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

Ying Cui, Zitian Wang, Yang Yang, Feng Yang, Lianghui Ding, Liang Qian

***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 signal-to-noise ratio (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 converge 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 adopt 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.

Caching in cache-enabled multi-tier wireless networks for the case of the same operator is considered in many works. Cache-enabled multi-tier wireless networks with fixed topologies are considered in some of them. For example, in [1]–[3], the authors consider the optimal content placement at small-BSs to minimize the expected downloading time for files at the macro-BS in a single macro-cell with multiple small-cells. Note that [1]–[3] do not capture the stochastic natures of channel fading and geographic locations of POAs and users, and the obtained results in [1]–[3] may not be applied to real networks. To address these limitations, large-scale cache-enabled multi-tier wireless networks are considered in some other works, using tools from stochastic geometry. For example, in [4]–[6], the authors consider caching the most popular files at each small-BS in large-scale cache-enabled small-cell networks or HetNets. In [7], the authors propose a partition-based combined caching design in a large-scale cluster-centric small-cell network. In [8] and [9], the authors consider random caching of a uniform distribution at small-BSs in a large-scale cache-enabled HetNet and a large-scale cache-enabled small-cell network, respectively. In [10], each macro-BS caches the most popular files and each small-BS randomly caches popular files in a large-scale cache-enabled HetNet. Note that the focuses in [4]–[10] are only on performance analysis of some simple caching designs, which may not provide performance guarantee. In [11]–[14], the authors consider random caching and focus on the analysis and optimization of the probability that the signal-to-interference plus noise ratio (SINR) of a typical user is above a threshold, in a large-scale cache-enabled multi-tier wireless network. In [11], the authors consider two architectures (i.e., an always-on architecture and a dynamic on-off architecture), and formulate the optimization problem for each architecture, which is convex. For each problem, the closed-form optimal solution is obtained. In [12], the authors consider two cooperative transmission schemes, and formulate the optimization problem under each scheme, which is nonconvex in the general case. For each problem, a stationary point is obtained using the standard gradient projection method. For [13], [14], in a special case where all tiers have the

Manuscript received June 23, 2017; revised December 3, 2017; accepted January 15, 2018. The work of Y. Cui was supported by National Science Foundation of China under Grant 61401272 and Grant 61521062 as well as Shanghai Key Laboratory Funding STCSM15DZ2270400. The work of Y. Yang is supported by the ERC project AGNOSTIC. This paper was presented in part at IEEE GLOBECOM 2017. The associate editor coordinating the review of this paper and approving it for publication was Z. Dawy. (Corresponding author: Ying Cui.) Y. Cui, Z. Wang, F. Yang, L. Ding and L. Qian are with the Department of Electronic Engineering, Shanghai Jiao Tong University, China (e-mail: cuiying@sjtu.edu.cn). Yang is with the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, L-1855 Luxembourg.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

same threshold, the optimization problem is convex and the optimal solution is obtained; in the general case, the problem is nonconvex. In [14], the nonconvex problem is simplified to a convex one and the optimal solution to the simplified convex problem is used as a sub-optimal solution to the original nonconvex problem. In [15], the authors propose a random caching design, and focus on the maximization of the cache hit probability. The optimization problem is convex and the closed-form optimal solution is obtained. Note that [11]–[15] focus only on the typical user and do not consider the resource sharing among multiple users.

Some works consider competitive caching among different POAs that would like to maximize their own interests, using game theory. For instance, in [16], the authors consider an Exact Potential Game among cache-enabled femto-BSs where each femto-BS maximizes the expected number of its served users, prove the existence of Nash equilibrium (NE) and propose a convergent algorithm to obtain an NE. In [17], [18], the authors consider Stackelberg games among content providers and network operators. Specifically, the content providers rent part of the network resources from the network operators for content delivery to get payment from users. Note that in [16]–[18], the authors consider cache-enabled wireless networks with fixed topologies, and the obtained results in [16]–[18] may not be applied to real networks. [19], [20] consider large-scale cache-enabled wireless networks. In particular, in [19], the authors consider a mean-field game among cache-enabled small-BSs where each small-BS minimizes its long run average cost in a large-scale cache-enabled single-tier wireless network, and obtain the unique mean field equilibrium. In [20], the authors consider a Stackelberg game among content providers and network operators in a large-scale cache-enabled multi-tier wireless network, but do not provide a convergent algorithm to find the Stackelberg 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 [21], we consider analysis and optimization of a hybrid caching design and a corresponding multicasting design in a large-scale cache-enabled HetNet. The hybrid caching 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. In our previous work [22], we consider analysis and optimization of a random caching design and a corresponding multicasting design in a large-scale cache-enabled single-tier network. The proposed random caching design in [22] can offer high spatial file diversity, ensuring good network performance over a wide range of system parameters, but can not be directly applied to HetNets.

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 study a large-scale two-tier wireless network capturing the stochastic nature of channel fading and the stochastic nature of geographic locations of POAs and users. 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. We derive tractable expressions for the successful transmission probabilities in the general region as well as the high signal-to-noise ratio (SNR) and high user density region (i.e., the asymptotic region), respectively, 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 [23], 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. In addition, in the special case of the same cache size, we develop a low-complexity algorithm to obtain a globally optimal solution by extending the method in [14].
- 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 adopt an iterative algorithm to obtain the NE. In general, it is quite difficult to guarantee that an iterative algorithm can converge to an NE of a game, especially for a large-scale wireless network. By carefully analyzing structural properties of the competitive caching design game, we provide a convergence condition for the considered iterative algorithm, which holds in most practical scenarios.
- Finally, by numerical simulations, we show that the proposed designs achieve significant gains over existing schemes in terms of the successful transmission probability and complexity. We also show the caching probabilities of the proposed designs, revealing that the proposed designs offer high spatial file diversity.

II. System Model

A. Network Model

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., macro and small BSs in a HetNet) or by two different operators (e.g., IEEE 802.11 APs of two owners).[1] The locations of

[1]The network model considered in this paper is similar to that in [21]. But here, we consider a random caching design which is more general and includes the hybrid caching design in [21] as a special case. In addition, different from [13], [14], 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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

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[2] frequency band with a bandwidth W (Hz). For example, in HetNets under LTE, macro BSs and small BSs owned by the same operator use the same spectrum. In addition, LTE in unlicensed spectrum (LTE-Unlicensed) allows a cellular network operator to offload some data traffic by accessing the unlicensed spectrum, such as the 5 GHz band used by IEEE 802.11 compliant Wi-Fi equipment. That is, BSs and Wi-Fi APs of different operators can operate on the same spectrum. In IEEE 802.11, Wi-Fi APs of different owners can also operate on the same spectrum. 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 either tier with distance D is attenuated by a factor $D^{-\alpha}$, where $\alpha > 2$ is the path loss exponent [13], [14]; for small-scale fading, Rayleigh fading channels are adopted [24].

Let $\mathcal{N} \triangleq \{1, 2, \cdots, N\}$ denote the set of N files in the two-tier network. For ease of illustration, as in [1], [2], [4]–[7], assume that all files have the same size.[3] In addition, for ease of illustration, as in [1], [2], [4]–[7], file popularity is assumed to be identical among all users.[4] We assume all users make file requests at the same time. This assumption can be relaxed when considering temporal file request aggregation and serving accumulated asynchronous file requests simultaneously at the cost of delay. 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 [13], [14]. Note that file popularity evolves at a slower timescale and learning methodologies can be employed to track the evolution of file popularity over time. 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. Combination $i \in \mathcal{I}_j$ can be characterized by an N-dimensional vector $\mathbf{x}_{j,i} \triangleq (x_{j,i,n})_{n\in\mathcal{N}}$, where $x_{j,i,n} = 1$ if file n is included in combination i of tier j and $x_{j,i,n} = 0$ otherwise. Note that there are K_j 1's in each $\mathbf{x}_{j,i}$. Denote $\mathcal{N}_{j,i} \triangleq \{n \in \mathcal{N} : x_{j,i,n} = 1\}$ as the set of K_j files contained in combination i of tier j. To further explain the notations, consider the following example. When $N = 3$ and $K_j = 2$, we have $I_j = \binom{3}{2} = 3$, $\mathcal{I}_j = \{1, 2, 3\}$, $\mathcal{N}_{j,1} = \{1, 2\}$, $\mathcal{N}_{j,2} = \{1, 3\}$ and $\mathcal{N}_{j,3} = \{2, 3\}$.

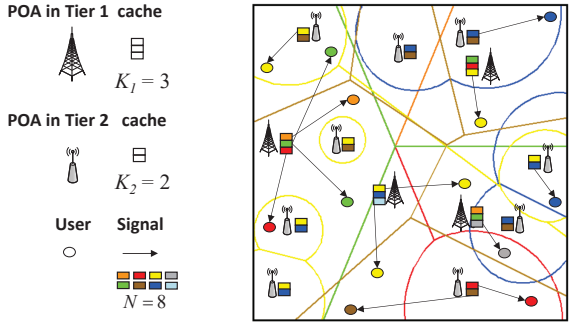


Fig. 1. Network model. Each file $n \in \mathcal{N}$ corresponds to a Voronoi tessellation (in the same color as the file), determined by the locations and transmission powers of all POAs storing this file.

