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 have 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 BS 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 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 threshold, 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.

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

S. Yang is with SUPELEC, Gif-sur-Yvette Cedex 91192, France (e-mail: sheng.yang@supelec.fr)
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]. Caching popular contents at the network edge (e.g., base stations (BSs) and dedicated access points) during off-peak hours has emerged as a promising solution, where highly skewed content popularity is exploited to alleviate the heavy burden on the backhaul networks and reduce latency in content delivery [4]–[6]. 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 [7]. Along the same theme, the current work considers caching for next-generation heterogeneous cellular networks (HCNs) 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 [8]. In [9], 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 are jointly optimized to minimize the communication overhead for content delivery. Coded caching in an erasure broadcast channel was then studied in [10] 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 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 [11], [12], 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 [11] and the theory of belief-propagation [12]. Recent advancements in wireless caching techniques have been summarized in various journal special issues and survey articles (see e.g., [13]).

It is also crucial to understand the performance gain that caching can bring 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 [14]–[21]. 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 aims 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 also can be wirelessly delivered to the user [18]. Based on the stochastic-geometry network models, the performance of cache-enabled D2D networks [14], [15] and HCNs [16], [17] 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 [20] 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 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 [20] 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 [18], [19], called placement probability. The strategy has also been considered in designing large-scale cache-enabled networks [18], [19]. 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 are optimized to maximize the hit probability for cellular networks in [18] and for D2D networks in [19]. 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 [21] 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. As the one in [20], the proposed strategy in [21] 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 [7]. 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 [22]. 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 receive 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 [22], the hit probability for cache-enabled HCNs are derived in closed form. The results reveal that the metric is determined by not only the physical-layer related parameters, including BS density, transmission power, and path-loss exponent, but also the content-related parameters, including content-popularity measures and placement probabilities. With uniform SIR threshold 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 to 1.

2) **Optimal Content Placement for Single-Tier HCNs.** We first consider the simple case of single-tier HCNs (e.g., the traditional cellular networks), which provides some useful insights for the general case treated subsequently. We derive the optimal placement probability for each file. Such probability is shown to have a structure that we name offset-popularity proportional (OPP) caching. Specifically, when the popularity measure is within a drive range, the optimal placement probability for a particular file is proportional to the square-root of the corresponding popularity measure, offset by a function of the SIR threshold, and scaled by a function of both the threshold and the cache capacity of each BS. Otherwise, the optimal placement probability is equal to one or zero depending on whether the measure is above or below the range.
3) **Optimal Content Placement for Multi-Tier HCNs.** For a multi-tier HCN, the placement probabilities form a $M \times K$ 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 on single-tier HCNs, a weighted sum (over tiers) of the placement probabilities for a particular file has the said structure of OPP caching. 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 insight 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 Section III and IV, respectively. Numerical results are provided in Section V followed by conclusions 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.

A. Network Topology

The spatial distributions of BSs and mobiles are modeled using the classic $K$-tier stochastic-geometry model for the HCN [22] described as follows. The network comprises of $K$ tiers of BSs modeled as $K$ independent homogeneous PPPs distributed in the plane. The $k$-th tier is denoted by $\Phi_k$ with the BS density and transmission power represented by $\lambda_k$ and $P_k$, respectively. All BSs and users are equipped with a single antenna. Assuming an interference-limited network, the transmission by a BS to an associated user is successful if the receive SIR exceeds a given threshold, denoted by $\beta_k$, identical for all links in the $k$-th tier.
Without loss of generality, consider a typical user located at the origin. The channel is narrow band and modeled in such a way that the signal power received at the user from a $k_0$-th tier BS located at $X_0 \in \mathbb{R}^2$ is given by $P_{k_0} h_{X_0} |X_0|^{-\alpha}$, where the random variable $h_{X_0} \sim \mathcal{CN}(0,1)$ models the Rayleigh fading and $\alpha > 2$ is the path-loss exponent. Based on the channel model, the interference power measured at the typical user, denoted by $I_0$, can be written as

$$I_0 = \sum_{k=1}^{K} \sum_{X \in \Phi_k \setminus X_0} P_k h_X |X|^{-\alpha},$$

(1)

