Content Caching and Channel Allocation in D2D-Assisted Wireless HetNets

WAEL JAFAAR, (Senior Member, IEEE), AMINA MSEDDI, (Student Member, IEEE), WESSAM AJIB, (Senior Member, IEEE), AND HALIMA ELBIAZE, (Senior Member, IEEE)

1Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada
2Department of Computer Science, University of Quebec at Montreal, Montreal, QC H2L 2C4, Canada

Corresponding author: Wael Jaafar (waeljaafar@sce.carleton.ca)

This work was supported by the National Sciences and Engineering Research Council of Canada (NSERC).

ABSTRACT The fifth generation cellular networks are required to provide very fast and reliable communications while dealing with the increase of users traffic. In heterogeneous networks (HetNets) assisted with device-to-device (D2D) communications, traffic can be offloaded to small base stations or to users to make content delivery faster and alleviate the traffic burden from the core network. In this paper, we aim to maximize the probability of successfully delivering files to users, referred to as the successful delivery probability, by jointly optimizing the caching placement and channel allocation in cache-enabled D2D-assisted HetNets. First, an analytical expression of the average content delivery delay is derived. The latter is subsequently used to formulate the joint optimization problem of cache placement and channel allocation. Due to the problem’s non-convexity, a linearization transformation is proposed, which allows to find the optimal solution. However, given the high complexity of the problem, we propose a low-complex heuristic approach for channel allocation and caching. Numerical results illustrate the efficacy of the proposed solutions and compare them to the conventional HetNet. Finally, the impact of several key parameters, e.g., transmit power, caching capacity, and QoS requirements, is investigated, which provides design guidelines for D2D-assisted HetNets.

INDEX TERMS HetNets, device-to-device, D2D, caching, channel allocation, successful delivery probability, SDP.

I. INTRODUCTION

With the fast growth of new heavy mobile applications such as live streaming, language processing, and augmented reality, future wireless networks are required to focus not only on communication and connectivity, but also on high caching in order to satisfy stringent application demands [1].

The proposal of heterogeneous networks (HetNets) using small cells and different radio technologies improves the energy efficiency (EE) and spectral efficiency (SE) performances [2]. However, the ultra-dense small cell deployment and coexistence/interaction with existing macro base stations (MBSs) creates interference issues, which require efficient radio resource allocation and interference management techniques [3]. Moreover, device-to-device (D2D) communications have been envisioned as an essential component of the future wireless networks. By allowing two users (UEs) to communicate directly without going through the cellular network, radio resources can be further exploited to enhance SE, transmission delay, and offload the backhaul’s traffic [4].

On one hand, many works leveraged the advantages of D2D in the HetNet architecture. For instance, Malandrino et al. solved in [5] the uplink and downlink traffic scheduling problem of a long term evolution (LTE) two-tier HetNet, with support of D2D communications. Authors of [6] studied the power allocation and cell selection problems to maximize EE for an LTE advanced HetNet with support for D2D communications and relaying. They showed that a higher number of UEs can perform D2D communications even for stringent quality-of-service (QoS) requirements. In [7], Huang et al. proposed a centrally controlled framework for a D2D communication underlaying a two-tier cellular network. The objective was to maximize the sum rate of the network with individual transmit power and rate constraints. The formulated problem encompasses cellular, dedicated and shared D2D modes, frequency sharing and power control.

The associate editor coordinating the review of this manuscript and approving it for publication was Ivan Wang-Hei Ho.
The authors in [8] designed an offloading scheme based on the usage of small base stations (SBSs) to offload traffic from the MBS, and also to use some UEs as relays to reach users not in the coverage areas of the SBSs. Similarly, Tsirypoulos et al. proposed in [9] a cooperation framework for heterogeneous devices to improve the spectrum access. Finally, the authors of [10] optimized the EE of D2D communications in a hybrid cellular network composed of millimeter wave and microwave cells, using power control and radio resource allocation, whereas Hao et al. investigated in [11] the EE-SE tradeoff for D2D communications underlying a HetNet. They proposed a two-stage solution to solve the power and spectrum allocation problems.

