $K = 1$, the hit probability for cache-enabled HCNs is

$$\mathcal{P} = \sum_{m=1}^{M} q_m \frac{p_m}{W(\beta)p_m + V(\beta)}, \tag{14}$$

where the functions $V(\cdot)$ and $W(\cdot)$ are given in (11) and (12).

Corollary 1 shows that the hit probability for single-tier cache-enabled networks is independent with BS density and transmit power, which is a well-known characteristic of interference-limited cellular networks. On the other hand, it is found to be monotone increasing with growing placement probabilities as the spatial content density increases.

Corollary 2 (Hit Probability for Multiple-tier HCNs with Uniform SIR Thresholds)**.** Given $\beta_k = \beta\ \forall k$, the hit probability for the cache-enabled HCNs is given as

$$\mathcal{P} = \sum_{m=1}^{M} q_m \frac{\sum_{k=1}^{K} p_{mk}\lambda_k P_k^{\delta}}{W(\beta)\sum_{k=1}^{K} p_{mk}\lambda_k P_k^{\delta} + V(\beta)\sum_{k=1}^{K}\lambda_k P_k^{\delta}}, \tag{15}$$

where functions $V(\cdot)$ and $W(\cdot)$ are given in (11) and (12), respectively.

Remark 1 (Effects of Large Cache Capacities)**.** Corollary 2 shows that the hit probability is a monotone increasing function of the placement probabilities $\{p_{mk}\}$ and converges to a constant, which is independent of the BS densities and transmission powers, as all the placement probabilities become ones, corresponding to the case of large cache capacities. At this limit, the cache-enable HCN is effectively the same as a traditional interference-limited HCN for general data services and the said independence is due to a uniform SIR threshold and is well known in the literature (see e.g., [10]).

IV. Optimal Tier-Level Content Placement

In this section, we maximize the hit probability derived for the cache-enabled HCNs in the preceding section over the placement probabilities.

A. Problem Formulation

The TLCP problem consists of finding the placement matrix $\mathbf{P}$ in (3) that maximizes the hit probability for HCNs as given in Theorem 1. Mathematically, the optimization problem can be formulated as follows:

$$\begin{aligned}\max_{\mathbf{P}} \sum_{m=1}^{M} q_m \sum_{k=1}^{K} \frac{p_{mk}\lambda_k P_k^{\delta}}{W(\beta_k)\sum_{i=1}^{K} p_{mi}\lambda_i P_i^{\delta} + V(\beta_k)\sum_{i=1}^{K}\lambda_i P_i^{\delta}} \\ \text{s.t.} \sum_{m=1}^{M} p_{mk} \le C_k, \forall k, \\ p_{mk} \in [0,1], \forall m,k,\end{aligned} \tag{P0}$$

where the first constraint from (2) is based on the BS cache capacity for each tier and the second constraint arises from the fact that p_{mk} is a probability.

It is numerically difficult to directly solve Problem P0, since it has a structure of "sum-of-ratios" with a non-convex nature and has been proved to be NP-complete. In order to provide useful insights and results for tackling the problem, the optimal content placement policies are first analyzed for the special case of single-tier HCNs and then extended to multi-tier HCNs.

B. Single-Tier HCNs

For the current case with $K = 1$, using Corollary 1, Problem P0 is simplified as:

$$\begin{aligned}\max_{\mathbf{p}} \sum_{m=1}^{M} q_m \frac{p_m}{W(\beta)p_m + V(\beta)} \\ \text{s.t.} \sum_{m=1}^{M} p_m \le C, \\ p_m \in [0,1], \forall m,\end{aligned} \tag{P1}$$

where p_m denotes the placement probability for file $\mathcal{F}_m$, the vector $\mathbf{p} = (p_1, p_2, \cdots, p_m)$, C denotes the cache capacity for single-tier HCNs.

It should be noted that Problem P1 has the same structure as that for the asymptotic (high SNR and high user density) case in [30] where the single-tier cache-enabled BSs are distributed as a PPP and random combination of files are cached in each BS with probability p_m. Nevertheless, it is still valuable to discuss this special case since it provides useful insights for solving the complex problem for multi-tier HCNs. Therefore, this paper focuses on these insights.

Algorithm 1 Computing the Optimal Lagrangian Multiplier u^* by a Bi-section Search.

initialize $u^0 \in [u^{(0,\min)}, u^{(0,\max)}] = [\frac{q_M V}{(W+V)^2}, \frac{q_1}{V}]$
repeat
  $u^{(\ell+1)} = u^{(\ell,\min)} + \frac{u^{(\ell,\max)} - u^{(\ell,\min)}}{2}$
  if $\sum_{m=1}^{M} p_m(u^{(\ell+1)}) < C$, $u^{(\ell+1,\max)} = u^{(\ell+1)}$
  else $u^{(\ell+1,\min)} = u^{(\ell+1)}$
until u converges

Problem P1 is convex since the objective function is convex and the constraints are linear and can thus be solved using the Lagrange method. The Lagrangian function can be written as

$$L(\mathbf{p}, u) = \sum_{m=1}^{M} q_m \frac{p_m}{W(\beta)p_m + V(\beta)} + u\left(C - \sum_{m=1}^{M} p_m\right), \tag{16}$$

where $u \ge 0$ denotes the Lagrangian multiplier. Using the Karush-Kuhn-Tucker (KKT) condition, setting the derivative of L in (16) to zero leads to the optimal placement probabilities as shown in Theorem 2 where the optimal Lagrange multiplier is denoted by u^*. Note that the capacity constraint is active at the optimal point, namely $\sum_{m=1}^{M} p_m^*(u^*) = C$. This result comes from the fact that the objective function of Problem P1 is a monotone-increasing function of $\{p_m\}$.

Theorem 2 (Optimal TLCP for Single-Tier HCNs)**.** For the single-tier cache-enabled HCN, given the optimal Lagrangian multiplier u^*, the optimal content placement probabilities, denoted by $\{p_m^*\}$, that solve Problem P1 are given as

$$p_m^*(u^*) = \begin{cases} 1, & q_m \ge T_1, \\ \frac{\sqrt{V(\beta)}}{\sqrt{u^* W(\beta)}}\sqrt{q_m} - \frac{V(\beta)}{W(\beta)}, & T_0 < q_m < T_1, \\ 0, & q_m \le T_0, \end{cases} \tag{17}$$

where the thresholds $T_1 = \frac{u^*(W(\beta)+V(\beta))^2}{V(\beta)}$ and $T_0 = u^* V(\beta)$, and the optimal Lagrange multiplier u^* satisfies the equality

$$\sum_{m=1}^{M} p_m^*(u^*) = C. \tag{18}$$

In addition, the optimal Lagrangian multiplier u^* in Theorem 2 can be found via a simple bisection search. Let D be the number of iterations needed to find the optimal Lagrange multiplier u^*. Clearly, the computational complexity of TLCP for single-tier HCNs is $O(DM)$. The corresponding algorithm is shown in Algorithm 1.

Remark 2 (Offset-Popularity Proportional Caching Structure)**.** As illustrated in Fig. 2, the optimal content placement in Theorem 2 has the mentioned *offset-popularity proportional* (OPP) structure described as follows. Specifically, if the popularity measure of a particular file is within the range $[T_0, T_1]$, the optimal placement probability, p_m, monotonically increases

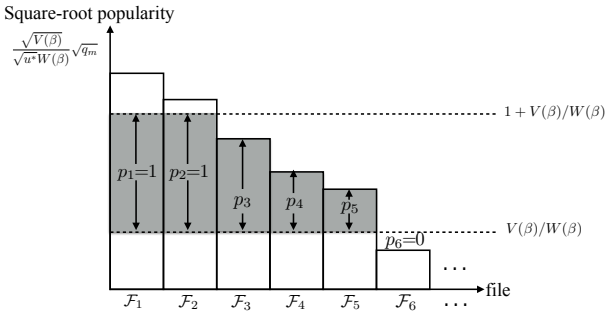


Figure 2: The structure of the optimal content-placement policy for single-tier HCNs.