B. Caching

To provide high spatial file diversity, we consider a random caching design in the cache-enabled two-tier network 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 $p_{j,i}$ satisfies

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

A random caching design in the tier j is specified by the caching distribution $\mathbf{p}_j \triangleq (p_{j,i})_{i\in\mathcal{I}_j}$. Let $\mathcal{I}_{j,n} \triangleq \{i \in \mathcal{I}_j : x_{j,i,n} = 1\}$ 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 $(\mathbf{p}_1, \mathbf{p}_2)$. In this paper, we focus on serving cached files at POAs to get first-order insights into the design of cache-enabled wireless networks, as in [13], [14], [21], [22], [25]. POAs may serve uncached files through other service machanisms, the investigation of which is beyond the scope of this paper.

Remark 1: Note that the random caching design considered in this paper is a generalization of the caching design caching the most popular files at each POA and the hybrid caching design proposed in [21]. In particular, by choosing the design parameters $(\mathbf{p}_1, \mathbf{p}_2)$ such that $T_{j,n} = 1$ for all $n = 1, 2, \cdots, K_j$

[2]POAs that do not operate on the same spectrum will not cause interference to each other, and can be considered separately.

[3]The results in this paper can be easily extended to the case of different file sizes by considering file combinations of the same total size, but formed by files of possibly different sizes.

[4]Consider R different file popularity distributions $\mathbf{a}_r = (a_{n,r})_{n\in\mathcal{N}}, r = 1, \cdots, R$ and each user randomly requests a file according to $\mathbf{a}_r$ with probability π_r, where $\sum_{r=1}^{R} \pi_r = 1$. Define $\bar{a}_n = \sum_{r=1}^{R} a_{n,r}\pi_r$ as the average popularity for file n. Later, we shall see that the results derived in this paper can be applied to the case of the nonidentical file popularity distributions, by replacing a_n with $\bar{a}_n$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

and $T_{j,n} = 0$ for all $n = K_j + 1, K_j + 2, \cdots, N$, where $j = 1, 2$, the proposed random caching design turns to the design caching the most popular files [7], [9]. In addition, by choosing the design paprameters $(\mathbf{p}_1, \mathbf{p}_2)$ in a certain manner, the proposed random caching design can reflect identical caching in the 1st tier, random caching in the 2nd tier and nonoverlapping caching across the two tiers, and hence incorporate the hybrid caching design in [21] as a special case. Therefore, by carefully designing $(\mathbf{p}_1, \mathbf{p}_2)$, the proposed random caching design can achieve better performance than the two designs. Later, we shall see the advantage of the proposed random design in Section VI.

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 rule: i) if file n is stored only in the jth tier, the user is associated with the nearest POA in the jth tier storing file 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 [13], [14]. As in [13], [14], [21], [22], we assume that the user association can be done through some signaling mechanisms.

Remark 2: Note that the content-centric user association considered in this paper is a generalization of the content-centric user association in [21]. In particular, Case ii) is not included in [21] due to the nonoverlapping caching constraint in [21].

We consider multicasting in the large-scale cache-enabled two-tier network. Consider a POA serving requests for k different files. Then, it transmits each of the k files only once to concurrently serve users requesting the same file at the same time, at a rate τ (bit/second) and over $\frac{1}{k}$ of the total bandwidth W using frequency division multiple access (FDMA).[5] As a matter of fact, both multicast and unicast may happen (with different probabilities). Without loss of generality, as in [21], 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 the wireless network, compared to the traditional connection-based unicast [24]. From the above illustration, we can see that the proposed multicasting design is also affected by the proposed caching design. Therefore, 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 $\overline{j}_0$ 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. For analytical tractability, as in [22] and [21], we assume all POAs are active for serving their own users, which corresponds to the worst-case interference strength for the typical user. In the general region, the performance obtained under this assumption provides a lower bound on the performance of the practical network where some void POAs may be shut down. In the high user density region, the performance obtained under this assumption is exact. Later, we will focus on joint and competitive caching designs in the high user density region. When u_0 requests file n and file n is transmitted by POA ℓ_0, the SINR of u_0 is given by (3). Note that, as in [7], [21], the transmitted symbols of file n from POA ℓ_0 are treated as the desired signal, while the transmitted symbols of file n from other POAs are regarded as interference. This is because the received signals from all the POAs transmitting file n may not be perfectly synchronized due to the large difference in distances from these POAs to u_0 [26]. In addition, note that we can use the signal and noise power over the whole frequency band in calculating $\mathrm{SINR}_{n,0}$ [21]. This is because under equal power allocation, the bandwidth for serving u_0 affects the signal, interference and noise power experienced at u_0 in the same manner and can be cancelled from the numerator and denominator in (3).

When $T_{j,n} > 0$ (i.e., u_0 may be associated with tier j), 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 BS 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 [21]. Let $A_{j,n}(\mathbf{p}_j, \mathbf{p}_{\overline{j}})$ 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 given by (4), where $q_j(\mathbf{p}_j, \mathbf{p}_{\overline{j}})$ 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.[6]

III. Performance Analysis

In this section, we first analyze the successful transmission probability in the general region. Then, we analyze the suc-

[5]Later, we can see that transmitting each of the k files within $\frac{1}{k}$ of a slot at rate $k\tau$, over the whole frequency band using time division multiple access (TDMA) will lead to the same successful transmission probability. In addition, allocating both frequency and time resources can be handled in the same manner.

[6]Note that the expression of the successful transmission probability in (4) is different from the performance metrics in [21] and [24].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

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

$$q\left(\mathbf{p}_1,\mathbf{p}_2\right)=\underbrace{\sum_{n\in\mathcal{N}}a_nA_{1,n}\left(\mathbf{p}_1,\mathbf{p}_2\right)\Pr\left[\frac{W}{K_{1,n,0}}\log_2\left(1+\mathrm{SINR}_{n,0}\right)\geq\tau\ \middle|\ j_0=1\right]}_{\triangleq q_1(\mathbf{p}_1,\mathbf{p}_2)}$$
$$+\underbrace{\sum_{n\in\mathcal{N}}a_nA_{2,n}\left(\mathbf{p}_2,\mathbf{p}_1\right)\Pr\left[\frac{W}{K_{2,n,0}}\log_2\left(1+\mathrm{SINR}_{n,0}\right)\geq\tau\ \middle|\ j_0=2\right]}_{\triangleq q_2(\mathbf{p}_1,\mathbf{p}_2)}. \tag{4}$$

cessful transmission probability in the asymptotic region.

A. Performance Analysis in General Region

In this subsection, we analyze the successful transmission probability in the general region (i.e., the general SNR and general user density region), using tools from stochastic geometry. First, the user association probability $A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\overline{j}})$ can be found in [13], [14] and is provided here for completeness:

$$A_{j,n}(\mathbf{p}_j,\mathbf{p}_{\overline{j}})=\frac{\lambda_jT_{j,n}}{\lambda_jT_{j,n}+\lambda_{\overline{j}}T_{\overline{j},n}\left(\frac{P_{\overline{j}}}{P_j}\right)^{\frac{2}{\alpha}}}\triangleq A_{j,n}(T_{j,n},T_{\overline{j},n}). \tag{5}$$

File load $K_{j,n,0}$ and SINR $\mathrm{SINR}_{n,0}$ are correlated in a complex manner in general, as POAs with larger association regions have higher file load and lower SINR (due to larger user to POA distances) [27]. For the tractability of the analysis, as in [21], [22], [27], the dependence is ignored, i.e., (6). To obtain the conditional p.m.f. of $K_{j,n,0}$ given $j_0=j$ by generalizing the methods for calculating the p.m.f. of file load in [21], we need the probability density function (p.d.f.) of the size of the Voronoi cell of BS ℓ_0 w.r.t. file $m\in\mathcal{N}_{j,i,-n}$ when ℓ_0 contains combination $i\in\mathcal{I}_{j,n}$, where $\mathcal{N}_{j,i,-n}\triangleq\mathcal{N}_{j,i}\setminus\{n\}$. However, since this p.d.f. is very complex and still unknown, we adopt the widely used approach in the existing literature [21], [22], [24], [27] and approximate this p.d.f. based on a tractable approximation of the p.d.f. of the size of the Voronoi cell to which a randomly chosen user belongs [28]. Under this approximation, the conditional p.m.f. of $K_{j,n,0}$ is given in the following lemma.

Lemma 1 (Conditional p.m.f. of $K_{j,n,0}$): The conditional p.m.f. of $K_{j,n,0}$ given $j_0=j$ is given by (7), where $k=1,\cdots,K_j$, and[7]

$$b_{j,m}\triangleq\left(1+\frac{a_m\lambda_u\widehat{A}_{j,m}\left(T_{j,m},T_{\overline{j},m}\right)}{3.5\lambda_j}\right)^{-3.5}, \tag{8}$$

$$\widehat{A}_{j,m}(T_{j,m},T_{\overline{j},m})\triangleq\frac{\lambda_j}{\lambda_jT_{j,m}+\lambda_{\overline{j}}T_{\overline{j},m}\left(\frac{P_{\overline{j}}}{P_j}\right)^{\frac{2}{\alpha}}}. \tag{9}$$

Proof: Please refer to Appendix A. ■

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

$$q\left(\mathbf{p}_1,\mathbf{p}_2\right)=q_1\left(\mathbf{p}_1,\mathbf{p}_2\right)+q_2\left(\mathbf{p}_2,\mathbf{p}_1\right), \tag{15}$$

where $q_j(\mathbf{p}_j,\mathbf{p}_{\overline{j}})$ is given by (10), $b_{j,m}$ is given by (8) and $f_{j,k}(T_{j,n},T_{\overline{j},n})$ is given by (11) with $\theta_{1,k}$, $\theta_{2,j,k}$ and $\theta_{3,j,k}$ given by (12), (13) and (14). Here, $B'(x,y,z)\triangleq\int_z^1u^{x-1}(1-u)^{y-1}\,\mathrm{d}u$ and $B(x,y)\triangleq\int_0^1u^{x-1}(1-u)^{y-1}\,\mathrm{d}u$ denote the complementary incomplete Beta function and the Beta function, respectively.

