When Exploiting Individual User Preference Is Beneficial for Caching at Base Stations

Dong Liu*, Chenyang Yang*, and Victor C.M. Leung†

*School of Electronic and Information Engineering, Beihang University, Beijing, China

†Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, BC, Canada

Email: {dliu, cyyang}@buaa.edu.cn, vleung@ece.ubc.ca

***Abstract*—Most of prior works optimize caching policies based on the following assumptions: 1) every user initiates request according to *content popularity*, 2) all users are with the same activity level, and 3) users are uniformly located in the considered region. In practice, these assumptions are often not true. In this paper, we explore the benefit of optimizing caching policies for base stations by exploiting *user preference* considering the spatial locality and different activity level of users. We obtain optimal caching policies, respectively minimizing the download delay averaged over all file requests and user locations in the network (namely *network average delay*), and minimizing the maximal weighted download delay averaged over the file requests and location of each user (namely *maximal weighted user average delay*), as well as minimizing the weighted sum of both. The analysis and simulation results show that exploiting heterogeneous user preference and activity level can improve user fairness, and can also improve network performance when users are with spatial locality.**

I. Introduction

By caching popular contents at base stations (BSs), user experience, network throughput, and energy efficiency can be improved remarkably [1–3].

To achieve high performance with limited cache size at wireless edge, optimizing proactive caching is critical by harnessing the knowledge of which and where the contents will be requested. In an early work [4], caching policy was optimized to minimize the average download delay assuming that the exact location where each user sends the file request is known *a priori*. Considering the uncertainty in where the users will send requests, a probabilistic caching policy maximizing the cache-hit probability was proposed in [5]. In the literature of wireless caching, the knowledge of which contents will be demanded is commonly interpreted as content popularity. As a result, most of prior works optimize caching policies based on *content popularity* [4–10].

However, as a demand statistic of multiple users, content popularity cannot reflect the demand statistic of each individual user. In fact, global content popularity observed at a large aggregation point (say a content server) cannot reflect local content popularity observed in a small region (say a campus [11] or a cell [12]), not to mention the preference of each user. These existing works implicitly assume that the preferences are identical among users in a region [4–9] or in a social group [10] and are equal to the content popularity. This inevitably degrades the caching gain, since the assumption is not true in practice.

In real-world networks, user preferences are heterogeneous, which can be learned from collaborative filtering (CF) based on users' rating or request history [13]. By assuming user preferences as Zipf distributions with different ranks, caching policy was optimized to minimize the average download delay in [14]. Yet the user locations were assumed unchanged during the period of content placement and content delivery and all the users were assumed to have identical activity level. In practice, the location of mobile users is neither known in advance as assumed in [1, 4, 7, 14], nor completely unknown (hence randomly distributed throughout the network) as assumed in [5, 6, 8, 9]. The data measured from mobile connections in [15, 16] showed that more than one third of the users visit only one cell and over 90% of the users travel across less than 10 BSs in one day, which indicates strong spatial locality of users. This suggests that the probability that a user is located in a cell when sending file request can be learned from the request history. Moreover, the activity level of users is highly heterogeneous, e.g., about 80% of the daily network traffic is generated by only 20% of the users [16].

In this paper, we analyze when optimizing caching policy with individual user preference is beneficial. Taking the spatial locality and different activity levels of users into account, we first derive the average delay for each user. We then minimize the network average delay, and show that exploiting user preference can improve network performance when users send requests with high probabilities in some cells. Noticing that user fairness issue appears when different users prefer the BSs to cache different files due to diverse user preference, we minimize the maximal weighted average delay among all the users, and show that caching policy can improve user fairness when user preferences are exploited.

The rest of the paper is organized as follows. In section II, we first introduce the system model, caching policy, and then connect content popularity with user preference. In section III, we optimize the caching policy with user preference, show when using user preference is beneficial, and use a toy example to help understand the impact of user preference heterogeneity. In section IV, simulation results are provided. Section V concludes the paper.

This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant 61731002, 61671036, and 61429101.

II. System Model and User Demand Statistics

We consider a cache-enabled wireless network with cell radius D, where N_b BSs serve N_u users. Each BS is equipped with N_t antennas and a cache with size N_c, and is connected to the core network via backhaul. The content library consists of N_f files each with size F that all the users in the considered region may request. Each user is allowed to associate with one of the three nearest BSs (called *neighboring BS set*) to download the requested file in order to increase the cache-hit probability. For example, when a user is located in the shaded area of Fig. 1, it can associate with BS_1, BS_2 or BS_3, where BS_1 is called the local BS of the user.[1] To avoid strong inter-cell interference inside the neighboring BS sets, the BSs within the neighboring BSs set use different frequency bands as shown in Fig. 1. To reflect the spatial locality of each user, we denote $\mathbf{A} = [a_{uj}]_{Nu \times N_b}$ as the location probability matrix, where a_{uj} is the probability that the uth user is located in the jth cell when it sends file request. Since the exact location of users in a cell is hard to predict, we assume that the user is uniformly located within a cell when it is located in the cell.