On the other hand, caching has attracted attention due to its ability to reduce the backhaul traffic and eliminate duplicate transmissions of popular content [1], [12]. Caching has been investigated for use in different network types, namely with macro cells, small cells, or D2D [13]. Peng et al. proposed in [14] content placement in MBSs to minimize the average download delay of files by users. Also, the authors of [15] maximized the cache hit probability when a UE can be served by several MBSs. Kreishah et al. investigated in [16] cooperative content placement among MBSs to minimize both the download and caching costs. Alternatively, Shanmugam et al. optimized in [17] cooperative content caching in a backhaul constrained HetNet, aiming to minimize the content download delay. Moreover, authors in [18] designed a caching method for small cells where users’ trajectories between cells is known beforehand. Also, the authors of [19] investigated the throughput-outage tradeoff of wireless networks, where clustered device caching via D2D communications is exploited. Zhang et al. optimized in [20] the D2D link scheduling and power allocation problems to maximize the system’s throughput. In addition, authors of [21] maximized the offloading gain of cache-enabled D2D networks by jointly optimizing caching and scheduling policies. Yi et al. proposed in [22] a traffic offloading framework with social-aware D2D content sharing and caching. In [23], the authors presented a caching algorithm to minimize the average transmission delay in a D2D-enabled macro-cell. The proposed algorithm performed better than popularity-based naive caching. This work was extended in [24] to unknown system files popularity and varying system parameters over time, where a non-parametric learner is used to estimate the intensity function of file requests. Authors of [25] proposed a caching policy for millimeter-wave (mmWave) D2D-based cellular networks, where content is strategically disseminated and exchanged among D2D users for maximized offloading gain, while Amer et al. designed in [26] a D2D caching framework with inter-cluster cooperation, where nodes of the same cluster cooperate through D2D communications, while nodes of different clusters exchange data through cellular transmissions. The formulated problem aimed to minimize the network’s average delay under caching, queuing and energy constraints. Sun et al. proposed in [27] mobility and delay-aware caching methods for D2D-enabled cellular networks. The aim is to minimize the average transmission and cache leasing costs. They showed that their successive convex approximation (SCA)-based and greedy algorithms consume less cost that benchmarks methods. In [28], Zhang et al. studied user preference and transmission coverage region aware D2D caching deployment methods, which aimed to maximize a cache utility function, while in [29], the authors investigated multi-winner auction based cache placement in order to minimize the traffic load and average content access delay of D2D-assisted cellular networks. Finally, authors of [30] investigated the segmented caching problem in D2D overlay networks, aiming to minimize the average download latency. They proposed a dynamic caching algorithm that outperforms random and popularity-based caching.

Caching has been also studied in integrated D2D-assisted multi-tier networks. For instance, Yang et al. investigated in [31] caching at the UEs, where each user can obtain the content from its own cache or from the cache of an SBS. The demonstrated significant energy savings compared to no cache-enabled UEs, whereas authors of [32] proposed and analyzed cache-based content delivery in a three-tier HetNet, where popular contents are disseminated in SBSs and in a part of the UEs. Li et al. developed in [33] a multi-tier cache-enabled HetNet framework, where MBS, SBSs and pico BSs coexist. They proposed an optimal distributed caching scheme to maximize the successful delivery probability and proved that the optimal solution depends on the cache sizes and BSs densities. In [34], the authors proposed a similar framework, but they aimed to minimize the average file transmission delay through caching and bandwidth allocation. Also, Quer et al. proposed in [35] a proactive caching strategy that takes into account user mobility and user interest classes for content. In our previous work [36], we optimized caching and bandwidth allocation to minimize the average transmission delay of a D2D-assisted HetNet. We divided the problem into bandwidth allocation and caching problems. The first problem was optimally solved. Then, given a bandwidth allocation strategy, a random search algorithm and a greedy algorithm were proposed to determine the caching policies. Also, we investigated in [37] the caching placement problem in a D2D-assisted HetNet, where we successfully derived the optimal solution that minimizes the average content delivery delay. In [38], the authors aimed to minimize the EE of D2D-aided HetNets through SBSs and D2D data offloading and SBSs power optimization. Finally, Fu et al. proposed in [39] cooperative content caching and delivery based on multicast. They solved the problem using a hybrid genetic algorithm (HGA) and demonstrated its superiority compared to conventional algorithms. The related works as summarized in Table 1.

