Joint Caching and Resource Allocation in D2D-Assisted Heterogeneous Networks

Wael Jaafar
Department of Computer Science
Université du Québec à Montréal
Montreal, Canada
jaafar.wael@uqam.ca

Wessam Ajib
Department of Computer Science
Université du Québec à Montréal
Montreal, Canada
ajib.wessam@uqam.ca

Halima Elbiaze
Department of Computer Science
Université du Québec à Montréal
Montreal, Canada
elbiaze.halima@uqam.ca

***Abstract*—Device-to-device (D2D) communications combined with Heterogeneous networks (Hetnets) has attracted growing interest. Indeed, Hetnets deploy small-cells within macro-cells in order to offload traffic and improve the overall network coverage and capacity. Whereas, D2D promotes the use of communications between users for content delivery without going through the small or macro bases stations. Hence, it reduces communication delays and improves the spectral efficiency. In this context, we aim in this paper at reducing the average transmission delay, defined as the average sum delays of contents transmission to satisfy users' requests in a macro-cell, by jointly optimizing caching placement and channel resource allocation, in cache-enabled Hetnet with D2D assistance. At first, a lower-bound expression of the average transmission delay is derived. Then, the optimization problem is formulated. Afterwards, we propose a sub-optimal random search algorithm and a low-complexity greedy algorithm that solve the problem. Finally, numerical results illustrate the performances of the proposed algorithms.**

***Index Terms*—Hetnet, D2D, content caching, content delivery, channel allocation.**

I. Introduction

As the standardization of 5G requirements by 3GPP is advancing rapidly, many aspects of the Non Stand Alone and Stand Alone 5G network are getting an increased attention. Hetnets are already popular, and are established as an incontestable component of future networks, where ultra dense small-cell are deployed and coexist with the Macro Base Stations (MBS) infrastructure. Hetnets provide better energy efficiency and spectral efficiency performances [1]. Moreover, the Device-to-Device (D2D) communication, defined as the direct communication between users (UEs) without going through the base stations (BSs), is seen as a promising technology to improve frequency reuse and throughput, and reduce the delay of wireless transmissions [2]-[3]. Recently, combining the two above networks (Hetnet and D2D) has led at proposing new traffic offloading schemes [4]-[5].

On the other hand, caching has attracted recent attention thanks to its ability to further reduce the backhaul traffic and eliminate duplicate transmissions of popular content [6],[7]. The coming of caching has moved the problem from intensive demands on the backhaul links into the evolution of the caching capability of the network. Consequently, Content-Centric Networking (CCN) has been developed to emphasize content placement and content retrieval by operating protocols on the PHY, MAC and network layers. Caching has been investigated for use in different network types (Macro-Cellular, Hetnets and D2D), for content placement, content delivery or both, and in a centralized or distributed manner [7]-[8].

A. Related Work

Although, research has improved the networking-caching framework for 5G, few work had investigated jointly the caching placement and caching delivery issues [9]-[12]. In [9], the authors optimized cooperative content caching in a backhaul constrained Hetnet, aiming at minimizing the content download delay. No D2D communication is supported in this paper. In [10], caching placement is performed at the SBS only or at the D2D-enabled users only. Joint design of the transmission and caching policies is investigated, with prior knowledge of user demands. [11] proposes a proactive caching strategy at both users and SBSs, considering user mobility and different classes of user's interests for content, aiming at minimizing the system's cost (e.g. energy, bandwidth) while serving all requests. Optimal caching policy was obtained using standard integer programming optimization tools. This work focused on caching and did not impose constraints on wireless channels, power or social-awareness. Authors in [12] propose a caching algorithm to minimize the average transmission delay in a macro-cell with D2D support. The proposed greedy algorithm performs better than popularity-based naive caching policy. However, no theoretical analysis was realized for this work, and it is limited to caching optimization only.

B. Contributions

Motivated by the shortcomings of the previous work, we propose in this paper a Multi-Tier D2D-assisted Hetnet, where caching placement at all tiers is considered jointly with content transmission. The main objective is to minimize the average transmission delay of the macro-cell, under channel resources and caching constraints.

The novelties are summarized as follows:

1) Even though we adopt a multi-tier cache-enabled Hetnet with D2D similar to [11], we do not focus only on optimizing the caching policy in small-cells and users, but we extend it to joint optimization of caching in the

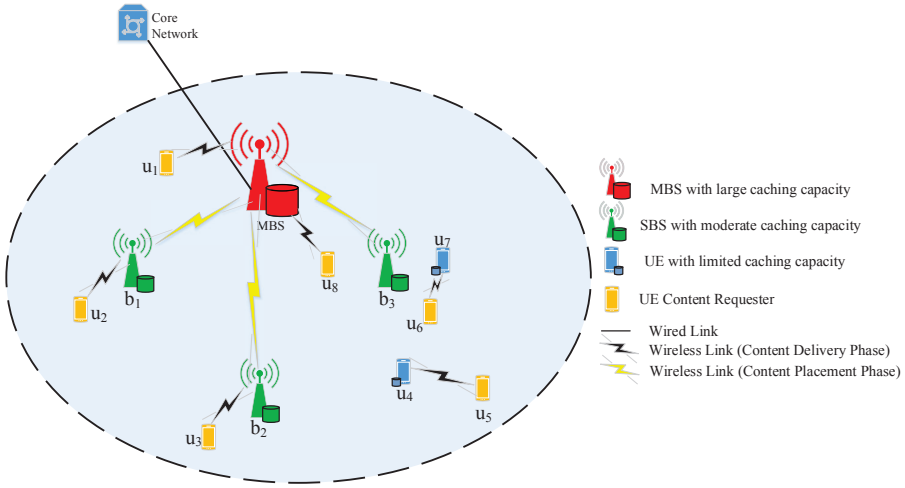

Fig. 1. System Model

macro-cell, small-cells and users within the macro-cell, and channel resource allocation.