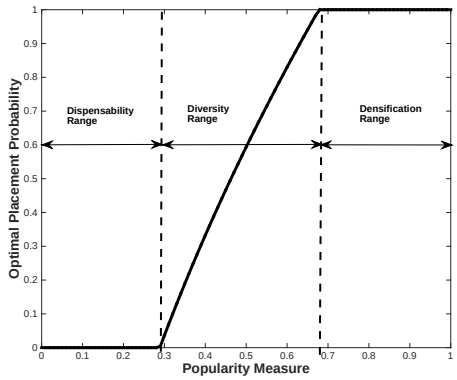


Figure 3: The effects of content popularity on the optimal placement probability.

with the *square root* of the popularity measure, i.e., $\sqrt{q_m}$. Otherwise, the probability is either 1 or 0 depending on whether the measure is above or below the range. Furthermore, the probability is offset by a function $V(\beta)/W(\beta)$ of the SIR threshold and scaled by a function of both the threshold and the cache capacity C.

Remark 3 (Effects of Content Popularity on Optimal Placement Probability)**.** The result in Theorem 2 shows that the optimal content placement probability is decided by its popularity measure. In particular, content files can be separated by defining three ranges of popularity measure, corresponding to placement probabilities of 0, $(0,1)$ and 1 as illustrated in Fig. 3, called the *dispensability*, *diversity*, and *densification* ranges, respectively. In the dispensability range, the files are highly unpopular and do not need to be cached in the network. In contrast, the files in the densification range are highly popular such that their spatial density should be maximized by caching the files at every BS. Last, files in the diversity range have moderate popularity and it is desirable to have all of them available in the network, corresponding to enhancing spatial content diversity. As a result, they are cached at different fractions of BSs.

Remark 4 (Effects of SIR threshold)**.** The SIR threshold β affects both the popularity thresholds (T_0 and T_1) in the optimal placement policy (see Theorem 2). It is observed from numerical results that both thresholds are monotone increasing functions of β.

Remark 5 (Effects of Lagrangian multiplier u^*)**.** The value of Lagrangian multiplier affects the popularity thresholds T_0 and T_1, and is determined by the capacity constraint equality, i.e., $\sum_{m=1}^{M} p_m^*(u^*) = C$. In the case that the requested cache unit is larger than the cache capacity, i.e., $\sum_{m=1}^{M} p_m(u) > C$, the Lagrangian multiplier u should be increased to enlarge the popularity thresholds and thus decrease the placement probabilities, and vice versa.

Problem P1 is considered purposely to help solve the general version and also for clarity in exposition. In particular, the insight from solving P1 is exploited to solve P2 in closed form given uniform SIR thresholds and develop a sub-optimal scheme for the case with non-uniform thresholds.

C. Multi-tier HCNs with Uniform SIR Thresholds

Consider a multi-tier HCN with uniform SIR thresholds for all tiers. Based on the hit probability in Corollary 2, Problem P0 for the current case is given as:

$$\begin{aligned} \max_{\mathbf{p}} \ & \sum_{m=1}^{M} q_m \frac{\sum_{k=1}^{K} p_{mk} z_k}{W \sum_{k=1}^{K} p_{mk} z_k + V'} \\ \text{s.t.} \ & \sum_{m=1}^{M} p_{mk} \leq C_k, \forall k, \\ & p_{mk} \in [0,1], \forall m, k, \end{aligned} \tag{\textbf{P2}}$$

where z_k and V' are constants defined as $z_k = \lambda_k P_k^{\delta}$ and $V' = V \sum_{k=1}^{K} z_k$. One can see that the problem is convex and can thus be solved numerically using a standard convex-optimization solver. However, the numerical approach may have high complexity if the content database is large and further yields little insight into the optimal policy structure. Thus, in the remainder of this section, a simple algorithm is developed for sequential computation of the optimal policy, which also reveals some properties of the policy structure.

To this end, define the tier-wise weighted sum of placement probabilities for each file as

$$g_m = \sum_{k=1}^{K} p_{mk} z_k, \qquad m = 1, 2, \cdots, M. \tag{19}$$

Using this definition, a relaxed version of Problem P2 can be rewritten as follows:

$$\begin{aligned} \max_{\{g_m\}} & \sum_{m=1}^{M} \frac{q_m g_m}{W g_m + V'} \\ \text{s.t.} & \sum_m g_m \le \sum_k C_k z_k, \forall k, \\ & 0 \le g_m \le \sum_k z_k, \forall m. \end{aligned} \tag{P3}$$

Comparing Problem P3 with P1 for the single-tier HCNs, one can see the two problems have identical forms. Thus, this allows Problem P3 to be solved following a similar procedure as P1, yielding the following proposition.

Proposition 1 (Weighted Sum of Optimal Placement Probabilities)**.** The weighted sum of the optimal placement probabilities for multi-tier HCNs with uniform SIR thresholds, denoted by g_m^*, is given as:

$$g_m^*(\eta^*) = \begin{cases} \sum_{k=1}^{K} \lambda_k P_k^{\delta}, & \text{if } q_m \ge T_1', \\ (\sqrt{q_m V'/\eta^*} - V')/W, & \text{if } T_0' < q_m < T_1', \\ 0, & \text{if } q_m \le T_0', \end{cases} \tag{20}$$

where $T_1' = \frac{\eta^*(W'+V')^2}{V'}$, $T_0' = \eta^* V'$, $W' = W\sum_{k=1}^{K} z_i$ and the optimal Lagrange multiplier η^* satisfies the following equality

$$\sum_{m=1}^{M} g_m^*(\eta^*) = \sum_{k=1}^{K} C_k z_k.$$

The value of η^* can be found using the bisection search in Algorithm 1. Then the optimal values for the weighted sum $\{g_m^*\}$ can be computed using Proposition 1.

Problem P3 is the relaxed version of P2 since the feasible region of P3 is larger than that of P2. Let $\{p_{mk}^*\}$ denote the optimal placement probabilities solving Problem P2 and $\{g_m^*\}$ the weighted sums solving Problem P3. The following proposition shows that the relaxation does not compromise the optimality of the solution.

Proposition 2. The solution of Problem P3 solves P2 in the sense that $\sum_{k=1}^{K} p_{mk}^* z_k = g_m^*, m = 1, 2, \ldots, M$.

Proof: See Appendix D. □

Next, based on the results in Propositions 1 and 2, the structure of the optimal placement policy is derived as shown in Theorem 3, which enables low-complexity sequential computation of the optimal placement probabilities.

Theorem 3 (Sequential Computation of Optimal Placement Probabilities)**.** One possible policy for optimal TLCP for the HCNs with uniform SIR thresholds is given as follows:

$$p_{mk}^* = \begin{cases} 1, & \text{if } q_m \ge T_1', \\ \min\left(\frac{1}{z_k}\sum_{j=1}^{k} \zeta_{mj}^* z_j - \frac{1}{z_k}\sum_{j=1}^{k-1} p_{mj}^* z_j, 1 \right), & \text{if } T_0' < q_m < T_1', \\ 0, & \text{if } q_m \le T_0', \end{cases} \tag{21}$$

Algorithm 2 Sequential Computation of Optimal Placement Probabilities for Multi-tier HCNs with Uniform SIR Thresholds.