The aforementioned works studied caching placement and delivery in D2D-aided HetNets in different ways. However, most of them relied on the assumption of slow fading channels during content delivery. Such an assumption is valid in a controlled fixed environment, however, due to shadowing.
| Ref. | Content caching location | Optimization Problem | Objective | Features | Methods |
|------|--------------------------|-----------------------|-----------|----------|---------|
| [14] | MBSs                     | ✓ - - -               | Min. avg. download delay | - Limited caching at MBSs - Backhaul delay awareness | - SCA-based algorithm |
| [15] | MBSs                     | ✓ - - -               | Max. cache hit prob. Min. avg. download delay | - BSs coverage awareness | - Heuristic bisection algorithm |
| [16] | MBSs                     | ✓ - - -               | Min. caching and download costs | - Collaborative caching - Limited caching - Non-coded or coded data | - Dynamic programming (non-coded data) - Linear programming (coded data) |
| [17] | MBSs + SBSs (helpers)    | ✓ - - -               | Min. download delay | - Low-rate backhaul - High storage capacity - Non-coded or coded data | - Greedy algorithm (non-coded data) - Linear programming (coded data) |
| [18] | SBSs                     | ✓ - - -               | Max. caching utility | - Mobility awareness | - Heuristic algorithm with bounded approximation ratio |
| [19] | UEs                      | ✓ - - -               | Throughput-outage tradeoff analysis | - Asynchronous content reuse - D2D communications in microwave or mmWave frequencies | - Decentralized random caching - Coded multicasting |
| [20] | UEs                      | - - ✓ ✓               | Max. throughput | - Randomly cached data - Limited storage capacity - Limited transmit power | - D2D link scheduling algorithm to select the largest sum of D2D links satisfying SINR and transmit power constraints - Optimal power allocation algorithm to maximize the transmission rate of links |
| [21] | UEs                      | ✓ ✓ ✓ ✓               | Max. successful offloading probability | - Collaborative D2D distance - D2D communication interference | - Optimized random scheduling policy - Low-complex bisection search based solution for D2D caching |
| [22] | UEs                      | ✓ ✓ ✓ ✓               | Max. utility function | - Social awareness - Incentivizing D2D caching and delivery | - Basic transformation method for joint resource management - Dynamic pricing scheme for utility design |
| [23] | UEs                      | ✓ - - -               | Min. avg. transmission delay | - Limited storage capacity - Varying channels and system parameters | - Greedy algorithm based on finding the best context-UE pairs (delay wise) iteratively |
| [24] | UEs                      | ✓ - - -               | Min. avg. transmission delay | - Unknown content popularity - Varying system parameters in time | - Non-parametric estimator learns the intensity function of file requests - Greedy algorithm based on finding the best context-UE pairs (delay wise) |
| [25] | UEs                      | ✓ - - -               | Max. offloading gain | - Min/Max D2D communications - Half or full-duplex D2D transmissions | - Optimized content partitioning and random caching in UEs |
| [26] | UEs                      | ✓ - - -               | Min. avg. delay | - Content delivery via direct D2D or two-hop cellular transmissions - Queuing and energy limitations - D2D inter-cluster cooperative content delivery | - Problem reformulation to min. of supermodular function s.t. uniform partition matrix constraints, which has a greedy algorithm solution |
| [27] | UEs                      | ✓ - - -               | Min. avg. transmission = caching leasing costs | - Mobility awareness - Derivation of avg. network cost and avg. file delivery delay | - SCA-based iterative caching algorithm |
| [28] | UEs                      | ✓ - - -               | Max. utility function | - User preference awareness - Cache utility function based on user preference and coverage region | - Dual decomposition method to get a near-optimal solution for logarithmic utility function |
| [29] | UEs                      | ✓ - - -               | Min. traffic load and avg. content access delay | - Mobility awareness - Auction-based caching | - Near-optimal caching policy based on semi-definite programming - Heuristic multi-winner repeated auction based caching policy |
| [30] | UEs                      | ✓ - - -               | Min. avg. download latency | - Limited storage capacity - Heterogeneous content sizes - Soft or D2D caching | - Iterative utility-based algorithm |
| [31] | HeNet (SBSs + UEs)       | ✓ - - -               | Min. energy consumption | - Limited storage capacity - Content popularity awareness - Cooperative cache access protocol | - Heuristic algorithm |
| [32] | HeNet (SBSs + UEs)       | - - - -               | Avg. ergodic rate and outage probability analysis | - Limited storage capacity - Push-and-cache enabled data access - Tier association protocol | - Stochastic geometry analysis |
| [33] | HeNet (MBSs + SBSs + UEs) | ✓ - - -               | Max. successful delivery prob. | - Limited storage capacity - High signal-to-noise ratio regime | - Optimal probabilistic caching using the standard interior point method |
| [34] | HeNet (MBSs + SBSs + UEs) | ✓ ✓ - -               | Min. avg. file transmission delay | - Limited cache/buffer capacity - Limited bandwidth - Heterogeneous content sizes | - Optimal bandwidth allocation through KKT conditions checking - Iterative and simulated annealing cache placement algorithms |
| [35] | HeNet (MBSs + SBSs + UEs) | ✓ - - -               | Min. system cost (energy/bandwidth) | - User mobility - User interest classes - Heterogeneous file popularity profiles | - Cyclic heuristic algorithm among devices for proactive caching |
| [36] | HeNet (MBSs + SBSs + UEs) | ✓ ✓ - -               | Min. avg. transmission delay | - Limited storage capacity - Limited bandwidth - Upper-bound derivation of avg. transmission delay | - Optimal bandwidth allocation through KKT conditions checking - Caching through random search - Caching using a greedy algorithm based on minimizing avg. transmission delay for each content |
| [37] | HeNet (MBSs + SBSs + UEs) | ✓ - - -               | Min. avg. content delivery delay | - Limited storage capacity - Upper bound derivation of avg. content delivery delay | - Optimal placement policy for a single file through KKT conditions checking - Heuristic algorithm to iteratively cache multiple files |
TABLE 1. (Continued.) Comparison of related works for content caching and delivery. (res. alloc.) refers to resource allocation, (sched.) to user scheduling, and (pow. alloc.) to power allocation.