From Theorem 1, we can see that in the general region, 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 $(\mathbf{p}_1,\mathbf{p}_2)$ jointly affect the successful transmission probability $q\left(\mathbf{p}_1,\mathbf{p}_2\right)$. The impacts of the physical layer parameters and the design parameters on $q\left(\mathbf{p}_1,\mathbf{p}_2\right)$ are coupled in a complex manner.

B. Performance Analysis in Asymptotic Region

The gain of multicasting over unicasting increases with user density [21]. 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. Due to analytical tractability, the high SNR region, where noise power is negligible, is widely studied [7]–[10], [13], [14] and the high user density region, where the gain of multicast over unicast achieves the maximum, is also considered when studying multicast performance [21], [22]. Note that in the rest of the paper, when considering the

[7]Note that $\widehat{A}_{j,m}\left(T_{j,m},T_{\overline{j},m}\right)=\frac{A_{j,m}\left(T_{j,m},T_{\overline{j},m}\right)}{T_{j,m}}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

$$\Pr\left[\frac{W}{K_{j,n,0}}\log_2\left(1+\mathrm{SINR}_{n,0}\right)\geq\tau\;\middle|\;j_0=j\right]\approx\sum_{k=1}^{K_j}\Pr\left[K_{j,n,0}=k\mid j_0=j\right]\Pr\left[\mathrm{SINR}_{n,0}\geq\left(2^{\frac{k\tau}{W}}-1\right)\;\middle|\;j_0=j\right]. \quad (6)$$

$$\Pr\left[K_{j,n,0}=k\mid j_0=j\right]\approx\sum_{i\in\mathcal{I}_{j,n}}\frac{p_{j,i}}{T_{j,n}}\sum_{\mathcal{X}\in\{\mathcal{S}\subseteq\mathcal{N}_{j,i,-n}:|\mathcal{S}|=k-1\}}\prod_{m\in\mathcal{X}}\left(1-b_{j,m}\right)\prod_{m\in\mathcal{N}_{j,i,-n}\setminus\mathcal{X}}b_{j,m} \quad (7)$$

$$q_j(\mathbf{p}_j,\mathbf{p}_{\bar{j}})=\sum_{n\in\mathcal{N}}a_n\sum_{k=1}^{K_j}\left(\sum_{i\in\mathcal{I}_{j,n}}p_{j,i}\sum_{\mathcal{X}\in\{\mathcal{S}\subseteq\mathcal{N}_{j,i,-n}:|\mathcal{S}|=k-1\}}\prod_{m\in\mathcal{X}}\left(1-b_{j,m}\right)\prod_{m\in\mathcal{N}_{j,i,-n}\setminus\mathcal{X}}b_{j,m}\right)f_{j,k}(T_{j,n},T_{\bar{j},n}). \quad (10)$$

$$f_{j,k}(x,y)\triangleq 2\pi\lambda_j\int_0^\infty d\exp\left(-\pi\lambda_j\left(\theta_{1,k}x+\theta_{2,j,k}y+\theta_{3,j,k}\right)d^2\right)\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^\alpha\frac{N_0}{P_j}\right)\mathrm{d}d. \quad (11)$$

$$\theta_{1,k}=\frac{2}{\alpha}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha}}\left(B'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{-\frac{k\tau}{W}}\right)-B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right)+1. \quad (12)$$

$$\theta_{2,j,k}=\frac{2\lambda_{\bar{j}}}{\alpha\lambda_j}\left(\sigma_{\bar{j}}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha}}\left(B'\left(\frac{2}{\alpha},1-\frac{2}{\alpha},2^{-\frac{k\tau}{W}}\right)-B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)\right)+\frac{\lambda_{\bar{j}}}{\lambda_j}\sigma_{\bar{j}}^{\frac{2}{\alpha}}. \quad (13)$$

$$\theta_{3,j,k}=\frac{2}{\alpha}\left(2^{\frac{k\tau}{W}}-1\right)^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right)+\frac{2\lambda_{\bar{j}}}{\alpha\lambda_j}\left(\sigma_{\bar{j}}\left(2^{\frac{k\tau}{W}}-1\right)\right)^{\frac{2}{\alpha}}B\left(\frac{2}{\alpha},1-\frac{2}{\alpha}\right). \quad (14)$$

asymptotic region (i.e., the high SNR and user density region), we assume $\frac{P_1}{N_0}\to\infty$ and $\frac{P_2}{N_0}\to\infty$ while fixing the power ratio, i.e., σ_1 (σ_2). In addition, in the high user density region where $\lambda_u\to\infty$, the discrete random variable $K_{j,n,0}\to K_j$ in distribution. From Theorem 1, we can derive the successful transmission probability in the asymptotic region.

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

$$q(\mathbf{p}_1,\mathbf{p}_2)=q_{1,\infty}(\mathbf{T}_1,\mathbf{T}_2)+q_{2,\infty}(\mathbf{T}_2,\mathbf{T}_1)\triangleq q(\mathbf{T}_1,\mathbf{T}_2), \quad (16)$$

where

$$q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})=\sum_{n\in\mathcal{N}}\frac{a_nT_{j,n}}{\theta_{1,K_j}T_{j,n}+\theta_{2,j,K_j}T_{\bar{j},n}+\theta_{3,j,K_j}}. \quad (17)$$

Here, $T_{j,n}$ is given by (2), and $\theta_{1,k}$, $\theta_{2,j,k}$ and $\theta_{3,j,k}$ are given by (12), (13) and (14).

Proof: When $\frac{P_1}{N_0}\to\infty$ and $\frac{P_2}{N_0}\to\infty$, we have $\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^\alpha\frac{N_0}{P_1}\right)\to 1$ and $\exp\left(-\left(2^{\frac{k\tau}{W}}-1\right)d^\alpha\frac{N_0}{P_2}\right)\to 1$. When $\lambda_u\to\infty$, we have $K_{j,n,0}\to K_j$ in distribution. Noting that $\int_0^\infty d\exp\left(-cd^2\right)\mathrm{d}d=\frac{1}{2c}$, we can solve integrals in (11). Thus, we can prove Corollary 1. ■

Note that $q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\bar{j}})=\lim_{\frac{P}{N_0}\to\infty,\lambda_u\to\infty}q_j(\mathbf{p}_j,\mathbf{p}_{\bar{j}})$ and $q_\infty(\mathbf{T}_1,\mathbf{T}_2)=\lim_{\frac{P}{N_0}\to\infty,\lambda_u\to\infty}q(\mathbf{p}_1,\mathbf{p}_2)$; 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 $\mathbf{p}_2$. In addition, the asymptotic successful transmission probability in Corollary 1 and the performance metric in [13], [14] 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. From (12), we see that θ_{1,K_j} is a function of $K_j\tau/W$ (which can be interpreted as the target bit rate (in bit/s/Hz), as shown in (4)) and the path loss exponent α. In most practical cases, $\theta_{1,K_j}>0$, as shown in Fig. 2. Thus, we consider $\theta_{1,K_1},\theta_{1,K_2}>0$ in the rest of the paper.

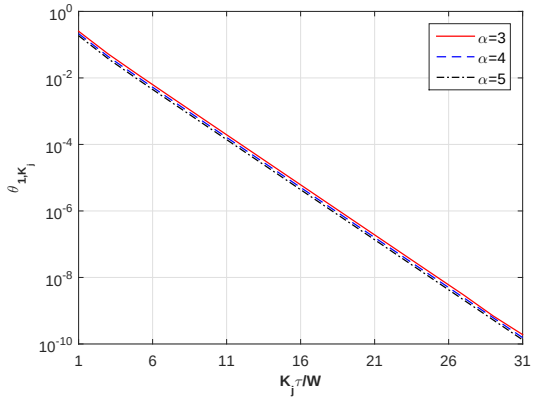


Fig. 2. θ_{1,K_j} versus $K_j\tau/W$ at different α.

