Cooperative Local Caching and File Sharing under Heterogeneous File Preferences

Yinghao Guo†, Lingjie Duan‡, and Rui Zhang†

†ECE Department, National University of Singapore, e-mail: {yinghao.guo, elezhang}@nus.edu.sg
‡Engineering System and Design Pillar, Singapore University of Technology and Design, e-mail: lingjie_duan@sutd.edu.sg

***Abstract*—Local caching with device-to-device (D2D) communications has been recently introduced as an effective scheme for reducing the average download time of the mobile terminals (MTs). The MTs first cache the files in their local memories and then exchange the files with each other within the vincinity via D2D communications. Prior works have largely overlooked MTs' heterogeneity in file preferences and assume unselfish caching behaviors of the MTs. In this work, we practically divide the MTs into different groups according to their invididual preferences over the files and propose optimal file caching strategies for self-interested MTs to reduce the average file download time. Assuming the knowledge of the social file preference for an intelligent group, we develop the optimal caching strategy for this group by formulating and solving a convex optimization problem. Closed-form solution for the problem is obtained, which is shown to follow a water-filling structure over the files. Finally, numerical examples are presented to show that the selfish caching of a group can be detrimental to both itself and the other intelligent groups.**

I. Introduction

It is estimated that the global mobile data traffic will exceed 24.3 exabytes per month by 2019 [1]. Moreover, the data traffic is increasingly concentrated to hotspot and the wireless network is getting more and more congested, especially during the peak hours [2]. As a solution, *local caching* [3] has been proposed recently, which exploits the memory available at the edges of the wireless network to pro-actively download files in off-peak hours and let the edges share the cached files locally with each other (e.g. via device-to-device (D2D) communications). This can significantly save the download time of the MTs compared with that of downloading the file from remote servers via the base stations (BSs).

Depending on which part of network edge to cache the popular files, the literature of local caching can be generally categorized into two types: *femto-caching* by employing the BSs or femto-cell access points [4]–[6] and *D2D-caching* by employing the MTs [7]–[9]. First, for local caching on the BS-side, [4] considers that femto-BSs cooperatively serve MTs by deterministically caching files for the coded and uncoded scenarios. The drawback of this approach is that the MT mobility information is assumed and the cache needs to be updated accordingly. In [5], an efficient caching protocol is proposed that allows the MTs to dynamically select BSs that can serve the MTs and adaptively adjust the quality of the video. [6] considers a cooperative broadcasting scheme where multiple BSs cooperatively serve MTs using multi-cast beamforming in the cloud radio access network. Second, for local caching on the MT-side, [7] proposes the approach of cooperative D2D-caching and derives the power scaling law of the network capacity with respect to the range of the D2D communications. The fairness issue of MTs in the cooperative D2D-caching is studied in [8] by considering the capacity-outrage trade-off. [9] discusses the effectiveness of cooperative D2D caching under different file request popularity, network size and cache size with multi-hop D2D communications.

Different from the above related works, the main contributions of this paper are summarized as follows:

- *Practical modeling for D2D-caching:* We practically model the MTs with heterogeneous file preference and categorize them into different groups according to their preferences over the files. Moreover, selfish MT behaviors are studied, which are largely overlooked in the prior studies that generally assume MTs are cooperative in nature and will cache unselfishly for the social optimum;
- *D2D-caching without the knowledge of social file preference:* As a benchmark case, MT groups do not know the social file preference and different groups randomly cache files according to their own preferences over the files, regardless of the other groups' preferences. The closed-form solution for the average download time in this case is obtained;
- *D2D-caching with the knowledge of social file preference:* With the knowledge of the social file preference, we formulate the caching problem for an intelligent group as an optimization problem to minimize the average download time. Closed-form optimal solution for the file caching distribution is obtained based on the Karush-Kuhn-Tucker (KKT) conditions, which is shown to have a water-filling structure. The optimal caching policy shows that the intelligent group should not only cache based on the group's own preference, but also exploit the social file preference;
- *Impacts of MTs' selfish behaviors:* By presenting a typical two-group numerical examples, we show that if only one group is intelligent and exploits the other group's file preference, the average download time of the group will decrease. However, if both groups are intelligent and try to exploit the other groups, each group's selfish caching may hurt both itself and the other group.

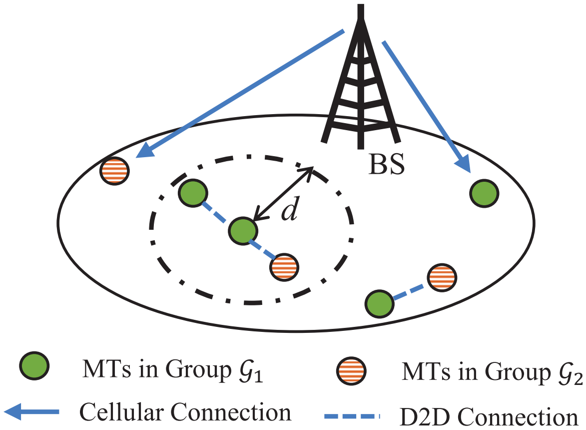

Fig. 1: Local caching and D2D communications enabled file sharing among MTs in two groups.

II. System Model

As illustrated in Fig. 1, we consider a large number of MTs are served by the wireless cellular network and these MTs can also leverage D2D communications to share cached files (e.g., videos) with each other. According to these MTs' file preferences in downloading, we divide them into K groups of MTs, denoted by the set $\mathcal{K} = \{1, 2, \cdots, K\}$ and Fig. 1 shows an example of $K = 2$. Within each group denoted by $\mathcal{G}_k,\ k \in \mathcal{K}$, we assume that the locations of the MTs in the group follow a two-dimensional Homogeneous Poisson Point Process (HPPP) with spatial density $\lambda_k,\ k \in \mathcal{K}$, which is independent from the other groups. We also denote the *social density* of the MTs as density of all the MTs, i.e. $\lambda_0 = \sum_{k\in\mathcal{K}} \lambda_k$. For a certain MT in each group, it can potentially acquire its requested files with *peer MTs* (if any) within the distance d, as denoted by the circle in Fig. 1. Here, d is the range of the D2D communications such as Bluetooth [10].[1] Then, due to our assumption of HPPP, the average number of peer MTs from group $\mathcal{G}_k$ in the range of d is $\mu_k = \pi d^2 \lambda_k,\ k \in \mathcal{K}$.