2) Differently from [9] and [12], we analyze theoretically the average transmission delay in a macro-cell, where the delay is averaged over the number of users within the macro-cell, their demands, and channels variation.
3) We derive a lower-bound for the average transmission delay in a macro-cell. The latter is expressed as a function of several parameters including the distances, transmit powers, caching and channel allocation probabilities.
4) The optimization problem aiming at minimizing the average transmission delay of a macro-cell is formulated. It results in a non-convex Mixed-Integer-Non-Linear-Program (MINLP) that is difficult to solve.
5) We propose to solve the problem using sub-optimal approaches where in the first, random search is performed by limiting the search space for caching policies. The second approach is greedy and it is based on a distance criterion that builds caching for a user of interest with minimum average delay to nodes within the macro-cell.
6) The complexity of the proposed approaches is studied, and shown to be very low compared to exhaustive search.
7) Numerical results are illustrated to demonstrate the efficacy of the proposed solutions.

The remaining of the paper is organized as follows. Section II presents the system model. Section III formulates the optimization problem. Section IV details the proposed solutions for the problem. Section V illustrates and discusses the numerical results. Finally, a conclusion closes the paper in section VI.

II. System Model

We assume a cellular network with one MBS, S Small BSs (SBSs) ($s \in \mathcal{S} \equiv \{1, \ldots, S\}$) and U D2D users ($u \in \mathcal{U} \equiv \{1, \ldots, U\}$), randomly spread within the coverage area of the MBS, as presented in Fig. 1.

Without loss of generality, we assume that W orthogonal frequency channels are available in the network, where resource $w \in \mathcal{W} = \{1, \ldots, W\}$ has bandwidth $b \in \mathcal{B} \equiv \{b_1, \ldots, b_W\}$, with $b_i \neq b_j$, $\forall i \neq j$. We define by $r_{u,w}$ the probability indicating that channel w is allocated to the transmission having user u as its receiver.

We assume that time is slotted, and in each time slot $t \in \mathbb{N}$, transmissions may occur simultaneously on orthogonal channels. We assume that one of two phases may occur: Content placement phase, in which the caching memories of the nodes within the network are updated, is periodically executed, and content delivery phase that emphasizes the transmissions of data to requesters. In this paper, we focus on the communication model of the content delivery phase, and the transmissions required for the content placement phase are not considered.

We assume that noise on all wireless channels is an independent identically distributed (i.i.d.) additive white Gaussian noise (AWGN) circularly symmetric with zero mean and variance N_0, and that the channel coefficients are i.i.d. in frequency. We assume also that the channel coefficients are constant within a time slot of duration τ, and varies along different time slots.

In our network, we assume that F files (or segments of files) belong to a library $\mathcal{F} \equiv \{0, 1, \ldots, F\}$ in the core network, where 0 means that no file is requested, and they have the same size of L bits [13]. Without loss of generality, the same size of files (or segments) can be obtained by segmenting large files into equally sized segments. The MBS, any SBS s and any user u have caching capacities of $C_0 \times L$ bits, $C_S \times L$ bits and $C_U \times L$ bits respectively, where $F \gg C_0 \gg C_S \gg C_U$.

The file popularity follows Zipf distribution, and the prob-

ability that a user u requests a file f can be given by:

$$q_f = \frac{i(f)^{-\beta}}{\sum_{j=0}^{F} i(j)^{-\beta}}, \forall f \in \mathcal{F}, \quad (1)$$

where $i(f)$ is the rank of file $\forall f \in \mathcal{F}\backslash\{0\}$, and $\beta \geq 0$ reflects how skewed the popularity distribution is, that means larger β exponents correspond to higher content reuse, i.e., the first few popular files account for the majority of requests. In the particular case of $f = 0$, q_0 reflects the probability that the node is idle. If $q_0 = \theta$ ($\theta \in [0,1]$), then the rank of $f = 0$ is $i(0) = 1/\sqrt[\beta]{\left(\frac{\theta}{1-\theta}\sum_{j=1}^{F} i(j)^{-\beta}\right)}$.

In this paper, we assume that a file cannot be cached at more than one node within the macro-cell. Moreover, caching at each node is realized according to a probability distribution $c_i = \{c_{i,1}, \ldots, c_{i,F}\}$, $i = 0, s, u$, where 0 is the index of the MBS, $s \in \mathcal{S}$, $u \in \mathcal{U}$, and $c_{i,f}$ is the probability that node i caches file f. To be noted that if $c_{i,f} \in \{0,1\}$, then caching is considered deterministic.

When a user $u \in \mathcal{U}$ requests a file $f \in \mathcal{F}\backslash\{0\}$ in a time slot t, it starts by checking its own cache. If the content is available locally, it is obtained directly without any delay. Otherwise, the MBS will identify the most adequate source, with minimum delay, from which the file request can be served. We assume that the MBS is the decision-making entity and it has knowledge of the nodes positions within its cell, the statistical knowledge of the channels and the nodes caching status. The MBS randomly (probability-based) allocates a channel resource to the transmitter node that has the content requested by the user u and at the same time provides the lowest macro-cell average transmission delay.

