RESEARCH ARTICLE

Optimal Caching Policy for D2D Assisted Cellular Networks With Different Cache Size Devices

**NADIA ABDOLKHANI[1], MOHSEN ESLAMI[2], JAVAD HAGHIGHAT[3],
AND WALAA HAMOUDA[1], (Senior Member, IEEE)**

[1]Department of Electrical and Computer Engineering, Concordia University, Montreal, QC H3G 1M8, Canada
[2]Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz 71557-13876, Iran
[3]Department of Electrical and Electronics Engineering, TED University, 06420 Ankara, Turkey

Corresponding author: Walaa Hamouda (walaa.hamouda@concordia.ca)

ABSTRACT This paper studies the problem of optimal cache placement to maximize the offloading probability in a device-to-device (D2D) enabled cellular network with small base stations (SBSs). Different from most existing works, we consider unequal users' equipment (UE) cache memory sizes and all wireless links are modeled as Nakagami-*m* fading. User preference for each UE and global popularity for SBS, as well as the higher priority of content request from neighboring UEs vs. SBS, are the main factors that make the problem formulation of our work different from that of existing works. It is assumed that each UE caches its desired content with the order of searching its cache, neighboring UEs' cache via D2D communications, and its serving SBS' cache. A close to optimal low complexity heuristic cache placement policy is proposed and it is shown that its performance reaches the optimal caching strategy.

INDEX TERMS Caching, cellular network, D2D.

I. INTRODUCTION

Since most mobile data is frequently requested and can be cached, a promising solution to reduce backhaul traffic is to cache popular content on the edge of the wireless network. Caching at the cellular edge, e.g., the base station (BS) or user equipment (UE), reduces backhaul network traffic load, user service latency, and backup network cost, while improving network power, energy efficiency, and user experience [1].

Besides caching, proximity-based device-to-device (D2D) communications have emerged as a promising solution for current and future cellular networks that reduce congestion in the core network, lower energy consumption, and maximizes spectrum utilization. One of the main challenges of D2D communications are interference reduction and management among devices due to their proximity [2].

While edge caching is considered to be a promising solution to enhance offloading, caching is still surrounded by challenges, i.e., cache placement and content delivery strategy [1], [3], where and which content should be cached due to limited cache size of UEs and SBSs, and how to

reliably deliver cached content from its stored location to a requesting UE.

Heterogeneous user activity level, different content request of users for the same content and unequal user cache size make the aforementioned challenges even more complicated. Most of the existing works, assume the cache size of devices to be equal, while according to the device model and other factors, this is an unrealistic assumption for real-world cellular networks [4].

Recently, extensive research has been conducted to address the aforementioned challenges. To optimize caching policy and increase the likelihood of offloading, a prior knowledge-based learning algorithm has been proposed to obtain user preference in [5]. In that work, caching at the edge of the wireless network, D2D communications, *equal user cache size*, and Rayleigh fading have been assumed. In [6], [7], and [8] the authors have maximized cache hit rate and throughput by optimizing cache policy at the edge of the wireless network. Furthermore, the aforementioned works ignored the user's preference and the cache size difference, while the telecommunication channel has been assumed to be Rayleigh fading.

In the literature, there has been work on caching that has assumed different cache sizes at either SBSs or UEs,

nevertheless, to the best of the author's knowledge no prior work has considered the problem of cache placement in a cellular network with different UE and SBS cache sizes. In [9], the cache hit maximization problem in a heterogeneous wireless network has been investigated where users don't cache data and no D2D communication is available. The authors in [4], [5], [6], [7], [8], [9], and [10] have considered a server connected to several users that have different cache sizes. The links between the server and each user and the links between devices have been assumed "error-free", coded caching is assumed, and their cost function is not a cache hit probability.

In this paper, we consider a cache-enabled small cell network with D2D communications. Assuming mmWave-band frequencies [8], [12], we consider Nakagami-*m* fading channel model, which is more suitable to model mm-wave frequencies that are expected to be used in 5G and beyond cellular networks, especially since we are assuming a small cell network. In terms of performance analysis, incorporating Nakagami-*m* fading requires a different approach to evaluate link connection success probability, i.e., the commonly used method of Laplace transform for Rayleigh fading is not applicable. Different from previous works, we consider individual user preference instead of content popularity and different user cache sizes in solving the offloading probability maximization. Furthermore, we propose a low complexity heuristic caching strategy that offers performance close to the optimal case.

The rest of the paper is organized as follows. Section II describes the system model. In Section III, the caching problem is formulated and the optimum caching strategy is presented. In Section IV numerical and simulation results are presented, while Section V concludes the paper.

II. SYSTEM MODEL

We consider a single cell of a cache-enabled small cell network, where each UE and SBS are equipped with cache memory. There are N_u users, each operating in either D2D or cellular mode. Furthermore, We assume that in the D2D communication mode, UEs communicate at different frequencies than in the cellular mode. Hence, no cross-tier interference. However, in D2D communication, we assume that all user pairs communicate with each other at the same frequency band, resulting in inter-user interference. In the sample scenario depicted in Fig.1, the first nearest UE to UE1 is UE2 which has UE1's requested file in its cache, and UE1 receives its requested file from UE2 in D2D mode. UE3 has its requested file cached in its device, while UE4's requested file is not cached in its own cache or its neighboring devices. Hence, it has to request the file in cellular mode from the SBS.

We consider a general probabilistic caching policy in the downlink network to accommodate heterogeneous user preferences and spatial locality by allowing each UE and SBS to cache files with a different probability distribution. Location of UEs and SBSs are modeled by two independent homogeneous Poisson Point Processes (PPP) Φ_u and Φ_b with

intensities λ_u and λ_b, respectively. Three pre-defined cache hit scenarios are:

1) Self request: occurs when the requested file is cached in its device.
2) D2D cache hit: when the requested file is not cached in a device but exists in nearby devices.
3) SBS cache hit: when the requested file is not cached in a device and its nearby devices, the file is fetched from devices associated with the SBS cache.