1. Compute η^* using Algorithm 1 and $\{g_m^*\}$ using Proposition 1
2. For $m = 1 : M$
   for $k = 1 : K$
   set p_{mk}^* according to (21)
   update $C_k' = C_k' - p_{mk}^*$
   end
   end

where

$$\zeta_{mj}^* = \frac{g_m^*}{\sum_{i=m}^{M} g_i^*} \left(C_k - \sum_{i=1}^{m-1} p_{ik}^* \right), \tag{22}$$

and g_m^* is as given in Proposition 1.

Proof: See Appendix E. □

A key observation of the policy structure in Theorem 3 is that p_{mk}^* depends only on $\{p_{ij}^*\}$ with $i < m$ and $j < k$. This suggests that the optimal placement probabilities can be computed *sequentially* as shown in Algorithm 2 and thus the computational complexity of Algorithm 2 is $O(DMK)$, where D is the number of iterations needed to find the optimal Lagrange multiplier.

One can observe from Proposition 1 that the optimal solution for Problem P2 is not unique. In other words, there may exist a set of placement probabilities different from that computed using Algorithm 2 but achieving the same hit probability.

D. Multi-tier HCNs with Non-Uniform SIR Thresholds

For the current case, the problem of optimal content placement is Problem P0. As the problem is non-convex, it is numerically complex to solve and also difficult to develop low-complexity algorithms by analyzing the optimal policy structure. Therefore, a low-complexity sub-optimal algorithm is proposed for content placement for the current case. The algorithm is designed based on approximating the hit probability in Theorem 1 by neglecting the effects of the placement probability of other tiers on the hit probability of the k-th tier. Specifically, given $z_i = \lambda_i P_i^{\delta}$ as defined previously and by replacing the term $\sum_i p_{mi} z_i$ with $p_{mk} z_k$, the hit probability in Theorem 1 can be approximated by $\widetilde{\mathcal{P}}$ given as

$$\begin{aligned} \widetilde{\mathcal{P}} &= \sum_{m=1}^{M}\sum_{k=1}^{K} \frac{q_m p_{mk} z_k}{W(\beta_k) p_{mk} z_k + V(\beta_k)\sum_{i=1}^{K} z_i} \\ &= \sum_{k=1}^{K} \underbrace{\sum_{m=1}^{M} \frac{q_m p_{mk}}{W(\beta_k) p_{mk} + \widetilde{V}(\beta_k)}}_{\widetilde{\mathcal{P}}_k}, \end{aligned} \tag{23}$$

where $\widetilde{V}(\beta_k) = V(\beta_k)\frac{\sum_{i=1}^{K} z_i}{z_k}$. Thus, $\widetilde{\mathcal{P}} = \sum_{k=1}^{K} \widetilde{\mathcal{P}}_k$ where $\{\widetilde{\mathcal{P}}_k\}$ are independent of each other. As a result, maximizing $\widetilde{\mathcal{P}}$ is equivalent to separate maximization of individual summation terms $\{\widetilde{\mathcal{P}}_k\}$. Therefore, Problem P0 can be approximated

by K single-tier optimization problems, each of which is written as:

$$\begin{aligned} \max_{\mathbf{p_k}} \ & \sum_{m=1}^{M} q_m \frac{p_{mk}}{W(\beta_k)p_{mk} + \widetilde{V}(\beta_k)} \\ \text{s.t.} \ & \sum_{m=1}^{M} p_{mk} \leq C_k, \\ & p_{mk} \in [0,1], \forall m. \end{aligned} \tag{P4}$$

Using the results in the case of single-tier HCNs in Theorem 2, we derive the sub-optimal content-placement policy as shown in the following proposition.

Proposition 3 (Sub-Optimal TLCP for Multi-Tier HCNs with Non-Uniform SIRs)**.** For the multi-tier cache-enabled HCNs with non-uniform SIR thresholds, the optimal TLCP placement probabilities, denoted by $\{\tilde{p}^*_{mk}\}$, that solve Problem P4 are given as

$$\tilde{p}^*_{mk}(u^*_k) = \begin{cases} 1, & \text{if } \ q_m \geq \widetilde{T}_{1k}, \\ \frac{\sqrt{\widetilde{V}(\beta_k)}}{\sqrt{u^*_k W(\beta_k)}}\sqrt{q_m} - \frac{\widetilde{V}(\beta_k)}{W(\beta_k)}, & \text{if } \ \widetilde{T}_{0k} < q_m < \widetilde{T}_{1k}, \\ 0, & \text{if } \ q_m \leq \widetilde{T}_{0k}, \end{cases} \tag{24}$$

where $\widetilde{T}_{1k} = \frac{u^*_k(W(\beta_k)+\widetilde{V}(\beta_k))^2}{\widetilde{V}(\beta_k)}$ and $\widetilde{T}_{0k} = u^*_k\widetilde{V}(\beta_k)$. The optimal dual variable u^*_k satisfies the equality

$$\sum_{m=1}^{M} p^*_{mk}(u^*_k) = C_k. \tag{25}$$

The above sub-optimal TLCP policy approximates problem P0 as K independent single-tier optimization problems. Thus, the corresponding computational complexity is $O(DMK)$, where D denotes the number of iterations needed to find the optimal Lagrange multiplier for each tier. In addition, the numerical results in the next section show that it can attain close-to-optimal performance.

V. Simulation Results

In this section, simulation is conducted to validate the optimality of the content-placement policies derived in the preceding section and to compare the performance of the strategy of TLCP with conventional ones. The benchmark strategies include the *"most popular" content placement* (MPCP) that caches the most popular contents in a greedy manner and the *hybrid content placement* (HCP) proposed in [27]. Our simulation is based on the following settings unless specified otherwise. The number of BS tiers is $K = 2$ and the path-loss exponent $\alpha = 3$. The BS transmission power for the two tiers are $P_1 = 46$ dBm and $P_2 = 30$ dBm, respectively. The SIR threshold for tier 1 is fixed at $\beta_1 = -4$ dB while the other β_2 is a variable.

A. Conditional Hit Probability

The conditional hit probability for a typical file $\mathcal{F}_m$ versus caching probability p_2 is shown in Fig. 4. The analytical results are computed numerically using Lemma 3 and the simulated ones are obtained from Monte Carlo simulation