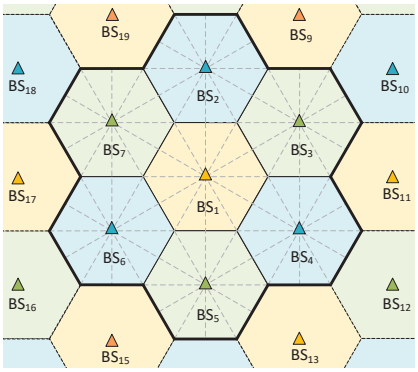


Fig. 1. Layout of the cache-enabled network. The considered region are surrounded by solid line. In this example, $N_b = 7$.

A. Caching Policy and Download Delay

To achieve better performance, we employ coded caching strategy [4, 9] where each file is encoded by rateless maximum distance separable coding so that a file can be retrieved by a user when F bits of the requested file is received by the user. Denote c_{bf} $(0 \le c_{bf} \le 1)$ as the fraction of the fth file cached at the bth BS, and b_u^l as the lth nearest BS of the uth user when the user receives file at the location of $\mathbf{x}_u = (x_{u1}, x_{u2})$.

When $\sum_{l=1}^{k-1} c_{b_u^l f} < 1$ and $\sum_{l=1}^{k} c_{b_u^l f} \ge 1$, the uth user needs to receive the fth file from the 1st, $\cdots$, kth nearest BSs[2] successively to retrieve the complete file. Then, the delay of the uth user that locating at $\mathbf{x}_u$ and downloading the fth file averaged over small-scale fading can be expressed as a piecewise function

$$t_u^f(\mathbf{x}_u) = \begin{cases} t_u^{f1}(\mathbf{x}_u), & c_{b_u^1 f} \ge 1 \\ t_u^{f2}(\mathbf{x}_u), & \sum_{l=1}^{1} c_{b_u^l f} < 1 \text{ and } \sum_{l=1}^{2} c_{b_u^l f} \ge 1 \\ t_u^{f3}(\mathbf{x}_u), & \sum_{l=1}^{2} c_{b_u^l f} < 1 \text{ and } \sum_{l=1}^{3} c_{b_u^l f} \ge 1 \\ t_u^{f4}(\mathbf{x}_u), & \sum_{l=1}^{3} c_{b_u^l f} < 1 \end{cases} \tag{1}$$

where

$$t_u^{fk}(\mathbf{x}_u) = F \sum_{l=1}^{k-1} c_{b_u^l f} \tau_{u b_u^l}(\mathbf{x}_u) + F\Big(1 - \sum_{l=1}^{k-1} c_{b_u^l f}\Big) \tau_{u b_u^k}(\mathbf{x}_u) \tag{2}$$

and $\tau_{u b_u^l}(\mathbf{x}_u)$ is the per-bit download delay of the uth user when downloading from its lth nearest BS averaged over small-scale fading.

We assume block Rayleigh fading channel, which is constant in each block and independently and identically distributed among blocks. Then, the per-bit download delay can be derived as [14]

$$\tau_{u b_u^l}(\mathbf{x}_u) = \frac{1}{\bar{R}_{u b_u^l}(\mathbf{x}_u)} \tag{3}$$

where $\bar{R}_{ub}(\mathbf{x}_u)$ is the achievable rate averaged over small-scale fading for the uth user that downloading from the bth BS. To unify the expression, we denote the download delay when the uth user downloading from the backhaul as $\tau_{u b_u^4}$. Since cache is intended for networks with stringent capacity backhaul [1], we assume that the download delay is limited by the backhaul bandwidth when the user downloads file from the backhaul. Then, we have $\tau_{u b_u^4} = \frac{1}{C_{\mathrm{bh,u}}}$, where $C_{\mathrm{bh,u}}$ is the backhaul bandwidth for the uth user.

To emphasize how to optimize caching policy exploiting user preference, we assume that each BS serves N_t users in the same time-frequency resource by zero-forcing beamforming with equal power allocation, then the average achievable rate can be expressed as

$$\bar{R}_{ub}(\mathbf{x}_u) = \mathbb{E}_h\left[W_u \log_2\left(1 + \frac{\frac{P_t}{N_t} h_{ub} r_{ub}^{-\alpha}}{\sum_{b' \in \Phi(b), b' \ne b} P_t h_{ub'} r_{ub'}^{-\alpha} + \sigma^2}\right)\right] \tag{4}$$

where W_u is the transmission bandwidth for the uth user, P_t is the transmit power of each BS, h_{ub} is the equivalent channel gain (including channel coefficient and beamforming) from the bth BS to the uth user, $r_{ub} = ||\mathbf{x}_u - \mathbf{x}_b||$ is the distance between the uth user and the bth BS, α is the pathloss exponent, Φ_b denotes the set of BSs that share the same frequency with the bth BS, σ^2 is the noise power, $\frac{P_t}{N_t} h_{ub} r_{ub}^{-\alpha}$ and $\sum_{b' \in \Phi(b), b' \ne b} P_t h_{ub'} r_{ub'}^{-\alpha}$ are the signal power and interference power, respectively.

B. Content Popularity and User Preference

We denote $\mathbf{p} = [p_1, \cdots, p_{N_f}]$ as *global content popularity*, where p_f is the probability that the fth file is requested by all users in the considered region. We denote $p_{f|j}$ as the *local content popularity* of the fth file in the jth cell, which is the probability that the fth file is requested by all users in the jth cell and reflects the user demands observed within a cell.