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

**B. Probabilistic Content Placement**

We consider a content database containing the total of $M$ files all having unit sizes. As illustrated in Fig. 1, the BSs from different tiers are assumed to have different cache capacities that are denoted by $C_k$ for the $k$-th tier with $k = 1, 2, \ldots, K$. We make the practical assumption that not all BSs have sufficient capacities for storing the whole database. We adopt a probabilistic content placement scheme similar to the one in [18] to randomly select files for caching at different tiers under their cache capacity constraints. Specifically, the $m$-th file, denoted by $F_m$, is cached at a tier-$k$ BS with a fixed probability $p_{mk}$ called a *placement probability*. The binary caching decisions for the BSs in the same tier are i.i.d. The placement probabilities,
are identical for all BSs in the same tier \( k, k = 1, \ldots, K \). We call it \textit{TL content placement}. Grouping the placement probabilities yields the following \textit{placement probability matrix}:

\[
P = \begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1K} \\
p_{21} & p_{22} & \cdots & p_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
p_{M1} & p_{M2} & \cdots & p_{MK}
\end{bmatrix}.
\]

The rows columns of \( P \) correspond to different files and the different tiers.

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 \( \sum_{m=1}^{M} q_m = 1 \). This set is a probability mass function such that the typical user requests file \( F_m \) with probability \( q_m \). Without loss of generality, it is assumed that the files are ordered in decreasing popularity, i.e., \( q_1 > q_2 > \cdots > q_M \).

\[\text{C. Content-Centric Cell Association}\]

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., [21]). For ease of exposition, we partition the HCN into \( M \times K \) effective tiers, called the \textit{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 \( F_m \), denoted by \( \Phi_{mk} \), while the remaining tier-\( k \) BSs are denoted by \( \Phi_{mk}^c \) with \( \Phi_{mk} \cup \Phi_{mk}^c = \Phi_k \). Due to the probabilistic content placement scheme, \( \Phi_{mk} \) and \( \Phi_{mk}^c \) 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 \( F_m \) and is served by a tier-\( k \) BS. Then condition on that the typical user requests file \( F_m \), the serving BS \( X_0 \) is given by

\[
\text{(Cell Association)} \quad X_0 = \arg \max_{X \in \bigcup_k \Phi_{mk}} P_X |X|^{-\alpha},
\]

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

\[ I_0(\mathcal{F}_m) = \sum_{k=1}^{K} \sum_{X \in \Phi_{mk} \setminus X_0} P_k h_X |X|^{-\alpha} + \]
\[ \sum_{k=1}^{K} \sum_{X \in \Phi_{mk}^c} P_k h_X |X|^{-\alpha}. \]  

(4)

**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., [18]). The hit probability quantifies the caching gain, i.e., the reduced fraction of requests requiring transmissions over the backhaul network. For the purpose of analysis, let \( \mathcal{P}_m \) denote the conditional hit probability given that the typical user requests file \( \mathcal{F}_m \). Then

\[ \mathcal{P} = \sum_{m=1}^{M} q_m P_m. \]  

(5)

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

\[ \mathcal{P}_m = \sum_{k=1}^{K} A_{mk} \mathcal{P}_{mk}. \]  

(6)

**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 straightforwardly modifying Lemmas 3 and 5 in [23] 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

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

(7)

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

*Proof:* See Appendix A.
The result in Lemma 1 shows that a 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 transmit power, aligned with intuition. In addition, it is shown that the placement probability and BS density have more dominant effects on determining association probability than transmit power, since \( \delta < 1 \).

**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

\[
    f_{m,k}(x) = \frac{2\pi p_{mk} \lambda_k}{A_{mk}} x \exp \left( -\pi \sum_{j=1}^{K} p_{mj} \lambda_j \left( \frac{P_j}{P_k} \right)^{\delta} x^2 \right),
\]

where \( A_{mk} \) is given in (7).

Next, we are ready to derive the hit probabilities using Lemmas 1 and 2. For ease of notation, let us define the following two functions:

\[
    V(\beta_k) = \beta_k^\delta \frac{\delta \pi \csc(\delta \pi)}{1 - \delta},
\]

\[
    W(\beta_k) = 1 + \frac{\delta \beta_k}{1 - \delta} 2F_1[1, 1 - \delta; 2 - \delta; -\beta_k] - V(\beta_k),
\]

where \( 2F_1[\cdot] \) denotes the Gauss hypergeometric function. 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 \( F_m \) is given as

\[
    P_m = \sum_{k=1}^{K} \frac{p_{mk} \lambda_k P_k^\delta}{W(\beta_k) \sum_{i=1}^{K} p_{mi} \lambda_i P_i^\delta + V(\beta_k) \sum_{i=1}^{K} \lambda_i P_i^\delta},
\]

where the functions \( V(\cdot) \) and \( W(\cdot) \) are defined in (9) and (10), respectively.

