Joint and Competitive Caching Designs in Large-Scale Multi-Tier Wireless Multicasting Networks

Zitian Wang, Zhehan Cao, Ying Cui
Shanghai Jiao Tong University, China

Yang Yang
Intel Deutschland GmbH, Germany

Abstract— **Caching and multicasting are two promising methods to support massive content delivery in multi-tier wireless networks. In this paper, we consider a random caching and multicasting scheme with caching distributions in the two tiers as design parameters, to achieve efficient content dissemination in a two-tier large-scale cache-enabled wireless multicasting network. First, we derive tractable expressions for the successful transmission probabilities in the general region as well as the high SNR and high user density region, respectively, utilizing tools from stochastic geometry. Then, for the case of a single operator for the two tiers, we formulate the optimal joint caching design problem to maximize the successful transmission probability in the asymptotic region, which is nonconvex in general. By using the block successive approximate optimization technique, we develop an iterative algorithm, which is shown to coverage to a stationary point. Next, for the case of two different operators, one for each tier, we formulate the competitive caching design game where each tier maximizes its successful transmission probability in the asymptotic region. We show that the game has a unique Nash equilibrium (NE) and develop an iterative algorithm, which is shown to converge to the NE under a mild condition. Finally, by numerical simulations, we show that the proposed designs achieve significant gains over existing schemes.**

Index Terms— **Cache, multicast, multi-tier wireless network, stochastic geometry, optimization, game theory, Nash equilibrium**

I. Introduction

The rapid proliferation of smart mobile devices has triggered an unprecedented growth of the global mobile data traffic. Multi-tier wireless networks have been proposed as an effective way to meet the dramatic traffic growth by deploying different tiers of point of attachments (POAs), e.g., base stations (BSs) or access points (APs) together, to provide better time or frequency reuse. In general, there are two scenarios, depending on whether different tiers are managed by the same operator. One typical example for the scenario of the same operator is deploying short range small-BSs together with traditional macro-BSs, i.e., heterogeneous wireless networks (Hetnets). One typical example for the scenario of different operators is deploying IEEE 802.11 APs of different owners. To further reduce the load of the core network, caching at POAs in multi-tier wireless networks is recognized as a promising approach.

A lot of literature considers optimal caching design in large-scale cache-enabled Hetnets for the case of the same operator. For example, in [1] and [2], the authors focus on the maximization of the probability that the signal-to-interference plus noise ratio (SINR) of a typical user is above a threshold, which does not reflect the resource sharing among users. In a special case where all tiers have the same threshold, the problem is convex and the optimal solution is obtained [1], [2]. In the general case, the problem is nonconvex. [2] simplifies the nonconvex problem to a convex one and uses the optimal solution to the simplified convex problem as a sub-optimal solution to the original nonconvex problem. In addition, some works consider competitive caching design games among different POAs [3], [4]. For instance, in [3], the authors consider an Exact Potential Game among cache-enabled femto-BSs, prove the existence of Nash equilibrium (NE) and propose a convergent algorithm to obtain a NE. In [4], the authors consider a mean-field game among cache-enabled small-BSs, and obtain the unique mean field equilibrium.

On the other hand, enabling multicast service at POAs in multi-tier wireless networks is an efficient way to deliver popular contents to multiple requesters simultaneously by effectively utilizing the broadcast nature of the wireless medium. In our previous work [5], we consider analysis and optimization of a hybrid caching and multicasting design in a large-scale cache-enabled Hetnet. The hybrid design requires the files stored at macro-BSs and pico-BSs to be nonoverlapping and the files stored at all macro-BSs to be identical. Thus, the spatial file diversity provided by the hybrid caching design is limited, which may cause network performance degradation at some system parameters. Note that [1]–[4] do not consider multicasting.

In summary, further studies are required to facilitate the design of practical cache-enabled multi-tier wireless multicasting networks for massive content dissemination. In this paper, we consider a random caching and multicasting design with caching distributions in the two tiers as the design parameters to provide high spatial file diversity, utilizing tools from stochastic geometry. Our main contributions are summarized below. For the case of a single operator for the two tiers, we formulate the optimal joint caching design problem to maximize the successful transmission probability in the asymptotic region, which is a nonconvex problem in general. By using the block successive approximate optimization technique [8], we develop an iterative algorithm to obtain a stationary point.

Specifically, by carefully choosing an approximation function, we obtain the closed-form optimal solution to the approximate optimization problem in each iteration. For the case of two different operators, one for each tier, we formulate the competitive caching design game where each tier maximizes its successful transmission probability in the asymptotic region. We show that the game has a unique NE and develop an iterative algorithm to obtain the NE. We also provide a convergence condition for the iterative algorithm, which holds in most practical scenarios. Finally, by numerical simulations, we show that the proposed designs achieve significant gains over existing schemes.

II. System Model

A. Network Model and Performance Metric

We consider a general large-scale two-tier downlink network consisting of two tiers of POAs, e.g., BSs or APs, as shown in Fig. 1. The two tiers can be managed by a single operator (e.g., Hetnet with BSs being POAs) or by two different operators (e.g., IEEE 802.11 APs of two owners).[1] The locations of the POAs in tier 1 and tier 2 are spatially distributed as two independent homogeneous Poisson point processes (PPPs) Φ_1 and Φ_2 with densities λ_1 and λ_2, respectively. The locations of the users are also distributed as an independent homogeneous PPP Φ_u with density λ_u. Each POA in the jth tier has one transmit antenna with transmission power P_j, where $j = 1, 2$. For notational convenience, we define $\sigma_1 \triangleq \frac{P_1}{P_2}$ and $\sigma_2 \triangleq \frac{P_2}{P_1}$. Each user has one receive antenna. All POAs are operating on the same frequency band with a bandwidth W (Hz). Consider a discrete-time system with time being slotted and study one slot of the network. Both path loss and small-scale fading are considered: for path loss, a transmitted signal from any tier with distance D is attenuated by a factor $D^{-\alpha}$, where $\alpha > 2$ is the path loss exponent; for small-scale fading, Rayleigh fading channels are adopted [6].