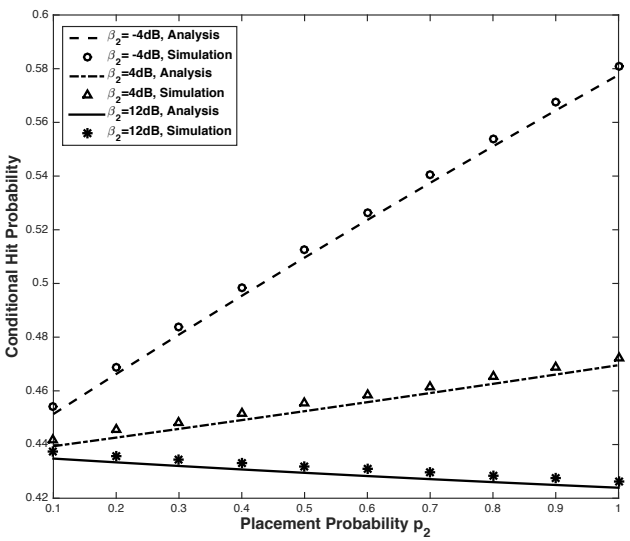
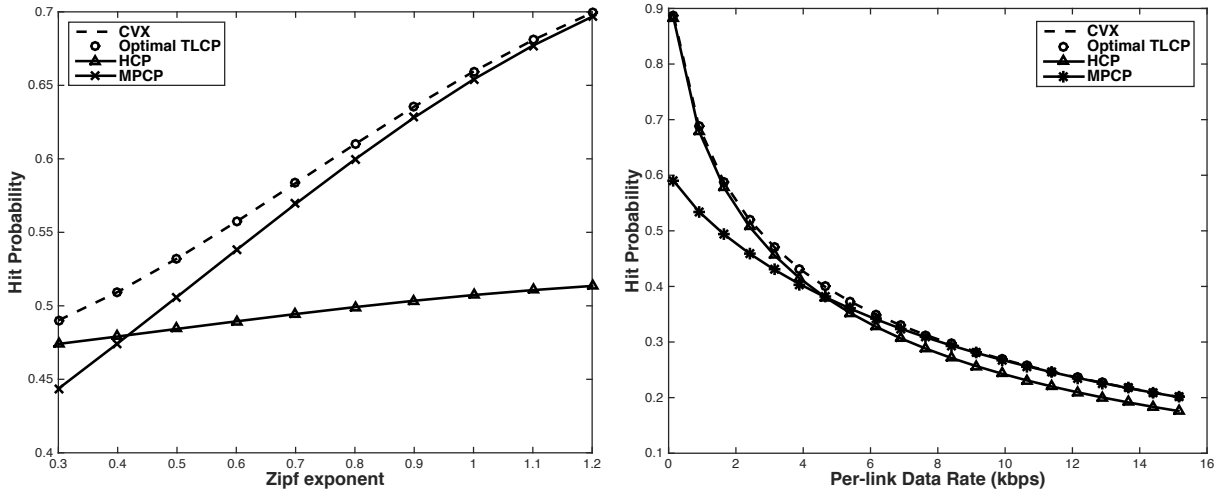


Figure 4: Conditional hit probability versus caching probability with $\lambda_2 = 5\lambda_1$, $p_1 = 1$.

using Matlab. First, it is observed that the simulated results match the analytical results well, which validates our analysis. In addition, the conditional hit probability increases with the growing placement probability p_2 if $\beta_2 = \beta_1$. However, it does not necessarily hold for the case $\beta_2 > \beta_1$, which shows that the effects of placement probability on the hit probability differ with SIR threshold. This is because increasing the placement probability increases the association probability of that tier (see Lemma 1) and thus decreases the conditional hit probability if that tier has smaller hit probability due to the larger SIR threshold. Meanwhile, it reduces the serving distance (see Lemma 2) and thus increases the conditional hit probability. The (final) effects of placement probability on the hit probability are determined by the absolute values of the above increment and decrement.

B. Optimal Content Placement

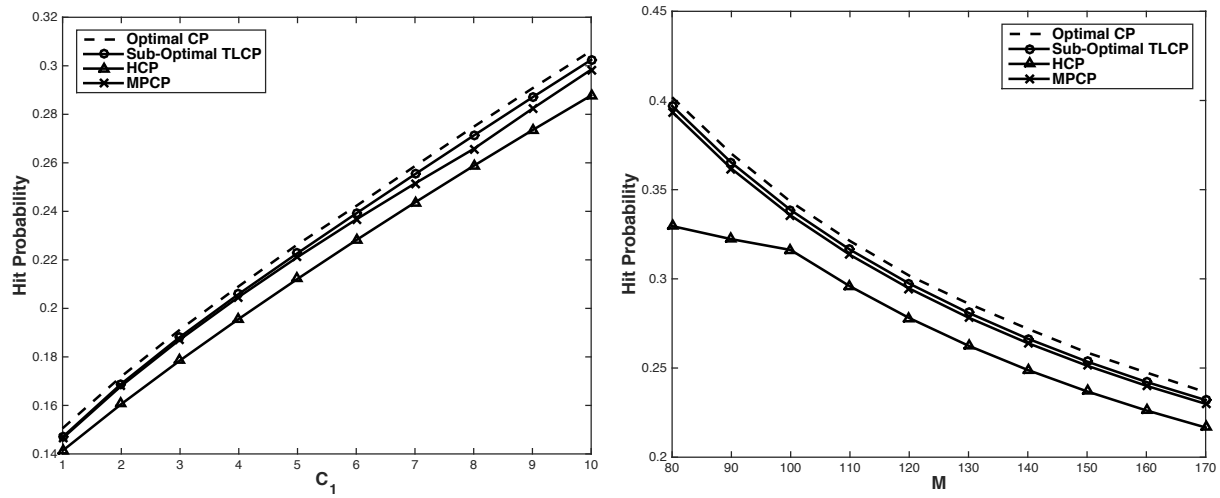
Fig. 5 compares the performance of the optimal TLCP proposed in this paper (Theorem 3) with MPCP and HCP. For MPCP, each macro-cell BS (or small-cell BS) caches the C_1 (or C_2) most popular files. For HCP, each macro-cell BS caches the C_1 most popular files while each small-cell BS caches the remaining files with optimal probabilistic content placement given in Theorem 2. First of all, Fig. 5 (a) shows that the hit probabilities under these three content placement policies increase as the content popularity becomes more skewed (a growing Zipf exponent γ), aligned with intuition. Next, TLCP is observed to achieve higher hit probability than MPCP and HCP due to the content densification and diversity (see Remark 3). Further, we observe that the gain over MPCP decreases with a growing γ since MPCP is a popularity-aware policy. In contrast, the gain over HCP increases with a growing γ. This is because, in the HCP, only Macro-cell tier caches the C_1 most popular files. In addition, the optimality of TLCP is verified by comparing the results given by the standard optimization tool CVX. Last, from Fig. 5 (b), it is observed



(a) Optimal hit probabilities under different content placement policies with $\beta_1 = \beta_2 = -4$ dB, $C_1 = 10$, $C_2 = 8$, $M = 20$.

(b) Optimal Hit probability for different throughput with $C_1 = 60$, $C_2 = 40$, $M = 120$

Figure 5: Optimal hit probabilities under different content placement policies with $\lambda_2 = 10\lambda_1$.



(a) Hit probability for different C_1 with $C_2 = 8$, $M = 20$.

(b) Hit probability for different M with $C_1 = 60$, $C_2 = 40$

Figure 6: Hit probabilities under different content placement policies with $\lambda_2 = 10\lambda_1$, $\beta_1 = -4$ dB, $\beta_2 = -2$ dB.

that the optimal hit probability increases as the per-link data rate reduces due to the reducing SIR threshold. In particular, the maximum hit probability (when the data rate approximates to 0) is less than 1 since the caching capacity is limited.

The hit probabilities under different CP policies, including the optimal CP, sub-optimal TLCP (see Proposition 3), MPCP, and HCP, versus different cache capacities and the number of contents are shown in Fig. 6(a) and Fig. 6(b), respectively. The optimal CP under this case (i.e., multi-tier HCNs with non-uniform SIR thresholds) is derived by adopting the dual methods for non-convex optimization problem in [32] since Problem P0 has the same structure as that in [32] and it satisfies the time-sharing condition (the proof is shown in Appendix F). Compared with the optimal CP, the sub-optimal TLCP provides close-to-optimal performance. It should be noted that the computational complexity of optimal solution is linear in the number of files M, but exponential in the number of BS tiers K, since it involves solving M nonconvex optimization problems, corresponding to the M tones, each with K variables. While the computational complexity of our proposed sub-optimal TLCP algorithm is linear with *both* M and K. In addition, besides the obvious monotone-increasing hit probability with cache capacity, we observe that the sub-optimal TLCP outperforms both the HCP and MPCP. Finally, it is shown that the hit probability increases with the growing cache capacity and decreases with the growing number of contents, which coincides with our intuition.