| Ref   | Content caching location | Optimization Problem | Objective | Features                                                                 | Methods                                      |
|-------|--------------------------|-----------------------|-----------|--------------------------------------------------------------------------|----------------------------------------------|
| [38]  | HetNet (SBSs + UEs)      | -                     | Max. energy efficiency | -D2D content delivery, SBSs power control, Derivation of EEE expression | Iterative Dinkelbach method                  |
| [39]  | HetNet (MBS + SBSs + UEs) | ✓ - ✓                 | Min. avg. download delay | Multicast content delivery, Cooperative caching within and among different HetNet levels | DGA based content caching                    |
| This work | HetNet (MBS + SBSs + UEs) | ✓ ✓ ✓                 | Max. successful delivery probability | Limited storage capacity, Limited bandwidth, User interest classes, Heterogeneous file popularity profiles, Fast varying channels, Upper-bound derivation of avg content delivery delay | Iterative linearization to find optimal solution, Geometry-based channel allocation algorithm, Iterative caching algorithm |

and/or mobility, the channel rapidly varies during transmissions. Moreover, many of these works assumed orthogonal transmissions, thus ignoring the interference management issue. In an attempt to overcome these shortcomings, we propose in this paper to investigate a multi-tier D2D-assisted HetNet, taking into account caching at all tiers, interference among simultaneous transmissions, and accounting for the effect of fast varying channels on content delivery delay.

The main objective is to maximize the successful delivery probability (SDP), defined as the probability of successfully delivering files to users, under channel resources, caching, and delay constraints. To the best of our knowledge, this is the first work that investigates the joint caching placement and delay constraints. To the best of our knowledge, this is the first work that investigates the joint caching placement and channel allocation problem for SDP maximization, given the effect of fast varying channels on content delivery delay.

The rest of the paper is organized as follows. In Section II, the system model is presented. Section III derives the content delivery delay and formulates the optimization problem. Section IV details the proposed solutions, while in Section V, numerical results are presented. Finally, Section VI closes the paper.

II. SYSTEM MODEL

In this section, we describe the communication model and the content caching and delivery model. Symbols used within the remaining of the paper are summarized in Table 2.