Let $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$ denote the set of N files in the two-tier network. For ease of illustration, assume that all files have the same size. Each file is of certain popularity, which is assumed to be identical among all users. Each user randomly requests one file, which is file $n \in \mathcal{N}$ with probability $a_n \in (0, 1)$, where $\sum_{n \in \mathcal{N}} a_n = 1$. Thus, the file popularity distribution is given by $\mathbf{a} \triangleq (a_n)_{n \in \mathcal{N}}$, which is assumed to be known apriori. In addition, without loss of generality (w.l.o.g.), assume $a_1 > a_2 > \ldots > a_N$. The two-tier network consists of cache-enabled POAs. In the jth tier, each POA is equipped with a cache of size $K_j < N$ to store different popular files out of N. We say every K_j different files form a combination. Thus, there are in total $I_j \triangleq \binom{N}{K_j}$ different combinations, each with K_j different files. Let $\mathcal{I}_j \triangleq \{1, 2, \cdots, I_j\}$ denote the set of I_j combinations, and let $\mathcal{N}_{j,i}$ denote the set of K_j files contained in combination i of tier j.

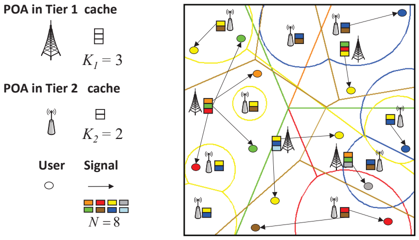


Fig. 1. Network model. Each file $n \in \mathcal{N}$ corresponds to a Voronoi tessellation (in the same color as the file), determined by the locations and transmission powers of all POAs storing this file.

B. Caching

To provide high spatial file diversity, we consider a random caching design where the caching distributions in the two tiers may be different, as illustrated in Fig. 1. The probability that combination $i \in \mathcal{I}_j$ is stored in each POA of tier j is $p_{j,i}$, where

$$0 \leq p_{j,i} \leq 1, \; i \in \mathcal{I}_j, \quad \sum_{i \in \mathcal{I}_j} p_{j,i} = 1. \tag{1}$$

A random caching design in tier j is specified by the caching distribution $\mathbf{p}_j \triangleq (p_{j,i})_{i \in \mathcal{I}_j}$. Let $\mathcal{I}_{j,n}$ denote the set of $I_{j,n} \triangleq \binom{N-1}{K_j-1}$ combinations containing file n. Let

$$T_{j,n} \triangleq \sum_{i \in \mathcal{I}_{j,n}} p_{j,i}, \;\; n \in \mathcal{N} \tag{2}$$

denote the probability that file n is stored at a POA in the jth tier. Therefore, the random caching design in the large-scale cache-enabled two-tier network is fully specified by the design parameters $(\mathbf{p}_1, \mathbf{p}_2)$.

C. Multicasting

Consider a user requesting file n. If file n is not stored in any tier, the user will not be served. Otherwise adopt the following user association rules: i) If file n is stored only in the jth tier, the user is associated with the nearest POA in the jth tier storing a combination $i \in \mathcal{I}_{j,n}$; ii) If file n is stored in both tiers, the user is associated with the POA which stores file n and provides the maximum long-term average received power (RP) (among all the POAs) [1], [2].

We consider multicasting in the large-scale cache-enabled two-tier network. Consider a POA schedules to serve requests for k different files. Then, it transmits each of the k files only once to concurrently serve users requesting the same file, at a rate τ (bit/second) and over $\frac{1}{k}$ of the total bandwidth W using frequency division multiple access (FDMA). As a matter of fact, both multicast and unicast may happen (with different probabilities). Without loss of generality, as in [5], we refer to this transmission as multicast. Note that, by avoiding transmitting the same file multiple times to multiple users, this content-centric multicast can improve the efficiency of the utilization of the wireless medium and reduce the load of

[1] The network model we considered in this paper is similar to that in [5]. But here, we consider a random caching design which is more general and includes the hybrid caching design in [5] as a special case. In addition, different from [1], [2], we specify the random caching design by the caching probabilities of file combinations, so as to investigate the file load distribution and the impact of multicasting.

$$\mathrm{SINR}_{n,0} = \frac{D_{j_0,\ell_0,0}^{-\alpha} \left|h_{j_0,\ell_0,0}\right|^2}{\sum_{\ell\in\Phi_{j_0}\setminus\ell_0} D_{j_0,\ell,0}^{-\alpha}\left|h_{j_0,\ell,0}\right|^2 + \sum_{\ell\in\Phi_{\bar{j}_0}} D_{\bar{j}_0,\ell,0}^{-\alpha}\left|h_{\bar{j}_0,\ell,0}\right|^2 \frac{P_{\bar{j}_0}}{P_{j_0}} + \frac{N_0}{P_{j_0}}}. \tag{3}$$

$$q_j(\mathbf{p}_j,\mathbf{p}_{\bar{j}}) = \sum_{n\in\mathcal{N}} a_n A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\bar{j}})\Pr\left[\frac{W}{K_{j,n,0}}\log_2\left(1+\mathrm{SINR}_{n,0}\right)\geq\tau \,\middle|\, j_0=j\right]. \tag{4}$$

$$\omega_{j,n,k}(\mathbf{p}_j,\mathbf{p}_{\bar{j}}) \triangleq \sum_{i\in\mathcal{I}_{j,n}} p_{j,i} \sum_{\mathcal{X}\in\left\{\mathcal{S}\subseteq\mathcal{N}_{j,i,-n}:|\mathcal{S}|=k-1\right\}} \prod_{m\in\mathcal{X}}\left(1-b_{j,m}\right) \prod_{m\in\mathcal{N}_{j,i,-n}\setminus\mathcal{X}} b_{j,m}. \tag{5}$$