VI. Conclusion

In this paper, we have studied the hit probability and the optimal content placement of the cache-enabled HCNs where the BSs are distributed as multiple independent PPPs and the files are probabilistically cached at BSs in different tiers with different BS densities, transmission powers, cache

capacities and SIR thresholds. Using stochastic geometry, we have analyzed the hit probability and shown that it is affected by both the physical layer and content-related parameters. Specifically, for the case where all the tiers have the uniform SIR thresholds, the hit probability increases with all the placement probabilities and converges to its maximum (constant) value as all the probabilities achieve one without considering the cache capacity constraint. Then, with the cache capacity constraint, the optimal content placement strategy has been proposed to maximize the hit probability for both single- and multi-tier HCNs. We have found that the placement probability for each file has the OPP caching structure, i.e., the optimal placement probability is linearly proportional to the square root of offset-popularity with truncation to enforce the range for the probability. On the other hand, for multi-tier HCNs with uniform SIR thresholds, interestingly, the weighted-sum of the optimal placement probabilities also has the OPP caching structure. Further, an optimal or a sub-optimal TLCP caching algorithm has been proposed to maximize the hit probability HCNs with uniform or non-uniform SIR thresholds, respectively.

The fundamental structure of the optimal content placement strategies proposed in this paper provides useful guidelines and insights for designing cache-enabled wireless networks. As a promising future direction, it would be very helpful to take BS cooperation and multicast transmissions into account for practical networks. In addition, coded caching can be used to further enhance network performance.

APPENDIX

A. Analysis of Backhaul Latency

Based on (9) in [33], the mean packet delay in propagation via a wired backhaul network can be approximated as

$$T_{uc} = \left(1 + 1.28\frac{\lambda_b}{\lambda_g}\right) c_1 + c_2, \tag{26}$$

where λ_b denotes the BS density, λ_g is the gateway density, c_1 and c_2 are constants related to the processing capability of a backhaul node. In cache-enabled HCNs, the typical user has to retrieve using the backhaul network a file that is not cached at BSs. It follows from (26) that the resultant backhaul latency is given as

$$T_c = (1-\mathcal{P})\left(1 + 1.28\frac{(1-\mathcal{P})\lambda_b}{\lambda_g}\right) c_1 + c_2, \tag{27}$$

where $\mathcal{P}$ is the hit probability for cache-enabled HCNs. From (27), we can see that the backhaul latency is a monotone-decreasing function of the hit probability as shown in Fig. 7. This shows that improving the hit probability of the radio access network reduces the burden on the backhaul network.

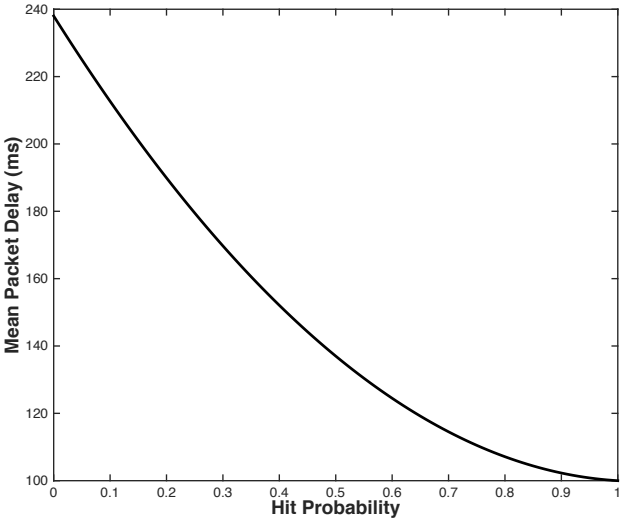


Figure 7: Mean packet delay in wired backhaul versus hit probability. Here, $\lambda_b = 10\lambda_g$, $c_1 = 10$ms, and $c_2 = 100$ms.

B. Proof of Lemma 1

Define $P_{\mathrm{r},m,k} = P_k R_{mk}^{-\alpha}$, which represents the received signal power due to transmissions by the BSs with file $\mathcal{F}_m$ in the k-th tier, where R_{mk} is the distance from the typical user to the nearest BS in content-centric tier Φ_{mk}. According to the content-centric cell association, the association probability A_{mk} is the probability that $P_{\mathrm{r},m,k} > P_{\mathrm{r},m,j}$, $\forall j$, $j \neq k$. Therefore,

$$\begin{aligned}
A_{mk} &= \mathbb{E}_{R_{mk}}\left[\mathbb{P}\left[P_{\mathrm{r},m,k}\left(R_{mk}\right) > \max_{j,j\neq k} P_{\mathrm{r},m,j}\right]\right] \\
&= \mathbb{E}_{R_{mk}}\left[\prod_{j=1,j\neq k}^{K} \mathbb{P}\left[P_{\mathrm{r},m,k}\left(R_{mk}\right) > P_{\mathrm{r},m,j}\right]\right] \\
&= \mathbb{E}_{R_{mk}}\left[\prod_{j=1,j\neq k}^{K} \mathbb{P}\left[R_{mj} > \left(P_j/P_k\right)^{1/\alpha} R_{mk}\right]\right] \\
&= \int_0^\infty \prod_{j=1,j\neq k}^{K} \mathbb{P}\left[R_{mj} > \left(P_j/P_k\right)^{1/\alpha} r\right] f_{R_{mk}}(r)\mathrm{d}r. \quad (28)
\end{aligned}$$

To derive A_{mk}, $\mathbb{P}\left[R_{mj} > \left(P_j/P_k\right)^{1/\alpha} r\right]$ and the probability density function (PDF) of R_{mk}, denoted by $f_{R_{mk}}(r)$, are calculated as follows.

$$\begin{aligned}
&\mathbb{P}\left[R_{mj} > \left(P_j/P_k\right)^{1/\alpha} r\right] \\
=&\mathbb{P}\left[\text{No BS with file } f_m \text{ closer than } \left(\left(P_j/P_k\right)^{1/\alpha} r\right) \text{ in the } j\text{th tier}\right] \\
=&\exp\left(-\pi p_{mj}\lambda_j \left(P_j/P_k\right)^{2/\alpha} r^2\right). \quad (29)
\end{aligned}$$

Further, $f_{R_{mk}}(r)$ is derived by taking the derivative of $1 - \mathbb{P}\left[R_{mk} > r\right]$ with respect to r,

$$f_{R_{mk}}\left(r\right) = \frac{\mathrm{d}\left[1 - \mathbb{P}\left[R_{mk} > r\right]\right]}{\mathrm{d}r} = 2\pi p_{mk}\lambda_k r \exp\left(-\pi p_{mk}\lambda_k r^2\right). \tag{30}$$

Last, the expression of A_{mk} is derived by substituting (29) and (30) into (28).