The table below presents variable notations used throughout the paper.

TABLE 1. Variable notations.

Variable	Notation
N_u	Number of users
λ_u	User density
λ_b	SBS density
N	Number of neighboring users UE can associate with
m	Fading parameter
Ω	Average power
$\lambda = \frac{m}{\Omega}$	Nakagami parameter
N_u^c	Cache size of the uth user
N_b^c	Cache size of the bth SBS
N_f	Number of files in content library
$c_{f,u}$	UE caching policy
$\hat{c}_f$	SBS caching policy
q_u	User preference
p	Global content popularity
ν	User activity level
γ	SINR
$n_u(i)$	Index of the ith nearest user to user$u$
P_u	UE transmit power
P_b	SBS transmit power
h	channel gain
r	Distance
α	Pathloss component
σ^2	Noise variance
$S(\gamma_0)$	Content downloading success probability
δ	Zipf parameter
P_{self}	UE self-request cache hit probability
P_{hit}^{D2D}	D2D cache hit probability
P_{hit}^{SBS}	SBS cache hit probability
V	particle velocity
pos	particle position

In D2D communications, to increase the cache-hit probability, each user is allowed to associate with one of the N-nearest UEs to download the requested file. For the sake of lower latency and required transmission power, the nearest user that has the requested file is selected. For a given user, if the requested file is not found in any of its Nth nearest users, that user requests its desired file from SBS in cellular mode.

We consider Nakagami-*m* fading channel model for D2D and cellular connections. The Nakagami-*m* probability density function (PDF) is given by

$$p(x) = 2(\frac{m}{\Omega})^m \frac{x^{2m-1}}{\Gamma(m)} exp\{-\frac{mx^2}{\Omega}\}, \quad m \geq \frac{1}{2}, \quad (1)$$

where *m* is the fading parameter, *x* is the random variable, Ω is the average power, $\Gamma(.)$ is the Gamma function, and $\lambda = \frac{m}{\Omega}$ is the Nakagami parameter.
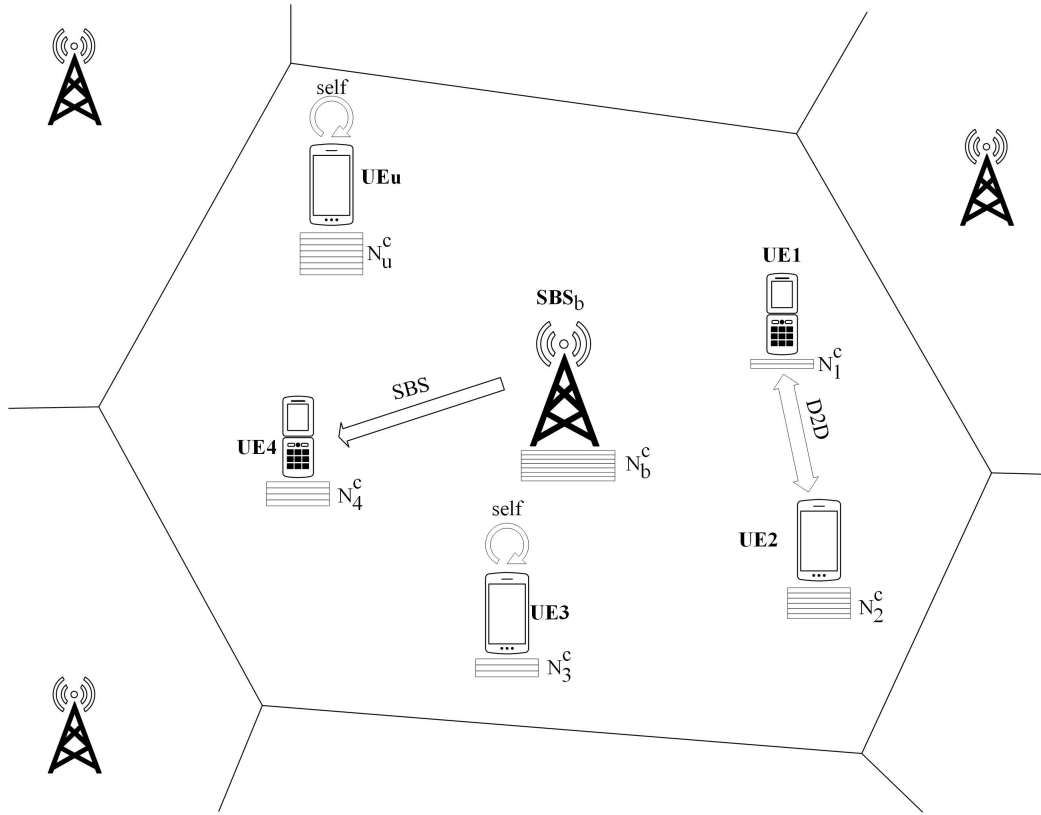
FIGURE 1. Cellular layout of a cache-and-D2D-enabled network with voronoi regions.

The uth user can cache at most N_u^c files and the SBS can cache at most N_b^c files from a content library consisting of N_f equal-sized files that all users in this cell can request. Denote $c_{f,u}$ as the probability that the uth user caches the fth file and $\hat{c}_f$ as the probability that the SBS caches the fth file. We assume $\sum_{f=1}^{N_f} c_{f,u} \leq N_u^c$ for $u = 1, \ldots, N_u$ and $\sum_{f=1}^{N_f} \hat{c}_f \leq N_b^c$ due to the cache storage limit. The caching policy can be optimized and updated at off-peak times due to the fact that the user preference changes much slower than the traffic load.