[1] The framework can be extended to neighboring BS sets with any number of BSs. We choose three only for illustration.

[2] To unify the expression, we refer the backhaul to as the 4th nearest "BS".

We denote $\mathbf{Q} = [\mathbf{q}_1^T, \cdots, \mathbf{q}_{N_u}^T]^T$ as *user preference* matrix, where $\mathbf{q}_u = [q_{1|u}, \cdots, q_{N_f|u}]$ is the preference of the uth user and $q_{f|u} \in [0,1]$ is the conditional probability that the uth user requests the fth file given that it requests a file. User preference reflects the demands of each individual user.

Based on the law of total probability, the global content popularity can be connected with user preference as

$$p_f = \sum_{u=1}^{N_u} s_u q_{f|u} \triangleq \sum_{u=1}^{N_u} q_{uf} \tag{5}$$

where s_u is the probability that the request is sent from the uth user, which reflects the *activity level* of the user, and q_{uf} is the joint probability that the requested file is the fth file and the request is sent from the uth user. We denote $\mathbf{s} = [s_1, \cdots, s_{N_u}]$ as the user activity level vector.

Further considering the user location probability $\mathbf{A}$, the local content popularity of the fth file in the jth cell can be connected with user preference as

$$p_{f|j} = \frac{\sum_{u=1}^{N_u} a_{uj} s_u q_{f|u}}{\sum_{f=1}^{N_f}\sum_{u=1}^{N_u} a_{uj} s_u q_{f|u}} = \frac{\sum_{u=1}^{N_u} a_{uj} q_{uf}}{\sum_{u=1}^{N_u} a_{uj} s_u} \tag{6}$$

Both $\mathbf{Q}$ and $\mathbf{s}$ can be learned by CF at a service gateway [17, 18], which are assumed perfect in the following analysis.

III. Caching Policy Optimization With User Preference

In practice, the exact location where each user sends the file request is unknown in advance when optimizing the caching policy. Therefore, we first derive the delay of each user averaged over its possible locations and file requests.

To derive the user average delay, we divide each cell into 12 sectors as shown in Fig. 1. In this way, the lth nearest BS of the uth user, i.e., b_u^l, does not depend on $\mathbf{x}_u$ any more given that the user is located in the ith sector of the jth cell. Then, based on the law of total expectation, the average delay of the uth user can be obtained by the following proposition

Proposition 1: The download delay of the uth user averaged over all its possible requests and locations is

$$\bar{t}_u = \sum_{j=1}^{N_b}\sum_{i=1}^{12}\sum_{f=1}^{N_f} \frac{a_{uj} q_{f|u}}{12} \max_{k=1,\cdots,4} \left\{ F\bar{\tau}_{uk} - F\sum_{l=1}^{k-1} c_{b_{ij}^l f}(\bar{\tau}_{uk} - \bar{\tau}_{ul}) \right\} \tag{7}$$

where $\bar{\tau}_{ul} \approx \frac{2\sqrt{3}}{W_u D^2}\int_0^D \int_0^{\frac{x_2}{\sqrt{3}}} \frac{1}{\log_2 \frac{k_x}{k_y} + \frac{1}{\ln 2}(\psi(\theta_x) - \psi(\theta_y))} \mathrm{d}x_1 \mathrm{d}x_2$ is the per-bit download delay of the uth user when downloading from the lth nearest BS averaged over the uth user's location, k_x, k_y, θ_x and θ_y are given in the Appendix.

Proof: See Appendix. ∎

A. Caching Policy Optimization

Network average delay is the delay averaged over the requests of all the users in the considered region. This is a performance metric from the network perspective and is widely used in literature [4, 10, 14], which can be expressed as $T = \sum_{u=1}^{N_u} s_u \bar{t}_u$.

To capture user fairness, we consider the weighted user average delay $\max_{u=1,\cdots,N_u}\{w_u \bar{t}_u\}$. Considering that the users with more file requests will suffer more if they have longer delay, we can set w_u as an increasing function of the user activity level s_u. As an illustration, we set $w_u = N_u s_u$ in the sequel. Then, the weighted user average delay can be expressed as $w_u \bar{t}_u = N_u s_u \bar{t}_u$.

To improve both network performance and user fairness, we formulate the following general optimization framework minimizing the weighted sum of these two metrics as

$$\min_{c_{bf}} \quad (1-\eta)T + \eta \max_{u=1,\cdots,N_u}\{N_u s_u \bar{t}_u\} \tag{8a}$$

$$s.t. \quad \sum_{f=1}^{N_f} c_{bf} \le N_c, \ \forall b \tag{8b}$$

$$0 \le c_{bf} \le 1, \ \forall f, b \tag{8c}$$

By changing the value of η from 0 to 1, we can obtain the caching policy from minimizing the network average delay (refer to as *Problem 1*) to minimizing the maximal weighted user average delay (refer to as *Problem 2*). By introducing auxiliary variables μ_{ij}^{uf} and ν, which are upper bounds of $\{\bar{\tau}_{ul} - \sum_{l=1}^{k-1} c_{b_{ij}^l f}(\bar{\tau}_{uk} - \bar{\tau}_{ul})\}_{k=1,\cdots,4}$ and $\{N_u s_u \bar{t}_u\}_{u=1,\cdots,N_u}$, respectively, we can convert the problem equivalently into