C. Proof of Lemma 3

In order to calculate the conditional hit probability, we first derive the probability that the typical user successfully receives

the requested file from its given serving BS X_k in the k-th tier, denoted by $\mathcal{P}_m(X_k,k)$, as follows.

$$\begin{aligned}
\mathcal{P}_m(X_k,k) &= \mathbb{P}\left[\frac{P_k h_{X_k}||X_k||^{-\alpha}}{I(\mathcal{F}_m)} > \beta_k\right] \\
&\overset{(a)}{=} \mathbb{E}_{I(\mathcal{F}_m)}\left[\exp\left(-\frac{\beta_k I(\mathcal{F}_m)||X_k||^{\alpha}}{P_k}\right)\right] \\
&\overset{(b)}{=} \prod_{i=1}^{K}\mathbb{E}_{\Phi_{mi}}\left\{\prod_{X\in\Phi_{mi}\setminus X_k}\mathbb{E}_{h_X}\left[\exp\left(-\frac{\beta_k P_i h_X||X_k||^{\alpha}}{P_k||X||^{\alpha}}\right)\right]\right\} \\
&\quad\cdot\prod_{i=1}^{K}\mathbb{E}_{\Phi^c_{mi}}\left\{\prod_{X\in\Phi^c_{mi}}\mathbb{E}_{h_X}\left[\exp\left(-\frac{\beta_k P_i h_X||X_k||^{\alpha}}{P_k||X||^{\alpha}}\right)\right]\right\} \\
&\overset{(c)}{=} \prod_{i=1}^{K}\mathbb{E}_{\Phi_{mi}}\left[\prod_{X\in\Phi_{mi}\setminus X_k}\left(1+\frac{\beta_k P_i||X_k||^{\alpha}}{P_k||X||^{\alpha}}\right)^{-1}\right] \\
&\quad\cdot\prod_{i=1}^{K}\mathbb{E}_{\Phi^c_{mi}}\left[\prod_{X\in\Phi^c_{mi}}\left(1+\frac{\beta_k P_i||X_k||^{\alpha}}{P_k||X||^{\alpha}}\right)^{-1}\right] \\
&\overset{(d)}{=} \prod_{i=1}^{K}\exp\left\{-\int_{\mathbb{R}^2\setminus \mathrm{b}(o,z_i)}\left[1-\left(1+\frac{\theta}{||X||^{\alpha}}\right)^{-1}\right]p_{mi}\lambda_i \mathrm{d}X\right\} \\
&\quad\cdot\prod_{i=1}^{K}\exp\left\{-\int_{\mathbb{R}^2}\left[1-\left(1+\frac{\theta}{||X||^{\alpha}}\right)^{-1}\right](1-p_{mi})\lambda_i \mathrm{d}X\right\} \\
&\overset{(e)}{=} \prod_{i=1}^{K}\exp\left\{-2\pi p_{mi}\lambda_i\int_{z_i}^{\infty}\left[1-\left(1+\frac{\theta}{r^{\alpha}}\right)^{-1}\right]r\mathrm{d}r\right\} \\
&\quad\cdot\prod_{i=1}^{K}\exp\left\{-2\pi(1-p_{mi})\lambda_i\int_{0}^{\infty}\left[1-\left(1+\frac{\theta}{r^{\alpha}}\right)^{-1}\right]r\mathrm{d}r\right\} \\
&= \prod_{i=1}^{K}\exp\left\{-2\pi p_{mi}\lambda_i\int_{z_i}^{\infty}\frac{r}{1+\frac{r^{\alpha}}{\theta}}\mathrm{d}r\right\} \\
&\quad\cdot\prod_{i=1}^{K}\exp\left\{-2\pi(1-p_{mi})\lambda_i\int_{0}^{\infty}\frac{r}{1+\frac{r^{\alpha}}{\theta}}\mathrm{d}r\right\} \\
&\overset{(f)}{=} \prod_{i=1}^{K}\exp\left[-\delta\pi p_{mi}\lambda_i\frac{\theta z_i^{\alpha(\delta-1)}}{1-\delta}{}_2F_1\left(1,1-\delta;2-\delta;-\frac{\theta}{z_i^{\alpha}}\right)\right] \\
&\quad\cdot\prod_{i=1}^{K}\exp\left[-\delta\pi(1-p_{mi})\lambda_i\theta^{\delta}B(\delta,1-\delta)\right] \\
&\overset{(g)}{=} \exp\left[-\sum_{i=1}^{K}\pi p_{mi}\lambda_i\left(\frac{P_i}{P_k}\right)^{\delta}Q\left(\beta_k\right)||X_k||^2\right] \\
&\quad\cdot\exp\left[-\sum_{i=1}^{K}V(\beta_k)\pi(1-p_{mi})\lambda_i\left(\frac{P_i}{P_k}\right)^{\delta}||X_k||^2\right], \qquad (31)
\end{aligned}$$

where (a) and (c) come from taking expectation with respect to $h\sim\exp(1)$; (b) follows from the expression of $I(\mathcal{F}_m)$ in (5); (d) follows from the probability generating functional of PPP, $\theta=\beta_k P_i||X_k||^{\alpha}/P_k$, and $\mathrm{b}(o,z_i)$ denotes a ball of radius z_i centered at the origin denoted by o. Note that the closest interferer in Φ_{mi} is at least at the distance $z_i=(P_i/P_k)^{1/\alpha}||X_k||$, while all the BSs in Φ^c_{mi} are the interferers. Equality (e) comes from converting from Cartesian to polar coordinates, (f) follows from replacing r^{α} with u and calculating the corresponding integral based on the formula (3.194.2) and (3.194.3) in [34], and finally (g) comes from the expressions of z_i, θ, $Q(\beta_k)$, and $V(\beta_k)$.

Averaging over the distance $||X_k||$, we have the hit probability for a user requesting for file $\mathcal{F}_m$ in the k-th tier:

$$\mathcal{P}_m(k)=\mathbb{E}_{||X_k||}\left[\mathcal{P}_m(X_k,k)\right]=\int_0^{\infty}\mathcal{P}_m(X_k,k)f_{||X_k||}(x)\mathrm{d}x. \qquad (32)$$

Last, the result of Lemma 3 is obtained by substituting (8) and (32) into (7) based on the law of total probability.

D. Proof of Proposition 2

Since Problem P2 is convex, it can be solved by using the Lagrange method. The corresponding partial Lagrangian function is

$$L(\mathbf{p},\mathbf{u})=\sum_{m=1}^{M}q_m\frac{\sum_{k=1}^{K}p_{mk}z_k}{W\sum_{k=1}^{K}p_{mk}z_k+V'}+\sum_{k=1}^{K}u_k(C_k-\sum_{m=1}^{M}p_{mk}), \qquad (33)$$

where $\mathbf{u}=(u_1,\ldots,u_K)\geq 0$ denotes the Lagrangian multiplier. Taking derivative of $L(\mathbf{p},\mathbf{u})$ with respect to p_{mk}, we have

$$\frac{\partial L}{\partial p_{mk}}=\frac{q_m V' z_k}{(V'+W\sum_{k=1}^{K}p_{mk}z_k)^2}-u_k. \qquad (34)$$

Thus, given the optimal Lagrangian multiplier $\mathbf{u}^*$, the optimal placement probability p^*_{mk} is expressed as

$$p^*_{mk}(u^*_k)=\begin{cases}1, & \text{if} \quad q_m\geq\frac{u^*_k(W\sum_{k=1}^{K}z_k+V')^2}{V'z_k},\\ \xi(u^*_k), & \text{if} \quad \frac{u^*_k V'}{z_k}<q_m<\frac{u^*_k(W\sum_{k=1}^{K}z_k+V')^2}{V'z_k},\\ 0, & \text{if} \quad q_m\leq\frac{u^*_k V'}{z_k},\end{cases} \qquad (35)$$

where $\xi(u^*_k)$ is the solution over p_{mk} of the equation

$$\frac{q_m V' z_k}{(V'+W\sum_{k=1}^{K}p^*_{mk}z_k)^2}=u^*_k. \qquad (36)$$

Thus, we have

$$\sum_{k=1}^{K}p^*_{mk}z_k=\left(\sqrt{q_m V' z_k/u^*_k}-V'\right)/W. \qquad (37)$$