Let $q_u = [q_{1,u}, \ldots, q_{N_f,u}]$ denote the user preference of the uth user, where $q_{f,u} \in [0, 1]$ is the probability that the fth file is requested by the uth user. Also, let $p = [p_1, \ldots, p_{N_f}]$ denote global content popularity, which is the probability of the file requests in the considered cell, where p_f is the average probability that the fth file is requested by the users in that cell. Furthermore, let the user activity level be $v = [v_1, \ldots, v_{N_u}]$, where v_u is the probability that the uth user sends a file request in the cell. Given that the global content popularity is the average of user preferences in a region, hence it is given by

$$p_f = \sum_{u=1}^{N_u} v_u q_{f,u}; \tag{2}$$

In practice, user preference q_u and activity level v can be learned through machine learning techniques [5].

III. CACHING POLICY OPTIMIZATION
In this section, we evaluate the offloading probability or equivalently cache hit probability, which is the probability that a user finds its requested file in the local cache. A user chooses the first nearest user that has the requested file (due to lower latency and required transmission power); it searches up to the ith nearest user in priority of lower distance to itself. If the requested file is not found in any of the ith nearest users, it requests its desired file mode from SBS in cellular.

A. CACHING POLICY
We consider self-request, D2D cache hit, and SBS cache hit cases for offloading.

1) Self request: In this case, p_{self} is the average probability that a user finds its requested file in its cache, given by

$$p_{self} = \frac{1}{N_u} \sum_{u=1}^{N_u} \sum_{f=1}^{N_f} q_{f,u} c_{f,u}. \tag{3}$$

2) D2D cache hit: In this case, the probability of success is defined as the probability that a user retrieves the requested file from the cache of a neighboring device. This is possible when the received signal-to-interference-plus-noise ratio (SINR) at the user that acquires the requested file, γ, is greater than a threshold denoted by γ_0. The uth user's received SINR downloading from the $n_u(i)$th user is

given by

$$\gamma_{n_u(i)} = \frac{P_{n_u(i)} h_{n_u(i)} r_{n_u(i)}^{-\alpha}}{\sum_{\hat{u} \in \Phi_u, \hat{u} \neq n_u(i)} P_{\hat{u}} h_{u,\hat{u}} r_{u,\hat{u}}^{-\alpha} + \sigma^2}, \qquad (4)$$

where $n_u(i)$ denotes the index of the ith nearest user to the uth user, $P_{n_u(i)}$ is the transmit power of the $n_u(i)$th user, $h_{n_u(i)}$ is the channel gain between uth user and $n_u(i)$th user, $r_{n_u(i)}$ is the distance between the two users, α is pathloss exponent, σ^2 is the noise variance, and Φ_u is the user set that shares the same frequency with the uth user. For evaluating the power of the interference term in our analysis, we have assumed that only the nearest user causes interference in D2D mode, which generates the strongest interference.

The D2D cache hit probability is given by (5), as shown at the bottom of the next page, where $c_{f,n_u(i)} \prod_{w=0}^{i-1}(1 - c_{f,w})$ is the probability that the ith nearest user has the fth file in its cache, while all users from the first nearest to the $i - 1$th nearest (including the UE's cache probability ($w = 0$)) do not have the content in their cache, and $\mathbb{P}(\gamma_{n_u(i)} > \gamma_0 | r_{n_u(i)}, h_{n_u(i)})$ is the success probability when user u downloads a file from user $n_u(i)$. $\mathbb{E}_{f,r,h}$ denotes the expectation over user request, distance, and channel fading, respectively.

Lemma 1: The success probability of the uth user downloading from the $n_u(i)$th user is given by

$$S_{n_u(i)}(\gamma_0) = \int_0^\infty \int_0^\infty \mathbb{P}(\gamma_{n_u(i)} > \gamma_0 | r_{n_u(i)}, h_{n_u(i)})$$
$$\times f_R(r_{n_u(i)}, \lambda_u) f_H(h_{n_u(i)}) dr_{n_u(i)} dh_{n_u(i)}$$
$$= \int_0^\infty G_{n_u(i)}(r_{n_u(i)}, \gamma_0) \times f_R(r_{n_u(i)}, \lambda_u) dr_{n_u(i)}; \quad (6)$$

where $G_{n_u(i)}(r_{n_u(i)}, \gamma_0)$ is given in (7), as shown at the bottom of the next page.

Proof: Let $F(\cdot)$ be CDF of the random variable x, and $\phi(\cdot)$ its corresponding characteristic function with real variable t, then [13],

$$F(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{Im\{\phi(t)e^{-jtx}\}}{t} dt. \qquad (8)$$

Furthermore, let $\zeta_0 = |h_{n_u(i)}|$, and $\zeta = |h_{u,\hat{u}}|$, respectively. For a real valued μ, we have [13]: $Pr(\mu\zeta - \zeta_0 > 0) = 1 - F(0)$, then

$$Pr(r_{n_u(i)}^{-\alpha}\zeta_0 - \mu \sum_{\hat{u}=1}^K r_{u,\hat{u}}^{-\alpha}\zeta > \frac{\sigma^2}{P_{n_u(i)}}) = 1 - F(\frac{\sigma^2}{P_{n_u(i)}}). \quad (9)$$