For simplicity, we consider the following assumptions:

1) Any file transmission is allocated one channel resource only.
2) A request is satisfied by one node only.
3) A transmission can occupy several successive time slots. In the last time slot, if the transmission ends before the time slot expiration, the transmitter frees the channel resource and keeps silent during the rest of the time slot.
4) The transmit powers of MBS, SBS s and user u are P_0, P_S and P_U respectively.
5) MBS (resp. SBS) can communicate with several users simultaneously using orthogonal channels.

III. Problem Formulation

In this section, we characterize at first the average transmission delay of the macro-cell and derive its lower-bound expression. Then, the problem is formulated.

A. Average Transmission Delay Characterization

In this paper, we adopt a similar definition of the transmission delay as [12]: number of time slots occupied to serve a content request. It is expressed by:

$$T = \min\left\{t' : L \leq \sum_{t=1}^{t'} \tau R(t)\right\}, \quad (2)$$

where $R(t)$ is the instantaneous channel's data rate in time slot t and τ is the time slot duration. Without loss of generality, for a given resource w, the data rate for a channel $i-j$ can be given by:

$$\begin{aligned} R_{ij}^{w}(t) &= b_w \log_2\left(1 + \frac{P_i |h_{ij}(t)|^2}{N_0}\right) \\ &= b_w \log_2\left(1 + \gamma_{ij}(t)\right) \end{aligned} \quad (3)$$

where P_i is the transmit power of node i, $h_{ij}(t) = h'_{ij}(t) d_{ij}^{-\alpha}$ is the Rayleigh channel coefficient capturing both short-scale (h'_{ij}) and long-scale (d_{ij}) fading, $|h_{ij}(t)|^2$ is the channel gain following an exponential distribution of zero mean and variance $d_{ij}^{-\alpha}$, $\gamma_{ij}(t) = \frac{P_i}{N_0}|h_{ij}(t)|^2$ is the signal-to-noise-ratio.

Knowing the channel distribution of $\gamma_{ij}(t)$, the average transmission delay of link $i-j$ ($i \in \mathcal{U}\cup\mathcal{S}\cup\{0\}\backslash\{j\}$, $j \in \mathcal{U}$), denoted by $\bar{T}_{ij}^{w}$, is defined similarly to [14] as follows:

$$\bar{T}_{ij}^{w} = \mathbb{E}\left[T_{ij}^{w}\right] = \sum_{T=1}^{+\infty} \mathbb{P}\left[T_{ij}^{w} > T\right] \quad (4)$$

where $\mathbb{E}$ is the expectation operator, T_{ij}^{w} is the achieved delay on transmission $i-j$ and $\mathbb{P}\left[T_{ij}^{w} > T\right]$ is the probability that T_{ij}^{w} is above T. The latter probability is determined by:

$$\begin{aligned} \mathbb{P}\left[T_{ij}^{w} > T\right] &= \mathbb{P}\left[\sum_{t=1}^{T} R_{ij}^{w}(t) < \frac{L}{\tau}\right] \\ &= \mathbb{P}\left[\sum_{t=1}^{T} \log_2\left(1 + \gamma_{ij}(t)\right) < \frac{L}{\tau b_w}\right] \\ &= A. \end{aligned} \quad (5)$$

Notice that if $\log_2(1+\gamma_{ij}(t)) < \frac{L}{\tau b_w T}$, $\forall t = 1, \ldots, T$, then $\sum_{t=1}^{T} \log_2(1+\gamma_{ij}(t)) < \frac{L}{\tau b_w}$ [14]. Hence,

$$A \geq \mathbb{P}\left[\log_2\left(1+\gamma_{ij}(t)\right) < \frac{L}{\tau b_w T}, \forall t\right] \quad (6)$$

and

$$\begin{aligned} \mathbb{P}\left[T_{ij}^{w} > T\right] &\geq \mathbb{P}\left[\log_2\left(1+\gamma_{ij}(t)\right) < \frac{L}{\tau b_w T}, \forall t\right] \\ &= \left(\mathbb{P}\left[\log_2\left(1+\gamma_{ij}\right) < \frac{L}{\tau b_w T}\right]\right)^{T} \quad (7) \\ &= \left(\mathbb{P}\left[\gamma_{ij} < 2^{\frac{L}{\tau b_w T}} - 1\right]\right)^{T} \quad (8) \end{aligned}$$

where in (7), dependency from t is released.

Since the cumulative distribution function (cdf) of γ_{ij} for a Rayleigh channel is given by:

$$F_{\gamma_{ij}}(z) = 1 - e^{-\frac{z}{\bar{\gamma}_{ij}}}, (z \geq 0), \quad (9)$$

where $\bar{\gamma}_{ij} = \frac{P_i}{N_0} d_{ij}^{-\alpha}$, then, the lower-bound of $\mathbb{P}\left[T_{ij}^w > T\right]$ is expressed by:

$$\mathbb{P}\left[T_{ij}^w > T\right] \geq g_{ij}^w(T), \tag{10}$$

where $g_{ij}^w(T) = \left(F_{\gamma_{ij}}\left(2^{\frac{L}{\tau b_w T}} - 1\right)\right)^T$. As a consequence, the average transmission delay of link $i-j$ is lower-bounded by:

$$\bar{T}_{ij}^w \geq G_{ij}^w \tag{11}$$

where $G_{ij}^w = \sum_{T=1}^{+\infty} g_{ij}^w(T)$.