$$f_{j,k}(x,y) \triangleq 2\pi\lambda_j\int_0^\infty d\exp\left(-\pi\lambda_j\left(\theta_{1,K_j}x+\theta_{2,j,K_j}y+\theta_{3,j,K_j}\right)d^2\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^\alpha\frac{N_0}{P_j}\right)\mathrm{d}d. \tag{6}$$

$$\theta_{1,k} =\frac{2}{\alpha}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha}}\left(B'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{-\frac{k\tau}{W}}\right)-B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right)+1. \tag{7}$$

$$\theta_{2,j,k} =\frac{2\lambda_{\bar{j}}}{\alpha\lambda_j}\left(\sigma_{\bar{j}}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha}}\left(B'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{-\frac{k\tau}{W}}\right)-B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right)+\frac{\lambda_{\bar{j}}}{\lambda_j}\sigma_{\bar{j}}^{\frac{2}{\alpha}}. \tag{8}$$

$$\theta_{3,j,k} =\frac{2}{\alpha}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)+\frac{2\lambda_{\bar{j}}}{\alpha\lambda_j}\left(\sigma_{\bar{j}}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right). \tag{9}$$

the wireless network, compared to the traditional connection-based unicast [6]. From the above illustration, we can see that the design parameters $(\mathbf{p}_1,\mathbf{p}_2)$ affect the performance of the random caching and multicasting design.

D. Performance Metric

In this paper, we study w.l.o.g. the performance of a typical user u_0, which is located at the origin. Suppose u_0 requests file n. Let j_0 denote the index of the tier with which u_0 is associated, and let $\bar{j}_0$ denote the other tier. Let $\ell_0 \in \Phi_{j_0}$ denote the index of the serving POA of u_0. We denote $D_{j,\ell,0}$ and $h_{j,\ell,0} \overset{d}{\sim} \mathcal{CN}(0,1)$ as the distance and the small-scale channel between POA $\ell \in \Phi_j$ and u_0, respectively. We assume the complex additive white Gaussian noise of power N_0 (evaluated over the entire frequency band) at u_0. When u_0 requests file n and file n is transmitted by POA ℓ_0, the SINR of u_0 is given by (3) [5] at the top of this page. When $T_{j,n} > 0$, let $K_{j,n,0} \in \{1,\cdots,K_j\}$ denote the number of different cached files requested by the users associated with POA $\ell_0 \in \Phi_j$. Note that $K_{j,n,0}$ is a discrete random variable, whose probability mass function (p.m.f.) depends on $\mathbf{a}$, λ_u and the design parameters $(\mathbf{p}_1,\mathbf{p}_2)$.

The file can be decoded correctly at u_0 if the channel capacity between POA ℓ_0 and u_0 is greater than or equal to τ. Requesters are mostly concerned about whether their desired files can be successfully received. Therefore, we adopt the probability that a randomly requested file by u_0 is successfully transmitted, referred to as the successful transmission probability, as the network performance metric [5]. Let $A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$ denote the probability that u_0 requesting file n is associated with tier j. By total probability theorem, the successful transmission probability under the considered scheme is

$$q\left(\mathbf{p}_1,\mathbf{p}_2\right) = q_1\left(\mathbf{p}_1,\mathbf{p}_2\right) + q_2\left(\mathbf{p}_2,\mathbf{p}_1\right), \tag{10}$$

where $q_j(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$ is given by (4) at the top of this page and represents the probability that a randomly requested file by u_0 is successfully transmitted from a POA in tier j, also referred to as the successful transmission probability of tier j.

III. Performance Analysis

A. Performance Analysis in General Region

In this subsection, we analyze the successful transmission probability $q(\mathbf{p}_1,\mathbf{p}_2)$ in the general region. The user association probability $A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$ and the cumulative distribution function (c.d.f.) of $\mathrm{SINR}_{n,0}$ can be found in [1], [2]. In this paper, we analyze the p.m.f. of file load $K_{j,n,0}$ by generalizing the method in [5], [6]. As in [5], [6], the dependence of the p.m.f. of $K_{j,n,0}$ and the c.d.f. of $\mathrm{SINR}_{n,0}$ is ignored. Then, based on $A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$, c.d.f. of $\mathrm{SINR}_{n,0}$ and p.m.f. of $K_{j,n,0}$, we can derive $q\left(\mathbf{p}_1,\mathbf{p}_2\right)$.[2]

Theorem 1 (Performance): The successful transmission probability is $q\left(\mathbf{p}_1,\mathbf{p}_2\right) = q_1\left(\mathbf{p}_1,\mathbf{p}_2\right) + q_2\left(\mathbf{p}_2,\mathbf{p}_1\right)$, where $q_j(\mathbf{p}_j,\mathbf{p}_{\bar{j}}) = \sum_{n\in\mathcal{N}} a_n \sum_{k=1}^{K_j} \omega_{j,n,k}(\mathbf{p}_j,\mathbf{p}_{\bar{j}}) f_{j,k}(T_{j,n},T_{\bar{j},n})$. Here, $\omega_{j,n,k}(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$ is given by (5) at the top of this page with $b_{j,m}$ given by

$$b_{j,m} \triangleq \left(1+\frac{a_m\lambda_u\widehat{A}_{j,m}(T_{j,m},T_{\bar{j},m})}{3.5\lambda_j}\right)^{-3.5}, \tag{11}$$

$\widehat{A}_{j,m}(T_{j,m},T_{\bar{j},m}) \triangleq \frac{\lambda_j}{\lambda_j T_{j,m}+\lambda_{\bar{j}}T_{\bar{j},m}\left(\frac{P_{\bar{j}}}{P_j}\right)^{\frac{2}{\alpha}}}$ and $\mathcal{N}_{j,i,-n} \triangleq \mathcal{N}_{j,i}\setminus\{n\}$, and $f_{j,k}(T_{j,n},T_{\bar{j},n})$ is given by (6) with $\theta_{1,k}$, $\theta_{2,j,k}$ and $\theta_{3,j,k}$ given by (7), (8) and (9) at the top of this page. $B^{'}\left(x,y,z\right) \triangleq \int_z^1 u^{x-1}\left(1-u\right)^{y-1}\mathrm{d}u$ and $B(x,y) \triangleq \int_0^1 u^{x-1}\left(1-u\right)^{y-1}\mathrm{d}u$ denote the complementary incomplete Beta function and the Beta function, respectively.