For the case where $h_{n_u(i)}$ has Nakagami-m distribution,

$$f(r_{u,\hat{u}}^{-\alpha}\zeta) = (r_{u,\hat{u}}^{\alpha})^m (\frac{m}{\Omega})^m \frac{\zeta^{m-1}}{\Gamma(m)} exp(-\frac{m}{\Omega} r_{u,\hat{u}}^{\alpha}\zeta), \qquad (10)$$

where $\lambda = \frac{r_{u,\hat{u}}^{\alpha} m}{\Omega}$. The characteristic function of $r_{u,\hat{u}}^{-\alpha}\zeta$ is given by

$$\phi(t) = (1 + (\frac{t}{r_{n_u(i)}^{\alpha}\lambda})^2)^{-\frac{m}{2}} \cdot \prod_{\hat{u}=1}^K (1 + (\frac{\mu t}{r_{u,\hat{u}}^{\alpha}\lambda})^2)^{-\frac{m}{2}}$$

$$\times exp\{j[m \tan^{-1}(\frac{t}{r_{n_u(i)}^{\alpha}\lambda}) - \sum_{\hat{u}=1}^K m \tan^{-1}(\frac{\mu t}{r_{u,\hat{u}}^{\alpha}\lambda})]\}. \qquad (11)$$

Then $Im\{\phi(t)e^{-jtx}\}$ is given in (12), as shown at the bottom of the next page, and by substituting (12) and (8) in (9) we reach (13), as shown at the bottom of the next page. ∎

To achieve the success probability in (6), the PDF of the distance of a UE from its ith nearest UE, $r_{n_u(i)}$, according to the PPP is given by [14]:

$$f_R(r_{n_u(i)}, \lambda_u) = \frac{2(\pi\lambda_u)^i}{(i - 1)!} r_{n_u(i)}^{2i-1} e^{-\pi\lambda_u r^2}. \qquad (14)$$

3) SBS cache hit: when a user's requested file has not been cached in its cache or via D2D mode, then it can download from the SBS cache. The received SINR of the uth user downloading from the SBS is given by

$$\gamma_{u,b} = \frac{P_b h_{u,b} r_{u,b}^{-\alpha}}{\sum_{\hat{b} \in \Phi_b, \hat{b} \neq b} P_{\hat{b}} h_{u,\hat{b}} r_{u,\hat{b}}^{-\alpha} + \sigma^2}, \qquad (15)$$

where P_b is the transmit power of SBS, $h_{u,b}$ is the channel gain between the uth user and the SBS, $r_{u,b}$ is the distance between the uth user and the bth SBS, and Φ_b is the SBS set that shares the same frequency with bth SBS. The SBS cache hit probability is then given by

$$P_{hit}^{sbs} = \frac{1}{N_u} \sum_{u=1}^{N_u} v_u \sum_{f=1}^{N_f} p_f [\hat{c}_f \prod_{w=0}^{N} (1 - c_{f,w})] \times S_{u,b}(\gamma_0). \quad (16)$$

Similar to Lemma 1, it can be shown that the success probability when the uth user downloads from the SBS is given by

$$S_{u,b}(\gamma_0) = \int_0^\infty \int_0^\infty \mathbb{P}(\gamma_{u,b} > \gamma_0 | r_{u,b}, h_{u,b})$$
$$\times f_R(r_{u,b}, \lambda_b) f_H(h_{u,b}) dr_{u,b} dh_{u,b}$$
$$= \int_0^\infty G_{u,b}(r_{u,b}, \gamma_0) \times f_R(r_{u,b}) dr_{u,b}, \qquad (17)$$

where $G_{u,b}(r_{u,b}, \gamma_0)$ is given in (18), as shown at the bottom of the next page. Also, $f_R(r_{u,b}, \lambda_b)$ in (17) is calculated similar to (14).

For the general case of N (assuming the requested file is stored in the ith nearest user), (6) and (17) are difficult to solve. Furthermore, for the analysis to be tractable we have assumed that only the nearest user (strongest interferer) causes interference in D2D mode. Therefore, we have closed form expressions for the case of $N = 1$, $m = 2$, $\alpha = 2$, and $K = 2$, where $S_{n_u(i)}$ is given in (19), as shown at the bottom of the next page. The expression for $S_{u,b}(\gamma_0)$ is similar to (19) and it can be obtained by replacing λ_u and P_u with λ_b and P_b, respectively.

Finally, the total cache hit probability is given by

$$p_{hit} = p_{self} + p_{hit}^{D2D} + p_{hit}^{SBS}$$
$$= \frac{1}{N_u} \sum_{u=1}^{N_u} \sum_{f=1}^{N_f} q_{f,u} c_{f,u} + \frac{1}{N_u} \sum_{u=1}^{N_u} v_u \sum_{f=1}^{N_f} q_{f,u}$$

$$\times \sum_{i=1}^{N} [c_{f,n_u(i)} \prod_{w=1}^{n_u(i)-1} (1 - c_{f,w})] S_{n_u(i)}(\gamma_0)$$

$$+ \frac{1}{N_u} \sum_{u=1}^{N_u} v_u \sum_{f=1}^{N_f} p_f [\hat{c}_f \prod_{w=1}^{N} (1 - c_{f,w})] \times S_{u,b}(\gamma_0). \tag{20}$$