**Proof**: See Appendix B.

Using Lemma 3 and the definition of hit probability in (5), we obtain the first main result of this paper.

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

\[
    P = \sum_{m=1}^{M} q_m \sum_{k=1}^{K} \frac{p_{mk} \lambda_k P_k^\delta}{W(\beta_k) \sum_{i=1}^{K} p_{mi} \lambda_i P_i^\delta + V(\beta_k) \sum_{i=1}^{K} \lambda_i P_i^\delta},
\]
where functions $V(\cdot)$ and $W(\cdot)$ are given in (9) and (10), 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 $\delta$; the other set contains content related parameters including the popularity measures $\{q_m\}$ and caching probabilities $\{p_{mk}\}$.

From Theorem 1, we can immediately 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 propositions.

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

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

where the functions $V(\cdot)$ and $W(\cdot)$ are given in (9) and (10).

Proposition 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.

**Proposition 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

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

where functions $V(\cdot)$ and $W(\cdot)$ are given in (9) and (10), respectively.

**Remark 1** (Effects of Large Cache Capacities). Proposition 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 transmit power, as the probabilities reach all 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., [22]).
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 in finding the placement matrix $P$ in (2) that maximizes the hit probability for HCNs as given in Theorem 1. Mathematically, the optimization problem can be formulated as follows:

$$\begin{align*}
\max_P & \quad \sum_{m=1}^{M} q_m \sum_{k=1}^{K} \frac{p_{mk} \lambda_k P_k^\delta}{W(\beta_k) \sum_{i=1}^{K} P_{mi} \lambda_i P_i^\delta + V(\beta_k) \sum_{i=1}^{K} \lambda_i P_i^\delta} \\
\text{s.t.} & \quad \sum_{m=1}^{M} p_{mk} \leq C_k, \forall k, \\
& \quad p_{mk} \in [0, 1], \forall m, k,
\end{align*}$$

(P0)

where the first constraint is on the BS cache capacity for each tier [18] and the second constraint arises for the fact that $p_{mk}$ is a probability. One can see that so long as the cache capacities $\{C_k\}$ are nonzero, there always exists a feasible non-trivial caching policy for the problem. Note that 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 insight 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 Proposition 1, Problem P0 is simplified as:

$$\begin{align*}
\max_P & \quad \sum_{m=1}^{M} q_m \frac{p_m}{W(\beta)p_m + V(\beta)} \\
\text{s.t.} & \quad \sum_{m=1}^{M} p_m \leq C, \\
& \quad p_m \in [0, 1], \forall m,
\end{align*}$$

(P1)

where $p_m$ denotes the placement probability for file $\mathcal{F}_m$, the vector $p = (p_1, p_2, \cdots, p_M)$, $C$ denotes the cache capacity for single-tier HCNs.
Problem P1 is convex since the objective function is concave and the constraints are linear and can thus be solved using the Lagrange method. The Lagrangian function can be written as

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

where \( u \geq 0 \) denotes the Lagrangian multiplier. Using the Karush-Kuhn-Tucker (KKT) condition, setting the derivative of \( L \) in (15) 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 results 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

\[ p_m^*(u^*) = \begin{cases} 1, & q_m \geq T_1, \\ \frac{\sqrt{V(\beta)} - V(\beta)}{\sqrt{u^*W(\beta)}} \sqrt{q_m}, & T_0 < q_m < T_1, \\ 0, & q_m \leq T_0, \end{cases} \]  

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

\[ \sum_{m=1}^{M} p_m^*(u^*) = C. \]  

Note that the optimal Lagrangian multiplier \( u^* \) in Theorem 2 can be found via a simple bisection search. 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) Caching structure. Specifically, if the popularity measure of a particular file is within the range \([T_0, T_1]\), the placement probability is proportional to the square root of the measure. Otherwise, the probability 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 \( C \).
Algorithm 1 Computing the Optimal Lagrangian Multiplier $u^*$ by a Bisection Search.

initialize $u^0 \in [u^{(0,\min)}, u^{(0,\max)}] = [\frac{qM}{(W+V)^2}, \frac{q}{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

Figure 2: Offset-Popularity Proportional Caching Structure.