$$\min_{c_{bf}, \mu_{ij}^{uf}, \nu} (1-\eta)\frac{F}{12}\sum_{j=1}^{N_b}\sum_{i=1}^{12}\sum_{f=1}^{N_f}\left(\sum_{u=1}^{N_u} a_{uj} q_{uf}\right)\mu_{ij}^{uf} + \eta\nu \tag{9a}$$

$$s.t. \quad \bar{\tau}_{uk} - \sum_{l=1}^{k-1} c_{b_{ij}^l f}(\bar{\tau}_{uk} - \bar{\tau}_{ul}) \le \mu_{ij}^{uf}, \forall \ i, j, u, f, k \tag{9b}$$

$$N_u \sum_{j=1}^{N_b}\sum_{i=1}^{12}\sum_{f=1}^{N_f} a_{uj} q_{uf} \mu_{ij}^{uf} \le \nu, \ \forall u \tag{9c}$$

$$\sum_{f=1}^{N_f} c_{bf} \le N_c, \ \forall b \tag{9d}$$

$$0 \le c_{bf} \le 1, \ \forall f, b \tag{9e}$$

which is a linear programming problem and can be solved by interior point method [19]. We refer the optimal caching policies for Problem 1 and Problem 2 to as *Policy* 1 and *Policy* 2, respectively.

B. Analysis for Special Cases

Since transmission and caching resource allocation operated in very different time-scales, to focus on the difference brought by exploiting user preference, we consider the special cases where transmission resources are identical for each user (i.e., $\bar{\tau}_{1l} = \cdots = \bar{\tau}_{N_u l} \triangleq \bar{\tau}_l$) in the following. Depending on whether the coverage areas of BSs are overlapped, we analyze Policy 1 and Policy 2 in two scenarios.

No matter the coverage of adjacent BSs overlap or not, we can obtain Corollaries 1 and 2 in the following.

Corollary 1: When each user sends request in uniform-distributed locations throughout the network, exploiting user preference cannot improve network average delay.

Proof: In this case, we have $a_{u1} = \cdots = a_{uN_b} = \frac{1}{N_b}$, $\bar{\tau}_{1l} = \cdots = \bar{\tau}_{N_u l}$, and $\mu_{ij}^{1f} = \cdots = \mu_{ij}^{N_u f} \triangleq \mu_{ij}^{f}$. Then, the first term in (9a) can be rewritten as

$$\frac{F}{12N_b}\sum_{j=1}^{N_b}\sum_{i=1}^{12}\sum_{f=1}^{N_f}\Big(\sum_{u=1}^{N_u} q_{uf}\Big)\mu_{ij}^{f} = \frac{F}{12N_b}\sum_{j=1}^{N_b}\sum_{i=1}^{12}\sum_{f=1}^{N_f} p_f \mu_{ij}^{f}$$

where we use the relation in (5). We can see that the network average delay only depends on global content popularity p_f. ∎

Corollary 2: When the location probabilities and preferences are identical for all users, Policies 1 and 2 are identical.

Proof: In this case, since $\bar{\tau}_{1l} = \cdots = \bar{\tau}_{N_u l}$, $a_{1j} = \cdots = a_{N_u j}$ for all j, and $q_{f|1} = \cdots = q_{f|N_u}$ for all f, we can see from (7) that the average delay of each user is identical, i.e. , $\bar{t}_1 = \cdots = \bar{t}_{N_u} \triangleq \bar{t}$. Then, both Problem 1 and Problem 2 are equivalent to minimizing $\bar{t}$. ∎

From the corollaries we can conclude that if the transmission resources are identical for all users, the gain of exploiting user preference in terms of network average delay will vanish without user spatial locality. If location distributions and preferences are further identical for all users, the maximal weighted user average delay can be minimized by simply minimizing the average network delay.

In sparse networks where the coverage of adjacent BSs do not overlap, the average delay of the uth user degenerates into

$$\bar{t}_u = F\sum_{j=1}^{N_b}\sum_{f=1}^{N_f} a_{uj} q_{f|u}(c_{jf}\bar{\tau}_1 + (1-c_{jf})\bar{\tau}_4) \qquad (10)$$

Corollary 3: When the coverage areas of BSs are non-overlapped, Policy 1 is to let each BS cache the most popular files according to local content popularity.

Proof: In this case, considering (6) and (10), we can obtain $T = F\sum_{j=1}^{N_b}(\sum_{u=1}^{N_u} a_{uj}s_u)(\sum_{f=1}^{N_f} p_{f|j}(c_{jf}\bar{\tau}_1 + (1-c_{jf})\bar{\tau}_4))$. Then, minimizing T is equivalent to minimizing $\sum_{f=1}^{N_f} p_{f|j}(c_{jf}\bar{\tau}_1 + (1-c_{jf})\bar{\tau}_4)$ for each cell, $j = 1, \cdots, N_b$, which can be rewritten as $\sum_{f=1}^{N_f} p_{f|j}\bar{\tau}_4 - \sum_{f=1}^{N_f} p_{f|j}c_{jf}(\bar{\tau}_4 - \bar{\tau}_1)$. Since $\bar{\tau}_4 > \bar{\tau}_1$, it is easy to see that the optimal caching policy is to let each cell cache the N_c complete files with the highest values of $p_{f|j}$. ∎

Corollary 4: When the coverage areas of BSs are non-overlapped and user preference is identical, Policy 1 and Policy 2 are the same.