For a request of file f by user u, MBS selects the node i that has f and allocates the channel resource w to achieve the lowest average transmission delay. The latter is defined by $\bar{D}_{u,f}^w$, and is given by (12) in the next page, where $\bar{T}_0$ is the average transmission delay on the backhaul link from the core network to the MBS, and $\mathcal{T}_{u,f}$ is the set of potential transmitters of file f to user u. $\mathbb{P}\left[\mathcal{T}_{u,f} = \emptyset\right]$ is the probability that no node is a potential transmitter of file f to user u, expressed by:

$$\mathbb{P}\left[\mathcal{T}_{u,f} = \emptyset\right] = \prod_{u'=1;u'\neq u}^{U} \left(1 - c_{u',f} q_0\right) \prod_{s=1}^{S} \left(1 - c_{s,f}\right)\left(1 - c_{0,f}\right), \tag{13}$$

while $\mathbb{P}\left[\mathcal{T}_{u,f} = \{i\}\right]$ is the probability that node i is the only potential transmitter of file f to user u, given by (14) below.

By combining (11) into (12), we obtain a lower-bound on $\bar{D}_{u,f}^w$, defined as $G_{u,f}^w$, and expressed by (15) in the next page.

Finally, the average transmission delay of the macro-cell is expressed by:

$$\bar{D} = \frac{1}{U} \sum_{u=1}^{U} \sum_{w=1}^{W} \sum_{f=1}^{F} r_{u,w} q_f \bar{D}_{u,f}^w, \tag{16}$$

and is lower-bounded by:

$$G = \frac{1}{U} \sum_{u=1}^{U} \sum_{w=1}^{W} \sum_{f=1}^{F} r_{u,w} q_f G_{u,f}^w. \tag{17}$$

B. Joint Optimization Problem Formulation

We formulate the joint optimization problem aiming at minimizing the macro-cell average transmission delay. Due to the complexity of the expression of $\bar{D}$ and similarly to [14], we opt for the lower-bound objective function expressed by G in the remaining of this paper.

The problem (**P1**) is formulated as follows:

$$\begin{aligned}
\min_{\mathbf{C},\mathbf{R}} \quad & G \\
\text{s.t.} \quad & c_1 : \sum_{w=1}^{W} r_{u,w} \leq 1, \ \forall u \in \mathcal{U} \\
& c_2 : \sum_{u=1}^{U} r_{u,w} \leq 1, \ \forall w \in \mathcal{W} \\
& c_3 : 0 \leq r_{u,w} \leq 1, \ \forall u \in \mathcal{U} \\
& c_4 : \sum_{w=1}^{W} \sum_{u=1}^{U} r_{u,w} \leq \min\left(U, W\right) \\
& c_5 : \sum_{f=1}^{F} c_{u,f} \leq C_u, \ \forall u \in \mathcal{U} \cup \mathcal{S} \cup \{0\} \\
& c_6 : \sum_{u=1}^{U} c_{u,f} + \sum_{s=1}^{S} c_{s,f} + c_{0,f} \leq 1, \ \forall f \in \mathcal{F} \backslash \{0\} \\
& c_7 : c_{u,f} \in \{0,1\}, \ \forall u \in \mathcal{U} \cup \mathcal{S} \cup \{0\}
\end{aligned}$$

where $\mathbf{C} = [c_{i,f}]_{(U+S+1)\times F}$ is the caching placement matrix and $\mathbf{R} = [r_{u,w}]_{U\times W}$ is the channel allocation matrix.

Constraints c_1-c_4 emphasize the random probability-based channel allocation to a communication. Meanwhile, c_5-c_7 express caching placement with memory limitation at node i ($i \in \mathcal{U} \cup \mathcal{S} \cup \{0\}$) and caching redundancy limitation within the macro-cell for any file f.

Since $\mathbf{C}$ is a matrix of binary variables, the feasible set of problem (P1) is non-convex. In addition, product relationships between elements of $\mathbf{C}$ in the objective function make the latter non-convex. (P1) is a mixed non-convex optimization problem, considered as NP-hard [15].

IV. Proposed Solutions

A. Complexity Analysis

Due to the non-convexity of problem (P1) and the numerous constraints, the global optimal solution can be obtained only by exhaustive search. However, exhaustive search is very difficult to implement due to the very large search space for parameters $\mathbf{C}$ and $\mathbf{R}$. Indeed, assuming that F=100, U=20, W=3, S=2, and that $\forall u, \forall w, r_{u,w} \in \{0, 0.1, \ldots, 0.9, 1\}$, the search space is $2^{F\times(1+S+U)} \times 11^{U\times W} = 2^{100\times 23} \times 11^{60}$.

Consequently, we propose in this paper efficient algorithms that provide feasible and low-complex solutions for joint caching and resource allocation. We propose at first a random search algorithm that provides sub-optimal performances, but realizable compared to exhaustive search. Then, we propose a low-complex distance-based algorithm that provides interesting performances compared to the previous one.