B. CACHE HIT OPTIMIZATION

Based on (20), the optimization problem is formulated as P_1 :

$$\max_{\{c_{f,u}, \hat{c}_f\}} \quad p_{self} + p_{hit}^{D2D} + p_{hit}^{SBS} \tag{21a}$$

$$\text{s.t. } C1: 0 \le c_{f,u} \le 1, \ f = 1, \ldots, N_f, \quad u = 1, \ldots, N_u \tag{21b}$$

$$C2: \sum_{f=1}^{N_f} c_{f,u} \le N_u^c, \quad , \tag{21c}$$

$$C3: \sum_{f=1}^{N_f} \hat{c}_f \le N_b^c. \tag{21d}$$

In problem $P1$, constraints C2 and C3 ensure that the cache content of each user device and SBS do not exceed their cache size. Note that (21) is a multivariate polynomial in terms of $c_{f,u}$s and $\hat{c}$s, and the determinant of its Hessian matrix is found to be infinite. Therefore, it is a non-convex optimization problem and we suggest to solve it using meta-heuristic methods. Accordingly, its numerical solution using particle swarm optimization (PSO) algorithm with adjustable search steps [15] is considered. In this algorithm, an initial velocity is assigned to each particle. These particles move in the problem space, and the results are calculated based on a predefined competency function, after each movement. As the algorithm progresses, these new locations determine the direction of congestion according to,

$$V(t+1) = V(t) + c_1 \times rand(t) \times (p_{best}(t) - pos(t))$$
$$+ c_2 \times rand(t) \times (g_{best}(t) - pos(t)) \tag{22}$$
$$pos(t+1) = pos(t) + V(t+1) \tag{23}$$

In (22) and (23), the best position the particle has ever achieved, also called the best particle nostalgia or its best solo experience, is denoted by p_{best}. Another parameter used by

$$p_{hit}^{D2D} = \mathbb{E}_{f,r,h} \Big[\sum_{i=1}^{N} [c_{f,n_u(i)} \prod_{w=0}^{i-1} (1 - c_{f,w})] \times \mathbb{P}(\gamma_{n_u(i)} > \gamma_0 | r_{n_u(i)}, h_{n_u(i)}) \Big]$$

$$= \frac{1}{N_u} \sum_{u=1}^{N_u} v_u \sum_{f=1}^{N_f} q_{f,u} \times \sum_{i=1}^{N} [c_{f,n_u(i)} \prod_{w=0}^{i-1} (1 - c_{f,w})] S_{n_u(i)}(\gamma_0), \tag{5}$$

$$G_{n_u(i)}(r_{n_u(i)}, \gamma_0) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin[m \tan^{-1}(\frac{t}{r_{n_u(i)}^\alpha \lambda}) - t\frac{\sigma^2}{P_u} - \sum_{\hat{u}=1, \hat{u} \neq n_u(i)}^{K} m \tan^{-1}(\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})]}{t.[1 + (\frac{t}{r_{n_u(i)}^\alpha \lambda})^2]^{\frac{m}{2}} \cdot \prod_{\hat{u}=1, \hat{u} \neq n_u(i)}^{K} [1 + (\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})^2]^{\frac{m}{2}}} dt \tag{7}$$

$$\text{Im}\{\phi(t)e^{-jtx}\} = \frac{\sin[m \tan^{-1}(\frac{t}{r_{n_u(i)}^\alpha \lambda}) - tx - \sum_{\hat{u}=1}^{K} m \tan^{-1}(\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})]}{[1 + (\frac{t}{r_{n_u(i)}^\alpha \lambda})^2]^{\frac{m}{2}} \cdot \prod_{\hat{u}=1}^{K} [1 + (\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})^2]^{\frac{m}{2}}} \tag{12}$$

$$Pr(r_{n_u(i)}^{-\alpha} \zeta_0 - \mu \sum_{\hat{u}=1}^{K} r_{u,\hat{u}}^{-\alpha} \zeta > \frac{\sigma^2}{P}) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin[m \tan^{-1}(\frac{t}{r_{n_u(i)}^\alpha \lambda}) - t\frac{\sigma^2}{P} - \sum_{\hat{u}=1, \hat{u} \neq n_u(i)}^{K} m \tan^{-1}(\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})]}{t.[1 + (\frac{t}{r_{n_u(i)}^\alpha \lambda})^2]^{\frac{m}{2}} \cdot \prod_{\hat{u}=1, \hat{u} \neq n_u(i)}^{K} [1 + (\frac{\mu t}{r_{u,\hat{u}}^\alpha \lambda})^2]^{\frac{m}{2}}} dt. \tag{13}$$

$$G_{u,b}(r_{u,b}, \gamma_0) = \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\sin[m \tan^{-1}(\frac{t}{r_{u,b}^\alpha \lambda}) - t\frac{\sigma^2}{P_b} - \sum_{\hat{b}=1, \hat{b} \neq b}^{K} m \tan^{-1}(\frac{\mu t}{r_{u,\hat{b}}^\alpha \lambda})]}{t.[1 + (\frac{t}{r_{u,b}^\alpha \lambda})^2]^{\frac{m}{2}} \cdot \prod_{\hat{b}=1, \hat{b} \neq b}^{K} [1 + (\frac{\mu t}{r_{u,\hat{b}}^\alpha \lambda})^2]^{\frac{m}{2}}} dt. \tag{18}$$

$$S_{n_u(i)}(\gamma_0) = \frac{1}{\lambda_u P_u \pi (-1 + \mu)^6 \mu^4} ((-1 + \mu)\mu^4 (4\lambda \sigma^2 \mu (1 + \mu(19 + 10\mu)) + \lambda_u P_u \pi (-1 + \mu)(1 + \mu(-4 + \mu(23 + 4\mu))))$$

$$- 6\mu^6 (\lambda_u P_u \pi (-1 + \mu)(1 + 3\mu) + 2\lambda \sigma^2 (3 + \mu(6 + \mu)))Log[\mu]) \tag{19}$$

the algorithm is the best position ever obtained by the particle mass, called g_{best}. Also, pos in (23) and V in (22) denote the position and the velocity of the particles, respectively. The initialization of the location and velocity of particles are given respectively by

$$x(0) = x_{min} + rand(x_{max} - x_{min}) \qquad (24)$$

$$V(0) = V_{min} + rand(V_{max} - V_{min}). \qquad (25)$$

where the rand function results in a random value between the maximum and minimum values of the particle location and velocity.