Remark 3 (Content Dispensability, Diversity, and Densification). Intuitively, content placement should follow the greedy approach of caching the files at a BS one by one with decreasing popularity. The result in Theorem 2 shows that the greedy approach is sub-optimal. Furthermore, the optimal policy structure therein reveals that 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 need not 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 at different fractions of BSs.
Remark 4 (Fundamentality of the Optimal Policy Structure.). It is interesting to find that the optimal content placement policies derived in [18] and [21] have the similar threshold based structure as that in Theorem 2. Specifically, all optimal policies are shown to contain two thresholds that partition on the popularity measure into three ranges (see Remark 3), yielding corresponding placement probabilities of 0, (0,1) and 1. However, the policy expression in [18] has no closed form due to a relatively complex cell-association rule, i.e., the typical user is associated with an arbitrary BS storing the requested file among those that can support successful transmissions but not necessarily have the file. On the other hand, due to the complex hybrid caching strategy, neither does the policy in [21] has a closed form except for approximate analysis for the extreme cases of high signal-to-noise ratio (SNR) and high user density. Interestingly, the resultant placement probabilities are also found to be proportional to the square-root of corresponding popularity measures as the current ones. Despite the differences in content-placement strategies in the prior and current works, their similarity in the policy structure suggests its being fundamental for general cache-enabled networks.

Remark 5 (Effect of SIR threshold). The SIR threshold $\beta$ affects both the popularity threshold ($T_0$ and $T_1$) in the optimal placement policy (see Theorem 2). It can be observed from numerical results that both thresholds are monotone increasing functions of $\beta$.

Remark 6 (Effects of Lagrangian multiplier $u^*$). The value of Lagrangian multiplier affects the popularity threshold $T_1$ and $T_0$ 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 units 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.

C. Multi-tier HCNs with Uniform SIR Thresholds

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

\[
\max_{P} \quad \sum_{m=1}^{M} \frac{q_m \sum_{k=1}^{K} p_{mk} z_k}{W \sum_{k=1}^{K} p_{mk} z_k + V'}
\]

subject to

\[
\sum_{m=1}^{M} p_{mk} \leq C_k, \forall k,
\]

\[
p_{mk} \in [0, 1], \forall m, k,
\]

where \( z_k \) and \( V' \) are constants defined as \( z_k = \lambda_k P_\delta_k \) 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 reveal some properties of the policy structure.

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

\[
g_m = \sum_{k=1}^{K} p_{mk} z_k, \quad m = 1, 2, \ldots, M. \quad (18)
\]

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

\[
\max \{ g_m \} \sum_{m=1}^{M} \frac{q_m g_m}{W g_m + V'}
\]

subject to

\[
\sum_{m=1}^{M} g_m \leq \sum_{k} C_k z_k, \forall k,
\]

\[
0 \leq g_m \leq \sum_{k} z_k, \forall m.
\]

(P3)

Comparing Problem P3 with P1 for the single-tier HCN, 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 3** (Weighted Sum of Optimal Placement Probabilities). The weighted sum of the optimal placement probabilities for multi-tier HCNs with identical SIR threshold, denoted by \( g^*_m \), is given as:

\[
g^*_m(\eta^*) = \begin{cases} 
\sum_{k=1}^{K} \lambda_k P^0_k, & \text{if } q_m \geq T'_1, \\
(\sqrt{q_m V'}/\eta^* - V')/W, & \text{if } T'_0 < q_m < T'_1, \\
0, & \text{if } q_m \leq T'_0,
\end{cases}
\]  

(19)

where \( T'_1 = \eta^*(W' + V') \), \( T'_0 = \eta^* V' \), \( W' = W \sum_{k=1}^{K} z_i \) and the optimal Lagrange multiplier \( \eta^* \) satisfies the following equality

\[
\sum_{m=1}^{M} g^*_m(\eta^*) = \sum_{k=1}^{K} c_k z_k.
\]

The value of \( \eta^* \) 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 3.

The reason that Problem P3 is the relaxed version of P2 is that 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 4.** The solution for 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 C.

Next, based on the results in Propositions 3 and 4, 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 HCN with uniform SIR thresholds is given as follows:

\[
p^*_{mk} = \begin{cases} 
1, & \text{if } q_m \geq T'_1, \\
\min \left( \frac{1}{z_k} \sum_{j=1}^{k} c_{mj}^* z_j - \frac{1}{z_k} \sum_{j=1}^{k-1} p^*_{mj} z_j, 1 \right), & \text{if } T'_0 < q_m < T'_1, \\
0, & \text{if } q_m \leq T'_0,
\end{cases}
\]  

(20)
Algorithm 2 Sequential Computation of Optimal Placement Probabilities for Multi-tier HCNs with Non-Uniform SIR Thresholds.