B. Proposed Random Search (RS)

Given the high complexity of problem (P1), we opt for a sub-optimal solution as follows: For a given caching policy within a search space, the resource allocation strategy $\mathbf{R}^*$ is determined by solving the following problem (**P2**):

$$\begin{aligned}
\min_{\mathbf{R}} \quad & \frac{1}{U} \sum_{u=1}^{U} \sum_{w=1}^{W} r_{u,w} \eta_{u,w} \\
\text{s.t.} \quad & c_1 : \sum_{w=1}^{W} r_{u,w} \leq 1, \ \forall u \in \mathcal{U} \\
& c_2 : \sum_{u=1}^{U} r_{u,w} \leq 1, \ \forall w \in \mathcal{W}
\end{aligned}$$

$$\bar{D}^w_{u,f} = (1-c_{u,f})\left[\mathbb{P}\left[\mathcal{T}_{u,f}=\emptyset\right]\left(\bar{T}_0+\bar{T}^w_{mu}\right)+\sum_{i\in\mathcal{U}\cup\mathcal{S}\cup\{0\}\setminus\{u\}}\mathbb{P}\left[\mathcal{T}_{u,f}=\{i\}\right]\bar{T}^w_{iu}\right] \quad (12)$$

$$\mathbb{P}\left[\mathcal{T}_{u,f}=\{i\}\right] = \begin{cases} c_{i,f}q_0\left[\prod_{u'=1;u'\neq\{u,i\}}^{U}(1-c_{u',f}q_0)\right]\left(\prod_{s=1}^{S}(1-c_{s,f})\right)(1-c_{0,f}) & i \text{ is a user} \\ c_{i,f}\prod_{u'=1;u'\neq u}^{U}(1-c_{u',f}q_0)\left(\prod_{s=1;s\neq i}^{S}(1-c_{s,f})\right)(1-c_{0,f}) & i \text{ is a SBS} \\ c_{i,f}\prod_{u'=1;u'\neq u}^{U}(1-c_{u',f}q_0)\left(\prod_{s=1}^{S}(1-c_{s,f})\right) & i \text{ is the MBS} \end{cases} \quad (14)$$

$$G^w_{u,f} = (1-c_{u,f})\left[\mathbb{P}\left[\mathcal{T}_{u,f}=\emptyset\right]\left(\bar{T}_0+G^w_{0u}\right)+\sum_{i\in\mathcal{U}\cup\mathcal{S}\cup\{0\}\setminus\{u\}}\mathbb{P}\left[\mathcal{T}_{u,f}=\{i\}\right]G^w_{iu}\right] \quad (15)$$

$$c_3: 0\leq r_{u,w}\leq 1,\ \forall u\in\mathcal{U}$$
$$c_4: \sum_{w=1}^{W}\sum_{u=1}^{U} r_{u,w}\leq \min(U,W)$$

where we define the matrix $\mathbf{\Gamma}=[\eta_{u,w}]_{U\times W}$, with $\eta_{u,w}=\sum_{f=1}^{F}q_f G^w_{u,f}$, for given user u and resource w.

It is straightforward to see that the objective function is linear and convex. By defining, then minimizing the Lagrange dual function $\mathcal{L}(\mathbf{R},\mathbf{v})$ given in (20) below, the optimal $\mathbf{R}^*$ is obtained, where $\forall u\in\mathcal{U}$, $\forall w\in\mathcal{W}$,

$$r^*_{u,w}=\frac{\frac{\lambda_w}{U}+\frac{\lambda_u}{W}+\frac{\lambda_0\min(U,W)}{UW}}{\lambda_w+\lambda_u+\lambda_0-\eta_{u,w}},\ \forall\lambda_w+\lambda_u+\lambda_0\neq\eta_{u,w} \quad (18)$$

and $\mathbf{v}=\{\lambda_u,\forall u\in\mathcal{U},\lambda_w,\forall w\in\mathcal{W},\lambda_0\}$. By setting $\lambda_0=-1$ and $\lambda_u=\lambda_w=-l$ where $l>0$, $r_{u,w}$ is simplified into:

$$r^*_{u,w}=\frac{l(U+W)+\min(U,W)}{UW\left(2l+1+\eta_{u,w}\right)},\ \forall u\in\mathcal{U},\ \forall w\in\mathcal{W}. \quad (19)$$

With this approach, the total search space reduces to the search space of the caching policies, that is $2^{F\times(1+S+U)}$ which is also very large. In order to make use of this solution, the search space must be further reduced to a reasonable size $Dim\propto U$. Thus, we assume that a number of caching policies $\mathbf{C}$ is randomly selected to form a set with size Dim. Within this set, the best caching policy and calculated resource allocation strategy ($\mathbf{C}$, $\mathbf{R}^*$), offering the lowest average delay, will be retained. Algorithm 1 details this procedure.

Algorithm 1 RS

```
Set G_0 = +∞, C_0 = [ ], R_0 = [ ] % Policy initialization
Set Dim % Size of caching policies search space
for i=1 to Dim do
    % Caching Matrix Creation %
    Randomly create C_i, such that constraints c_5 − c_7
    of (P1) are satisfied with equality.
    % Matrix R_i building %
    Build R_i associated to C_i using (19)
    Calculate G using (17)
    % Policy optimization %
    if G < G_0 do
        Set G_0 = G, C_0 = C_i, R_0 = R_i
    end if
end for
```

C. Proposed Greedy Criterion-Based Algorithm (GCBA)

To further reduce the search space of caching policies compared to the previous solution, we propose to construct the caching matrices $\mathbf{C}$ by following a greedy procedure:

1) At the beginning, all caches are empty. We start by choosing a user of interest u.
2) Identify node i_1 such that $i_1=\underset{i'\in\mathcal{U}\cup\mathcal{S}\cup\{0\}\setminus\{u\}}{\operatorname{argmin}} G^w_{i'u}$, i.e. node with the lowest transmission delay to u.
3) Fill i_1's memory with C_i files (i.e. set $c_{i_1,f}$=1, such that $\sum_{f=1}^{F}c_{i_1,f}=C_{i_1}$).
4) If not all files are cached within node i_1, then identify node i_2 such that $i_2=\underset{i'\in\mathcal{U}\cup\mathcal{S}\cup\{0\}\setminus\{u,i_1\}}{\operatorname{argmin}} G^w_{i'u}$. Afterwards, fill i_2's memory with files while respecting its maximum capacity C_{i_2}.
5) Repeat step 4) for the nodes set $\mathcal{U}\cup\mathcal{S}\cup\{0\}\setminus\{u,i_1,i_2\}$ until all files in $\mathcal{F}\setminus\{0\}$ are cached in the network or all cache memories of nodes in the network are full, i.e. matrix $\mathbf{C}$ is complete.

The presented procedure is a simple solution to minimize the average delay for one user only, in an egoistic manner. One way to generalize it to all users within the macro-cell, is to execute it for all or a set of users, then cross-correlate the

$$\mathcal{L}(\mathbf{R},\mathbf{v}) = \sum_{u=1}^{U}\sum_{w=1}^{W} r_{u,w}\eta_{u,w} - \sum_{w=1}^{W}\lambda_w\left(\sum_{u=1}^{U} r_{u,w} - 1\right) - \sum_{u=1}^{U}\lambda_u\left(\sum_{w=1}^{W} r_{u,w} - 1\right) - \lambda_0\left(\sum_{u=1}^{U}\sum_{w=1}^{W} r_{u,w} - \min\left(U,W\right)\right)$$
$$= \sum_{u=1}^{U}\sum_{w=1}^{W} r_{u,w}\left(\eta_{u,w} - \lambda_w - \lambda_u - \lambda_0\right) + \frac{\lambda_w}{U} + \frac{\lambda_u}{W} + \frac{\lambda_0 \min\left(U,W\right)}{UW}. \quad (20)$$

resulting caching matrices in order to determine an adequate caching matrix that can serve all nodes. The complete solution, called GCBA, is presented in Algorithm 2. In the i_t^{th} loop of Algorithm 2, the caching policy $\mathbf{C}$ is built with $(1+S+U)$ operations. Then, $\mathbf{R}^*$ is obtained using (19). The search space of Algorithm 2 is $I_{max} \times (1+S+U)$, where I_{max} is the number of users of interest to consider.

Algorithm 2 GCBA

Set $\mathcal{U}' \subset \mathcal{U}$; Set I_{max} % Size of $\mathcal{U}'$
Set $\mathbf{C}_0 = \text{zeros}(1+S+U, F)$ % Overall Caching Matrix
Set $\mathbf{x} = [x_0,\ x_s, \forall s \in \mathcal{S},\ x_u, \forall u \in \mathcal{U}]$ % Vector indicating if a node's caching memory is full %
Set $i_t = 0$ % Number of iterations in while loop %
while $i_t < I_{max}$ **do**
    Set $\mathbf{C} = \text{zeros}(1+S+U, F)$ % User u based
    Caching Matrix %
    Randomly select a user of interest $u \in \mathcal{U}'$
    % Matrix $\mathbf{C}$ building %
    Set $\mathcal{I} = \{u\}$
    Set $\mathcal{F}' = \mathcal{F} \backslash \{0, 1, \ldots, C_U\}$
    while $\sum_{i=1}^{1+S+U} x_i < 1+S+U$ **do**
        Select $i = \underset{i' \in \mathcal{U}\cup\mathcal{S}\cup\{0\}\backslash\mathcal{I}}{\text{argmin}} G_{i'u}^{w}$
        Fill i's cache memory by setting $c_{i,f}$=1, f=1,..., C_i.
        Set $\mathcal{I} = \mathcal{I} \cup \{i\}$
        Set $\mathcal{F}' = \mathcal{F}' \backslash \{C_U+1, \ldots, C_i\}$
    end while
    $\mathbf{C}_0 = \mathbf{C}_0 + \mathbf{C}$
    Set $\mathcal{U}' = \mathcal{U}' \backslash \{u\}$
    $i_t = i_t + 1$
end while
Create $\mathbf{C}_0'$ such that $c'_{i,f}$=1 when i= $\underset{i'=1,\ldots,1+S+U}{\text{argmax}}\ \mathbf{C}_0(i', f)$, with respect to constraints $c_5 - c_7$ of problem (P1)
% Matrix $\mathbf{R}$ building %
Build $\mathbf{R}$ associated to $\mathbf{C}_0'$ using (19)
Calculate G associated to ($\mathbf{C}_0'$, $\mathbf{R}$) using (17)

V. Numerical Results

In what follows, we assume that one MBS occupies the center of a circular region, defined by a radius of 100 meters. Two SBSs are located at the vertical axis within 50 meters from the MBS. Locations of 60 users are randomly generated in the circular region as shown in Fig. 2. The following