Fig. 3 verifies Theorem 1 and Corollary 1, and demonstrates the accuracy of the approximation adopted. Fig. 3 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., $\lambda_u\geq 3\times 10^{-5}$).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

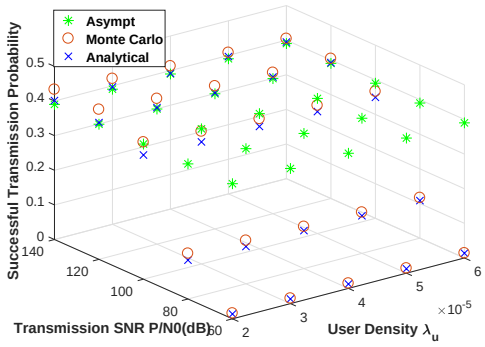


Fig. 3. Successful transmission probability versus SNR $\frac{P}{N_0}$ and user density λ_u. $N = 24$, $K_1 = 6$, $K_2 = 4$, $p_{1,i} = \frac{1}{\binom{24}{6}}$ for all $i = 1, 2, \cdots, \binom{24}{6}$, $p_{2,i} = \frac{1}{\binom{24}{4}}$ for all $i = 1, 2, \cdots, \binom{24}{4}$, $\lambda_1 = 5 \times 10^{-7}$, $\lambda_2 = 3 \times 10^{-6}$, $P_1 = 10^{1.5} P$, $P_2 = P$, $\alpha = 4$, $W = 2 \times 10^7$, $\tau = 35 \times 10^4$ and $a_n = \frac{n^{-\gamma}}{\sum_{n \in \mathcal{N}} n^{-\gamma}}$ with $\gamma = 1$. According to 3GPP release, Macro transmission power is 46 dBm, Pico transmission power is 30dBm and noise power is -104 dBm. That is, transmission SNR at the macro tier is 46-(-104)=150 dB and transmission SNR at the pico tier is 30-(-104)=134 dB. The transmission SNR range 60 dB-140dB is reasonable.

In the asymptotic region, from [21], 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

$$\mathcal{T}_j \triangleq \left\{ \mathbf{T}_j \;\middle|\; 0 \le T_{j,n} \le 1, n \in \mathcal{N}, \sum_{n \in \mathcal{N}} T_{j,n} = K_j \right\}. \quad (18)$$

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. Note that based on the designs in the asymptotic region, we can obtain promising designs in the general region, by using the method proposed in our previous work [21], [22]. We omit the details due to page limitation. In addition, we focus on offline joint and competitive caching designs. Note that file popularity evolves at a slower timescale and careful caching designs can provide reasonable performance over a certain period (say several hours or a few days depending on the changing rate of file popularity).

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 iterative algorithm to obtain a stationary point. The joint caching design can be obtained by the centralized controller of the operator.

Algorithm 1 Stationary Point of Problem 1 Based on the Standard Gradient Projection Method

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: For all $n \in \mathcal{N}$, compute $\bar{T}_{j,n}(t+1)$ according to $\bar{T}_{j,n}(t+1) = T_{j,n}(t) + \epsilon(t) \frac{\partial q_\infty(\mathbf{T}_1(t), \mathbf{T}_2(t))}{\partial T_{j,n}(t)}$.
3: For all $n \in \mathcal{N}$, compute $T_{j,n}(t+1)$ according to $T_{j,n}(t+1) = \min\left\{ \left[\bar{T}_{j,n}(t+1) - \nu_j^*\right]^+, 1 \right\}$, where ν_j^* satisfies $\sum_{n \in \mathcal{N}} \min\left\{ \left[\bar{T}_{j,n}(t+1) - \nu_j^*\right]^+, 1 \right\} = K_j$.
4: Set $t = t + 1$ and go to Step 2.

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):

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

where $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ is given by (16) and $\mathcal{T}_j$ is given by (18).

Problem 1 maximizes a differentiable (nonconcave in general) function over a convex set, and it is thus nonconvex in general. Note that Problem 1 in this paper and Problem 0 in [14] are mathematically equivalent, although this paper and [14] have different scopes. In the following subsection, we propose an efficient algorithm to solve Problem 1. In contrast, [14] simplifies the nonconvex problem to a convex one, and uses an optimal solution to the simplified problem as a sub-optimal solution to the original problem, which may not provide performance guarantee.

B. Algorithm

Recall that Problem 1 is to maximize a differentiable (nonconcave in general) function over a convex set. We can obtain a stationary point of Problem 1 using the gradient projection method with a diminishing stepsize [29, pp. 227], as summarized in Algorithm 1 for completeness. In Algorithm 1, the diminishing stepsize $\epsilon(t)$ satisfies $\epsilon(t) \to 0$ as $t \to \infty$, $\sum_{t=1}^{\infty} \epsilon(t) = \infty$ and $\sum_{t=1}^{\infty} (\epsilon(t))^2 < \infty$. In addition, Step 3 is the projection of $\bar{T}_{j,n}(t+1)$ onto set $\mathcal{T}_j$. It is shown in [29, pp. 229] that the sequence $\{(\mathbf{T}_1(t), \mathbf{T}_2(t))\}$ generated by Algorithm 1 converges to a stationary point of Problem 1. Note that a 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. However, the rate of convergence of Algorithm 1 is strongly dependent on the choices of stepsize $\epsilon(t)$. If it is chosen improperly, it may take a large number of iterations for Algorithm 1 to meet some convergence criterion.

To address the above problem, in this subsection we propose an iterative algorithm to obtain a stationary point of Problem 1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

more efficiently. This algorithm is based on the block successive upper-bound minimization algorithm originally proposed in [23]. 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 $\overline{j}$th tier, and fix the caching distribution of the $\overline{j}$th tier, where $j = ((t+1) \bmod 2) + 1$.

For notational convenience, we define

$$\tilde{q}_\infty\left(\mathbf{T}_j, \mathbf{T}_{\overline{j}}\right) \triangleq \begin{cases} q_\infty\left(\mathbf{T}_j, \mathbf{T}_{\overline{j}}\right), & j = 1, \\ q_\infty\left(\mathbf{T}_{\overline{j}}, \mathbf{T}_j\right), & j = 2. \end{cases} \tag{19}$$

At iteration t, choose $g_j(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t))$ to be an approximate function of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$, where $g_j(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t))$ is given by (20). Note that the first concave component function of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$, i.e., $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$ is left unchanged, and only the second nonconcave (actually convex) component function, i.e., $q_{\overline{j},\infty}(\mathbf{T}_{\overline{j}}(t), \mathbf{T}_j)$ is linearized at $\mathbf{T}_j = \mathbf{T}_j(t)$. This choice of the 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, and the resulting algorithm typically converges much faster than Algorithm 1, 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 2. Specifically, $g_j(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t))$ is strictly concave on $\mathcal{T}_j$ for any given $(\mathbf{T}_1(t), \mathbf{T}_2(t)) \in \mathcal{T}_1 \times \mathcal{T}_2$, and satisfies

$$g_j\left(\mathbf{T}_j(t); \mathbf{T}_1\left(t\right), \mathbf{T}_2\left(t\right)\right) = \tilde{q}_\infty\left(\mathbf{T}_j\left(t\right), \mathbf{T}_{\overline{j}}\left(t\right)\right), \quad (\mathbf{T}_1(t), \mathbf{T}_2(t)) \in \mathcal{T}_1 \times \mathcal{T}_2, \tag{21}$$

$$g_j\left(\mathbf{T}_j; \mathbf{T}_1\left(t\right), \mathbf{T}_2\left(t\right)\right) \le \tilde{q}_\infty\left(\mathbf{T}_j, \mathbf{T}_{\overline{j}}\left(t\right)\right), \quad \mathbf{T}_j \in \mathcal{T}_j,\ \left(\mathbf{T}_1\left(t\right), \mathbf{T}_2\left(t\right)\right) \in \mathcal{T}_1 \times \mathcal{T}_2. \tag{22}$$

Note that (21) holds since $g_j(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t))$ and $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$ have the same value at the point where $g_j(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t))$ is defined, i.e., $(\mathbf{T}_1, \mathbf{T}_2) = (\mathbf{T}_1(t), \mathbf{T}_2(t))$; (22) holds since $q_{\overline{j},\infty}(\mathbf{T}_{\overline{j}}, \mathbf{T}_j)$ is a convex function of $\mathbf{T}_j$ for any given $\mathbf{T}_{\overline{j}} \in \mathcal{T}_{\overline{j}}$. The conditions in (21) and (22) imply that $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is a tight lower-bound of $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$. The differentiability of $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ guarantees that the first-order behavior of $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is the same as $\tilde{q}_\infty(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$ locally. At each iteration t, we update the caching distribution of the jth tier given the caching distribution of the $\overline{j}$th tier by solving the following problem, and fix the caching distribution of the $\overline{j}$th tier, where $j = ((t+1) \bmod 2) + 1$.

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

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

where $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is given by (20).

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, as shown in the following lemma.

Lemma 2 (Optimal Solution to Problem 2): For all $j = ((t+1) \bmod 2)+1$, the optimal solution to Problem 2 is given by (23), where $[x]^+ \triangleq \max\{x, 0\}$ and $\nu_j^*(t)$ is the Lagrange multiplier that satisfies $\sum_{n\in\mathcal{N}} T_{j,n}(t+1) = K_j$.

Note that $\nu_j^*(t)$ can be efficiently obtained by using bisection search. The details of the proposed iterative algorithm are summarized in Algorithm 2. Based on the conditions in (21) and (22), we can show the convergence and optimality of Algorithm 2.

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

Proof: We show that the conditions in Theorem 2 (a) of [23] hold. i) By noting that (21) and (22) hold and $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is continuous and differentiable, we know that Assumption 2 in [23] is satisfied and $q_\infty\left(\mathbf{T}_1, \mathbf{T}_2\right)$ is regular at any point in $\mathcal{T}_1 \times \mathcal{T}_2$ [23]. ii) Since $g_j\left(\mathbf{T}_j; \mathbf{T}_1(t), \mathbf{T}_2(t)\right)$ is strictly concave on $\mathcal{T}_j$ for any given $(\mathbf{T}_1(t), \mathbf{T}_2(t)) \in \mathcal{T}_1 \times \mathcal{T}_2$, Problem 2 has a unique solution for any given $(\mathbf{T}_1(t), \mathbf{T}_2(t)) \in \mathcal{T}_1 \times \mathcal{T}_2$. Therefore, by Theorem 2 (a) in [23], we can prove Theorem 2. ■

Algorithm 2 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 2.
4: For all $n \in \mathcal{N}$, set $T_{\overline{j},n}(t+1) = T_{\overline{j},n}(t)$.
5: Set $t = t + 1$ and go to Step 2.