Proof: In this case, $\bar{t}_u = a_{uj}F\sum_{j=1}^{N_b}\sum_{f=1}^{N_f} p_f(c_{jf}\bar{\tau}_1 + (1-c_{jf})\bar{\tau}_4)$. Then, both minimizing $\max_{u=1,\cdots,N_u}\{N_u s_u \bar{t}_u\}$ and minimizing $T = \sum_{u=1}^{N_u} s_u\bar{t}_u$ are equivalent to minimizing $\sum_{j=1}^{N_b}\sum_{f=1}^{N_f} p_f(c_{jf}\bar{\tau}_1 + (1-c_{jf})\bar{\tau}_4)$. Therefore, Policy 1 and Policy 2 are the same. ∎

From Corollary 3 and Corollary 4, we can conclude that if the transmission resources are identical for users and the cells are not overlapped, using local content popularity will be enough to obtain the minimal network average delay as used in [12, 18]. Otherwise, user preference should be exploited to minimize the network average delay. If user preference is further identical, the maximal weighted user average delay can be minimized by simply minimizing the average network delay. Otherwise, caching policies should be designed more sophisticatedly to address user fairness issue.

C. Numerical Examples

To understand the behavior of Policy 1 and Policy 2, and analyze the impact of heterogeneous user preference, we present a simple numerical example as shown in Fig. 2.

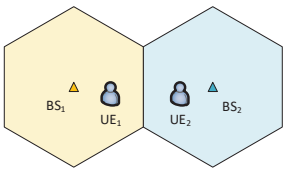


Fig. 2. A toy example, $N_b = N_u = 2$. The total number of files is $N_f = 3$ with size $F = 1$ and each BS can cache $N_c = 1$ file. Global content popularity is $\mathbf{p} = [0.46, 0.30, 0.24]$ and activity level is $\mathbf{s} = [0.6, 0.4]$.

Suppose each user can either associate with BS_1 or BS_2 to download files, and the average per-bit download delay when downloading from the nearest BS, second nearest BS and the backhaul is $[\bar{\tau}_{u1}, \bar{\tau}_{u2}, \bar{\tau}_{u4}] = [1, 2, 3]$ for both user equipments (UEs). We compare two cases with homogeneous and heterogeneous user preference, respectively, where both $\mathbf{Q}^{\text{hom}}$ and $\mathbf{Q}^{\text{het}}$ satisfy (5) with given $\mathbf{p}$ and $\mathbf{s}$.

$$\mathbf{Q}^{\text{hom}} = \begin{bmatrix} 0.46 & 0.30 & 0.24 \\ 0.46 & 0.30 & 0.24 \end{bmatrix},\ \mathbf{Q}^{\text{het}} = \begin{bmatrix} 0.75 & 0.25 & 0 \\ 0.02 & 0.38 & 0.60 \end{bmatrix}$$

For homogeneous user preference $\mathbf{Q}^{\text{hom}}$, we can obtain the results of Policy 1 $\mathbf{C}_1$ and the minimized network average delay T^*, the results of Policy 2 $\mathbf{C}_2$ and the minimized maximal weighted user average delay $\max\{N_u s_1 \bar{t}_1^{\dagger}, N_u s_2 \bar{t}_2^{\dagger}\}$ as

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ T^* = s_1\bar{t}_1^* + s_2\bar{t}_2^* = 1.07 + 0.77 = 1.84$$

$$\mathbf{C}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},\ \max\{N_u s_1 \bar{t}_1^{\dagger}, N_u s_2 \bar{t}_2^{\dagger}\} = 2.13$$

In this case, the *cache interests* of both users are exactly the opposite, i.e., UE_1 prefers its local BS (i.e., BS_1) to cache its most preferable file (i.e., file 1) and its neighboring BS (i.e., BS_2) to cache its second preferable file (i.e., file 2), while UE_2 prefers BS_2 to cache file 1 and BS_1 to cache file 2 according to its own preference. Since UE_1 has higher activity level, both Policies 1 and 2 let BSs cache file according to UE_1's cache interest and $\mathbf{C}_1 = \mathbf{C}_2$, which agrees with Corollary 2.

For heterogeneous user preference $\mathbf{Q}^{\text{het}}$, we can obtain

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ T^* = s_1\bar{t}_1^* + s_2\bar{t}_2^* = 0.90 + 0.71 = 1.61$$

$$\mathbf{C}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.57 & 0.43 \end{bmatrix},\ \max\{N_u s_1 \bar{t}_1^{\dagger}, N_u s_2 \bar{t}_2^{\dagger}\} = 1.63$$

In this case, UE_1 prefers BS_1 to cache file 1 and BS_2 to cache file 2, while UE_2 prefers BS_2 to cache file 3 and BS_1

to cache file 2. As a result, Policy 1 lets each BS cache the most preferable file of its local user, i.e. BS_1 caches file 1 and BS_2 caches file 3. As the user with higher activity level, UE_1 has the maximal weighted average delay (i.e., $0.90 > 0.71$). Hence, Policy 2 is more prone to let BSs cache the files preferred by UE_1, i.e., let BS_2 cache 0.57 part of file 2 and 0.43 part of file 3. We can see that both the average network delay and the maximal weighted average delay decrease compared to the case with homogeneous user preference, which can be explained from the following different perspective.