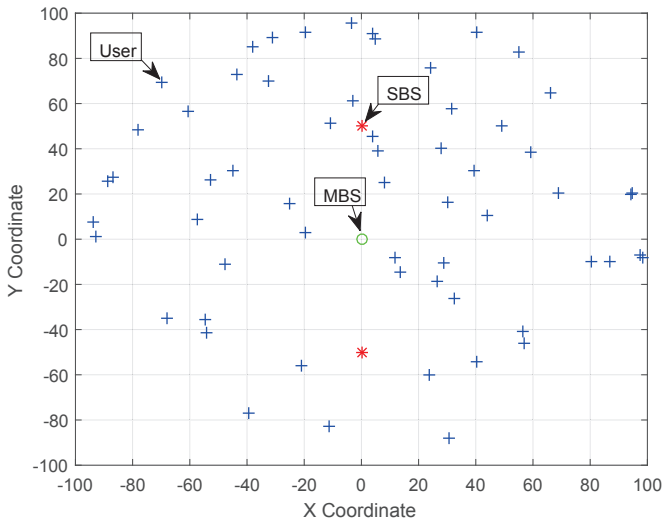

Fig. 2. Macro-cell Scenario: 1 MBS, 2 SBSs and 60 Users

parameters are fixed along this section: path-loss exponent $\alpha = 4$, number of channels is $W = 3$ where $\mathcal{B} = \{1, 2, 3\}$ (MHz), the number of files in the library is $F = 200$, the files length is $L = 1000$ bits, the SNRs are $\frac{P_0}{N_0} = 2\frac{P_S}{N_0} = 4\frac{P_U}{N_0}$, the time slot duration is $\tau = 0.01$ sec, and the average delay on the backhaul is $\bar{T}_0 = 50$ time slots (when not indicated).

In Fig. 3, we illustrate the macro-cell average delay as a function of transmit SNR ($\frac{P_U}{N_0}$) using the conventional Heterogeneous network (without D2D assistance) called "w.o D2D", the RS algorithm[1], and the GCBA algorithm. We assume that the caching capacities of nodes in the network are $C_U = 1$, $C_S = 2$ and $C_0 = 5$, the probability of an idle node is $q_0 = \theta = 0.4$, and the Zipf distribution popularity exponent $\beta = 0.5$. For $\bar{T}_0 = 10$, RS outperforms GCBA and "w.o. D2D" for almost all SNR values. Indeed, RS explores more caching scenarios, and hence has a better chance to minimize further the macro-cell average delay. Meanwhile, GCBA achieves similar average delay as "w.o. D2D" at very low SNR, and as SNR increases, its performance improves and gets very close to that of RS. That means, the proposed greedy algorithm achieves a sub-optimal solution for the problem with very low complexity. To be noted that GCBA is more interesting for SNR values above 6 dB in this case. However,

[1]It is to be noted that the total caching space is of dimension $2^{F\times(1+S+U)} = 2^{200\times 63}$. Hence, we propose to explore a sub-space limited to a fixed dimension $Dim = 2U$.

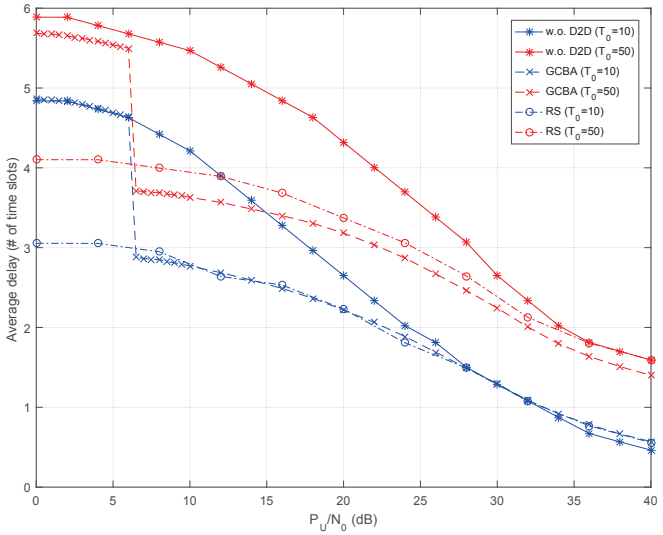

Fig. 3. Macro-cell Average Delay versus SNR (for different $\bar{T}_0$)

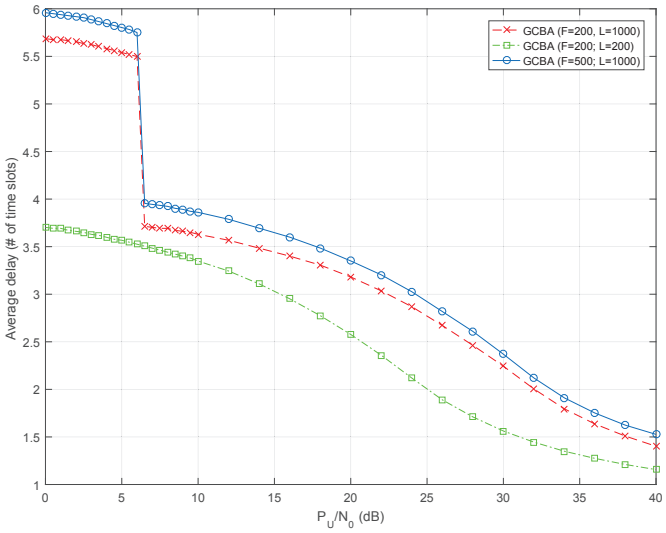

Fig. 5. Macro-cell Average Delay versus SNR (for different F and L)