Different from Algorithm 1, Algorithm 2 does not rely on a stepsize. Thus, Algorithm 2 may have more robust convergence performance than Algorithm 1, as we shall illustrate later in Fig. 4.

In the rest of this subsection, we consider a special case where $K_1 = K_2 \triangleq K$. In this case, $\lambda_1 P_1^{\frac{2}{\alpha}} \theta_{1,K_1} = \lambda_2 P_2^{\frac{2}{\alpha}} \theta_{2,2,K_2} \triangleq \mu_{1,K}$, $\lambda_1 P_1^{\frac{2}{\alpha}} \theta_{2,1,K_1} = \lambda_2 P_2^{\frac{2}{\alpha}} \theta_{1,K_2} \triangleq \mu_{2,K}$ and $\lambda_1 P_1^{\frac{2}{\alpha}} \theta_{3,1,K_1} = \lambda_2 P_2^{\frac{2}{\alpha}} \theta_{3,2,K_2} \triangleq \mu_{3,K}$. In addition, $q_\infty(\mathbf{T}_1, \mathbf{T}_2)$ can be further simplified as

$$q_\infty(\mathbf{T}_1, \mathbf{T}_2) = \sum_{n\in\mathcal{N}} a_n \frac{\lambda_1 P_1^{\frac{2}{\alpha}} T_{1,n} + \lambda_2 P_2^{\frac{2}{\alpha}} T_{2,n}}{\mu_{1,K} T_{1,n} + \mu_{2,K} T_{2,n} + \mu_{3,K}}, \tag{24}$$

which is a concave function of $(\mathbf{T}_1, \mathbf{T}_2)$. Thus, Problem 1 becomes a convex optimization problem, and a (globally)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

$$g_j\left(\mathbf{T}_j;\mathbf{T}_1(t),\mathbf{T}_2(t)\right) \triangleq q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\overline{j}}(t)) + q_{\overline{j},\infty}(\mathbf{T}_{\overline{j}}(t),\mathbf{T}_j(t)) + \sum_{n\in\mathcal{N}} \frac{\partial q_{\overline{j},\infty}\left(\mathbf{T}_{\overline{j}}(t),\mathbf{T}_j(t)\right)}{\partial T_{j,n}}\left(T_{j,n}-T_{j,n}(t)\right)$$
$$= q_{j,\infty}(\mathbf{T}_j,\mathbf{T}_{\overline{j}}(t)) + q_{\overline{j},\infty}(\mathbf{T}_{\overline{j}}(t),\mathbf{T}_j(t)) - \sum_{n\in\mathcal{N}} \frac{a_n\theta_{2,\overline{j},K_{\overline{j}}}T_{\overline{j},n}(t)\left(T_{j,n}-T_{j,n}(t)\right)}{\left(\theta_{1,K_{\overline{j}}}T_{\overline{j},n}(t)+\theta_{2,\overline{j},K_{\overline{j}}}T_{j,n}(t)+\theta_{3,\overline{j},K_{\overline{j}}}\right)^2}. \quad (20)$$

$$T_{j,n}(t+1) = \min\left\{\left[\frac{1}{\theta_{1,K_j}}\sqrt{\frac{a_n\left(\theta_{2,j,K_j}T_{\overline{j},n}(t)+\theta_{3,j,K_j}\right)}{\nu_j^*(t)+\frac{a_n\theta_{2,\overline{j},K_{\overline{j}}}T_{\overline{j},n}(t)}{\left(\theta_{1,K_{\overline{j}}}T_{\overline{j},n}(t)+\theta_{2,\overline{j},K_{\overline{j}}}T_{j,n}(t)+\theta_{3,\overline{j},K_{\overline{j}}}\right)^2}}} - \frac{\theta_{2,j,K_j}T_{\overline{j},n}(t)+\theta_{3,j,K_j}}{\theta_{1,K_j}}\right]^+, 1\right\},\ n\in\mathcal{N}. \quad (23)$$

optimal solution can be obtained by standard convex optimization methods such as interior-point methods. However, when N is very large, standard convex optimization methods may not scale very well. Motivated by [14], by exploring structural properties of Problem 1 in this case, we develop a low-complexity algorithm to obtain an optimal solution. The method consists of two stages. In the first stage, we solve a relaxed version of Problem 1 to obtain a system of linear equations of an optimal solution to Problem 1. This stage is the same as that in [14], and is included for completeness. Denote $\mathbf{R} \triangleq (R_n)_{n\in\mathcal{N}}$, where $R_n \triangleq P_1^{\frac{2}{\alpha}}\lambda_1 T_{1,n} + P_2^{\frac{2}{\alpha}}\lambda_2 T_{2,n}$. Specifically, Problem 1 can be relaxed as follows.

Problem 3 (Relaxed Version of Problem 1 When $K_1 = K_2 = K$ [14]):

$$\max_{\mathbf{R}} \quad \sum_{n\in\mathcal{N}} a_n \frac{R_n}{\theta_{1,K}R_n+\mu_{3,K}},$$
$$s.t. \quad 0 \le R_n \le P_1^{\frac{2}{\alpha}}\lambda_1 + P_2^{\frac{2}{\alpha}}\lambda_2,\ n\in\mathcal{N},$$
$$\sum_{n\in\mathcal{N}} R_n = (P_1^{\frac{2}{\alpha}}\lambda_1 + P_2^{\frac{2}{\alpha}}\lambda_2)K.$$

Let $\mathbf{R}^*$ denote the optimal solution to Problem 3.

The optimal soluton to Problem 3 is given by [14, Proposition 3], i.e.,

$$R_n^* = \min\left\{\left[\frac{1}{\theta_{1,K}}\left(\sqrt{\frac{a_n\mu_{3,K}}{\nu^*}}-\mu_{3,K}\right)\right]^+, P_1^{\frac{2}{\alpha}}\lambda_1 + P_2^{\frac{2}{\alpha}}\lambda_2\right\} \quad n\in\mathcal{N}, \quad (25)$$

where ν^* is the Lagrange multiplier that satisfies

$$\sum_{n\in\mathcal{N}} R_n^* = \left(P_1^{\frac{2}{\alpha}}\lambda_1 + P_2^{\frac{2}{\alpha}}\lambda_2\right)K. \quad (26)$$

Note that ν^* can be efficiently obtained by using bisection search. In addition, by Proposition 4 in [14], we know that the optimal solution to Problem 3 and an optimal solution to Problem 1 satisfy a system of linear equations:

$$P_1^{\frac{2}{\alpha}}\lambda_1 T_{1,n}^* + P_2^{\frac{2}{\alpha}}\lambda_2 T_{2,n}^* = R_n^*,\ n\in\mathcal{N}. \quad (27)$$

In the second stage, we solve the system of linear equations given in (27) to obtain an optimal solution $(\mathbf{T}_1^*,\mathbf{T}_2^*)$ to Problem 1. In our case, we can easily show that

$$T_{j,n}^* = \frac{R_n^*}{P_1^{\frac{2}{\alpha}}\lambda_1 + P_2^{\frac{2}{\alpha}}\lambda_2},\ n\in\mathcal{N},\ j=1,2 \quad (28)$$

is a solution to the system of linear equations in (27). This stage is different from that in [14], as we can directly obtain $T_{j,n}^*$ using the closed-form expression in (28), due to $K_1 = K_2$. The details are summarized in Algorithm 3. Note that the complexity of Algorithm 3 is close to that of one iteration of Algorithm 2. Thus, the complexity of Algorithm 3 is much lower than that of Algorithm 2.

Algorithm 3 Globally Optimal Solution

1: Obtain R_n^* by (25) and (26).
2: Compute $T_{j,n}^*$ by (28), $n\in\mathcal{N}$, $j=1,2$.

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 an 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 an NE of the game and adopt an iterative algorithm to obtain an NE. The competitive caching design can be obtained by the centralized controllers of the two operators, allowing message passing between them.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

A. Game Formulation

In this subsection, we formulate the competitive caching design for the 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., $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$. Each tier j competes against the other tier $\overline{j}$ 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., $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$.

Problem 4 (Competitive Caching Game): For all $j = 1, 2$, we have

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

where $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is given by (17) and $\mathcal{T}_j$ is given by (18). Let $\mathcal{G}$ denote the game.

A solution of a game, i.e., an NE, is reached when each player, given the strategy profiles of the others, does not get any performance increase by unilaterally changing his own strategy [30]. An 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 an NE of game $\mathcal{G}$ if

$$q_{j,\infty}(\mathbf{T}_j^\dagger, \mathbf{T}_{\overline{j}}^\dagger) \geq q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}}^\dagger), \ \mathbf{T}_j \in \mathcal{T}_j, \ j = 1, 2. \quad (29)$$

By Definition 1, we know that an NE of game $\mathcal{G}$ is given by the following problem.

Problem 5 (NE of Game $\mathcal{G}$): For all $j = 1, 2$, we have

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

B. Nash Equilibrium

In this subsection, we characterize an NE of game $\mathcal{G}$. First, we show the existence and uniqueness of the NE of game $\mathcal{G}$.