[2] Due to page limitation, the proofs are omitted. Please refer to [7] for the detailed proofs.

From Theorem 1, we can see that the impacts of the physical layer parameters α, W, λ_1, λ_2, λ_u, $\frac{P_1}{N_0}$, $\frac{P_2}{N_0}$ and the design parameters $(\mathbf{p}_1, \mathbf{p}_2)$ on $q(\mathbf{p}_1, \mathbf{p}_2)$ are coupled in a complex manner.

B. Performance Analysis in Asymptotic Region

Note that the gain of multicasting over unicasting increases with user density [5]. In this subsection, to obtain design insights into caching and multicasting, we analyze the asymptotic successful transmission probability in the high SNR and high user density region. Let $\frac{P_1}{N_0} \to \infty$ and $\frac{P_2}{N_0} \to \infty$ while fixing the power ratio, i.e., σ_1 (σ_2). In addition, when $\lambda_u \to \infty$, $K_{j,n,0} \to K_j$ in distribution. From Theorem 1, we have the following corollary.

Corollary 1 (Asymptotic Performance): When $\frac{P}{N_0} \to \infty$ and $\lambda_u \to \infty$,[3]

$$q(\mathbf{p}_1, \mathbf{p}_2) = q_{1,\infty}(\mathbf{T}_1, \mathbf{T}_2) + q_{2,\infty}(\mathbf{T}_2, \mathbf{T}_1) \triangleq q_\infty(\mathbf{T}_1, \mathbf{T}_2), \tag{12}$$

where

$$q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\bar{j}}) = \sum_{n \in \mathcal{N}} \frac{a_n T_{j,n}}{\theta_{1,K_j} T_{j,n} + \theta_{2,j,K_j} T_{\bar{j},n} + \theta_{3,j,K_j}}. \tag{13}$$

Here, $T_{j,n}$ is given by (2), and $\theta_{1,k}$, $\theta_{2,j,k}$ and $\theta_{3,j,k}$ are given by (7), (8) and (9).

Note that the asymptotic successful transmission probability in Corollary 1 and the performance metric in [1], [2] have different meanings, although they share similar forms. From Corollary 1, we can see that in the high SNR and high user density region, the impact of the physical layer parameters α, W, λ_j and σ_j, captured by $\theta_{1,j}$, θ_{2,j,K_j} and θ_{3,j,K_j}, and the impact of the design parameters $(\mathbf{p}_1, \mathbf{p}_2)$ on $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ can be easily separated. In most practical cases, $\theta_{1,K_1}, \theta_{1,K_2} > 0$. Thus, we consider $\theta_{1,K_1}, \theta_{1,K_2} > 0$ in the rest of the paper. Fig. 2 verifies Theorem 1 and Corollary 1, and demonstrates the accuracy of the approximation adopted. Fig. 2 also indicates that $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ provides a simple and good approximation for $q(\mathbf{p}_1, \mathbf{p}_2)$ in the high SNR (e.g., $\frac{P}{N_0} \geq$ 120 dB) and the high user density region (e.g., $\lambda_u \geq 3 \times 10^{-5}$).

In the asymptotic region, from [5], we know that the constraints on $(\mathbf{p}_1, \mathbf{p}_2)$ in (1) and (2) can be equivalently rewritten as $(\mathbf{T}_1, \mathbf{T}_2) \in \mathcal{T}_1 \times \mathcal{T}_2$, where $\mathcal{T}_j$ is defined as

$$\mathcal{T}_j \triangleq \left\{ \mathbf{T}_j \;\middle|\; 0 \leq T_{j,n} \leq 1, n \in \mathcal{N}, \sum_{n \in \mathcal{N}} T_{j,n} = K_j \right\}. \tag{14}$$

To obtain design insights into caching in large-scale multi-tier wireless multicasting networks, in Section IV and Section V, we focus on the joint and competitive caching designs in the asymptotic region, respectively.

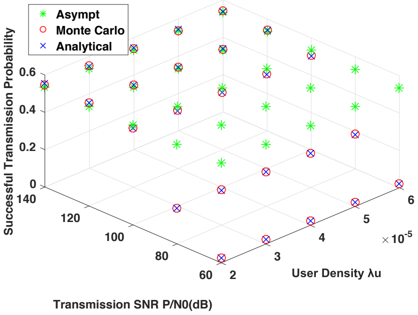


Fig. 2. Successful transmission probability versus SNR $\frac{P}{N_0}$ and user density λ_u. $N = 10$, $K_1 = 3$, $K_2 = 2$, $p_{1,i} = \frac{1}{\binom{10}{3}}$ for all $i = 1, 2, \cdots, \binom{10}{3}$, $p_{2,i} = \frac{1}{\binom{10}{2}}$ for all $i = 1, 2, \cdots, \binom{10}{2}$, $\lambda_1 = 5 \times 10^{-7}$, $\lambda_2 = 3 \times 10^{-6}$, $P_1 = 10^{1.5} P$, $P_2 = P$, $\alpha = 4$, $W = 20 \times 10^6$, $\tau = 35 \times 10^4$ and $a_n = \frac{n^{-\gamma}}{\sum_{n \in \mathcal{N}} n^{-\gamma}}$ with $\gamma = 1$.

IV. Joint Caching Design

In this section, we consider the case that the two tiers of POAs are managed by a single operator, e.g., as in a Hetnet. We first formulate the optimal joint caching design problem to maximize the successful transmission probability in the asymptotic region. Then, we develop an algorithm to obtain a stationary point.