In our results, the PSO parameters used are given by: $c_1 = 1.2$, $c_2 = 2.8$, $x_{min} = 0.2$, $x_{max} = 0.9$, *swarmsize* $= 500$, *population* $= rand(N_u + 1, N_f, swarmsize)$, where N_u is the number of users and N_f is the number of files in the content library. To incorporate constraints C1 and C2 from (21c) and (21d) in the PSO algorithm, (21a) can be written as

$$p_{hit} = p_{self} + p_{hit}^{D2D} + p_{hit}^{SBS} - a_1 \times |N_u^c - \sum_{f=1}^{N_f} c_{f,u}|$$

$$- a_2 \times |N_b^c - \sum_{f=1}^{N_f} \hat{c}_f| \qquad (26)$$

It is noted that in (26), one can achieve the desired results by setting $a_1 = 0.09$, and $a_2 = 0.1$. Based on [16], if c_1 and c_2 are selected such that the condition in (27) holds, the system guarantees convergence,

$$0 \leq \frac{1}{2}(c_1 + c_2) \leq 1 \qquad (27)$$

where in our case, this holds true and hence assuring convergence. Also, the results were confirmed with c_1 and c_2 values other than the ones mentioned before. The same results were achieved but with larger number of iterations. For instance, for $c_1 = 1.2$ and $c_2 = 2.8$ the results converged around 20 iterations, for $c_1 = 2.2$ and $c_2 = 1.8$ the results converged around 30 iterations, and for $c_1 = 2.2$ and $c_2 = 1.6$ the results converged around 40 iterations.

In the next section, we present numerical results for the proposed solution.

IV. SIMULATION RESULTS

In this section, we present numerical results of the proposed caching policy. Specifically, we investigate the difference between two optimal solutions and compare with several other scenarios, i.e., equal and unequal UE cache size, given that UEs are equipped with unequal cache sizes in reality.

For numerical evaluation, a single cell with radius $R = 40$m and $N_u = 10$ users has been considered. User density of $\lambda_u = 2 \times 10^{-3}/m^2$ and SBS density of $\lambda_b = 2 \times 10^{-4}/m^2$ have been assumed. Transmit power has been set to $P_u = 10^{-2}$ watts for users, and $P_b = 43$ dBm for SBS; noise variance $\sigma^2 = 10^{-3}$ watts/Hz, and $\gamma_0 = 0$ dB. It is assumed that content library has $N_f = 20$ files. The users'
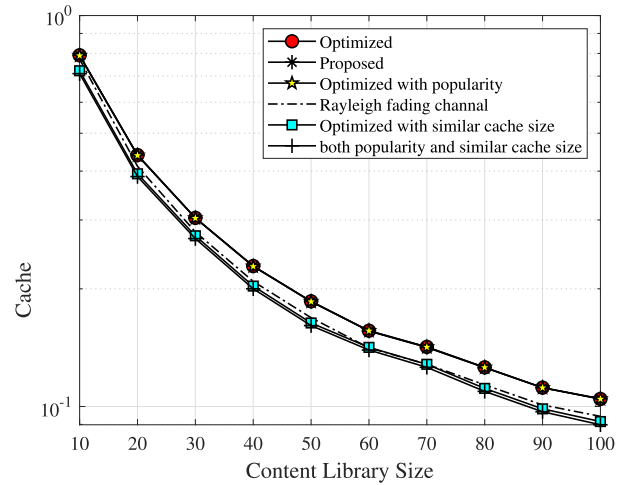
FIGURE 2. Optimal Cache Hit Probability vs. Content Library Size. Showing the effect of considering different cache sizes for different users' equipment. For the case of "optimized with false assumption", cache sizes are not equal, while the optimization is performed assuming equal caches.

cache capacity, N_u^c, is randomly selected and the SBS cache capacity is set to $N_b^c = 10$.

Individual user preference follows Zipf distribution with parameter $\delta = 0.2$, and $q_{f,u} = \frac{(A(f))^{-\delta}}{\sum_{i=1}^{N_f} i^{-\delta}}$, where $A(f)$ is a random permutation of $[1, \ldots, N_f]$. Furthermore, the user activity level, v_u, is considered a uniform random where $0 \leq v_u \leq 1$. Finally, in all results we consider Nakagami fading channels with $\lambda = 2$.

In Figures 2 and 3, we compare four caching strategies for the D2D assisted cellular network:

1) Optimal strategy, which is found by solving the optimization problem of Section III.B.

2) Proposed strategy, where the files with least popularity are discarded for each user.

3) Proposed strategy with Rayleigh fading channel, and optimization is performed using the success probability achieved.

4) Optimized with popularity, where the content distribution is calculated using popularity, while the optimization is performed assuming individual user preference.

5) Optimized with similar cache size, where cache sizes are not equal, while the optimization is performed assuming they are equal; i.e., average cache size of UEs has been used as the mistaken common cache size of UE's. The results for this case are obtained as follows:

(i) All devices are assumed to have similar cache sizes which is the average of the randomly generated unequal cache sizes used for the results in the case of "Proposed scheme".