Lemma 3 (Existence and Uniqueness of NE of Game $\mathcal{G}$): There exists a unique NE of game $\mathcal{G}$.

Proof: First, we use Proposition 20.3 in [31] to prove the existence of NE of game $\mathcal{G}$. It is obvious that for all $j = 1, 2$, the set of admissible strategies $\mathcal{T}_j$ is nonempty, compact and convex, and the utility function $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is a continuous function of $(\mathbf{T}_1, \mathbf{T}_2)$. Since $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is strictly concave on $\mathcal{T}_j$ for any given $\mathbf{T}_{\overline{j}} \in \mathcal{T}_{\overline{j}}$, then $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is quasi-concave on $\mathcal{T}_j$ for any given $\mathbf{T}_{\overline{j}} \in \mathcal{T}_{\overline{j}}$. Thus, by Proposition 20.3 in [31], we know that there exists at least one NE of game $\mathcal{G}$. Next, we prove the uniqueness of NE by Theorem 2 in [32]. By the first-order strict concavity condition, we know that a strictly concave function must be diagonally strictly concave [32]. Thus, by Theorem 2 in [32], we know that there exists a unique NE of game $\mathcal{G}$. ■

We now obtain the closed-form expression of the unique NE of game $\mathcal{G}$. Since $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is strictly concave on $\mathcal{T}_j$ for any given $\mathbf{T}_{\overline{j}} \in \mathcal{T}_{\overline{j}}$, Problem 5 is convex and Slater's condition is satisfied, implying that strong duality holds. Using KKT conditions, we can solve Problem 5 and show Lemma 4.

Lemma 4 (NE of Game $\mathcal{G}$): Game $\mathcal{G}$ has a unique NE $(\mathbf{T}_1^\dagger, \mathbf{T}_2^\dagger)$ which is given by (30), where for all $j = 1, 2$, $\nu_j^\dagger$ is the Lagrange multiplier that satisfies $\sum_{n\in\mathcal{N}} T_{j,n}^\dagger = K_j$.

Remark 3: The file popularity distribution $\mathbf{a}$ and the physical layer parameters (captured in θ_{1,K_j}, θ_{2,j,K_j} and θ_{3,j,K_j}) jointly affect $\nu_j^\dagger$. Given $\nu_j^\dagger$ and $T_{\overline{j},n}^\dagger$, $n \in \mathcal{N}$, the physical layer parameters (captured in θ_{1,K_j}, θ_{2,j,K_j} and θ_{3,j,K_j}) affect the caching probabilities of all the files in the same way, while the popularity of file n (i.e., a_n) only affects the caching probability of file n (i.e., $T_{j,n}^\dagger$) [22].

Algorithm 4 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 $\mathbf{T}_j(t+1) = \arg\max_{\mathbf{T}_j \in \mathcal{T}_j} q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}}(t))$.
4: Set $\mathbf{T}_{\overline{j}}(t+1) = \mathbf{T}_{\overline{j}}(t)$.
5: Set $t = t + 1$ and go to Step 2.

C. Algorithm

In this subsection, we adopt 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.

Problem 6 (Optimization at Iteration t): For player $j = ((t+1) \bmod 2) + 1$, we have

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

Similar to Problem 5, using KKT conditions, we can obtain the closed-form expression of the unique optimal solution to Problem 6. We present it below for completeness.

Lemma 5 (Optimal Solution to Problem 6): For all $j = ((t+1) \bmod 2)+1$, the optimal solution to Problem 6 is given by (31), where $\nu_j^\dagger(t)$ is the Lagrange multiplier that satisfies $\sum_{n\in\mathcal{N}} T_{j,n}(t+1) = K_j$.

Note that Problem 5 and Problem 6 share similar forms. Thus, the NE of game $\mathcal{G}$ in Lemma 4 and the solution to Problem 6 in Lemma 5 share similar forms. Based on the optimal solution to Problem 6, at iteration t, we update the strategy of player j, and fix the strategy of player $\overline{j}$, where $j = ((t+1) \bmod 2) + 1$. The details for obtaining the NE of game $\mathcal{G}$ is summarized in Algorithm 4.

Although iterative algorithms based on alternating optimizations are widely used in solving games, there is in general no guarantee that an iterative algorithm can converge to an NE of a game, especially for a game with a complex utility function for each player in a large-scale two-tier wireless network. By carefully analyzing structural properties of the competitive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

$$T_{j,n}^{\dagger} = \min\left\{\left[\frac{1}{\theta_{1,K_j}}\sqrt{\frac{a_n\left(\theta_{2,j,K_j}T_{\overline{j},n}^{\dagger}+\theta_{3,j,K_j}\right)}{\nu_j^{\dagger}}}-\frac{\theta_{2,j,K_j}T_{\overline{j},n}^{\dagger}+\theta_{3,j,K_j}}{\theta_{1,K_j}}\right]^{+},1\right\},\ n\in\mathcal{N},\ j=1,2. \tag{30}$$

$$T_{j,n}(t+1) = \min\left\{\left[\frac{1}{\theta_{1,K_j}}\sqrt{\frac{a_n\left(\theta_{2,j,K_j}T_{\overline{j},n}(t)+\theta_{3,j,K_j}\right)}{\nu_j^{\dagger}(t)}}-\frac{\theta_{2,j,K_j}T_{\overline{j},n}(t)+\theta_{3,j,K_j}}{\theta_{1,K_j}}\right]^{+},1\right\},\ n\in\mathcal{N}. \tag{31}$$

caching design game, we provide a convergence condition for Algorithm 4.

Theorem 3 (Convergence of Algorithm 4): 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 (12), (13) and (14), Algorithm 4 converges to the unique NE of game $\mathcal{G}$, i.e., $(\mathbf{T}_1(t),\mathbf{T}_2(t))\rightarrow(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})$ as $t\rightarrow\infty$, for all $\mathbf{T}_j(1)\in\mathcal{T}_j$, $j=1,2$, where $(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})$ is given by Lemma 4.

Proof: Please refer to Appendix B. ■

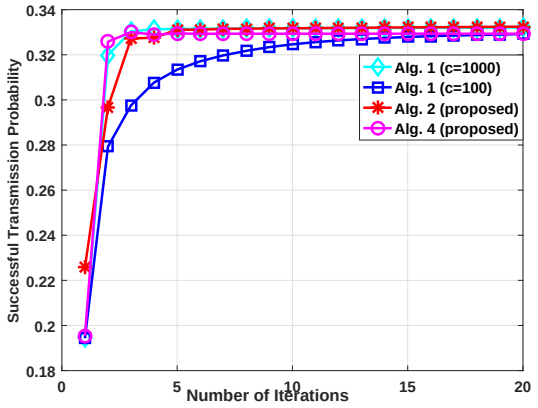
Note that the convergence condition given in Theorem 3 can be easily satisfied in most cases we are interested in (which will be shown in Fig. 4). As the unique NE of game $\mathcal{G}$, i.e., $(\mathbf{T}_1^{\dagger},\mathbf{T}_2^{\dagger})$ is a feasible point of Problem 1 with promising performance (which will be shown in Fig. 6), Algorithm 4 can also be treated as a suboptimal solution of Problem 1.

VI. Numerical Results

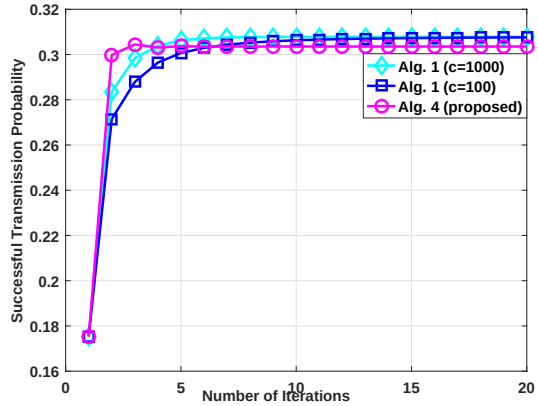
In this section, we first illustrate the convergence and complexity of the proposed algorithms. Then, we compare the successful transmission probabilities and caching probabilities of the proposed algorithms with those of existing solutions. In the simulation, we consider the asymptotic region (i.e., high SNR and high user density region) where we propose the joint caching and competitive caching designs, and choose $W=20\times10^6$, $\tau=4\times10^4$, $N=500$, $\alpha=4$, $\lambda_1=5\times10^{-7}$, $\lambda_2=3\times10^{-6}$ and $P_1=10^{1.6}P_2$. 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.

A. Convergence and Complexity

In this subsection, we show the convergence and complexity of the proposed algorithms. Fig. 4 illustrates the successful transmission probability versus the number of iterations when $K_1\neq K_2$ and $K_1=K_2$. From Fig. 4, we can observe that the rate of convergence of Algorithm 1 is strongly dependent on the choices of stepsize $\epsilon(t)$. In addition, Algorithm 2 and Algorithm 4 have more robust convergence performance than Algorithm 1, as they do not rely on a stepsize. Fig. 5 illustrates the computing time versus the cache size K_j and the Zipf exponent γ when $K_1\neq K_2$ and $K_1=K_2$. From Fig. 5, we can observe that the computing times of all the algorithms do not change much with K_j or γ, and the computing times



(a) $K_1=55$ and $K_2=35$.



(b) $K_1=K_2=35$.