The benefit of the D2D communications is that it exempts an MT from resorting to the heavily-loaded BS for downloading during the peak hours and thus improves the quality of service (QoS) in terms of latency. However, its efficiency highly depends on the caching strategies of all K groups of MTs, which pre-cache files in their local memories during the off-peak hours (e.g., at night). The general procedure of local caching consists of two phases: *caching phase* and *sharing phase*, which are explained as follows:

A. Caching Phase

During the off-peak hours, MTs cache files into their memories. There are mainly two approaches for file caching, namely *deterministic caching* and *random caching*. Deterministic caching is generally the approach used in femtocaching, where the locations of the BS are fixed. However, this approach is not practical for D2D caching due to the mobility of the MTs and the combinatorial structure of the files. In this paper, we consider random caching for our D2D caching scheme. We define the set of popular files as $\mathcal{F} = \{1, 2, \cdots, F\}$ and assume each MT caches one file.[2] Each MT in group $k \in \mathcal{K}$ caches file $f \in \mathcal{F}$ based on independent random sampling from the *group caching distribution* $\boldsymbol{c}_k = [c_{k,1}, c_{k,2}, \cdots, c_{k,F}]$, which is defined as the probability mass function (PMF) over the files $\mathcal{F}$ with $\sum_{f\in\mathcal{F}} c_{k,f} = 1$. We also integrate the group caching distributions of the other groups $\mathcal{K} \setminus \{k\}$ as $\boldsymbol{c}_{-k} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{k-1}, \boldsymbol{c}_{k+1}, \cdots, \boldsymbol{c}_K]$. Then, given all the group caching distributions ($c_{f,k},\ \forall k \in \mathcal{K},\ f \in \mathcal{F}$) and the densities of the groups ($\lambda_k,\ \forall k \in \mathcal{K}$), we denote the *social caching distribution* as the weighted sum of the group caching distribution with respect to the density of each group:

$$C_f = \frac{1}{\lambda_0} \sum_{k\in\mathcal{K}} \lambda_k c_{k,f},\ \ f \in \mathcal{F}. \tag{1}$$

Actually, C_f has its operational meaning in our D2D local caching scheme that it denotes the average availability of a certain file $f \in \mathcal{F}$ in the range of d for a certain MT.

B. Sharing Phase

During the peak hours, MTs make request to the files based on their preference over the files. Because we divide the MTs into groups by their interests, the MTs within the group $\mathcal{G}_k$ have the common *group request distribution* $\boldsymbol{r}_k = [r_{k,1}, r_{k,2}, \cdots, r_{k,F}]$, which is another set of PMFs defined on $\mathcal{F}$ with $\sum_{f\in\mathcal{F}} r_{k,f} = 1$. In a certain realization, each MT in the group $\mathcal{G}_k,\ k \in \mathcal{K}$ makes independent request to a certain file $f \in \mathcal{F}$ based on a random sampling from $\boldsymbol{r}_k$. We also denote $\boldsymbol{r}_{-k} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_{k-1}, \boldsymbol{r}_{k+1}, \cdots, \boldsymbol{r}_K]$ as the group request distributions of the other groups $\mathcal{K} \setminus \{k\}$. Finally, for the MTs in all groups, we denote the average probability that file $f \in \mathcal{F}$ is requested by all the MTs as the *social request distribution*, which is a weighted sum of the group request distribution with respect to their densities:

$$R_f = \frac{1}{\lambda_0} \sum_{k\in\mathcal{K}} \lambda_k r_{k,f},\ \ f \in \mathcal{F}. \tag{2}$$

The MT can obtain the requested file from its own on-board cache, its peer MTs (in all K groups) via uni-cast D2D communications, or from the BS. Due to the uni-cast property of the D2D transmission, the possibility that the transmitted files are received by multiples MTs is excluded. To avoid any interference, we assume orthogonal transmission such that two D2D links in the vicinity will not interfere with each other, and such communications also will not interfere with the cellular network's downlink transmissions either by operating on orthogonal bands. Furthermore, we consider the static state of the file caching and downloading, which is an one-shot decision problem and the file diffusion and the change of download distribution over time are not considered.

[1] The typical range of the most common Class 2 Bluetooth device is 10 m [10].

[2] We assume each MT only caches a single file for the tractability of analysis. While, our results can be extended to multi-file caching, which offers essentially the same insights.

Under a given realization of file request distribution for a particular MT, we are ready to specify the transmission protocol via D2D communications and cellular downlink. If an MT in group $\mathcal{G}_k,\ k \in \mathcal{K}$ requests a certain file $f \in \mathcal{F}$, the protocol for obtaining the file is given as follows:

- *On-board caching:* If the file $f \in \mathcal{F}$ is already cached in the MT, it can obtain the file with zero delay;
- *D2D-caching:* If the file is not found in its cache, MT requests the file from its peer MTs within the range d, both from its own group or the other groups, and downloads the file via D2D communications with delay $\tau_f^{(D)} > 0$. If multiple peer MTs response to the request, choose one randomly;
- *BS downloading:* Finally, if the above request fails, MT downloads the file from the BS with delay $\tau_f^{(B)} > 0$.

Because the D2D communications is over a short-distance link and the data rate is high, the *transmission delay* of downloading with D2D communications is lower than that from the BS. Furthermore, because the D2D communications is over a direct link between the MTs, the *routing delay* is also much lower that of downloading from the remote BS due to the network congestion, especially during the peak hour [11]. Hence, due to the above two reasons, we assume that the delay for downloading a certain file $f \in \mathcal{F}$ from the BS is larger than that of downloading form peer MTs with D2D communications, i.e. $\tau_f^{(B)} > \tau_f^{(D)}$.

C. Average Download Time for D2D-caching

With the above specified transmission protocol, we are ready to analyze the average download time of each group and the society. We assume the MTs in group $\mathcal{G}_k,\ k \in \mathcal{K}$, initiate their file requests with equal probability. We then specify the *group average download time* of the MTs in group $\mathcal{G}_k$ under the given caching distribution of the other groups $\boldsymbol{c}_{-k}$ with the following theorem.