A. COMMUNICATION MODEL

The network consists of one MBS, S SBSs and U UEs, where SBS s ∈ S ≡ {s₁, . . . , s₅} and UE u ∈ U ≡ {u₁, . . . , u₅}, such that U > S > 1. Both SBSs and UEs are deployed randomly within the MBS’s coverage area, as illustrated in Fig. 1.

We assume that the MBS and SBSs can share the same frequency channels. Moreover, D2D communications reuse the same cellular channels as the BSs [40]. We assume that intra-cell interference occur when channels are reused at the same time. The number of orthogonal frequency channels is set to W, each with bandwidth B, and such that w ∈ W ≡ {w₁, . . . , w₆}. We define by rₓ,w the binary variable indicating whether channel w is allocated to the transmission having user u as its receiver or not, such that ∑ᵢ₌₁⁶ rₓ,w = 1, i.e., a transmission is allocated a single channel. Moreover, we assume that time is divided into slots (TSs), and in each TS t ∈ N, transmissions may occur for one of the two following purposes: 1) Off-peak content placement, where caching memories of the nodes are updated during low traffic periods, and 2) content delivery to satisfy users’ requests. In our system, the MBS, SBS s and user u transmit signals using powers Pₘ, Pₛ and Pₚ, respectively, such that Pₘ > Pₛ > Pₚ. Also, we assume that wireless channels are modeled as Rayleigh channels that are constant within a TS of duration τ, but vary from a TS to another. Finally, the communications experience additive white Gaussian noise (AWGN) with zero mean and variance σ₀²).

B. CONTENT CACHING AND DELIVERY MODELS

1) CONTENT CACHING

We assume that F files of the same size L kilobits (kb) belong to a library F ≡ {1, . . . , F}, which is fully stored in the
We assume that the MBS, SBSs and UEs have caching capacities $C_m$, $C_s$, and $C_u$ Mb respectively, such that $FL > C_m > C_s > C_u$, $\forall s \in S$ and $u \in U$.

We assume that user $u$ belongs to a specific class of interest $k \in K = \{1, \ldots, K\}$ that provides a specific ranking order of the file popularity. We introduce the probability that user $u$ belongs to class $k$, defined as $p_k^u \in [0, 1]$, where $\sum_{k=1}^{K} p_k^u = 1$, $\forall u \in U$. The definition of $p_k^u$ is motivated by the fact that the user’s interests may vary in time due to its environment, behavior or events [35]. For each class $k$, the file popularity follows Zipf distribution and the probability that user $u$ in class $k$ requests file $f$ can be given by

$$q_f^k = \eta(f, k)\beta / \left( \sum_{f'=1}^{F} \eta(f', k)\beta \right), \quad \forall f \in F, \forall k \in K,$$

where $\eta(f, k)$ (resp. $\eta(f', k)$) is the rank of file $f$ (resp. of file $f'$) for a user in class $k$, and $\beta \geq 0$ reflects how skewed the popularity distribution is, that means larger $\beta$ exponents correspond to higher content reuse, i.e., the first few popular files account for the majority of requests.