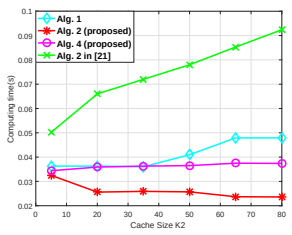
Fig. 4. Successful transmission probability versus the number of iterations. The stepsize for Algorithm 1 is $\epsilon(t)=\frac{c}{2+t^{0.55}}$. We choose the same initial point for all the algorithms shown in Fig. 4

of the proposed algorithms are shorter than that of Algorithm 2 in [21] which is to obtain an asymptotically optimal hybrid caching design. These observations demonstrate the advantage of the proposed algorithms in terms of complexity.

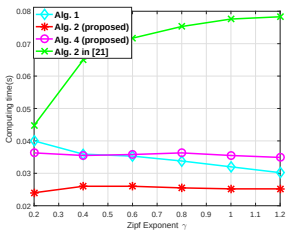
B. Successful Transmission Probabilities and Caching Probabilities

In this subsection, we compare the successful transmission probabilities and caching 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

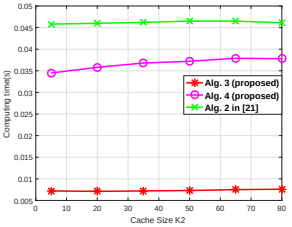
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications



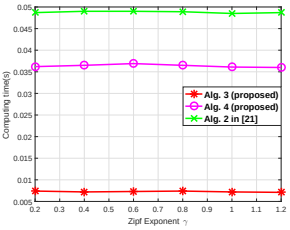
(a) Cache size K_2 (K_1) at $\gamma = 0.55$, $K_1 = K_2 + 20$.



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



(c) Cache size K_2 (K_1) at $\gamma = 0.55$, $K_1 = K_2$.



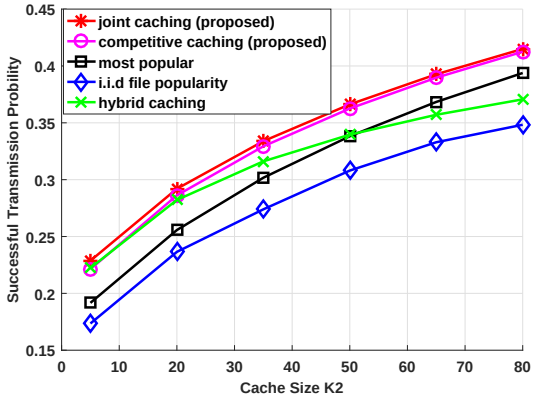
(d) Zipf exponent γ at $K_1 = K_2 = 35$.

Fig. 5. Computing time versus the cache size or Zipf exponent γ. The stepsize for Algorithm 1 is $\epsilon(t) = \frac{c}{2+t^{0.55}}$. For Algorithm 1, each point corresponds to the minimum computing time by choosing the optimal parameter $c \in \{500, 1000, 1500, 2000, 2500\}$.

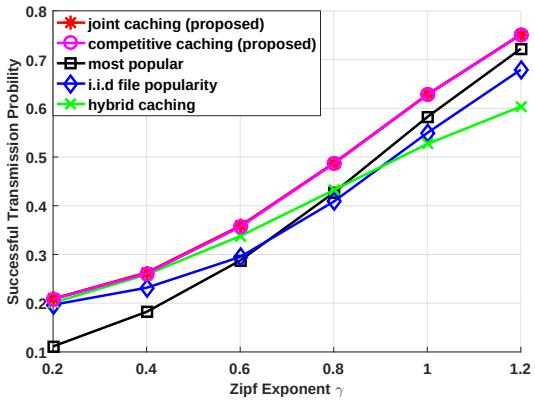
POA in tier j stores the K_j most popular files [4]–[6]. Baseline 2 (i.i.d. file popularity) refers to the design in which each POA in tier j randomly stores K_j files, in an i.i.d. manner with file n being selected with probability a_n [33]. Baseline 3 (hybrid caching) refers to the hybrid caching design obtained by Algrithm 2 in [21]. The three baseline schemes also adopt the same multicasting scheme as in our design.

Fig. 6 illustrates the successful transmission probability versus the cache size K_j and the Zipf exponent γ, respectively. From Fig. 6, we can observe that as K_j and γ increase, 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 or γ is large, the two proposed designs reduce to the most popular caching design. When K_j or γ is small, the two proposed designs perform similarly to 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.

Fig. 7 illustrates the caching probabilities for the proposed joint caching design, competitive caching design and the hybrid caching design. Recall that under the hybrid caching design, the files stored in the two tiers are non-overlapping, while the proposed joint caching design and competitive caching design allow a file to be stored in the two tiers. By comparing Fig. 7 (a) and Fig. 7 (b) (or Fig. 7 (c) and Fig. 7 (d)), we can see that when $\frac{\lambda_1}{\lambda_2}$ is above some threshold (see Fig. 7 (a) and Fig. 7(c)), POAs of tier 1 cache the most popular files under the hybrid caching design; when $\frac{\lambda_1}{\lambda_2}$ is below some threshold (see Fig. 7 (b) and Fig. 7(d)), POAs of tier 2 cache



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



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

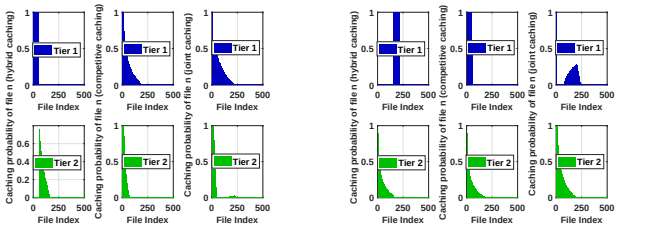
Fig. 6. Successful transmission probability versus the cache size or Zipf exponent γ.

the most popular files under the hybrid caching design. The reason is as follows. When $\frac{\lambda_1}{\lambda_2}$ is large enough ($\frac{\lambda_1}{\lambda_2}$ is small enough), POAs of tier 1 (tier 2) can offer relatively higher received signal powers, thus providing a larger successful transmission probability. In addition, by comparing Fig. 7 (a) and Fig. 7 (c) (or Fig. 7 (b) and Fig. 7 (d)), we can see that under the joint caching design and the hybrid caching design, when K_1 is larger, POAs of tier 1 cache more popular files. Finally, by comparing the caching probabilities under the three designs, we can see that the joint caching design and the competitive caching design offer much higher spatial file diversity (i.e. storing more popular files), leading to higher successful transmission probabilities.

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 derived tractable expressions for the successful transmission probabilities in the general region and the asymptotic region, respectively. Then, we formulated the optimal joint caching design problem in the asymptotic region. We develop an iterative algorithm, which is shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications



(a) $\lambda_1 = 1.3 \times 10^{-6}$, $\lambda_2 = 2.2 \times 10^{-6}$, $K_2 = 35$, $K_1 = 55$.

(b) $\lambda_1 = 9 \times 10^{-9}$, $\lambda_2 = 3.5 \times 10^{-6}$, $K_2 = 35$, $K_1 = 55$.

(c) $\lambda_1 = 1.3 \times 10^{-6}$, $\lambda_2 = 2.2 \times 10^{-6}$, $K_1 = K_2 = 35$.

(d) $\lambda_1 = 9 \times 10^{-9}$, $\lambda_2 = 3.5 \times 10^{-6}$,$K_1 = K_2 = 35$.

Fig. 7. Caching probabilities for the files in $\{5, 10, \cdots, 500\}$ of joint caching design, competitive caching design and hybrid caching design at $\alpha = 4$ and $\gamma = 0.55$.

to converge to a stationary point. Next, we formulated the competitive caching design game in the asymptotic region, obtained the unique NE of the game and adopted 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, in terms of successful transmission probability and complexity.

This paper opens up several directions for future research. For instance, the proposed analysis and optimization framework for two-tier large-scale cache-enabled wireless multicasting networks can be extended to study general multi-tier large-scale cache-enabled wireless multicasting networks. In addition, the proposed analysis and optimization framework can be utilized to study other performance metrics, such as the mean successful transmission rate and cache hit probability. Finally, a possible direction for future research is to consider temporal file request aggregation to increase multicast opportunities at the cost of delay.

Appendix A: Proof of Lemma 1

Let random variable $Y_{j,m,n,i} \in \{0, 1\}$ denote whether file $m \in \mathcal{N}_{j,i} \setminus \{n\}$ is requested from ℓ_0 when ℓ_0 contains combination $i \in \mathcal{I}_{j,n}$. When ℓ_0 contains combination $i \in \mathcal{I}_{j,n}$, we have $K_{j,n,0} = 1 + \sum_{m \in \mathcal{N}_{j,i,-n}} Y_{j,m,n,i}$. For analytical tractability, as in [22], assume $Y_{j,m,n,i}$, $m \in \mathcal{N}_{j,i} \setminus \{n\}$ are independent. By Appendix C of [22], we have (32).

Similar to Appendix B in [34], we have (34).

$$\Pr[Y_{j,m,n,i} = 0] \approx \left(1 + \frac{a_m \lambda_u A_{j,m} \left(T_{j,m}, T_{\overline{j},m}\right)}{3.5 T_{j,m} \lambda_j}\right)^{-3.5} = b_{j,m}. \quad (34)$$

By substituting (34) into (32), we can prove Lemma 1.