1. Compute $\eta^*$ using Algorithm 1 and $\{g_m^*\}$ using Proposition 3.
2. For $m = 1 : M$
   
   for $k = 1 : K$
   
   set $p_{mk}^*$ according to (20)
   
   update $C_k' = C_k - p_{mk}^*$
   
   end

end

where

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

and $g_m^*$ is as given in Proposition 3.

Proof: See Appendix D.

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.

One can observe from Proposition 3 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 $\tilde{P}$ given as

$$\tilde{P} = \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{q_m p_{mk} z_k}{W(\beta_k) p_{mk} z_k + V(\beta_k) \sum_{i=1}^{K} z_i}$$

$$= \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{q_m p_{mk}}{W(\beta_k) p_{mk} + V(\beta_k)}$$

(22)

where $V(\beta_k) = V(\beta_k) \sum_{i=1}^{K} z_i$. Thus, $\tilde{P} = \sum_{k=1}^{K} \tilde{P}_k$ where $\{\tilde{P}_k\}$ are independent of each other.

As a result, maximizing $\tilde{P}$ is equivalent to separate maximization of individual summation terms $\{\tilde{P}_k\}$. As a result, Problem P0 can be approximated by $K$ single-tier optimization problems, each of which is written as:

$$\max_{p_k} \sum_{m=1}^{M} q_m \frac{p_{mk}}{W(\beta_k) p_{mk} + V(\beta_k)}$$

s.t. $\sum_{m=1}^{M} p_{mk} \leq C_k$,

$$p_{mk} \in [0, 1], \forall m.$$  (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 5** (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 $\{\bar{p}_{mk}^*\}$, that solve Problem P4 are given as

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

(23)

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

$$\sum_{m=1}^{M} p_{mk}^*(u_k^*) = C_k.$$  (24)

The above sub-optimal TLCP policy can attain close-to-optimal performance, as shown in the next section.
V. Simulation Results

In this section, simulation is conducted to validate the optimality of the content-placement policies derived in the preceding section and compare the performance of the strategy of TL content placement 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 [21]. 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 $\beta_2$ is a variable.

A. Conditional Hit Probability

Fig. 4 shows the conditional hit probability for a typical file $F_m$ versus caching probability $p_2$ under different SIR thresholds $\beta_2$. The analytical results are computed numerically using Lemma 3 and the simulated ones are obtained from Monto Carlo simulation using Matlab. It is observed that the simulated results match well the analytical results, which validates our analysis.
In addition, it is shown that 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 effect of placement probability on the hit probability differs 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) effect of placement probability on the hit probability is determined by the absolute values of the above increment and decrement.

In Fig. 5, the conditional hit probabilities varying with BS density are plotted for different cases accounting for different transmit power and placement probabilities. First of all, it is observed that the conditional hit probability is independent of the BS density and transmit power under the condition that $\beta_1 = \beta_2$ which is consistent with the well-known independent nature in HCNs (see e.g., [22]). Next, the conditional hit probability is shown to be a monotone-decreasing function of the growing BS density and transmit power if the placement probability is strictly smaller.
than 1. The reason is that the resulted increment in interference is greater than that in signal if $p < 1$, but is equal if $p = 1$.

**B. Optimal Content Placement**

Fig. 6 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, it is found that the hit probability under these three content placement policies increasing as the content popularity becomes more concentrate (a growing $\gamma$), aligned with intuition. Next, TOPP 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 $\gamma$ since MPCP is popularity-aware policy. In contrast, the gain over HCP increases with a growing $\gamma$. This is because, IN the HCP, only Macro-cell tier caches the $C_1$ most popular files. Last, the optimality of TLCP is verified by comparing the results given by the standard optimization tool CVX.
Figure 7: Hit probabilities under different content placement policies. Here, $\lambda_2 = 10\lambda_1$, $\beta_1 = -4$ dB, $\beta_2 = -2$ dB, $C_2 = 8$, $M = 20$.