Note that (37) holds for all k. Thus, $\xi(u^*_k)$ in Eq. (35) satisfies the following equation by denoting $\eta^*=u^*_k/z_k$

$$\sum_{k=1}^{K}p^*_{mk}z_k=\left(\sqrt{q_m V'/\eta^*}-V'\right)/W. \qquad (38)$$

According to the KKT conditions, the dual variable u_k satisfies the following equation:

$$u_k\left(C_k-\sum_{m=1}^{M}p^*_{mk}\right)=0. \qquad (39)$$

Thus, the alternative multiplier η^* satisfies

$$\eta^*\left(\sum_{k=1}^{K}C_k z_k-\sum_{m=1}^{M}\sum_{k=1}^{K}p^*_{mk}z_k\right)=0. \qquad (40)$$

If $\eta^* = 0$ ($u_k^* = 0$), according to Eq. (35), all files should be cached with probability 1 which conflicts with our assumption of limited cache capacity. Thus, we have

$$\sum_{k=1}^{K} C_k z_k = \sum_{m=1}^{M} \sum_{k=1}^{K} p_{mk}^* z_k. \tag{41}$$

According to Eq. (35), Eq. (38) and Eq. (41), we have $\sum_{k=1}^{K} p_{mk}^* z_k = g_m^*, m = 1, 2, \ldots, M$.

E. Proof of Theorem 3

In order to prove that p_{mk}^* given in Theorem 3 is the optimal placement probability for Problem P2, we need to follow the following two steps: (1) $\sum_{k=1}^{K} p_{mk}^* z_k = g_m^*$ where g_m^* is given in Proposition 1. (2) $\sum_{m=1}^{M} p_{mk}^* = C_k$. This is because the objective function of Problem P2 monotonically increases with the growing placement probability and thus the optimal content probabilities satisfy the relaxed constraint with equality.

(1) Proof $\sum_{k=1}^{K} p_{mk}^* z_k = g_m^*$:

To this end, we first prove $\sum_{k=1}^{K} \zeta_{mj}^* z_k = g_m^*$. When $q_m \in (T_0', T_1')$, we have $\sum_{i=m}^{M} \sum_{k=1}^{K} p_{ik}^* z_k = \sum_{i=m}^{M} g_i^*$ (since $g_m^* = \sum_{k=1}^{K} p_{ik}^* z_k$). On the other hand, $\sum_{i=m}^{M} \sum_{k=1}^{K} p_{ik}^* z_k = \sum_{k=1}^{K} \sum_{i=m}^{M} p_{ik}^* z_k = \sum_{k=1}^{K} C_k' z_k$. Thus, we have the following equation:

$$\sum_{i=m}^{M} g_i^* = \sum_{k=1}^{K} C_k' z_k. \tag{42}$$

Based on (42), we have

$$\sum_{k=1}^{K} \zeta_{mj}^* z_k = \sum_{k=1}^{K} \frac{c_k' g_m^* z_k}{\sum_{i=m}^{M} g_i^*} = \frac{g_m^* \sum_{k=1}^{K} c_k' z_k}{\sum_{i=m}^{M} g_i^*} = g_m^*. \tag{43}$$

Further, according to the expression of p_{mk}^* in (21), we have

$$\sum_{k=1}^{K} p_{mk}^* z_k = \sum_{k=1}^{K} \zeta_{mj}^* z_k = g_m^*. \tag{44}$$

Thus, $\sum_{k=1}^{K} p_{mk}^* z_k = g_m^*$.

(2) Proof $\sum_{m=1}^{M} p_{mk}^* = C_k$:

$$\begin{aligned} \sum_{m=1}^{M} p_{mk}^* &= \sum_{m=1}^{M-1} p_{mk}^* + p_{Mk}^* = \sum_{m=1}^{M-1} p_{mk}^* + C_k' \\ &= \sum_{m=1}^{M-1} p_{mk}^* + C_k - \sum_{m=1}^{M-1} p_{mk}^* = C_k. \end{aligned} \tag{45}$$

F. Proof of Time-sharing Condition

In order to prove the Problem (P0) in this paper can be solved by the dual methods for nonconvex optimization problem in [32], the following two steps are needed:

(1) Prove Problem (P0) has the same structure as (4) in [32]:

By observation, the optimization problem (P0) can be rewritten as follows:

$$\begin{aligned} \max \ & \sum_{m=1}^{M} f_m(\mathbf{P}_m) \\ \text{s.t.} \ & \sum_{m=1}^{M} h_m(\mathbf{P}_m) \leq \mathbf{C} \end{aligned} \tag{46}$$

where $\mathbf{P}_m = (p_{m1}, \cdots, p_{mk})$, $f_m(\mathbf{P}_m) = \sum_{k=1}^{K} q_m \frac{p_{mk} \lambda_k P_k^{\delta}}{W(\beta_k) \sum_{i=1}^{K} p_{mi} \lambda_i P_i^{\delta} + V(\beta_k) \sum_{i=1}^{K} \lambda_i P_i^{\delta}}$, $h_m(\mathbf{P}_m) = [p_{m1}, \cdots, p_{mk}]^T$ and $\mathbf{C} = [C_1, \cdots, C_K]^T$. Comparing Eq. (46) and (4) in [32], we find that they have the same structure.

(2) Prove Problem (P0) satisfies the time-sharing condition introduced in [32]:

Let $\mathbf{P}_{mx}^*$ and $\mathbf{P}_{my}^*$ be optimal solutions to optimization problem (P0) with the constraint $\mathbf{C} = \mathbf{C}_x$ and $\mathbf{C} = \mathbf{C}_y$, respectively. To prove Problem (P0) satisfies the time-sharing condition introduced in [32], we need to construct a feasible cache placement strategy $\mathbf{P}_{mz}$, such that the hit probability is at least $\upsilon \sum_{m=1}^{M} f_m(\mathbf{P}_{mx}^*) + (1-\upsilon) \sum_{m=1}^{M} f_m(\mathbf{P}_{my}^*)$ with cache capacity at most $\upsilon \mathbf{C}_x + (1-\upsilon)\mathbf{C}_y$ for all υ between zero and one.

In our system, each BS corresponds to a cache placement strategy. Thus, such a feasible cache placement strategy $\mathbf{P}_{mz}$ can be constructed by dividing the whole plane into two parts: the BSs in υ portion of which have $\mathbf{P}_{mz} = \mathbf{P}_{mx}^*$ and the BSs in $(1-\upsilon)$ portion of which have $\mathbf{P}_{mz} = \mathbf{P}_{my}^*$. Obviously, the resulting $\mathbf{P}_{mz}$ satisfies the cache capacity constraint $\upsilon \mathbf{C}_x + (1-\upsilon)\mathbf{C}_y$ and its hit probability achieves $\upsilon \sum_{m=1}^{M} f_m(\mathbf{P}_{mx}^*) + (1-\upsilon) \sum_{m=1}^{M} f_m(\mathbf{P}_{my}^*)$. Therefore, the cache placement optimization problem satisfies the time-sharing condition.