A. Optimization Problem Formulation

In this subsection, we formulate the optimal joint caching design problem to maximize the successful transmission probability $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ by optimizing the caching distributions of the two tiers, i.e., $(\mathbf{T}_1, \mathbf{T}_2)$.

Problem 1 (Joint Caching Design):

$$\begin{aligned} q_\infty^* \triangleq \max_{\mathbf{T}_1, \mathbf{T}_2} &\quad q_\infty(\mathbf{T}_1, \mathbf{T}_2) \\ \text{s.t.} &\quad \mathbf{T}_j \in \mathcal{T}_j, \end{aligned} \tag{19}$$

where $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ is given by (12) and $\mathcal{T}_j$ is given by (14).

Problem 1 maximizes a differentiable (nonconcave in general) function over a convex set, and it is thus nonconvex in general. Note that Problem 1 and Problem 0 in [2] are mathematically equivalent, although this paper and [2] have different scopes. In the following subsection, we propose an efficient algorithm to solve Problem 1. In contrast, [2] simplifies the nonconvex problem to a convex one, and uses the optimal solution to the simplified problem as a sub-optimal solution to the original problem, which does not provide performance guarantee.

B. Algorithm Design

We can obtain a stationary point of Problem 1 using the gradient projection method (GPM) with a diminishing stepsize. However, the rate of convergence of GPM is strongly dependent on the choices of stepsize. If it is chosen improperly, it may take a large number of iterations for GPM to meet some convergence criterion. To address this problem, in this subsection we propose an iterative algorithm to obtain a stationary point of Problem 1 more efficiently. Note that a

[3] Note that when $\lambda_u \to \infty$ (corresponding to the full file load case), q_j and q become functions of $\mathbf{T}_1$ and $\mathbf{T}_2$ instead of $\mathbf{p}_1$ and $\mathbf{p}_2$.

$$g_j\left(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t)\right) \triangleq q_{j,\infty}\left(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t)\right) + q_{\bar{j},\infty}\left(\mathbf{T}_{\bar{j}}(t), \mathbf{T}_j(t)\right) + \sum_{n\in\mathcal{N}} \frac{\partial q_{\bar{j},\infty}\left(\mathbf{T}_{\bar{j}}(t), \mathbf{T}_j(t)\right)}{\partial T_{j,n}}\left(T_{j,n} - T_{j,n}(t)\right)$$

$$= q_{j,\infty}\left(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t)\right) + q_{\bar{j},\infty}\left(\mathbf{T}_{\bar{j}}(t), \mathbf{T}_j(t)\right) - \sum_{n\in\mathcal{N}} \frac{a_n \theta_{2,\bar{j},K_{\bar{j}}} T_{\bar{j},n}(t)\left(T_{j,n} - T_{j,n}(t)\right)}{\left(\theta_{1,K_{\bar{j}}} T_{\bar{j},n}(t) + \theta_{2,\bar{j},K_{\bar{j}}} T_{j,n}(t) + \theta_{3,\bar{j},K_{\bar{j}}}\right)^2}. \quad (15)$$

$$g_j\left(\mathbf{T}_j(t), \mathbf{T}_1(t), \mathbf{T}_2(t)\right) = \tilde{q}_\infty\left(\mathbf{T}_j(t), \mathbf{T}_{\bar{j}}(t)\right), \ \forall\left(\mathbf{T}_1(t), \mathbf{T}_2(t)\right) \in \mathcal{T}_1 \times \mathcal{T}_2. \quad (16)$$

$$g_j\left(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t)\right) \le \tilde{q}_\infty\left(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t)\right), \ \forall \mathbf{T}_j \in \mathcal{T}_j, \ \forall\left(\mathbf{T}_1(t), \mathbf{T}_2(t)\right) \in \mathcal{T}_1 \times \mathcal{T}_2. \quad (17)$$

$$g_j'\left(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t); \mathbf{d}_j\right)\Big|_{\mathbf{T}_j = \mathbf{T}_j(t)} = \tilde{q}_\infty'\left(\mathbf{T}_j, \mathbf{T}_{\bar{j}}; \mathbf{d}\right)\Big|_{\mathbf{T}_j = \mathbf{T}_j(t), \mathbf{T}_{\bar{j}} = \mathbf{T}_{\bar{j}}(t)}, \ \forall \mathbf{d}_j \text{ with } \mathbf{T}_j(t) + \mathbf{d}_j \in \mathcal{T}_j. \quad (18)$$

stationary point is a point that satisfies the necessary optimality conditions of a nonconvex optimization problem, and it is the classic goal in the design of iterative algorithms for nonconvex optimization problems. This algorithm is based on the block successive upper-bound minimization algorithm originally proposed in [8]. It alternatively updates $\mathbf{T}_1$ and $\mathbf{T}_2$ by maximizing an approximate function of $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$, which is successively refined so that eventually the iterative algorithm can converge to a stationary point of Problem 1. Specifically, at iteration t, we update the caching distribution of the jth tier by maximizing the approximate function of $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ given the caching distribution of the $\bar{j}$th tier, and fix the caching distribution of the $\bar{j}$th tier.