Theorem 2.1: Given the group caching distribution $\boldsymbol{c}_k$ and that of the other groups $\boldsymbol{c}_{-k}$, the average download time for MTs in group $\mathcal{G}_k,\ k \in \mathcal{K}$, is

$$\begin{aligned}&\mathbb{E}[\mathcal{T}_k(\boldsymbol{c}_k;\boldsymbol{c}_{-k})]\\&=\sum_{f\in\mathcal{F}} r_{k,f}\bar{c}_{k,f}\left(\left[1-e^{-\mu_0 C_f}\right]\tau_f^{(D)}+e^{-\mu_0 C_f}\tau_f^{(B)}\right), \quad (3)\\&=\sum_{f\in\mathcal{F}} r_{k,f}\bar{c}_{k,f}\left(\tau_f^{(D)}+e^{-\mu_0 C_f}\tau_f'\right), \quad (4)\end{aligned}$$

where $e^{-\mu_0 C_f}$, with C_f given in (1), denotes the probability that file $f \in \mathcal{F}$ is not found in the peer MTs within the range d, $\tau_f' = \tau_f^{(B)} - \tau_f^{(D)}$, $\bar{c}_{k,f} = 1 - c_{k,f}$ denotes the complement of the group caching distribution, and $\mu_0 = \pi d^2 \lambda_0$ denotes the average number of peer MTs from all groups within the range d.

Proof: Consider one arbitrary MT $i \in \mathcal{G}_k,\ k \in \mathcal{K}$ and denote its peer MTs in group $\mathcal{G}_k$ as $\mathcal{N}_k^{(i)} \subseteq \mathcal{G}_k,\ k \in \mathcal{K}$. It follows that $|\mathcal{N}_k^{(i)}| = N_k^{(i)}$ is a Poisson random variable with mean $\mu_k = \pi d^2 \lambda_k,\ k \in \mathcal{K}$ and its PMF is given by

$$\mathbb{P}[N_k^{(i)}] = \frac{\mu_k^{N_k^{(i)}}}{N_k^{(i)}!} e^{-\mu_k},\ N_k^{(i)} = 0, 1, \cdots, \infty.$$

Then, given the requested file $f \in \mathcal{F}$ and the peers $\mathcal{N}_k^{(i)} \subseteq \mathcal{G}_k,\ k \in \mathcal{K}$, the average file download time for MT $i \in \mathcal{G}_k$ is

$$\begin{aligned}&\mathbb{E}[\mathcal{T}_k^{(i)}(\boldsymbol{c}_k;\boldsymbol{c}_{-k})|f,\{\mathcal{N}_k^{(i)}\}] = c_{k,f}\cdot 0 + (1-c_{k,f})\\&\times\left(\Big[\prod_{k\in\mathcal{K}}(1-c_{k,f})^{N_k^{(i)}}\Big]\tau_f^{(B)}+\Big[1-\prod_{k\in\mathcal{K}}(1-c_{k,f})^{N_k^{(i)}}\Big]\tau_f^{(D)}\right)\\&=\bar{c}_{k,f}\left(\Big[1-\prod_{k\in\mathcal{K}}\bar{c}_{k,f}^{N_k^{(i)}}\Big]\tau_f^{(D)}+\Big[\prod_{k\in\mathcal{K}}\bar{c}_{k,f}^{N_k^{(i)}}\Big]\tau_f^{(B)}\right),\end{aligned}$$

where $[\prod_{k\in\mathcal{K}}(1-c_{k,f})^{N_k^{(i)}}]$ denotes the probability that file $f \in \mathcal{F}$ is not found in caches of the peer MTs from all groups. Then, the average download time of MT $i \in \mathcal{G}_k$ can be obtained with the theorem of iterated expectation as

$$\begin{aligned}&\mathbb{E}[\mathcal{T}_k^{(i)}(\boldsymbol{c}_k;\boldsymbol{c}_{-k})]\\&=\mathbb{E}_{N_1^{(i)}}[\mathbb{E}_{N_2^{(i)}}[\cdots\mathbb{E}_{N_K^{(i)}}[\mathbb{E}_f[\mathbb{E}[\mathcal{T}_k^{(i)}(\boldsymbol{c}_k;\boldsymbol{c}_{-k})|f,\{\mathcal{N}_k^{(i)}\}]]]]\cdots]]\\&=\sum_{N_1^{(i)}=0}^{\infty}e^{-\mu_1}\frac{\mu_1^{N_1^{(i)}}}{N_1^{(i)}!}\sum_{N_2^{(i)}=0}^{\infty}e^{-\mu_2}\frac{\mu_2^{N_2^{(i)}}}{N_2^{(i)}!}\cdots\sum_{N_K^{(i)}=0}^{\infty}e^{-\mu_K}\frac{\mu_K^{N_K^{(i)}}}{N_K^{(i)}!}\\&\times\bar{c}_{k,f}\left(\Big[1-\prod_{k\in\mathcal{K}}\bar{c}_{k,f}^{N_k^{(i)}}\Big]\tau_f^{(D)}+\Big[\prod_{k\in\mathcal{K}}\bar{c}_{k,f}^{N_k^{(i)}}\Big]\tau_f^{(B)}\right)\\&\overset{(a)}{=}\sum_{f\in\mathcal{F}} r_{k,f}\bar{c}_{k,f}\left(\Big[1-e^{-\mu_0 C_f}\Big]\tau_f^{(D)}+e^{-\mu_0 C_f}\tau_f^{(B)}\right),\end{aligned}$$

where (a) is obtained by utilizing (1) and the Taylor expansion for exponential function $e^x = \sum_{i=0}^{\infty} x^i/i!$.

Finally, because all the MTs in group $\mathcal{G}_k$ initiate their traffic with equal probability, the average download time of the group is equal to the average download time of the arbitrary MT $i \in \mathcal{G}_k$, i.e. $\mathbb{E}[\mathcal{T}_k(\boldsymbol{c}_k;\boldsymbol{c}_{-k})] = \mathbb{E}[\mathcal{T}_k^{(i)}(\boldsymbol{c}_k;\boldsymbol{c}_{-k})]$.