(ii) Using the similar cache sizes, we computed the system's caching strategy. Since in practical systems, the cache sizes of different devices are not equal then that caching strategy has to fit the actual cache sizes of the devices.
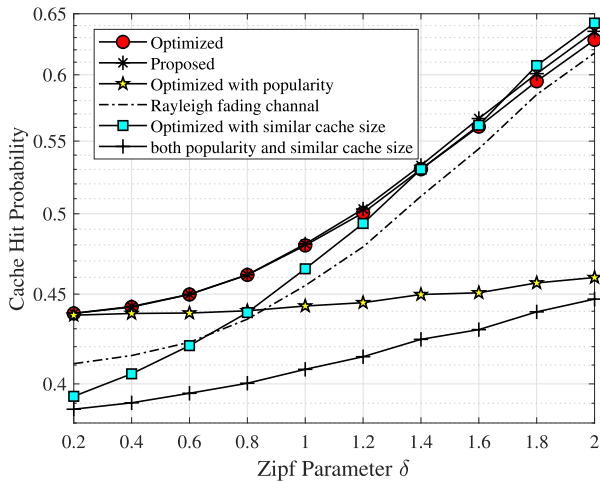
FIGURE 3. Optimal Cache Hit Probability vs. Zipf Parameter δ. Showing effect of considering different cache sizes for different users' equipment. For the case of "optimized with similar cache size", cache sizes are not equal, while the optimization is performed assuming equal caches.

FIGURE 4. Success Probability for accessing the nearest neighboring device and also serving SBS vs. User and SBS density, respectively.

(iii) We used an algorithm that compares the caching strategy computed with similar cache sizes to the actual cache sizes of the devices and adapt the strategy to fit the actual cache sizes.

(iv) Finally, we used the new strategy to compute the new cache hit probability of the system where the cache sizes were unequal. The results in Fig. 2 clearly show that the assumption of "Different Cache Size Devices" improves the performance.

In the case the cache strategy takes lower capacity than the UE's cache capacity, we have unused cache capacity which is a waste of the already existing resources. In the case of the cache strategy with larger file size than the cache capacity, we have to forego the less popular files and only cache the most popular ones. Hence, extra computation and waste of resources.

6) Optimized with both popularity and similar cache size, which is a combination of scenarios 4 and 5.

Figure 2 shows the cache probability versus content library size. By increasing content library size, while fixing cache sizes, the cache hit probability reduces. In Figure 3, cache hit probability is plotted for different Zipf distribution parameters, δ. Note that by increasing δ, the Zipf PDF becomes narrower, i.e., the number of files with relatively high probability reduces. In this case, any wise strategy only caches the files with a higher probability. Hence, the probability of offloading popular files increases. In both aforementioned figures, and in terms of cache hit probability, the proposed scheme offers almost the same performance as the optimal one. Furthermore, caching with the proposed strategy has lower complexity. Also, it is shown that the assumption of "Unequal Cache Size Devices" improves the performance since it represents the actual system.

In Fig. 4, the success probability for accessing the nearest neighboring device and also serving BS is plotted versus user and SBS density, respectively. Due to incorporating the line
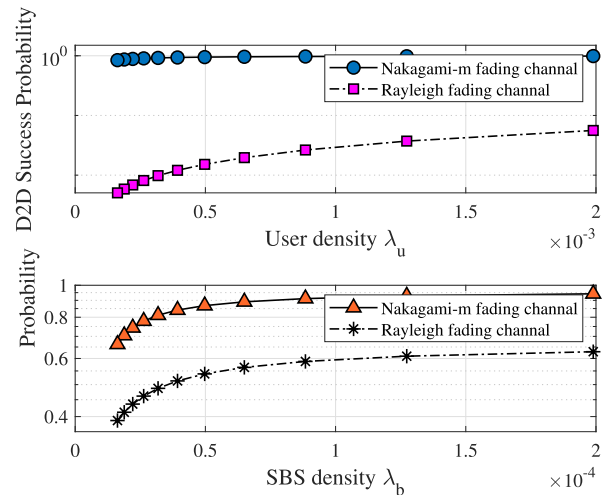
of sight component, the Nakagami-*m* fading channel has a much higher success probability in transferring files among devices compared to that of Rayleigh fading and accordingly higher probability of cache-hit. This is most beneficial in our network as it represents the case of small cell networks in 5G, where mm-Wave frequencies are expected to be used and with smaller cells the probability of having LoS is high. However, since cache placement is done with the knowledge of success probability, the cache-hit probability curves for Rayleigh fading in prior figures are very close to that of Nakagami-*m* fading.

V. CONCLUSION

We mathematically formulated the optimal cache placement problem to maximize offloading probability in a D2D-enabled cellular network with SBSs. To the best of our knowledge, no prior work has considered unequal UE cache size in solving the cache placement problem, while this is a realistic assumption in many practical scenarios. Our results showed that solving the cache placement problem with an equal cache size assumption, in a network where cache sizes are actually unequal, leads to cache hit probability reduction.

REFERENCES

[1] D. Liu and C. Yang, "Caching at base stations with heterogeneous user demands and spatial locality," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1554–1569, Feb. 2019.

[2] S. Joshi and R. K. Mallik, "Analysis of dedicated and shared device-to-device communication in cellular networks over Nakagami-*m* fading channels," *IET Commun.*, vol. 11, no. 10, pp. 1600–1609, Jul. 2017.

[3] J. Chuan, B. Bai, X. Wu, and H. Zhang, "Optimizing content placement and delivery in wireless distributed cache systems through belief propagation," *IEEE Access*, vol. 8, pp. 100684–100701, 2020.

[4] Y. Lin, Z. Lin, P. Chen, Z. Chen, and L. Wu, "On consideration of content and memory sizes in 5G D2D-assisted caching networks," *IEEE Access*, vol. 8, pp. 52759–52773, 2020.

[5] B. Chen and C. Yang, "Caching policy for cache-enabled D2D communications by learning user preference," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6586–6601, Dec. 2018.

[6] Q. Li, Y. Zhang, A. Pandharipande, X. Ge, and J. Zhang, "D2D-assisted caching on truncated Zipf distribution," *IEEE Access*, vol. 7, pp. 13411–13421, 2019.