When user preference become heterogeneous, the most preferable files of users located in different cells differ. File diversity (i.e., caching different files at different BSs) can be naturally achieved by letting each BS cache the most preferable file of its local user, which increases the cache-hit probability. On the contrary, when user preference is identical, there will be no file diversity if each BS caches the most preferable file of its local user, and to achieve file diversity, the cache interest of UE_2 has to be sacrificed.

With given content popularity, the skewness of both users' preferences increase (i.e., the shape of probability distribution $[0.75, 0.25, 0]$ and $[0.02, 0.38, 0.60]$ are more "*peaky*" than $[0.46, 0.30, 0.24]$) when user preferences are less similar, which means that the file requests of users become less uncertain. Analogously to the widely recognized result that the performance of content popularity based caching policies improves with the skewness of popularity, the performance of user preference based caching policies improves with the skewness of user preference.

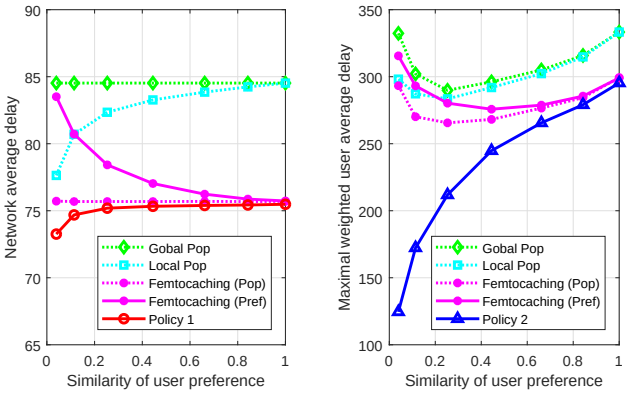
IV. Simulation Results

In this section, we compare the performance of the proposed caching policies with prior works that are based on content popularity, and analyze the impact of various factors by simulation.

We consider $N_b = 7$ cells each with radius $D = 250$ m as shown in Fig. 1, and $N_u = 100$ users. Each BS is with four antennas and with transmit power 46 dBm. The pathloss is modeled as $35.5 + 37.6 \log_{10}(r_{ub})$. The backhaul bandwidth and the downlink transmission bandwidth for each user are set as $C_{\mathrm{bh},u} = 2$ Mbps and $W_u = 5$ MHz, respectively. The probability distribution for the users located in different cells when sending requests is modeled as Zipf distribution with skewness parameter $\delta_a = 1$ based on the measured data in [15]. To analyze the impacts of user preference and activity level and fairly compare with prior works, we generate user preferences satisfying the relation in (5) with different level of cosine similarity as defined in [17]. To reduce simulation time, we consider $N_f = 100$ files in total each with size of $F = 30$ MB. We assume that each BS can cache 10% of the total files, i.e., $N_c = 10$. The global content popularity and the activity level are modeled as Zipf distribution with the skewness parameter $\delta_p = 0.6$ and $\delta_s = 0.4$, respectively.

The following baselines are compared with Policy 1 and Policy 2, where the activity levels and user preferences are implicitly assumed identical when designing caching policies for baselines 1) and 3):

1) *"Global Pop"*: Each BS caches the N_c most popular files according to the global content popularity p_f.
2) *"Local Pop"*: Each BS caches the N_c most popular files according to the local content popularity within its cell $p_{f|j}$ given by (6). This is the method used in [12, 18].
3) *"Femtocaching (Pop)"*: This is the caching policy proposed in [4] minimizing the network average delay, which is based on global content popularity assuming that user location is fixed.
4) *"Femtocaching (Pref)"*: We modify the caching policy in [4] to exploit user preference by simply replacing the global content popularity p_f by user preference $q_{f|u}$.



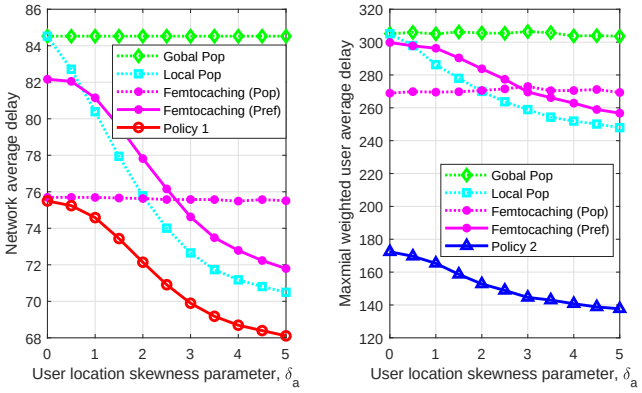
(a) Network performance (b) User fairness

Fig. 3. Impact of user preference similarity.

In Fig. 3(a), we show the impact of user preference similarity on the network average delay (in seconds). It is shown that "Local Pop" can reduce network average delay compared with "Global Pop" when user preferences are heterogeneous. The network average delay of "Femtocaching (Pref)" is even higher than that of "Femtocaching (Pop)" when user preference is less similar. This is because "femtocaching" method does not consider the uncertainty of user location, which has large impact when user preference is less similar. The network average delay of Policy 1 is the lowest as expected, which increases with the preference similarity. This coincides with the results of numerical example in Section III-B.