Theorem 2.1 is thus proved. ■

Remark 2.1: From the results in Theorem 2.1, we can observe that, for a certain group $\mathcal{G}_k$, its average download time relies on its own group request and caching distribution $r_{k,f}$ and $c_{k,f}$, as well as social density λ_0 and the social caching distribution C_f, but regardless of the specific caching distribution of the other groups $\boldsymbol{c}_{-k}$ or their densities. We can also observe that, for a certain file $f \in \mathcal{F}$, the expected download time in $\mathbb{E}[\mathcal{T}_k(\boldsymbol{c}_k;\boldsymbol{c}_{-k})]$ for the MTs in group $\mathcal{G}_k$ monotonically decreases with respect to group caching distribution $c_{k,f}$ and the social caching distribution C_f. This is because if the MT or its peer MTs cache the file with higher probability, the requested file is more available to the MT via D2D sharing and the average download time will decrease. Moreover, from (4), it is also observed the effect of the difference τ_f' is exponentially vanishing with respect to the social density λ_0 and caching distribution C_f.

Given group $\mathcal{G}_k$'s average download time, the *social average download time* is obtained by the weighted sum of group average download time with respect to their densities as

$$\mathbb{E}[\mathcal{T}(\{\boldsymbol{c}_k\})] = \sum_{k\in\mathcal{K}} \frac{\lambda_k}{\lambda_0} \sum_{f\in\mathcal{F}} r_{k,f}\bar{c}_{k,f} \times \left(\left[1-e^{-\mu_0 C_f}\right]\tau_f^{(D)} + e^{-\mu_0 C_f}\tau_f^{(B)}\right), \tag{5}$$

where the weights for different groups are their own spatial densities λ_k.

Given the above derivation of average download time, in the following we aim to answer the following question: how should a group optimally cache files to minimize average download time by taking advantage of the social file preference information?

III. Problem Formulation and Optimal Group Caching Solution

In this section, we first discuss a benchmark case where each MT group $\mathcal{G}_k,\ k\in\mathcal{K}$ only knows its own group's file request distribution $\boldsymbol{r}_k$ and does not know the preference information about the other groups. In this case, an MT group only cache files to match its own group preference, (i.e., $c_{k,f}=r_{k,f},\ f\in\mathcal{F},\ k\in\mathcal{K}$). By substituting each group's $\boldsymbol{c}_{-k}=\boldsymbol{r}_{-k}$ to $\mathbb{E}[\mathcal{T}_k(\boldsymbol{c}_k;\boldsymbol{c}_{-k})]$ in Theorem 2.1 and $\mathbb{E}[\mathcal{T}(\{\boldsymbol{c}_k\})]$ in (5), the group $\mathcal{G}_k$'s and social average download time reduces to $\mathbb{E}[\mathcal{T}_k(\boldsymbol{r}_k;\boldsymbol{r}_{-k})]$ and $\mathbb{E}[\mathcal{T}(\{\boldsymbol{r}_k\})]$, respectively.

Next, we assume the social file request distribution R_f given in (2) is public information and can be accessible to certain groups when making caching decisions.[3] We seek to understand if a particular group is intelligent and self-interested, how will this group optimize its group cache distribution to minimize its own group's average download time. We also want to understand how such selfish behaviors would impact the society.

For the purpose of exposition, we assume group $\mathcal{G}_k$ is the intelligent group and the other groups are not. As the other groups are not intelligent (e.g., unaware of social file preference) and they cache with $\boldsymbol{c}_{-k}=\boldsymbol{r}_{-k}$, we formulate this caching problem for group $\mathcal{G}_k$ as the optimization problem of minimizing its individual group's average download time, i.e.,

$$\begin{aligned}(\mathrm{P1}):\ \min_{\boldsymbol{c}_k}.\ & \mathbb{E}[\mathcal{T}_k(\boldsymbol{c}_k;\boldsymbol{r}_{-k})]\\ \mathrm{s.t.}\ & \sum_{f\in\mathcal{F}} c_{k,f}\le 1, & (6a)\\ & 0\le c_{k,f}\le 1,\ f\in\mathcal{F}. & (6b)\end{aligned}$$

Notice that the objective function is convex in the defined probability simplex and all the constraints of the problem are affine; thus, problem (P1) is a convex optimization problem and it is always feasible. In the following theorem, we give the closed-form solution for the optimal group caching distribution by utilizing Karush-Kuhn-Tucker (KKT) conditions [12].

[3]This is possible when the MTs are under the same social platform, where the statistics of the social preference is available.

Theorem 3.1: The intelligent group $\mathcal{G}_k$'s optimal group caching probability for any file $f\in\mathcal{F}$ for problem (P1) is

$$c_{k,f}^* = \left[1-\frac{1}{\mu_k}\left\{\mathcal{W}\left(\frac{(\gamma^*-r_{k,f}\tau_f^{(D)})e^{1+\mu_0}}{r_{k,f}\tau_f' e^{\mu_0\bar{R}_f-\mu_k\bar{r}_{k,f}}}\right)-1\right\}\right]_0^1. \tag{7}$$

where $[\cdot]_0^1=\min\{1,\max\{0,\cdot\}\}$, $\bar{R}_f=1-R_f$, $\bar{r}_{k,f}=1-r_{k,f}$, $\mathcal{W}(\cdot)$ is the Lambert W function [13] and γ^*, which is the optimal dual variable of constraint (6a), denotes the water level that satisfies $\sum_{f\in\mathcal{F}} c_{k,f}^*=1$.

Proof: Please refer to Appendix 24 for the details of the proof. ■

In Theorem 3.1, the allocation of probability can be interpreted as water-filling over different files with γ^* being the water level and the optimal probability allocation $c_{k,f}^*$ follows from (7). Based on the above optimal solution, we propose the optimal algorithm for obtaining the group caching distribution of the intelligent group $\mathcal{G}_k$ based on bi-section in Table I.

TABLE I: Water-filling Algorithm Based on Bi-section for Solving the Problem (P1).

Algorithm I

INPUT: Group and social file request distribution $\{r_{k,f}\}$ and $\{R_f\}$, D2D and BS download time for different files $\{\tau_f^{(D)}\}$ and $\{\tau_f^{(B)}\}$, average number of MTs of its own group μ_k and the society μ_0 in the range of d, desired precision δ_γ.
OUTPUT: The optimal group caching distribution $\{c_{k,f}^*\}$

1) **Initialize:** $\gamma^{(l)}:=\max_f\{r_{k,f}\tau_f^{(D)}\}$, $\gamma^{(h)}:=\infty$;
2) **Repeat:**
   i. $\gamma:=\frac{1}{2}(\gamma^{(l)}+\gamma^{(h)})$;
   ii. $c_{k,f}=\left[1-\frac{1}{\mu_k}\left\{\mathcal{W}\left(\frac{(\gamma-r_{k,f}\tau_f^{(D)})e^{1+\mu_0}}{r_{k,f}\tau_f' e^{\mu_0\bar{R}_f-\mu_k\bar{r}_{k,f}}}\right)-1\right\}\right]_0^1$,
   iii. If $\sum_{f\in\mathcal{F}} c_{k,f}<1$, set $\gamma^{(h)}:=\gamma$;
   iv. Else, set $\gamma^{(l)}:=\gamma$;
   v. $\Delta_\gamma:=|\gamma^{(l)}-\gamma^{(h)}|$;
3) **Until:** the condition $\Delta_\gamma>\delta_\gamma$ is violated.