[7] Y. Lan, X. Wang, D. Wang, Z. Liu, and Y. Zhang, "Task caching, offloading, and resource allocation in D2D-aided fog computing networks," *IEEE Access*, vol. 7, pp. 104876–104891, 2019.

[8] W. Wu, N. Zhang, N. Cheng, Y. Tang, K. Aldubaikhy, and X. Shen, "Beef up mmWave dense cellular networks with D2D-assisted cooperative edge caching," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3890–3904, Apr. 2019.

[9] S. Krishnendu, B. N. Bharath, and V. Bhatia, "Cache enabled cellular network: Algorithm for cache placement and guarantees," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1550–1554, Dec. 2019.

[10] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Device-to-device coded-caching with distinct cache sizes," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2748–2762, May 2020.

[11] Q. Wang, Y. Cui, S. Jin, J. Zou, C. Li, and H. Xiong, "Optimization-based decentralized coded caching for files and caches with arbitrary sizes," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2090–2105, Apr. 2020.

[12] L. Yang, F.-C. Zheng, W. Wen, and S. Jin, "Analysis and optimization of random caching in mmWave heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10140–10154, Sep. 2020.

[13] Q. T. Zhang, "Outage probability in cellular mobile radio due to Nakagami signal and interferers with arbitrary parameters," *IEEE Trans. Veh. Technol.*, vol. 45, no. 2, pp. 364–372, May 1996, doi: 10.1109/25.492911.

[14] A. H. Imani, M. Eslami, J. Haghighat, and M. Noor, "Effect of fading on the k-coverage of wireless sensor networks," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 7, p. e3994, 2020.

[15] M. F. Pervej, L. T. Tan, and R. Q. Hu, "Artificial intelligence assisted collaborative edge caching in small cell networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–7.

[16] A. Engelbrecht, *Computational Intelligence: An Introduction*, vol. 7. Chichester, U.K.: Wiley, 2004.

NADIA ABDOLKHANI received the B.Sc. and M.Sc. degrees in electrical engineering from the Shiraz University of Technology, Shiraz, Iran, in 2017 and 2021, respectively. She is currently pursuing the Ph.D. degree with Concordia University, Montreal, Canada. Her research interests include wireless and cellular communications, caching at wireless networks, D2D communications, and MIMO communications.

MOHSEN ESLAMI received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tehran, Tehran, Iran, in 2000 and 2002, respectively, and the Ph.D. degree in electrical engineering from the University of Alberta, Edmonton, AB, Canada, in 2009. He was a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, University of Alberta, from 2009 to 2011. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Shiraz University of Technology. His research interests include wireless communications, multiple antenna systems, cooperative communications, and mm-Wave communications.

JAVAD HAGHIGHAT received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Tehran, in 2000 and 2002, and the Ph.D. degree in electrical engineering from Concordia University, Montreal, Canada, in 2007. He is currently with the Department of Electrical and Electronics Engineering, TED University, Ankara, Turkey, as an Associate Professor. His research interests include source coding, channel coding, cooperative communications, and wireless sensor networks.

WALAA HAMOUDA (Senior Member, IEEE) received the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from Queen's University, Kingston, ON, Canada, in 1998 and 2002, respectively. Since July 2002, he has been with the Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada, where he is currently a Professor. Since June 2006, he is the Concordia University Research Chair Tier I in wireless communications and networking. His current research interests include machine-to-machine communications, the IoT, 5G and beyond technologies, single/multiuser multiple-input multiple-output communications, space-time processing, cognitive radios, and wireless networks. He received numerous awards, including the Best Paper Awards of the GC in 2021, ICC in 2021, 2019 ICCSPA, IEEE WCNC in 2016, ICC in 2009, and the IEEE Canada Certificate of Appreciation in 2007 and 2008. He served(ing) as the Co-Chair of the IoT and the Sensor Networks Symposium of the GC'22; the TPC Co-Chair of the ITC/ADC 2022 Conference; the track Co-Chair: Antenna Systems, Propagation, and RF Design, IEEE Vehicular Technology Conference (VTC Fall'20); the Tutorial Chair of IEEE Canadian Conference in Electrical and Computer Engineering (CCECE 2020); the General Co-Chair of the IEEE SmartNets 2019 Conference; the Co-Chair of the MAC and Cross Layer Design Track of the IEEE (WCNC) 2019, the Wireless Communications Symposium of the IEEE ICC'18, and the Ad-hoc, Sensor, and Mesh Networking Symposium of the IEEE GC'17; the Technical Co-Chair of the Fifth International Conference on Selected Topics in Mobile and Wireless Networking (MoWNet'2016); the Track Co-Chair: Multiple Antenna and Cooperative Communications, IEEE Vehicular Technology Conference (VTC Fall'16); the Co-Chair: ACM Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks (ACMPE-WASUN'14); and the Technical Co-Chair of the Wireless Networks Symposium, 2012 Global Communications Conference, the Ad hoc, Sensor, and Mesh Networking Symposium of the 2010 ICC, and the 25th Queen's Biennial Symposium on Communications. He also served as the Track Co-Chair of the Radio Access Techniques of the 2006 IEEE VTC Fall 2006 and the Transmission Techniques of the IEEE VTC-Fall 2012. From September 2005 to November 2008, he was the Chair of the IEEE Montreal Chapter in Communications and Information Theory. He is an IEEE ComSoc Distinguished Lecturer. He served as an Associate Editor for the IEEE COMMUNICATIONS LETTERS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and *IET Wireless Sensor Systems*. He is currently serves as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE WIRELESS COMMUNICATIONS LETTERS.

• • •