In Fig. 3(b), we show the impact of user preference similarity on the maximal weighted user average delay (in seconds). We can see that Policy 2 can reduce 60% of the maximal weighted user average delay compared with "Global Pop". Similar to Fig. 3(a), the maximal weighed download delay of "Femtocaching (Pref)" is higher than "Femtocaching (Pop)". The maximal weighted user average delay of Policy 2 is the lowest, which increases with the preference similarity. The explanations are similar to those for numerical results in Section III-B.

In Fig. 4, we show the impact of spatial locality on the two performance metrics. We can see that the benefit of exploiting user preference increases with spatial locality of users. Without spatial locality (i.e., $\delta_a = 0$), the network performance does



(a) Network performance (b) User fairness

Fig. 4. Impact of spatial locality of users. The similarity of user preference is 0.1.

not benefit from exploiting user preference while user fairness can still be improved when comparing Policies 1 and 2 with "Femtocaching (Pop)", respectively.

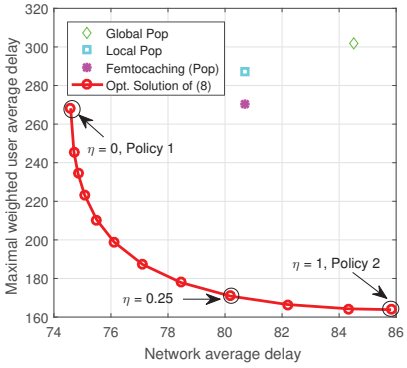


Fig. 5. Tradeoff between performance and user fairness. The similarity of user preference is 0.1.

In Fig. 5, we show the tradeoff between network average download and maximal weighted average download delay by solving problem (8) with different values of η. It is shown that when η is set between 0 and 0.25, the optimal solution achieves lower network average delay and better user fairness than the baseline policies at the same time.

V. Conclusion and Discussion

In this paper, we strived to investigate when and how optimizing caching policy with user preference is beneficial. We showed that the network average delay can be reduced when users are with spatial locality, and user fairness can be improved when user preferences are heterogeneous. Simulation results showed that network performance and user fairness can even be improved at the same time compared with prior works by exploiting heterogeneous user preference and activity level with spatial locality. With given content popularity, the performance gain comes from the facts that cache-hit probability can be improved with less sacrifice of users' cache interests and user demands become less uncertain with more heterogeneous user preference.

It is worthy to mention that learning individual preference of a large number of users can be more computational complex than learning content popularity, and informing the predicted user preference to BSs may incur overhead. In practice, user preference can be learned not very frequently (say each day) at a service gateway or even at a content server that has abundant computing resource. Nevertheless, to harness the benefit of user preference based caching policy, it is worthwhile to investigate how to reduce the complexity and overhead.

Appendix

Based on the law of total expectation, the average delay of the uth user can be expressed as

$$\bar{t}_u = \mathbb{E}_{f,\mathbf{x}_u}\left[t_u^f(\mathbf{x}_u)\right] = \sum_{f=1}^{N_f}\sum_{j=1}^{N_b}\sum_{i=1}^{12}\frac{a_{uj}}{12}\mathbb{E}_{\mathbf{x}_u}\left[t_u^f(\mathbf{x}_u)|ij\right] \quad (11)$$

where $\mathbb{E}_{\mathbf{x}_u}\left[t_u^f(\mathbf{x}_u)|ij\right]$ is the average delay of the user conditioned on that it is located at the ith sector of the jth cell and requesting the fth file, and $\frac{a_{uj}}{12}$ is the probability that the uth user is located at the ith sector of the jth cell. Further considering (1), we can obtain

$$\mathbb{E}_{\mathbf{x}_u}\left[t_u^f(\mathbf{x}_u)|ij\right] = \mathbb{E}_{\mathbf{x}_u}\left[t_u^{fk}(\mathbf{x}_u)|ij\right] \quad (12)$$

when $\sum_{l=1}^{k-1} c_{b_{ij}^l f} < 1$ and $\sum_{l=1}^{k} c_{b_{ij}^l f} \geq 1$, where b_{ij}^l is the lth nearest BS when the user is located in the ith sector of the jth cell. From (2), we can obtain

$$\mathbb{E}_{\mathbf{x}_u}\left[t_u^{fk}(\mathbf{x}_u)|ij\right] = F\bar{\tau}_{uk} - F\sum_{l=1}^{k-1} c_{b_{ij}^l f}(\bar{\tau}_{uk} - \bar{\tau}_{ul}) \quad (13)$$

where $\bar{\tau}_{ul} \triangleq \mathbb{E}_{\mathbf{x}_u}[\tau_{ub_{ij}^l}(\mathbf{x}_u)|ij]$ is the per-bit delay from the lth nearest BS averaged over user location given that the user is located in the ith sector of the jth cell.