For notational convenience, we define

$$\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}) \triangleq \begin{cases} q_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}), \ j = 1, \\ q_\infty(\mathbf{T}_{\bar{j}}, \mathbf{T}_j), \ j = 2. \end{cases} \quad (20)$$

At iteration t, choose $g_j(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t))$ to be an approximate function of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t))$, where $g_j(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t))$ is given by (15) at the top of this page. Note that the first concave component function of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t))$, i.e., $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t))$ is left unchanged, and only the second nonconcave (actually convex) component function, i.e., $q_{\bar{j},\infty}(\mathbf{T}_{\bar{j}}(t), \mathbf{T}_j)$ is linearized at $\mathbf{T}_j = \mathbf{T}_j(t)$. This choice of approximate function is beneficial from several aspects. Firstly, it can guarantee the convergence of the algorithm to a stationary point of Problem 1, which will be seen in Theorem 2. Secondly, the partial concavity of the original objective function is preserved as much as possible, and the resulting algorithm typically converges much faster than GPM, where all component functions are linearized and no partial concavity is exploited. Thirdly, it yields a closed-form optimal solution to the optimization problem at each iteration, which will be explained in Lemma 1. Specifically, $g_j(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t))$ is strictly concave on $\mathcal{T}_j$ for any given $\left(\mathbf{T}_1(t), \mathbf{T}_2(t)\right) \in \mathcal{T}_1 \times \mathcal{T}_2$, and satisfies[4] (16), (17) and (18) (shown at the top of this page), where $\mathbf{d} = (\mathbf{d}_j, \mathbf{0})$ when $j = 1$, and $\mathbf{d} = (\mathbf{0}, \mathbf{d}_j)$ when $j = 2$. The conditions in (16) and (17) imply that $g_j(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t))$ is a tight lower bound of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t))$. The condition in (18) guarantees that the first order behavior of $g_j(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t))$ is the same as $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\bar{j}}(t))$ locally. At each iteration t, we update the caching distribution of the jth tier given the caching distribution of the $\bar{j}$th tier by solving the following problem, where $j = ((t+1) \bmod 2) + 1$, and fix the caching distribution of the $\bar{j}$th tier.

Algorithm 1 Stationary Point of Problem 1 Based on BSUM

```
1: Initialize $t = 1$ and choose any $\mathbf{T}_j(1) \in \mathcal{T}_j$ (e.g., $T_{j,n}(1) = \frac{K_j}{N}$ for all $n \in \mathcal{N}$), $j = 1, 2$.
2: Compute $j = ((t+1) \bmod 2) + 1$.
3: For all $n \in \mathcal{N}$, compute $T_{j,n}(t+1)$ according to Lemma 1.
4: For all $n \in \mathcal{N}$, set $T_{\bar{j},n}(t+1) = T_{\bar{j},n}(t)$.
5: Set $t = t + 1$ and go to Step 2.
```

Problem 2 (Optimization at Iteration t): For tier $j = ((t+1) \bmod 2) + 1$, we have

$$\mathbf{T}_j(t+1) = \underset{\mathbf{T}_j}{\operatorname{argmax}} \quad g_j\left(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$$
$$\text{s.t.} \quad \mathbf{T}_j \in \mathcal{T}_j,$$

where $g_j\left(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is given by (15).

Problem 2 is a convex optimization problem and Slater's condition is satisfied, implying that strong duality holds. Using KKT conditions, we can obtain the closed-form optimal solution to Problem 2.

Lemma 1 (Optimal Solution to Problem 2): The optimal solution to Problem 2 is given by (21) at the top of the next page, where $[x]^+ \triangleq \max\{x, 0\}$ and ν_j^* is the Lagrange multiplier that satisfies $\sum_{n\in\mathcal{N}} T_{j,n}(t+1) = K_j$.

Note that ν_j^* can be efficiently obtained by using bisection search. The details are summarized in Algorithm 1. Based on the conditions in (16), (17) and (18), we show the convergence of Algorithm 1.

Theorem 2 (Convergence of Algorithm 1): The sequence $\left\{q_\infty(\mathbf{T}_1(t), \mathbf{T}_2(t))\right\}$ generated by Algorithm 1 is convergent, and every limit point of $\left\{(\mathbf{T}_1(t), \mathbf{T}_2(t))\right\}$ is a stationary point of Problem 1.

Different from GPM, Algorithm 1 does not rely on a step-size. Thus, Algorithm 1 may have more robust convergence performance than GPM, as we shall illustrate later in Fig. 3.

[4] Note that the directional derivative of function $r : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is a convex set, at point $\mathbf{x}$ in direction $\mathbf{d}$ is defined by $r'(\mathbf{x}; \mathbf{d}) \triangleq \liminf_{\delta \downarrow 0} \frac{r(\mathbf{x}+\delta\mathbf{d}) - r(\mathbf{x})}{\delta}$. In addition, in (18), the directional derivative $g_j'(\mathbf{T}_j, \mathbf{T}_1(t), \mathbf{T}_2(t); \mathbf{d}_j)$ is only with respect to $\mathbf{T}_j$. (17) is hold since $q_{\bar{j},\infty}(\mathbf{T}_{\bar{j}}, \mathbf{T}_j)$ is a convex function of $\mathbf{T}_j$ for any given $\mathbf{T}_{\bar{j}} \in \mathcal{T}_{\bar{j}}$.

$$T_{j,n}(t+1)=\min\left\{\left[\frac{1}{\theta_{1,K_j}}\sqrt{\frac{a_n(\theta_{2,j,K_j}T_{\bar{j},n}(t)+\theta_{3,j,K_j})}{\nu_j^*+\frac{a_n\theta_{2,\bar{j},K_{\bar{j}}}T_{\bar{j},n}(t)}{\left(\theta_{1,K_{\bar{j}}}T_{\bar{j},n}(t)+\theta_{2,\bar{j},K_{\bar{j}}}T_{j,n}(t)+\theta_{3,\bar{j},K_{\bar{j}}}\right)^2}}}-\frac{\theta_{2,j,K_j}T_{\bar{j},n}(t)+\theta_{3,j,K_j}}{\theta_{1,K_j}}\right]^+,1\right\},\ n\in\mathcal{N}. \tag{21}$$

$$T_{j,n}^{\dagger}=\min\left\{\left[\frac{1}{\theta_{1,K_j}}\sqrt{\frac{a_n(\theta_{2,j,K_j}T_{\bar{j},n}^{\dagger}+\theta_{3,j,K_j})}{\nu_j^{\dagger}}}-\frac{\theta_{2,j,K_j}T_{\bar{j},n}^{\dagger}+\theta_{3,j,K_j}}{\theta_{1,K_j}}\right]^+,1\right\},\ n\in\mathcal{N},\ j=1,2. \tag{22}$$

V. Competitive Caching Design

In this section, we study the scenario that the two tiers of POAs are managed by two different operators, e.g., IEEE 802.11 APs of two owners. The two different operators have their own interests and thus cannot be jointly managed. Besides, one operator may be sacrificed in order to achieve the maximum total utility. Therefore, we propose a game theoretic approach and adopt the NE as a desirable outcome. We first formulate the competitive caching design for the two different operators within the framework of game theory. Then, we characterize a NE of the game and develop an algorithm to obtain a NE.