Next, we give an illustrative example to show the impacts of different parameters on the optimal group caching distribution $c_{k,f}^*$ in (7) for the intelligent group $\mathcal{G}_k$.

Example 3.1: For the example, there are $K=2$ groups and $F=5$ files in total. The densities of MT in group 1 and 2 are $\lambda_1=0.15$ and $\lambda_2=0.05$, respectively. The range of D2D communications is $d=10$ m. We separately examine the effects of $r_{1,f}$, $r_{2,f}$, $\tau_f^{(D)}$ and τ_f' while keeping the others equal across the 5 files. First, for the default values of these parameters that are equal across the 5 files: The group request distributions are $r_{1,f}=1/5,\ f\in\mathcal{F}$ and $r_{2,f}=1/5,\ f\in\mathcal{F}$, respectively; the download time for different files from the BS and with D2D communications are $\tau_f^{(D)}=10$ ms and $\tau_f'=75$ ms.[4] Next, for the random generation of these

[4]The field test for the latencies of major US wireless service providers is given in [15]. Here, we adopt the typical latency of 75 ms.

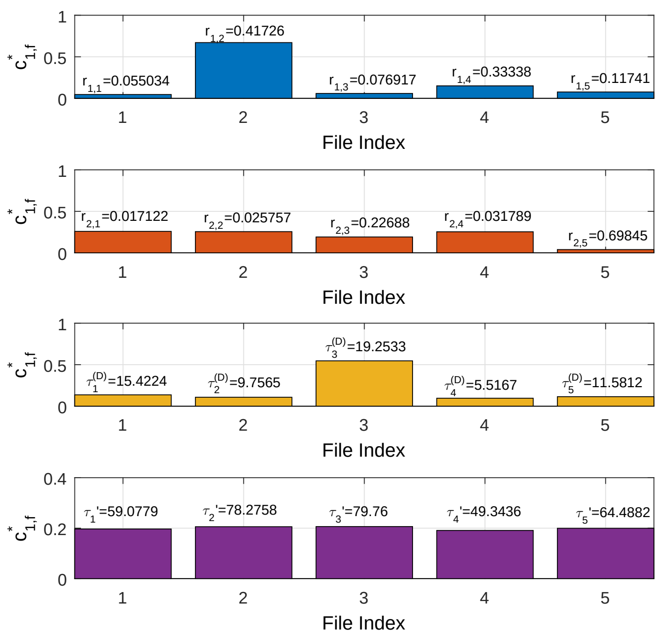

Fig. 2: Optimal file caching probabilities $c_{1,f}^*,\ f=1,\cdots,5$ for the intelligent group $\mathcal{G}_1$.

parameters: The PMFs $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are randomly generated with $\sum_{f\in\mathcal{F}} r_{1,f} = 1$ and $\sum_{f\in\mathcal{F}} r_{2,f} = 1$ for all the files $f \in \mathcal{F}$; the delays for different files by downloading via D2D communications and from BS are generated by uniform distributions on $[5,15]$ ms and $[50,100]$ ms, respectively. All the randomly generated values are shown on top of the bar for its corresponding optimal $c_{1,f}^*$.

The result is shown in Fig. 2. It can be observed from the first two sub-figures that $c_{1,f}^*$'s are monotonically increasing and decreasing with respect to $r_{1,f}$ and $r_{2,f}$, respectively. This is intuitive because if the group request probability $r_{1,f}$ is high, caching the file $f\in\mathcal{F}$ with a high probability will decrease the download time and if group 2 caches the files with high $r_{2,f}$, there is less need for group 1 to cache the file. It can also be observed from the last two sub-figures that the file caching probability is monotonically increasing with $\tau_f^{(D)}$ and τ'_f. This is also intuitive because if the delay of a file is large, caching the file with a high probability will decrease the average download time significantly. While comparing $\tau_f^{(D)}$ and τ'_f, we can see that the effect of $\tau_f^{(D)}$ is much more significant than τ'_f. This is because τ'_f is exponentially vanishing with respect to the average file availability C_f and social density λ_0, as has been shown in (4).

IV. Numerical Results

In this section, we show the impacts of selfish actions by intelligent group on the average download delays of itself and the other groups. The intelligent group exploits the social caching distribution as shown in Theorem 3.1. The range of D2D communications in the simulation is $d = 10$ m. There are $F = 100$ files in the library for $K = 2$ groups, where

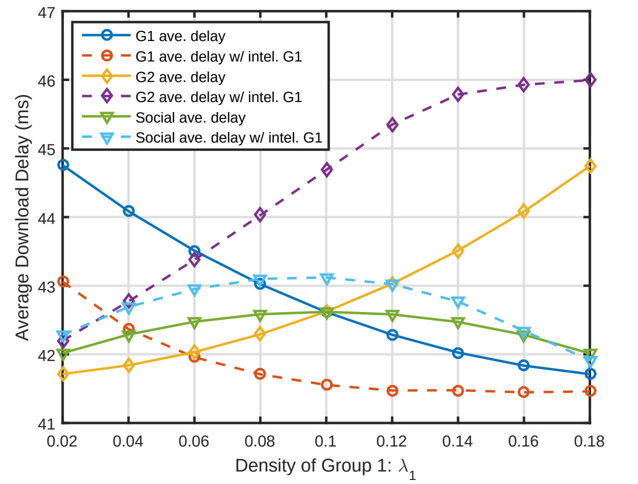

Fig. 3: Average download delays with intelligent $\mathcal{G}_1$ and non-intelligent $\mathcal{G}_2$.

for all the files $f \in \mathcal{F}$, the delays are $\tau_f^{(D)} = 10$ ms and $\tau_f^{(B)} = 75$ ms. We consider a fixed social density $\lambda_0 = 0.2$ with $\lambda_1 + \lambda_2 = \lambda_0$. For the group request distribution, we assume that it follows a Zipf distribution [16]:

$$r_{1,f} = \frac{f^{-\gamma_r}}{H(\gamma_r, 1, F)},\ f\in\mathcal{F}, \tag{8}$$

where $H(\gamma_r, x, y) = \sum_{i=x}^{y} i^{-\gamma_r}$ defines the harmonic sum from x to y and γ_r is denoted as the *Zipf exponent*, whose value is $\gamma_r = 0.4$ in our simulation. Furthermore, the group request distribution $\boldsymbol{r}_2$ of group 2 follows the same distribution, while it is assumed to be a random permutation of $\boldsymbol{r}_1$.