### TABLE 2. Notations.

| Symbol | Description | Symbol | Description |
|--------|-------------|--------|-------------|
| $A_i$  | A set of polygons | $q_f^k$ | Probability to request file $f$ by UE in class $k$ |
| $B$    | Bandwidth of one frequency channel | $Q_{u,f}$ | Avg. popularity of file $f$ for UE $u$ |
| $C$    | Set of all caching within the network | $Q_f$ | Overall popularity of file $f$ in the network |
| $C_m, C_s, C_u$ | Caching capacity of MBS, SBS $s$, and UE $u$ | $r_{u,w}$ | Binary indicator of assigning channel $w$ to transmission with receiver UE $u$ |
| $C_i, c_i,f$ | Binary caching status matrix, binary indicator of caching file $f$ in node $i$ | $R$ | Reuse factor of channels |
| $d_{i,j}$ | Distance between nodes $i$ and $j$ | $R_{ij}(t)$ | Data rate of transmission $i - j$ in TS $t$ |
| $D_{th}$ | Successful delivery delay threshold | $S, S'$ | Set and number of SBSs |
| $D_{u,f}$ | Avg. delay to deliver file $f$ to UE $u$ | $SDP_f$ | Successful delivery probability of file $f$ |
| $F, F'$ | Set and total number of files | $\text{SINR}_{ij}(t)$ | SINR of transmission $i - j$ in TS $t$ |
| $G_{i,j}(I_j)$ | Upper-bound of avg. content delivery delay for transmission $i - j$ | $T_{i,j}, T_{i,j'}$ | Content delivery delay and avg. delay for transmission $i - j$ |
| $h_{i,j}(t)$ | Channel coefficient of transmission $i - j$ | $T_0$ | Avg. delivery delay on the backhaul link |
| $I_j$ | Set of interferers on receiver $j$’s reception | $U, U'$ | Set and number of UEs |
| $K, K'$ | Set and number of UE classes | $W, W'$ | Set and number of frequency channels |
| $P_{m}, P_{s}, P_{c}$ | Transmit power of MBS, SBS, UE, and node $i$ | $X, x_{u,f}$ | Binary success delivery matrix, binary indicator of successful delivery of file $f$ to UE $u$ |
| $p_u^k$ | Probability of UE $u$ to belong to class $k$ | $\alpha$ | Path-loss exponent |
| $p_i^j$ | Perimeter of $j^{th}$ polygon in set $A_i$ | $\tau$ | Duration of one TS |
| $p_i$ | Avg. polygon perimeter in set $A_i$ | $\eta(f, k)$ | Rank of file $f$ for UEs in class $k$ |
| $P_{j,1}, P_{j,2}$ | Polygons in set $A_i$ | $\sigma_0^k$ | Noise power |

![FIGURE 1. System model.](image-url)
For the sake of simplicity, we assume that a file cannot be cached at more than one node within the network. Indeed, redundant caching may lead to a fast saturation of storage resources, hence slowing down caching updates when old files need to be replaced by new ones. Moreover, caching at node \( i \in C = U \cup S \cup \{MBS\} \) is represented by a binary vector \( \mathbf{c}_i = [c_{i,1}, \ldots, c_{i,F}] \), where \( c_{i,f} \) indicates whether node \( i \) caches file \( f \) or not [13].

2) CONTENT DELIVERY

When user \( u \) requests file \( f \), it starts by checking its own cache memory. If the content is available locally, it is obtained directly without any delay. Otherwise, the request is forwarded to the system controller (typically located within the MBS), and the latter decides which device (among MBS, SBSs, UEs, or core network) will serve user \( u \)’s request. For proper operation, it is assumed that the controller, e.g., MBS, has knowledge of the nodes’ caching status and of the statistics of the channel states within its cell. Also, we assume that a file delivery occupies one channel resource only, but can span over several channel varying TSs and is realized by a single source node only. Finally, only the MBS and SBSs can deliver simultaneously different contents to UEs using orthogonal channels, while a D2D transmitter can serve only one user\(^1\).

III. CONTENT DELIVERY DELAY EXPRESSION AND PROBLEM FORMULATION

In this section, we define the system’s average content delivery delay and derive its upper-bound expression. The latter is then used to formulate the SDP optimization problem.

A. CONTENT DELIVERY DELAY AND UPPER-BOUND DERIVATION

The transmission delay between two nodes \( i \) and \( j \) is defined as the minimum number TSs to transmit a given file \( f \) from \( i \) to \( j \). For a time varying channel, the average content delivery delay can be written as

\[
T_{ij} = \min \left\{ T; L \leq \sum_{t=1}^{T} \tau_{R_{ij}(t)} \right\},
\]

where \( R_{ij}(t) \) is the channel’s data rate in TS \( t \). Without loss of generality, the data rate can be expressed by

\[
R_{ij}(t) = B \log_2 \left( 1 + \frac{P_{t} |h_{ij}(t)|^2}{\sigma_i^2 + \sum_{r \in I_j} P_{r} |h_{ir}(t)|^2} \right) = B \log_2 \left( 1 + \text{SINR}_{ij}(t) \right),
\]