A. Game Formulation

In this subsection, we formulate the competitive caching design for two different operators within the framework of game theory. We consider a strategic noncooperative game, where the two operators are the players. The utility function of player j is the successful transmission probability for tier j, i.e., $q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})$. Each tier j competes against the other tier $\bar{j}$ by choosing its caching distribution $\mathbf{T}_j$ (i.e., strategy or action) in the set of admissible strategies $\mathcal{T}_j$ to maximize its utility function, i.e., $q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})$.

Problem 3 (Competitive Caching Game): For all $j=1,2$, we have

$$\begin{aligned}\max_{\mathbf{T}_j}\quad & q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})\\ \text{s.t.}\quad & \mathbf{T}_j\in\mathcal{T}_j,\end{aligned}$$

where $q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})$ is given by (13) and $\mathcal{T}_j$ is given by (14). Let $\mathcal{G}$ denote the game.

A solution, i.e., a NE, of game $\mathcal{G}$ is defined as follows.

Definition 1 (Nash Equilibrium of Game $\mathcal{G}$): A (pure) strategy profile $(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})\in\mathcal{T}_1\times\mathcal{T}_2$ is a NE of game $\mathcal{G}$ if

$$q_{j,\infty}(\mathbf{T}_j^{\dagger},\mathbf{T}_{\bar{j}}^{\dagger})\geq q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}}^{\dagger}),\ \forall\mathbf{T}_j\in\mathcal{T}_j,\ j=1,2. \tag{23}$$

B. Algorithm Design

In this subsection, we first characterize a NE of game $\mathcal{G}$.

Lemma 2 (NE of Game $\mathcal{G}$): Game $\mathcal{G}$ has a unique NE which is given by (22) at the top of this page, where $\nu_j^{\dagger}$ is the Lagrange multiplier that satisties $\sum_{n\in\mathcal{N}}T_{j,n}^{\dagger}=K_j$.

Then, we develop an iterative algorithm to obtain the NE of game $\mathcal{G}$. It alternatively updates $\mathbf{T}_1$ while $\mathbf{T}_2$ is fixed and $\mathbf{T}_2$ while $\mathbf{T}_1$ is fixed by solving the following problem at each iteration t.

Algorithm 2 Nash Equilibrium of Game $\mathcal{G}$

```
1: Initialize $t=1$ and choose any $\mathbf{T}_j(1)\in\mathcal{T}_j$ (e.g., $T_{j,n}(1)=\frac{K_j}{N}$ for all $n\in\mathcal{N}$), $j=1,2$.
2: Compute $j=((t+1)\bmod 2)+1$.
3: Compute $\mathbf{T}_j(t+1)=\underset{\mathbf{T}_j\in\mathcal{T}_j}{\operatorname{argmax}}\,q_{j,\infty}\left(\mathbf{T}_j,\mathbf{T}_{\bar{j}}(t)\right)$.
4: Set $\mathbf{T}_{\bar{j}}(t+1)=\mathbf{T}_{\bar{j}}(t)$.
5: Set $t=t+1$ and go to Step 2.
```

Problem 4 (Optimization at Iteration t): For player $j=((t+1)\bmod 2)+1$, we have

$$\begin{aligned}\mathbf{T}_j(t+1)=\underset{\mathbf{T}_j}{\operatorname{argmax}}\quad & q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}}(t))\\ \text{s.t.}\quad & \mathbf{T}_j\in\mathcal{T}_j.\end{aligned}$$

The optimal solution to Problem 4 has the same form as (22) except that $T_{j,n}^{\dagger}$ and $T_{\bar{j},n}^{\dagger}$ are replaced by $T_{j,n}(t+1)$ and $T_{\bar{j},n}(t)$, respectively. Based on the optimal solution to Problem 4, at iteration t, we update the strategy of player $j=((t+1)\bmod 2)+1$, and fix the strategy of player $\bar{j}$. The details for obtaining the NE of game $\mathcal{G}$ is summarized in Algorithm 2.

Next, we provide a convergence condition for Algorithm 2.

Theorem 3 (Convergence of Algorithm 2): If

$$\max\left\{1,\left\|1-\frac{\theta_{1,K_1}}{\theta_{3,1,K_1}}\right\|\right\}\max\left\{1,\left\|1-\frac{\theta_{1,K_2}}{\theta_{3,2,K_2}}\right\|\right\}<4,$$

where $\theta_{1,k}$, $\theta_{2,j,k}$ and $\theta_{3,j,k}$ are given by (7), (8) and (9), Algorithm 2 converges to the unique NE of game $\mathcal{G}$ for all $\mathbf{T}_j(1)\in\mathcal{T}_j$, $j=1,2$, i.e., $(\mathbf{T}_1(t),\mathbf{T}_2(t))\rightarrow(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})$ as $t\rightarrow\infty$, where $(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})$ is given by Lemma 2.

Note that the convergence condition given in Theorem 3 can be easily satisfied in most cases we are interested in, which will be shown in Fig. 3.