Appendix B: Proof of Theorem 3

We prove the convergence of Algorithm 4 by verifying the conditions in Theorem 1 of [35]. i) It can be easily seen that for all $j = 1, 2$, $q_{j,\infty}(\mathbf{T}_j, \mathbf{T}_{\overline{j}})$ is strictly concave on $\mathcal{T}_j$ for any given $\mathbf{T}_{\overline{j}} \in \mathcal{T}_{\overline{j}}$ and is second-order Fréchet differentiable [36]. ii) By Lemma 5, we know that there exists an optimal solution to Problem 6 for any given $\mathbf{T}_{\overline{j}}(t) \in \mathcal{T}_{\overline{j}}$. iii) To guarantee the convergence of Algorithm 4, it remains to show that (33) holds for any $\mathbf{T}_j(t) \in \mathcal{T}_j$ and $((t+1) \bmod 2) + 1 = \overline{j}$. Here, $\|\cdot\|_2$ denotes the spectral norm [36]. By Corollary 1, we have

$$\begin{aligned}
\zeta &\overset{(a)}{=} \left\| \operatorname{diag}\left(\left(\frac{1}{4} \left(1 - \frac{\theta_{1,K_j} T_{j,n}(t+2)}{\theta_{2,j,K_j} T_{\overline{j},n}(t+1) + \theta_{3,j,K_j}} \right) \right.\right. \right. \\
&\quad \left.\left.\left. \times \left(1 - \frac{\theta_{1,K_{\overline{j}}} T_{\overline{j},n}(t+1)}{\theta_{2,\overline{j},K_{\overline{j}}} T_{j,n}(t) + \theta_{3,\overline{j},K_{\overline{j}}}} \right) \right)_{n \in \mathcal{N}} \right) \right\|_2 \\
&\overset{(b)}{=} \max_{n \in \mathcal{N}} \left| \frac{1}{4} \left(1 - \frac{\theta_{1,K_j} T_{j,n}(t+2)}{\theta_{2,j,K_j} T_{\overline{j},n}(t+1) + \theta_{3,j,K_j}} \right) \right. \\
&\quad \left. \times \left(1 - \frac{\theta_{1,K_{\overline{j}}} T_{\overline{j},n}(t+1)}{\theta_{2,\overline{j},K_{\overline{j}}} T_{j,n}(t) + \theta_{3,\overline{j},K_{\overline{j}}}} \right) \right| \\
&\overset{(c)}{\leq} \frac{1}{4} \max_{n \in \mathcal{N}} \left| 1 - \frac{\theta_{1,K_j} T_{j,n}(t+2)}{\theta_{2,j,K_j} T_{\overline{j},n}(t+1) + \theta_{3,j,K_j}} \right| \\
&\quad \times \max_{n \in \mathcal{N}} \left| 1 - \frac{\theta_{1,K_{\overline{j}}} T_{\overline{j},n}(t+1)}{\theta_{2,\overline{j},K_{\overline{j}}} T_{j,n}(t) + \theta_{3,\overline{j},K_{\overline{j}}}} \right| \\
&\overset{(d)}{\leq} \frac{1}{4} \max\left\{1, \left| 1 - \frac{\theta_{1,K_j}}{\theta_{3,j,K_j}} \right|\right\} \max\left\{1, \left| 1 - \frac{\theta_{1,K_{\overline{j}}}}{\theta_{3,\overline{j},K_{\overline{j}}}} \right|\right\} < 1,
\end{aligned}$$

where (a) is obtained by the definition of second-order derivative, (b) is obtained by the definition of spectral norm, (c) is obtained based on the formula $\max|x_n y_n| \leq \max|x_n| \cdot \max|y_n|$, $n \in \mathcal{N}$ and (d) is obtained due to

$$1 - \frac{\theta_{1,K_j}}{\theta_{3,j,K_j}} \leq 1 - \frac{\theta_{1,K_j} T_{j,n}(t+2)}{\theta_{2,j,K_j} T_{\overline{j},n}(t+1) + \theta_{3,j,K_j}} \leq 1, \quad (35)$$

$$1 - \frac{\theta_{1,K_{\overline{j}}}}{\theta_{3,\overline{j},K_{\overline{j}}}} \leq 1 - \frac{\theta_{1,K_{\overline{j}}} T_{\overline{j},n}(t+1)}{\theta_{2,\overline{j},K_{\overline{j}}} T_{j,n}(t) + \theta_{3,\overline{j},K_{\overline{j}}}} \leq 1. \quad (36)$$

Therefore, by Theorem 1 of [23], we can prove Theorem 3.

References

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2018.2807445, IEEE Transactions on Communications

$$\Pr\left[K_{j,n,0}=k \mid j_0=j\right] \approx \sum_{i\in\mathcal{I}_{j,n}} \frac{p_{j,i}}{T_{j,n}} \sum_{\mathcal{X}\in\{\mathcal{S}\subseteq\mathcal{N}_{j,i,-n}:|\mathcal{S}|=k-1\}} \prod_{m\in\mathcal{X}} \left(1-\Pr[Y_{j,m,n,i}=0]\right) \prod_{m\in\mathcal{N}_{j,i,-n}\setminus\mathcal{X}} \Pr[Y_{j,m,n,i}=0]. \quad (32)$$

$$\Bigg\| (\nabla^2_{\mathbf{T}_j^2} q_{j,\infty}\left(\mathbf{T}_j,\mathbf{T}_{\overline{j}}\right))^{-1} \nabla^2_{\mathbf{T}_j\mathbf{T}_{\overline{j}}} q_{j,\infty}\left(\mathbf{T}_j,\mathbf{T}_{\overline{j}}\right)\Big|_{\mathbf{T}_j=\mathbf{T}_j(t+2),\mathbf{T}_{\overline{j}}=\mathbf{T}_{\overline{j}}(t+1)}$$
$$\times (\nabla^2_{\mathbf{T}_{\overline{j}}^2} q_{\overline{j},\infty}\left(\mathbf{T}_{\overline{j}},\mathbf{T}_j\right))^{-1} \nabla^2_{\mathbf{T}_{\overline{j}}\mathbf{T}_j} q_{\overline{j},\infty}\left(\mathbf{T}_{\overline{j}},\mathbf{T}_j\right)\Big|_{\mathbf{T}_j=\mathbf{T}_j(t),\mathbf{T}_{\overline{j}}=\mathbf{T}_{\overline{j}}(t+1)}\Bigg\|_2 = \zeta < 1. \quad (33)$$

[2] 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.

[3] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016, pp. 1–6.

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

[5] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

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

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

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

[9] S. Tamoor-ul Hassan, M. Bennis, P. H. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," in *Wireless Communication Systems (ISWCS), 2015 International Symposium on*. IEEE, 2015, pp. 765–769.

[10] J. Zhang, X. Zhang, M. A. Imran, B. Evans, and W. Wang, "Energy efficiency analysis of heterogeneous cache-enabled 5g hyper cellular networks," in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016, pp. 1–6.

[11] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Transactions on Vehicular Technology*, 2016.

[12] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," *arXiv preprint arXiv:1701.05761*, 2017.

[13] 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.

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

[15] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*. IEEE, 2017, pp. 1–6.

[16] Y. Tan, Y. Yuan, T. Yang, Y. Xu, and B. Hu, "Femtocaching in wireless video networks: Distributed framework based on exact potential game," in *Proc. IEEE ICCC*. IEEE, 2016, pp. 1–6.

[17] K. Poularakis, G. Iosifidis, and L. Tassiulas, "A framework for mobile data offloading to leased cache-endowed small cell networks," in *Mobile Ad Hoc and Sensor Systems (MASS), 2014 IEEE 11th International Conference on*. IEEE, 2014, pp. 327–335.

[18] F. Shen, K. Hamidouche, E. Baştuğ, and M. Debbah, "A stackelberg game for incentive proactive caching mechanisms in wireless networks," *arXiv preprint arXiv:1609.02596*, 2016.

[19] 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.

[20] J. Li, J. Sun, Y. Qian, F. Shu, M. Xiao, and W. Xiang, "A commercial video-caching system for small-cell cellular networks using game theory," *IEEE Access*, vol. 4, pp. 7519–7531, 2016.

[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. 16, no. 1, pp. 250–264, Jan 2017.

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

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

[24] 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.

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

[26] S. H. Chae, J. Y. Ryu, T. Q. Quek, and W. Choi, "Cooperative transmission via caching helpers," in *Global Communications Conference (GLOBECOM), 2015 IEEE*. IEEE, 2015, pp. 1–6.

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

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

[29] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[30] G. Scutari, D. Palomar, and S. Barbarossa, "Competitive design of multiuser mimo systems based on game theory: A unified view," *IEEE J. Select. Areas Commun.*, vol. 26, no. 7, 2008.

[31] M. J. Osborne and A. Rubinstein, *A course in game theory*. MIT press, 1994.

[32] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

[33] B. B. Nagaraja and K. G. Nagananda, "Caching with unknown popularity profiles in small cell networks," in *Global Communications Conference (GLOBECOM), 2015 IEEE*. IEEE, 2015, pp. 1–6.

[34] 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, 2013.

[35] S. Li and T. Başar, "Distributed algorithms for the computation of noncooperative equilibria," *Automatica*, vol. 23, no. 4, pp. 523–533, 1987.

[36] D. G. Luenberger, *Optimization by vector space methods*. John Wiley & Sons, 1969.