The hit probability under the optimal CP, sub-optimal TLCP (see Proposition 5), MPCP, and HCP policies versus cache capacity is shown in Fig. 7. The optimal CP under this case is derived by adopting the dual methods for nonconvex optimization problem in [24] since Problem $P_0$ has the same structure as that in [24]. Compared with the optimal CP, the sub-optimal TLCP provides the close-to-optimal performance. 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.

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, transmit power, 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 same SIR threshold, 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 identical SIR threshold, interestingly, the weighted-sum of the optimal placement probability also has the OPP caching structure. Further, a optimal or sub-optimal TLCP caching algorithm has been proposed to maximize the hit probability HCNs with identical or general SIR threshold, respectively.

The fundamental structure of the optimal content placement strategies proposed in this paper provides useful guidelines and insight for designing the 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 the network performance.

**APPENDIX**

**A. Proof of Lemma 1**

Denote $P_{r,m,k} = P_k R_{mk}^{-\alpha}$ as the received power from the BSs with file $F_m$ in the $k$-th tier, where $R_{mk}$ is the distance from the typical user to the nearest BS in content-centric tier $\Phi_{mk}$. According to the content-centric cell association, the association probability $A_{mk}$ is the probability that $P_{r,m,k} > P_{r,m,j}$, $\forall j, j \neq k$. Therefore,

$$A_{mk} = \mathbb{E}_{R_{mk}} \left[ \mathbb{P} \left[ P_{r,m,k} (R_{mk}) > \max_{j \neq k} P_{r,m,j} \right] \right]$$

$$= \mathbb{E}_{R_{mk}} \left[ \prod_{j=1,j \neq k}^{K} \mathbb{P} [P_{r,m,k} (R_{mk}) > P_{r,m,j}] \right]$$

$$= \mathbb{E}_{R_{mk}} \left[ \prod_{j=1,j \neq k}^{K} \mathbb{P} [R_{mj} > (P_j/P_k)^{1/\alpha} R_{mk}] \right]$$

$$= \int_0^{\infty} \prod_{j=1,j \neq k}^{K} \mathbb{P} [R_{mj} > (P_j/P_k)^{1/\alpha} r] f_{R_{mk}}(r) dr. \quad (25)$$
To derive $A_{mk}$, $\mathbb{P}[R_{mj} > (P_j/P_k)^{1/\alpha} r]$ and the probability density function (PDF) of $R_{mk}$, denoted by $f_{R_{mk}}(r)$, are calculated as follows.

$$
P[R_{mj} > (P_j/P_k)^{1/\alpha} r] = \mathbb{P} [\text{No BS with file } f_m \text{ closer than } ((P_j/P_k)^{1/\alpha} r) \text{ in the } j\text{th tier}] = \exp \left( -\pi p_{mj} \lambda_j (P_j/P_k)^{2/\alpha} r^2 \right). \quad (26)
$$

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

$$
f_{R_{mk}}(r) = \frac{d [1 - \mathbb{P}[R_{mk} > r]]}{dr} = 2\pi p_{mk} \lambda_k r \exp \left( -\pi p_{mk} \lambda_k r^2 \right). \quad (27)
$$

Last, the expression of $A_{mk}$ is derived by substituting (26) and (27) into (25).

**B. 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 $X_0$ in the $k$-th tier, denoted by $P_m(X_0, k)$, as follows.

$$
P_m(X_0, k) = \mathbb{P} \left[ \frac{P_k h_{X_0} X_0^{\alpha}}{I_0(F_m)} > \beta_k \right] = \mathbb{E}_{I_0(F_m)} \left[ \exp \left( -\beta_k I_0(F_m) |X_0|^{\alpha} / P_k \right) \right] \quad (a)
$$

$$
= \prod_{i=1}^{K} \mathbb{E}_{\Phi_{mi}} \left\{ \prod_{X \in \Phi_{mi} \setminus X_0} \mathbb{E}_{h_X} \left[ \exp \left( -\beta_k P_i h_X |X_0|^{\alpha} / P_k |X|^{\alpha} \right) \right] \right\} \quad (b)
$$

$$
= \prod_{i=1}^{K} \mathbb{E}_{\Phi_{mi}} \left[ \prod_{X \in \Phi_{mi} \setminus X_0} \mathbb{E}_{h_X} \left[ \exp \left( -\beta_k P_i h_X |X_0|^{\alpha} / P_k |X|^{\alpha} \right) \right] \right] \quad (c)
$$