VI. Numerical Results

In the simulation, we choose $W=20\times10^6$, $\tau=4\times10^4$, $N=500$, $\alpha=4$, $\lambda_1=5\times10^{-7}$, $\lambda_2=3\times10^{-6}$, and $P_1-P_2=16$dB. We assume that the popularity follows Zipf distribution, i.e., $a_n=\frac{n^{-\gamma}}{\sum_{n\in\mathcal{N}}n^{-\gamma}}$, where γ is the Zipf exponent. First, we show the convergence and complexity

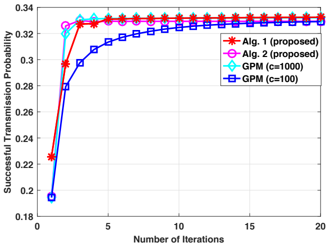


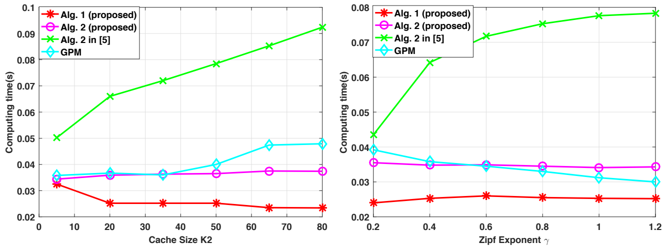
Fig. 3. Successful transmission probability versus number of iterations at $K_2 = 35$, $K_1 = 55$, stepsize for GPM is $\epsilon(t) = \frac{c}{2+t^{0.55}}$. we choose the same initial point for all the algorithms shown in Fig. 3

of the proposed algorithms. Fig. 3 illustrates the successful transmission probability versus the number of iterations. From Fig. 3, we can observe that the rate of convergence of GPM is strongly dependent on the choices of stepsize $\epsilon(t)$. In addition, Algorithm 1 and Algorithm 2 have more robust convergence performance than GPM, as they do not rely on a stepsize. Fig. 4 illustrates the computing time versus the cache size K_j and the Zipf exponent γ. From Fig. 4, we can observe that the computing times of all the algorithms do not change much with the K_j or γ, and the computing times of the proposed algorithms are shorter than that of Algorithm 2 in [5] which is to obtain an asymptotically optimal hybrid caching design. These observations demonstrate the advantage of the proposed algorithms.

Next, we compare the successful transmission probabilities of the proposed joint and competitive caching designs with those of three baselines. Baseline 1 (most popular) refers to the design in which each POA in tier j stores the K_j most popular files. Baseline 2 (i.i.d. file popularity) refers to the design in which each POA in tier j randomly stores K_j files, in an i.i.d. manner with file n being selected with probability a_n. Baseline 3 (hybrid caching) refers to the hybrid caching design obtained by Algrithm 2 in [5] (which is a feasible solution to Problem 1). The three baseline schemes also adopt the same multicasting scheme as in our design. Fig. 5 illustrates the successful transmission probability versus the cache size K_j and the Zipf exponent γ, respectively. From Fig. 5, we can observe that as K_j and γ increases, the successful transmission probability of each scheme increases. We can also observe that the two proposed designs outperform all the three baseline schemes. In addition, we can see that when K_j is large or γ is large, the two proposed designs reduce to the most popular caching design. When K_j is small or γ is small, the two proposed designs perform similarly as the hybrid caching design. These observations show that the two proposed designs can well adapt to the changes of the system parameters and can wisely utilize storage resources.

VII. Conclusion

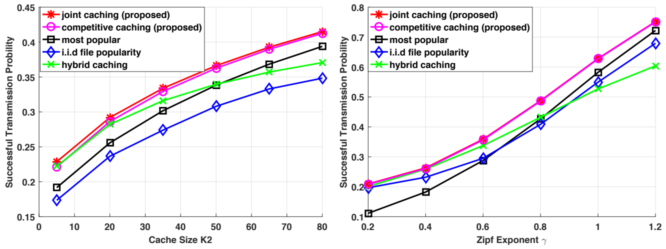
In this paper, we considered a random caching and multicasting scheme in a two-tier large-scale cache-enabled wireless multicasting network, operated by a single operator or two different operators. First, we formulated the optimal joint caching design problem in the asymptotic region. We develop an iterative algorithm, which is shown to coverage to a stationary point. Next, we formulated the competitive caching design game in the asymptotic region, obtained the unique NE of the game and developed an iterative algorithm, which is shown to converge to the NE under a mild condition. Finally, by numerical simulations, we showed that the two proposed designs achieve significant gains over existing schemes.



(a) cache size K_2 (K_1) at $\gamma = 0.55$, $K_1 = K_2 + 20$.

(b) Zipf exponent γ at $K_2 = 35$, $K_1 = 55$.

Fig. 4. Computing time versus the cache size or Zipf exponent γ at stepsize for GPM is $\epsilon(t) = \frac{c}{2+t^{0.55}}$. For GPM, each point corresponds to the minimum computing time by choosing the optimal parameter $c \in \{500, 1000, 1500, 2000, 2500\}$.



(a) cache size K_2 (K_1) at $\gamma = 0.55$, $K_1 = K_2 + 20$.

(b) Zipf exponent γ at $K_2 = 35$, $K_1 = 55$.

Fig. 5. Successful transmission probability versus the cache size or Zipf exponent γ.

References

[1] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in n-tier heterogeneous networks," *arXiv preprint arXiv:1612.04030*, 2016.

[2] J. Wen, K. Huang, S. Yang, and V. O. Li, "Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement," *arXiv preprint arXiv:1612.05506*, 2016.

[3] Y. Tan, Y. Yuan, T. Yang, Y. Xu, and B. Hu, "Femtocaching in wireless video networks: Distributed framework based on exact potential game," in *Communications in China (ICCC), 2016 IEEE/CIC International Conference on*. IEEE, 2016, pp. 1–6.

[4] H. Kim, J. Park, M. Bennis, S.-L. Kim, and M. Debbah, "Ultra-dense edge caching under spatio-temporal demand and network dynamics," *arXiv preprint arXiv:1703.01038*, 2017.

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

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

[7] Z. Wang, Z. Cao, Y. Cui, and Y. Yang, "Joint and competitive caching designs in large-scale multi-tier wireless multicasting networks," 2017. [Online]. Available: http://iwct.sjtu.edu.cn/personal/yingcui/publicans.html.

[8] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.