REFERENCES

[1] "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," Cisco, San Jose, CA, USA., Tech. Rep. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, Feb. 2016.

[3] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

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

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

[6] H. Sarkissian, "The business case for caching in 4G LTE networks," Wireless 2020, Tech. Rep., 2014.

[7] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 70–76, Aug. 2016.

[8] S. Traverso, M. Ahmed, and M. Garetto, "Temporal locality in today's content caching: Why it matters and how to model it," *ACM SIGCOMM*, vol. 43, no. 5, pp. 6–12, Oct. 2013.

[9] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[10] H. Dhillon, R. Ganti, F. Baccelli, and J. Andrews, "Modeling and analysis of K-tier downlink heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Apr. 2012.

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

[12] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[13] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6407–6422, Nov 2016.

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

[15] 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 Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

[16] B. Liu, H. Zhang, H. Ji, and X. Li, "A novel joint transmission and caching optimizing scheme in multi-relay networks: Video service quality assurance scheme," *Int. J. Commun. Syst.*, pp. 1–14, 2017.

[17] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, and V. W. S. Wong, "Cross-layer optimization of fast video delivery in cache-enabled relaying networks," in *2015 IEEE GLOBECOM*, Dec 2015, pp. 1–7.

[18] A. Liu and V. K. N. Lau, "Cache-enabled opportunistic cooperative mimo for video streaming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 390–402, Jan. 2014.

[19] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: Technical misconceptions and business barriers." [Online]. Available: http://arxiv.org/pdf/1602.00173v2.pdf

[20] S. Krishnan and H. Dhillon, "Effect of user mobility on the performance of device-to-device networks with distributed caching." [Online]. Available: https://arxiv.org/pdf/1604.07088v1.pdf

[21] M. Afshang and H. Dhillon, "Optimal geographic caching in finite wireless networks." [Online]. Available: https://arxiv.org/pdf/1603.01921v1.pdf

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

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

[24] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. of IEEE ICC*, London, 8-12 Jun. 2015.

[25] D. Malak and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks," in *IEEE Globecom Workshops*, Austin, TX, Dec. 2014, pp. 863–868.

[26] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, "Cooperative caching and transmission design in cluster-centric small cell networks." [Online]. Available: https://arxiv.org/pdf/1601.00321v1.pdf

[27] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–15, Oct. 2016.

[28] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2012.

[29] H. S. Dhillon and J. G. Andrews, "Downlink rate distribution in heterogeneous cellular networks under generalized cell selection," *IEEE Wireless Commun. Lett.*, vol. 3, no. 1, pp. 42–45, Feb. 2014.

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

[31] H.-S. Jo, Y. J. Sang, P. Xia, and J. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, Oct. 2012.

[32] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, Jul. 2006.

[33] G. Zhang, T. Q. S. Quek, M. Kountouris, A. Huang, and H. Shan, "Fundamentals of heterogeneous backhaul design-analysis and optimization," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 876–889, Feb. 2016.

[34] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.

Juan Wen received the B.Eng and the Ph. D. degree in telecommunications engineering at Xidian University, Xi'an, Shaanxi, China. Since Apr. 2016, she has been a postdoctoral researcher in the Dept. of Electrical and Electronics Engineering (EEE) at the University of Hong Kong. She was a visiting student at the University of Toronto from Sep. 2013 to Aug. 2014. Her research interests focus on the analysis and design of wireless networks using stochastic geometry. She was a recipient of the Best Paper Award at IEEE/CIC ICCC 2013.

Kaibin Huang (M'08-SM'13) received the B.Eng. (first-class hons.) and the M.Eng. from the National University of Singapore, respectively, and the Ph.D. degree from The University of Texas at Austin (UT Austin), all in electrical engineering.

Since Jan. 2014, he has been an assistant professor in the Dept. of Electrical and Electronic Engineering (EEE) at The University of Hong Kong. He used to be a faculty member in the Dept. of Applied Mathematics (AMA) at the Hong Kong Polytechnic University (PolyU) and the Dept. of EEE at Yonsei University. His research interests focus on the analysis and design of wireless networks using stochastic geometry and multi-antenna techniques.

He frequently serves on the technical program committees of major IEEE conferences in wireless communications. Most recently, he served as the lead chairs for the Wireless Comm. Symp. of IEEE Globecom 2017 and the Comm. Theory Symp. of IEEE GLOBECOM 2014 and the TPC Co-chairs for IEEE PIMRC 2017 and the IEEE CTW 2013. Currently, he is an editor for IEEE Transactions on Green Communications and Networking, and IEEE Transactions on Wireless Communications. He was an editor for IEEE Journal on Selected Areas in Communications (JSAC) series on Green Communications and Networking in 2015-2016, for IEEE Wireless Communications Letters in 2011-2016, and for IEEE/KICS Journal of Communication and Networks in 2009-2015. He has edited a JSAC 2015 special issue on communications powered by energy harvesting. Dr. Huang received the 2015 IEEE ComSoc Asia Pacific Outstanding Paper Award, Outstanding Teaching Award from Yonsei, Motorola Partnerships in Research Grant, the University Continuing Fellowship from UT Austin, and a Best Paper Award from IEEE GLOBECOM 2006 and PolyU AMA in 2013.

Sheng Yang (M'07) received the B.E. degree in electrical engineering from Jiaotong University, Shanghai, China, in 2001, and both the engineer degree and the M.Sc. degree in electrical engineering from Telecom ParisTech, Paris, France, in 2004, respectively. In 2007, he obtained his Ph.D. from Universit de Pierre et Marie Curie (Paris VI). From October 2007 to November 2008, he was with Motorola Research Center in Gif-sur-Yvette, France, as a senior staff research engineer. Since December 2008, he has joined CentraleSuplec where he is currently an associate professor. From April 2015, he also holds an honorary faculty position in the department of electrical and electronic engineering of the University of Hong Kong (HKU). He received the 2015 IEEE ComSoc Young Researcher Award for the Europe, Middle East, and Africa Region (EMEA). He is an editor of the IEEE transactions on wireless communications.

Victor O.K. Li (S'80-M'81-F'92) received SB, SM, EE and ScD degrees in Electrical Engineering and Computer Science from MIT in 1977, 1979, 1980, and 1981, respectively. He is Chair Professor of Information Engineering, Cheng Yu-Tung Professor in Sustainable Development, and Head of the Department of Electrical and Electronic Engineering at the University of Hong Kong (HKU). He has also served as Assoc. Dean of Engineering and Managing Director of Versitech Ltd., the technology transfer and commercial arm of HKU. He served on the board of China.com Ltd., and now serves on the board of Sunevision Holdings Ltd., listed on the Hong Kong Stock Exchange. Previously, he was Professor of Electrical Engineering at the University of Southern California (USC), Los Angeles, California, USA, and Director of the USC Communication Sciences Institute. His research is in the technologies and applications of information technology, including clean energy and environment, social networks, wireless networks, and optimization techniques. Sought by government, industry, and academic organizations, he has lectured and consulted extensively around the world. He has received numerous awards, including the PRC Ministry of Education Changjiang Chair Professorship at Tsinghua University, the UK Royal Academy of Engineering Senior Visiting Fellowship in Communications, the Croucher Foundation Senior Research Fellowship, and the Order of the Bronze Bauhinia Star, Government of the Hong Kong Special Administrative Region, China. He is a Registered Professional Engineer and a Fellow of the Hong Kong Academy of Engineering Sciences, the IEEE, the IAE, and the HKIE.