$$
= \prod_{i=1}^{K} \mathbb{E}_{\Phi_{mi}} \left[ \prod_{X \in \Phi_{mi} \setminus X_0} \left( 1 + \frac{\beta_k P_i |X_0|^{\alpha}}{P_k |X|^{\alpha}} \right)^{-1} \right] \quad (d)
$$

$$
= \prod_{i=1}^{K} \exp \left\{ -2\pi p_{mi} \lambda_i \int_{z_i}^\infty \frac{r}{1 + \frac{r^2}{\theta}} dr \right\} \quad (d)
$$

$$
= \prod_{i=1}^{K} \exp \left\{ -2\pi (1 - p_{mi}) \lambda_i \int_0^\infty \frac{r}{1 + \frac{r^2}{\theta}} dr \right\} \quad (d)
$$
\begin{equation}
\prod_{i=1}^{K} \exp \left[ -\delta \pi p_{mi} \lambda_i \frac{\theta z_i^{\alpha(\delta-1)}}{1-\delta} \right]
\cdot \prod_{i=1}^{K} \exp \left[ -\delta \pi (1-p_{mi}) \lambda_i \theta^\delta B(\delta, 1-\delta) \right]
\exp \left[ -\sum_{i=1}^{K} \pi p_{mi} \lambda_i \left( \frac{P_i}{P_k} \right)^\delta \right]
\cdot \exp \left[ -\sum_{i=1}^{K} V(\beta_k) \pi (1-p_{mi}) \lambda_i \left( \frac{P_i}{P_k} \right)^\delta \right], \tag{28}
\end{equation}

where (a) and (c) come from taking expectation with respect to $h \sim \exp(1)$; (b) follows from the expression of $I_0(F_m)$ in (4); (d) follows from the probability generating functional of PPP, converting from Cartesian to polar coordinates, and $\theta = \beta_k P_i |X_0|^\alpha / P_k$. Note that the first integration limits are from $z_i$ to $\infty$ since the closest interferer in $\Phi_{mi}$ is at least at the distance $z_i = (P_i / P_k)^{1/\alpha} |X_0|$, while the second integration limits are from 0 to $\infty$ because all the BSs in $\Phi_{mi}$ are the interferers. Equality (e) follows from replacing $r^\alpha$ with $u$ and calculating the corresponding integral based on the formula (3.194.2) and (3.194.3) in [25], and finally (f) comes from the expressions of $z_i$, $\theta$, $Q(\beta_k)$, and $V(\beta_k)$.

Averaging over the distance $|X_0|$, we have the hit probability for a user requesting for file $F_m$ in the $k$-th tier:

$$P_m(k) = \mathbb{E}_{X_0} [P_m(X_0, k)] = \int_0^\infty P_m(X_0, k) f_{X_0}(x) \, dx. \tag{29}$$

Last, the result of Lemma 3 is obtained by substituting (7) and (29) into (6) based on the law of total probability.

C. Proof of Proposition 4

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

$$L(p, u) = \sum_{m=1}^{M} q_m \sum_{k=1}^{K} p_{mk} z_k + V' + \sum_{k=1}^{K} u_k (C_k - \sum_{m=1}^{M} p_{mk}), \tag{30}$$

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

$$\frac{\partial L}{\partial p_{mk}} = \frac{q_m V' z_k}{(V' + W \sum_{k=1}^{K} p_{mk} z_k)^2} - u_k. \tag{31}$$
Thus, given the optimal Lagrangian multiplier $u^*$, the optimal placement probability $p^*_{mk}$ is expressed as

$$
p^*_{mk}(u^*_k) = \begin{cases} 
1, & \text{if } q_m \geq \frac{u^*_k(W \sum_{k=1}^{K} z_k + V')^2}{V' z_k}, \\
\xi(u^*_k), & \text{if } \frac{u^*_k V'}{z_k} < q_m < \frac{u^*_k(W \sum_{k=1}^{K} z_k + V')^2}{V' z_k}, \\
0, & \text{if } q_m \leq \frac{u^*_k V'}{z_k},
\end{cases} \quad (32)
$$

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

$$
\frac{q_m V' z_k}{(V' + W \sum_{k=1}^{K} p^*_{mk} z_k)^2} = u^*_k. \quad (33)
$$

Thus, we have

$$
\sum_{k=1}^{K} p^*_{mk} z_k = \left( \sqrt{q_m V' z_k / u^*_k} - V' \right) / W. \quad (34)
$$