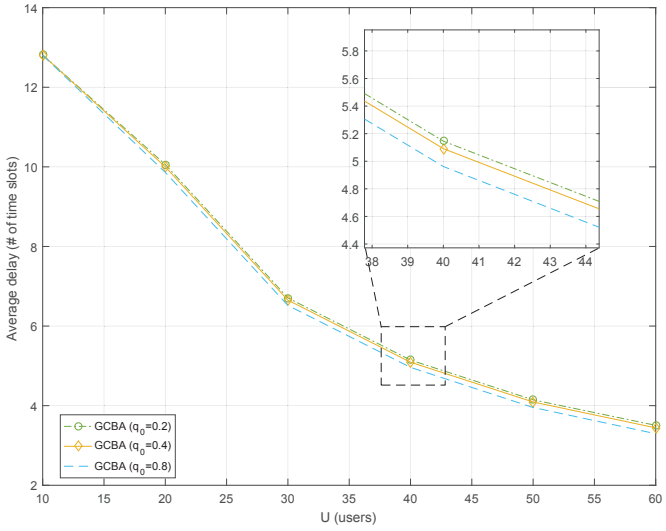

Fig. 4. Macro-cell Average Delay versus Number of Users (for different q_0)

for $\bar{T}_0 = 50$ and for SNR above 6 dB, GCBA achieves better performances than RS and "w.o. D2D". This can be due to the efficiency of the selected greedy criterion (transmission delay based), where for higher SNR values, the caching policy favors the D2D users. Finally, as $\bar{T}_0$ increases, the average delay degrades proportionally. This is expected since not all files can be cached within the network, and $F - (UC_U + SC_S + C_0)$ files have to be served directly from the core network through the backhaul link.

In Fig. 4, we investigate the impact of the number of users within the macro-cell on the average delay, with different $q_0 = \theta$ values. We set here $\frac{P_U}{N_0} = 15$ dB, and we keep the same remaining settings as Fig. 3. As the number of users increase, the average delay improves. Indeed, with more users in the network, more caching capacity is available and hence more D2D transmissions with low delays are exploited. With an increasing q_0, more D2D users are available to transmit, and hence the average delay is improved. However, the achieved gain is very small, about 0.2 time slots when q_0 increases from 0.2 to 0.8 at $U = 40$ users.

We study in Fig. 5 the influence of the number of files F and their length L on the average delay, for different SNR values. The parameters are fixed as for Fig. 3. As F increases, the average delay degrades. This is expected, since for the same caching capability of the system, more files will have to be obtained through the backhaul link. When L decreases, the files are smaller and hence they take less time to be transmitted. Consequently, the macro-cell average delay improves significantly. Notice that for SNR below 6 dB and L=200, the gap in performances seen in the other curves disappears. Indeed, for small-sized files, caching in users is favored, and hence more D2D transmissions are achieved with low delays.

VI. Conclusion

In this paper, we proposed joint multi-tier caching and resource allocation in a D2D-assisted heterogeneous network. The main objective is to minimize the macro-cell transmission delay averaged over the number of users within the macro-cell, their file demands and the channels variations. We derived at first a lower-bound expression for the average delay. Then, we formulated the optimization problem. Due to the problem's complexity and non-convexity, we have proposed a random search (RS) method and a greedy criterion-based algorithm (GCBA) that solve the problem. It has been shown that GCBA provides interesting performances that outperforms other schemes, while keeping a low complexity. Finally, the impact of parameters such as the backhaul average delay, the

probability of an idle user, the number of files within the core network library and the file size have been investigated.

REFERENCES

[1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[2] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, 4^{th} quarter 2014.

[3] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. J. P. C. Rodrigues, "5g d2d networks: Techniques, challenges, and future prospects," *IEEE Systems J.*, vol. PP, no. 99, pp. 1–15, 2018.

[4] W. Cao, G. Feng, S. Qin, and M. Yan, "Cellular offloading in heterogeneous mobile networks with d2d communication assistance," *IEEE Trans. on Veh. Tech.*, vol. 66, no. 5, pp. 4245–4255, May 2017.

[5] Y. Huang, A. A. Nasir, S. Durrani, and X. Zhou, "Mode selection, resource allocation, and power control for d2d-enabled two-tier cellular network," *IEEE Trans. on Commun.*, vol. 64, no. 8, pp. 3534–3547, Aug 2016.

[6] C. Wang, Y. He, F. R. Yu, Q. Chen, and L. Tang, "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges," *IEEE Commun. Surveys Tutorials*, vol. 20, no. 1, pp. 7–38, Firstquarter 2018.

[7] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tutorials*, pp. 1–1, 2018.

[8] 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.

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

[10] M. Gregori, J. Gmez-Vilardeb, J. Matamoros, and D. Gndz, "Wireless content caching for small cell and d2d networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[11] I. Pappalardo, G. Quer, B. D. Rao, and M. Zorzi, "Caching strategies in heterogeneous networks with d2d, small bs and macro bs communications," in *2016 IEEE Int. Conf. on Commun. (ICC)*, May 2016, pp. 1–6.

[12] Y. Li, M. C. Gursoy, and S. Velipasalar, "A delay-aware caching algorithm for wireless d2d caching networks," in *IEEE Conf. on Compu. Commun.s Worksh. (INFOCOM WKSHPS)*, May 2017, pp. 456–461.

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

[14] X. Peng, J. C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *IEEE Global Commun. Conf. (GLOBECOM)*, Dec 2015, pp. 1–6.

[15] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 1, pp. 248–257, January 2013.