Since the average delay increases with the distance between user and BS, we have $\bar{\tau}_{uk} > \bar{\tau}_{u(k-1)}$. Further considering the expressions of (12) and (13), similar to the proof of [4, Lemma 6], we can rewrite (12) as

$$\mathbb{E}_{\mathbf{x}_u}\left[t_u^f(\mathbf{x}_u)|ij\right] = \max_{k=1,\cdots,4}\{\mathbb{E}_{\mathbf{x}_u}\left[t_u^{fk}(\mathbf{x}_u)|ij\right]\} \quad (14)$$

Due to the symmetry of the network topology, $\bar{\tau}_{ul}$ does not depend on i and j but only depend on l and u. Without loss of generality, we derive the average download delay from the three nearest BSs when the uth user located in the shadow area in Fig. 1. From (3), by taking the expectation over $\mathbf{x}_u$ within the shadow area, we have

$$\bar{\tau}_{ul} = 2\sqrt{3}W_u^{-1}D^{-2}\int_0^D\int_0^{\frac{x_{u2}}{\sqrt{3}}}\bar{R}_{ul}(\mathbf{x}_u)^{-1}\mathrm{d}x_{u1}\mathrm{d}x_{u2} \quad (15)$$

Since the interference term in (4) is a weighted sum of Gamma distributed random variables $h_{ub'} \sim \mathbb{G}(N_t, 1/N_t)$, with different values of weight $r_{ub'}^{-\alpha}$, $\bar{R}_{ul}(\mathbf{x}_u)$ has no closed-form expression and the computation requires a $|\Phi_l|$-fold numerical integration that is of high complexity. To reduce computational complexity, we obtain an approximate $\bar{\tau}_{ul}$ for high signal-to-noise ratio (SNR) region.

When $\frac{P_t}{\sigma^2} \to \infty$, we can neglect the impact of σ^2 and (4) can be derived as

$$\begin{aligned}\bar{R}_{ub}(\mathbf{x}_u) &= W_u \mathbb{E}_h[\log_2 X] - W_u \mathbb{E}_h[\log_2 Y] \\ &\approx W_u \mathbb{E}_h[\log_2 \hat{X}] - W_u \mathbb{E}_h[\log_2 \hat{Y}] \\ &= W_u \left(\log_2 \tfrac{k_x}{k_y} + \tfrac{1}{\ln 2}(\psi(\theta_x) - \psi(\theta_x)) \right) \end{aligned} \quad (16)$$

where $X = h_{ub} r_{ub}^{-\alpha} + N_t \sum_{b' \in \Phi(b), b' \neq b} h_{ub'} r_{ub'}^{-\alpha}$, $Y = N_t \sum_{b' \in \Phi(b), b' \neq b} h_{ub'} r_{ub'}^{-\alpha}$, and we approximate X and Y as Gamma distributed random variables $\hat{X} \sim \mathbb{G}(k_x, \theta_x)$ and $\hat{Y} \sim \mathbb{G}(k_y, \theta_y)$, respectively, which is accurate as shown in [20]. The last equation is from $\mathbb{E}[\ln \hat{X}] = \psi(k_x) + \ln(\theta_x)$. By matching the first two moments of X and $\hat{X}$, we can obtain $k_x = \frac{\left(r_{ub}^{-\alpha} + N_t \sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-\alpha}\right)^2}{r_{ub}^{-2\alpha} + N_t \sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-2\alpha}}$, $\theta_x = \frac{r_{ub}^{-2\alpha} + N_t \sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-2\alpha}}{r_{ub}^{-\alpha} + N_t \sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-\alpha}}$, $k_y = \frac{N_t \left(\sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-\alpha}\right)^2}{\sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-2\alpha}}$, and $\theta_y = \frac{\sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-2\alpha}}{\sum_{b' \in \Phi(b), b' \neq b} r_{ub'}^{-\alpha}}$.

Then, by substituting (16) into (15) and then into (13) and further considering (14) and (11), Proposition 1 can be proved.

REFERENCES

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

[2] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[3] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sept. 2016.

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

[5] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *proc. IEEE ICC*, 2015.

[6] 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, July 2016.

[7] Z. Chen, J. Lee, T. Q. S. 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, May 2017.

[8] X. Li, X. Wang, K. Li, Z. Han, and V. C. Leung, "Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design," *IEEE Trans. Wireless Commun.*, early access, 2017.

[9] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[10] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching under heterogeneous file preferences," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 444–457, Jan. 2017.

[11] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Computer networks*, vol. 53, no. 4, pp. 501–514, 2009.

[12] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[13] M. D. Ekstrand, J. T. Riedl, J. A. Konstan *et al.*, "Collaborative filtering recommender systems," *Foundations and Trends® in Human–Computer Interaction*, vol. 4, no. 2, pp. 81–173, 2011.

[14] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in Fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, early access, 2017.

[15] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE INFOCOM*, 2011.

[16] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Topics Comput*, vol. 3, no. 1, pp. 95–106, Mar. 2015.

[17] B. Chen and C. Yang, "Caching policy optimization for D2D communications by learning user preference," in *Proc. IEEE VTC Spring*, 2017.

[18] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: moving from cloud to edge," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 36–42, Sept. 2016.

[19] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[20] D. Jaramillo-Ramrez, M. Kountouris, and E. Hardouin, "Coordinated multi-point transmission with imperfect CSI and other-cell interference," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1882–1896, Apr. 2015.