Note that (34) holds for all $k$. Thus, $\xi(u^*_k)$ in Eq. (32) satisfies the following equation by denoting $\eta^* = u^*_k / z_k$

$$
\sum_{k=1}^{K} p^*_{mk} z_k = \left( \sqrt{q_m V' / \eta^*} - V' \right) / W. \quad (35)
$$

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

$$
u_k \left( C_k - \sum_{m=1}^{M} p^*_{mk} \right) = 0. \quad (36)
$$

Thus, the alternative multiplier $\eta^*$ satisfies

$$
\eta^* \left( \sum_{k=1}^{K} C_k z_k - \sum_{m=1}^{M} \sum_{k=1}^{K} p^*_{mk} z_k \right) = 0. \quad (37)
$$

If $\eta^* = 0$ ($u^*_k = 0$), according to Eq. (32), all files should be cached with probability 1 which is conflicted with our assumption of limited cache capacity. Thus, we have

$$
\sum_{k=1}^{K} C_k z_k = \sum_{m=1}^{M} \sum_{k=1}^{K} p^*_{mk} z_k. \quad (38)
$$

According to Eq. (32), Eq. (35) and Eq. (38), we have $\sum_{k=1}^{K} p^*_{mk} z_k = g^*_m, m = 1, 2, \ldots, M$. 


D. Proof of Theorem 3

In order to proof $p^*_{mk}$ 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 3. (2) $\sum_{m=1}^{M} p^*_{mk} = C_k$. This is because the objective function of Problem P2 monotone increases with the growing placement probability and thus the the optimal content probabilities satisfy the relaxed constraint with equation.

(1) Proof $\sum_{k=1}^{K} p^*_{mk} z_k = g^*_m$:

To this end, we first proof $\sum_{k=1}^{K} \zeta_{mj} z_k = g^*_m$. 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, $\sum_{i=m}^{M} \sum_{k=1}^{K} p^*_{ik} z_k = \sum_{i=m}^{K} \sum_{k=1}^{M} p^*_{ik} z_k = \sum_{k=1}^{K} C'_k z_k$. Thus, we have the following equation:

$$\sum_{i=m}^{M} g^*_i = \sum_{k=1}^{K} C'_k z_k. \quad (39)$$

Based on (39), we have

$$\sum_{k=1}^{K} \zeta_{mj} z_k = \sum_{k=1}^{K} \frac{C'_k g^*_m z_k}{\sum_{i=m}^{M} g^*_i} = g^*_m \sum_{k=1}^{K} \frac{C'_k z_k}{\sum_{i=m}^{M} g^*_i} = g^*_m. \quad (40)$$

Further, according to the expression of $p^*_{mk}$ in (20), we have

$$\sum_{k=1}^{K} p^*_{mk} z_k = \sum_{k=1}^{K} \zeta_{mj} z_k = g^*_m. \quad (41)$$

Thus, $\sum_{k=1}^{K} p^*_{mk} z_k = g^*_m$.

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

$$\sum_{m=1}^{M} p^*_{mk} = \sum_{m=1}^{M-1} p^*_{mk} + p^*_{MK} = \sum_{m=1}^{M-1} p^*_{mk} + C'_k = \sum_{m=1}^{M-1} p^*_{mk} + C_k - \sum_{m=1}^{M-1} p^*_{mk} = C_k. \quad (42)$$

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 backbone networks: challenges and research advances,” IEEE Netw., vol. 28, no. 6, pp. 6–11, Nov. 2014.

[4] 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.
[5] H. Sarkissian, “The business case for caching in 4G LTE networks,” Wireless 2020, Tech. Rep., 2014.
[6] 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.
[7] E. Bastug, M. Bennis, and M. Debbah, “Living on the edge: The role of proactive caching in 5G wireless networks,” IEEE Commun. Mag., vol. 52, no. 8, pp. 82–89, Aug. 2014.
[8] 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.
[9] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” IEEE Trans. Inf. Theory, vol. 60, no. 5, pp. 2856–2867, May 2014.
[10] 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.
[11] 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.
[12] 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.
[13] 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
[14] 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
[15] M. Afshang and H. Dhillon, “Optimal geographic caching in finite wireless networks.” [Online]. Available: https://arxiv.org/pdf/1603.01921v1.pdf
[16] 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.
[17] 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.
[18] B. Blaszczyszyn and A. Giovanidis, “Optimal geographic caching in cellular networks,” in Proc. of IEEE ICC, London, 8-12 Jun. 2015.
[19] 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.
[20] 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
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.