where \( P_{t} \) is the transmit power of node \( i \in C, h_{ij}(t) = h_{ij}^{(T)}(t) \) is the channel coefficient capturing both short-scale \( h_{ij}^{(T)}(t) \) and long-scale \( (d_{ij}/d_0)^{-\alpha} \) fading, with \( d_0 \) is a reference distance, \( \alpha \) is the path-loss exponent, \( |h_{ij}(t)|^2 \) is the channel gain following an exponential distribution of zero mean and variance \( (\frac{d_{ij}}{d_0})^{-\alpha} \), \( \sum_{r \in I_j} P_{r} |h_{ir}(t)|^2 \) is the interference signal, and \( I_j \) is the set of interferers. The average content delivery delay of link \( i - j \), denoted by \( \bar{T}_{ij} \), is defined by [14]

\[
\bar{T}_{ij} = \mathbb{E} [T_{ij}] = \sum_{t=1}^{+\infty} \mathbb{P} [T_{ij} > T]
\]

where \( \mathbb{E} \) is the expectation operator and \( \mathbb{P} [T_{ij} > T] \) is the probability that \( T_{ij} \) is above \( T \). This probability is given by

\[
\mathbb{P} [T_{ij} > T] = \mathbb{P} \left[ \sum_{t=1}^{T} \log_2 \left( 1 + \text{SINR}_{ij}(t) \right) < \frac{L}{\tau B} \right] = \mathbb{P} \left[ \sum_{t=1}^{T} Y_t < \frac{L}{\tau B} \right] = \mathbb{P} \left[ Z_T < \frac{L}{\tau B} \right].
\]

where \( Y_t = \log_2 \left( 1 + \text{SINR}_{ij}(t) \right) \) and \( Z_T = \sum_{t=1}^{T} Y_t \).

Theorem 1: \( \mathbb{P} [T_{ij} > T] \) is upper-bounded by:

\[
\mathbb{P} [T_{ij} > T] \leq \xi_0(T) := \min_{t>0} \left\{ \frac{1}{\mathbb{E}[\text{SINR}_{ij}(t)]} \left( 1 + \sum_{r \in I_j} \mathbb{E}[\text{SINR}_{ir}(t)] \right) \right\}^T,
\]

when \( I_j = \emptyset \), and by

\[
\mathbb{P} [T_{ij} > T] \leq \xi_1(T, I_j) := \xi_0(T) \left( 1 + \sum_{r \in I_j} \mathbb{E}[\text{SINR}_{ir}(t)] \right)^	op,
\]

when \( I_j \neq \emptyset \), where \( \mathbb{E}[\text{SINR}_{ij}(t)] = \sum_{r \in I_j} \mathbb{E}[\text{SINR}_{ir}(t)] \). I_{ij} is the size of set \( I_j \), and \( \Gamma(a, x) = \int_x^{+\infty} s^{a-1} e^{-s} ds \) is the incomplete gamma function.

Proof: Refer to Appendix A.

Consequently, the average content delivery delay of link \( i - j \) is upper-bounded by

\[
\bar{T}_{ij} \leq G_{ij}(I_j) := \sum_{T=1}^{+\infty} g_{ij}(T, I_j),
\]

where \( g_{ij}(T, I_j) \) is the average delay of delivering file \( f \) to user \( u \). For the sake of simplicity and expression tractability, we assume that when interference occurs, it is dominated by the strongest interferer. Consequently, \( \bar{T}_{0} \) can be upper-bounded by (9), as shown at the bottom of the next page, where \( I_0 \) is the set of potential transmitters of file \( f \) to user \( u \), \( \bar{T}_{0} \) is the average transmission delay on the backhaul link, \( \mathbb{P} \left[ T_{u,f} = T \right] \) is the probability that no node is a potential transmitter of file \( f \) to user \( u \), and \( \mathbb{P} \left[ I_j = \emptyset \right] \) is the probability that no node is interfering on user \( u \)’s communication, and \( \mathbb{P} \left[ I_j = \{y\} \right] \) is the probability

---

\(^1\)For the sake of simplicity and tractability, we ignore D2D multicasting in this work, which will be left for a future work.
that the transmission of node $y$ interferes on user $u$ reception. Specifically,

$$
\mathbb{P}\left[ T_{u,f} = \{x\} \right] = c_{x,f} \prod_{x' \in \mathcal{C} \setminus \{u\}} \left(1 - c_{x',f}\right),
$$