First, we consider the case where only group 1 is intelligent and group 2 is not. Under the above setup, we consider the following metrics for groups $\mathcal{G}_1$, $\mathcal{G}_2$ and society:

1. **Group 1 average delay** *(G1 ave. delay)*: $\mathbb{E}[\mathcal{T}_1(\boldsymbol{r}_1;\boldsymbol{r}_2)]$;
2. **Group 2 average delay** *(G2 ave. delay)*: $\mathbb{E}[\mathcal{T}_2(\boldsymbol{r}_2;\boldsymbol{r}_1)]$;
3. **Group 1 average delay with intelligent Group 1** *(G1 ave. delay w/ intel. G1)*: $\mathbb{E}[\mathcal{T}_1(\boldsymbol{c}_1^*;\boldsymbol{r}_2)]$;
4. **Group 2 average delay with intelligent Group 1** *(G2 ave. delay w/ intel. G1)*: $\mathbb{E}[\mathcal{T}_2(\boldsymbol{r}_2;\boldsymbol{c}_1^*)]$;
5. **Social average delay** *(Social ave. delay)*: $\mathbb{E}[\mathcal{T}(\boldsymbol{r}_1,\boldsymbol{r}_2)]$;
6. **Social average delay with intelligent Group 1** *(Social ave. delay w/ intel. G1)*: $\mathbb{E}[\mathcal{T}(\boldsymbol{c}_1^*,\boldsymbol{r}_2)]$.

The simulation result is shown in Fig. 3. It can be observed that the average download delay $\mathbb{E}[\mathcal{T}_1(\boldsymbol{c}_1^*;\boldsymbol{c}_2)]$ of intelligent group 1 decreases, while the average download delay $\mathbb{E}[\mathcal{T}_2(\boldsymbol{c}_2;\boldsymbol{c}_1^*)]$ of non-intelligent group 2 increases. Hence, this shows that if an intelligent group can exploit the social information, it can indeed decrease its average download time, while this may increase that of the other group. We can also observe that the average delays of both groups monotonically decrease with higher group density. This is obvious because higher group density leads to larger probability for D2D file sharing within the group. Compared with the benchmark case (metrics 1, 2 and 5 in Fig. 3) , the social average delay

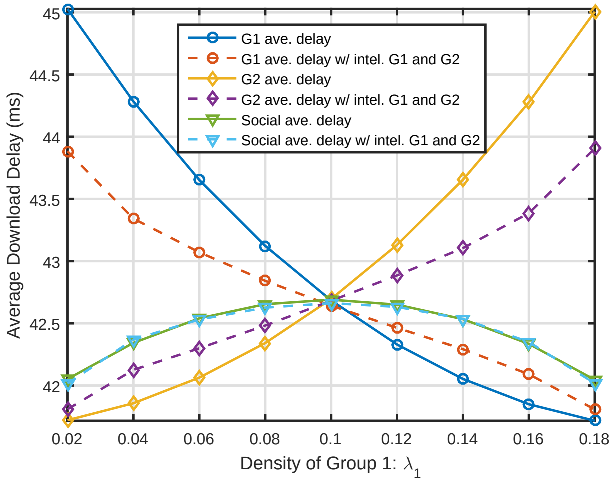

Fig. 4: Average download delays with intelligent $\mathcal{G}_1$ and intelligent $\mathcal{G}_2$.

increases with the selfish caching decisions of intelligent group 1. This is because the optimal strategy of the intelligent group is not necessarily socially optimal.

Next, we consider the case where both group 1 and group 2 are intelligent and investigate the impact on both themselves and society. For this setup, in addition to the average download delays for the benchmark cases (metrics 1, 2, and 5), we further compare the following metrics:

7. **Group 1 average delay with intelligent Group 1 and Group 2** *(G1 ave. delay w/ intel. G1 and G2)*: $\mathbb{E}[\mathcal{T}_1(\boldsymbol{c}_1^*; \boldsymbol{c}_2^*)]$;
8. **Group 2 average delay with intelligent Group 1 and Group 2** *(G2 ave. delay w/ intel. G1 and G2)*: $\mathbb{E}[\mathcal{T}_2(\boldsymbol{c}_2^*; \boldsymbol{c}_1^*)]$;
9. **Social average delay with intelligent Group 1 and Group 2** *(Social ave. delay w/ intel. G1 and G2)*: $\mathbb{E}[\mathcal{T}(\boldsymbol{c}_1^*, \boldsymbol{c}_2^*)]$.

The simulation result is shown in Fig. 4. It can be observed that if both groups are intelligent, the average delays of the groups might be worse than the benchmark case (metrics 1, 2, and 5 in Fig. 4), where both groups cache according to their actual file preferences. Hence, the intelligent actions of a group may be detrimental to the group itself and the other group. Furthermore, comparing the result with Fig. 3, we can observe that the average download time of group 1 increases due to the intelligent group 2 (versus non-intelligent group 2 in Fig. 3). Furthermore, the intelligent group's delay in the high group density regime is higher than that of the benchmark case for each of the two groups. This is because in the high density regime, cache according to the actual file preference is close to optimal for an intelligent group, while the optimal strategy of the other intelligent group will increase its delay.

V. Conclusions

This paper studies the D2D and local caching enabled file caching with heterogeneous file preferences among different groups. We practically categorize the MTs into different groups according to their preferences over the files and investigate the trade-off between the average download time of different groups. The problem is formulated as a convex optimization problem for minimizing the average download time of a self-interested group. Closed-form solution is obtained for its optimal caching strategy. We also provide extensive numerical examples to show that a group's selfish caching may be detrimental to itself and the other groups. In summary, our work provides essential insights in understanding the intelligent behaviors of different group of MTs with heterogeneous file preferences and their impacts on the individual and social download time under local caching and file sharing.

Appendix A
Proof for Theorem 3.1

First, it can be proved that the original problem (P1) can be equivalently converted to the following problem

$$
\begin{aligned}
(\mathrm{P1}'):\ & \underset{\bar{\boldsymbol{c}}_k}{\mathtt{min.}}\ \ \mathcal{T}_k(\boldsymbol{c}_k; \boldsymbol{r}_{-k}) \\
& \mathtt{s.t.}\ \ \sum_{f\in\mathcal{F}} \bar{c}_{k,f} \geq F-1, && \text{(9a)}\\
& \qquad 0 \leq \bar{c}_{k,f} \leq 1,\ \ f\in\mathcal{F}, && \text{(9b)}
\end{aligned}
$$

where $\bar{\boldsymbol{c}}_k = [\bar{c}_{i,1}, \bar{c}_{i,2}, \cdots, \bar{c}_{k,f}]$. By introducing the dual variable for the average memory constraint, the partial Lagrangian for the problem (P1) is

$$
\begin{aligned}
&\mathcal{L}_k\left(\bar{\boldsymbol{c}}_k, \gamma\right) = \sum_{f\in\mathcal{F}} r_{k,f}\bar{c}_{k,f}\Big(\tau_f^{(D)} \\
&+ e^{-\mu_0} e^{\mu_k \bar{c}_{k,f}} e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}} \tau_f'\Big) - \gamma\Big(\sum_{f\in\mathcal{F}} \bar{c}_{k,f} - F + 1\Big) \\
&= \sum_{f\in\mathcal{F}} \Bigg\{ r_{k,f}\bar{c}_{k,f}\Big(\tau_f^{(D)} + e^{-\mu_0} e^{\mu_k \bar{c}_{k,f}} e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}} \tau_f'\Big) \\
&- \gamma \bar{c}_{k,f} \Bigg\} + (F-1)\gamma, \qquad (10)
\end{aligned}
$$

where $\gamma \geq 0$ is the dual variable for constraints (6b). The dual function is then given as

$$
\begin{aligned}
g_k(\gamma) = & \underset{\bar{\boldsymbol{c}}_k}{\mathtt{min.}}\ \ \mathcal{L}_k\left(\bar{\boldsymbol{c}}_k, \gamma\right) \\
& \mathtt{s.t.}\ \ 0 \leq \bar{c}_{k,f} \leq 1,\ \ f\in\mathcal{F} \qquad (11)
\end{aligned}
$$

The associated dual problem is then defined as

$$
\underset{\gamma\geq 0}{\mathtt{max.}}\ \ g_k(\gamma). \qquad (12)
$$

Because the primal problem is convex optimization problem and satisfies the Slater's condition [12], the duality gap between the primal and dual problem is zero. Hence, the problem can be optimally solved in the dual domain by first optimizing the Lagrangian $\mathcal{L}_k\left(\bar{\boldsymbol{c}}_k, \gamma\right)$ for a given γ and then optimize the dual function $g_k(\gamma)$ with respect to γ. For a given γ, by the *Lagrange dual decomposition* [12], the dual function can be decomposed into a series of sub-problems each for a given file $f \in \mathcal{F}$ as follows

$$(\text{P1}''): \quad \min_{\bar{c}_{k,f}} \; r_{k,f}\bar{c}_{k,f}\left(\tau_f^{(D)} + e^{-\mu_0}e^{\mu_k \bar{c}_{k,f}}e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}\tau'_f\right) - \gamma \bar{c}_{k,f}$$

$$\text{s.t.} \quad 0 \le \bar{c}_{k,f} \le 1. \tag{13}$$

First, according to the KKT condition, the following equations should be satisfied in the optimal solution

$$\gamma^* \left(\sum_{f \in \mathcal{F}} \bar{c}^*_{k,f} - F + 1\right) = 0, \tag{14}$$

$$0 \le \bar{c}^*_{k,f} \le 1, \; f \in \mathcal{F}. \tag{15}$$

Then, the Lagrangian for sub-problem of file $f \in \mathcal{F}$ is

$$\begin{aligned} &\mathcal{L}_{k,f}(\bar{c}_{k,f}, \nu_{k,f,l}, \nu_{k,f,r}) \\ &= r_{k,f}\bar{c}_{k,f}\left(\tau_f^{(D)} + e^{-\mu_0}e^{\mu_k \bar{c}_{k,f}}e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}\tau'_f\right) \\ &- \gamma \bar{c}_{k,f} - \nu_{k,f,l}\bar{c}_{k,f} + \nu_{k,f,r}(\bar{c}_{k,f} - 1), \end{aligned} \tag{16}$$

where $\nu_{k,f,l} \ge 0$ is the dual variable for $\bar{c}_{k,f} \ge 0$, $\nu_{k,f,r} \ge 0$ is the dual variable for $\bar{c}_{k,f} \le 1$. Then, by taking the partial derivative of the Lagrangian with respect to $\bar{c}_{k,f}$, we obtain

$$\begin{aligned} \frac{\partial \mathcal{L}_{k,f}}{\partial \bar{c}_{k,f}} &= r_{k,f}\tau_f^{(D)} + e^{-\mu_0} r_{k,f}\tau'_f \\ &\times e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}\left(e^{\mu_k \bar{c}_{k,f}} + \mu_k \bar{c}_{k,f} e^{\mu_k \bar{c}_{k,f}}\right) \\ &- \gamma - \nu_{k,f,l} + \nu_{k,f,r}. \end{aligned} \tag{17}$$

Then, by the KKT conditions, we can obtain the following system of equations:

$$\frac{\partial \mathcal{L}_{k,f}}{\partial \bar{c}_{k,f}} = 0, \tag{18}$$

$$\nu^*_{k,f,l}\bar{c}^*_{k,f} = 0, \; \nu^*_{k,f,r}(\bar{c}^*_{k,f} - 1) = 0, \tag{19}$$

$$\nu^*_{k,f,l}, \nu^*_{k,f,r} \ge 0. \tag{20}$$

With (18), we can obtain that

$$\begin{aligned} &(1 + \mu_k \bar{c}^*_{k,f})e^{1 + \mu_k \bar{c}^*_{k,f}} \\ &= \frac{(\gamma + \nu^*_{k,f,l} - \nu^*_{k,f,r} - r_{k,f}\tau_f^{(D)})e^{1+\mu_0}}{r_{k,f}\tau'_f e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}} \end{aligned} \tag{21}$$

Hence, it can be obtained that

$$\bar{c}^*_{k,f} = \frac{1}{\mu_k}\left\{\mathcal{W}\left(\frac{(\gamma + \nu^*_{k,f,l} - \nu^*_{k,f,r} - r_{k,f}\tau_f^{(D)})e^{1+\mu_0}}{r_{k,f}\tau'_f e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}}\right) - 1\right\} \tag{22}$$

where $\mathcal{W}(\cdot)$ is the Lambert W function [13]. For notational convenience, in the following, we denote

$$\begin{aligned} &\tilde{c}_{k,f}(\gamma, \nu_{k,f,l}, \nu_{k,f,r}) \\ &= \frac{1}{\mu_k}\left\{\mathcal{W}\left(\frac{(\gamma + \nu_{k,f,l} - \nu_{k,f,r} - r_{k,f}\tau_f^{(D)})e^{1+\mu_0}}{r_{k,f}\tau'_f e^{\mu_0 \bar{R}_f - \mu_k \bar{r}_{k,f}}}\right) - 1\right\}. \end{aligned} \tag{23}$$

Because the Lambert function is monotonically increasing function, the function $\tilde{c}_{k,f}(\gamma, \nu_{k,f,l}, \nu_{k,f,r})$ is monotonically increasing with respect to $\nu_{k,f,l}$ and decreasing with respect to $\nu_{k,f,r}$.

Then, we discuss the three cases of $\tilde{c}_{k,f}(\gamma, \nu_{k,f,l}, \nu_{k,f,r})$ on the regions of $(-\infty, 0]$, $(0, 1)$ and $[1, \infty)$.

- First, for the case of $(-\infty, 0]$, assume that $\bar{c}^*_{k,f} > 0$ when $\tilde{c}_{k,f}(\gamma, 0, 0) \le 0$. When $\bar{c}^*_{k,f} > 0$, $\nu^*_{k,f,l} = 0$ should be satisfied due to the complementary slackness condition in (19). Hence, $\tilde{c}_{k,f}(\gamma, 0, \nu_{k,f,r}) > 0$. Moreover, because $\tilde{c}_{k,f}(\gamma, 0, \nu^*_{k,f,r})$ is monotonically decreasing with respect to $\nu_{k,f,r}$ and $\nu_{k,f,r} \ge 0$, we obtain $\tilde{c}_{k,f}(\gamma, 0, 0) \ge \tilde{c}_{k,f}(\gamma, 0, \nu^*_{k,f,r}) > 0$. This contradicts the assumption. Moreover, due to the primal feasibility $\bar{c}^*_{k,f} \ge 0$ in (15), when $\tilde{c}_{k,f}(\gamma, 0, 0) \le 0$, $\bar{c}^*_{k,f} = 0$ is the optimal solution;
- Next, for the case of $(0, 1)$, $\nu_{k,f,r} = \nu_{k,f,l} = 0, \; f \in \mathcal{F}$ due to the complementary slackness condition in (19). Hence, $\bar{c}^*_{k,f} = \tilde{c}_{k,f}(\gamma, 0, 0)$;
- Finally, for the case of $[1, \infty)$, similar to the first case, it can be proved by contradiction that, when $\tilde{c}_{k,f}(\gamma, 0, 0) \ge 1$, $\bar{c}^*_{k,f} = 1$.

By summarizing the above three cases, we can obtain that the optimal complimentary file caching probability for file $f \in \mathcal{F}$ of group $i \in \mathcal{K}$ is

$$\bar{c}^*_{k,f} = \left[\tilde{c}_{k,f}(\gamma, 0, 0)\right]_0^1. \tag{24}$$

Theorem 3.1 is proved.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2014-2019 white paper," 2015. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.pdf

[2] I. F. I. T. . Media, *Mobile Broadband Access at Home: The Business Case for Femtocells, UMA and IMS-VCC Dual Mode Solutions*. Informa Telecoms & Media, 2008.

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

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

[5] D. Bethanabhotla, G. Caire, and M. Neely, "Adaptive video streaming for wireless networks with multiple users and helpers," *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 268–285, Jan 2015.

[6] H. Zhou, M. Tao, E. Chen, and W. Yu, "Content-centric multicast beamforming in cache-enabled cloud radio access networks," 2015. [Online]. Available: http://arxiv.org/abs/1504.05663

[7] N. Golrezaei, A. Dimakis, and A. Molisch, "Scaling behavior for Device-to-Device communications with distributed caching," *IEEE Trans. Info. Theory*, vol. 60, no. 7, pp. 4286–4298, Jul. 2014.

[8] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," 2013. [Online]. Available: http://arxiv.org/abs/1312.2637

[9] S. Gitzenis, G. Paschos, and L. Tassiulas, "Asymptotic laws for joint content replication and delivery in wireless networks," *IEEE Trans. Info. Theory*, vol. 59, no. 5, pp. 2760–2776, May. 2013.

[10] Bluetooth SIG Inc., "Bluetooth low energy, bluetooth development portal," 2014. [Online]. Available: https://developer.bluetooth.org/TechnologyOverview/Pages/BLE.aspx

[11] S. Mumtaz and J. Rodriguez, *Smart Device to Smart Device Communication*. Springer, 2014.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[13] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational mathematics*, vol. 5, no. 1, pp. 329–359, 1996.
[14] Y. Guo, L. Duan, and R. Zhang, "Cooperative local caching and file sharing under heterogeneous file preferences," 2015. [Online]. Available: http://arxiv.org/abs/1510.04516
[15] FierceWireless, "3G/4G wireless network latency: Comparing Verizon, AT&T, Sprint and T-Mobile in February 2014," 2014. [Online]. Available: http://www.fiercewireless.com/special-reports/3g4g-wireless-network-latency-comparing-verizon-att-sprint-and-t-mobile-feb
[16] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *Proc. 18th Annual Joint Conf. IEEE Comput. and Commun. Societies (INFOCOM '99)*, vol. 1, Mar. 1999, pp. 126–134.