(10)

$$
\mathbb{P}\left[ T_{u,f} = \emptyset \right] = \prod_{x \in \mathcal{C} \setminus \{u\}} \left(1 - c_{x,f}\right),
$$

(11)

and

$$
\mathbb{P}\left[ \mathcal{I}_u = \{y\} \right] = \sum_{w=1}^{W} \sum_{u' \in \mathcal{U} \setminus \{u,x,y\}} \left( r_{u,w} r_{u',w} \sum_{k=1}^{K} \sum_{f'=1}^{F} p_{k} q_{f'} \right) \times \mathbb{P}\left[ T_{u,f'} = \{y\} \right],
$$

(12)

where the first two sums in (12) correspond to the channels allocated to the users and the next two sub-sums correspond to the class and the file requested by user $u' \neq u$. Finally, $\mathbb{P}\left[ \mathcal{I}_u = \emptyset \right]$ is written as

$$
\mathbb{P}\left[ \mathcal{I}_u = \emptyset \right] = \sum_{w=1}^{W} r_{u,w} \prod_{u' \in \mathcal{U} \setminus \{u\}} \left(1 - r_{u',w}\right).
$$

(13)

### B. PROBLEM FORMULATION

We formulate the joint caching and channel allocation problem aiming to maximize the SDP performance as follows:

$$\max \quad \text{SDP} = \frac{1}{U} \sum_{j=1}^{U} \sum_{u=1}^{U} \sum_{k=1}^{K} p_{k} d_{k} x_{u,j}$$

s.t. $\sum_{j=1}^{U} c_{i,j} \leq C_i / L$, \quad $\forall i \in \mathcal{C}$,

$$\sum_{i=1}^{K} c_{i,j} \leq 1$, \quad $\forall j \in \mathcal{F}$,

$$\sum_{w=1}^{W} r_{u,w} = 1$, \quad $\forall u \in \mathcal{U}$,

$$\sum_{j=1}^{U} r_{u,j} \leq R$, \quad $\forall w \in \mathcal{W}$,

$$x_{u,j} \in \{0,1\}$, \quad $\forall u \in \mathcal{U}$,

$$c_{i,j} \in \{0,1\}$, \quad $\forall i \in \mathcal{C}$, \quad $\forall j \in \mathcal{F}$,

$$r_{u,w} \in \{0,1\}$, \quad $\forall u \in \mathcal{U}$, \quad $\forall w \in \mathcal{W}$ (P1.a)

(14)

Finally, if $\mathcal{I}_u = \emptyset$, $\mathbb{P}\left[ T_{u,f} = \emptyset \right]$ is written as

$$\mathbb{P}\left[ T_{u,f} = \emptyset \right] = \sum_{w=1}^{W} r_{u,w} \prod_{u' \in \mathcal{U} \setminus \{u\}} \left(1 - r_{u',w}\right).$$

(15)

The defined optimization problem is non-linear due to the several product terms within the average content delivery delay expression in (P1.a), and is non-convex due to the binary constraints. Nevertheless, in what follows, we propose a method to transform it into an ILP and solve it optimally.

### IV. PROPOSED SOLUTIONS

We start in this section by proposing an approach to solve the problem optimally. Then, we develop a heuristic method that provides sub-optimal solutions but with significantly lower complexity.

### A. OPTIMAL SOLUTION DESIGN

In order to make problem (P1) tractable, we propose to substitute iteratively the products of the binary variables in function $G_{u,f}$ by new single variables [35], [41], as shown at the bottom of the next page. The resulting new problem (P2), as shown at the bottom of the next page, where in (P2.a), as shown at the bottom of the next page, $G_{u,f}$ is expressed by (14), as shown at the bottom of the next page, and in (P2.b)–(P2.e), as shown at the bottom of the next page, we have:

- $\Omega_{u,w,c_i}^x = \delta_{u,w} \phi_i \chi_i$, $\forall u \in \mathcal{U}$, $\forall w \in \mathcal{W}$ (P1.a)
- $\Lambda_{u,w,c_i}^x = \gamma_{u,w,c_i}$, $\forall u \in \mathcal{U}$, $\forall w \in \mathcal{W}$ (P1.b)
- $\phi_i = 1 - c_i$, $\forall i \in \mathcal{C}$ (P1.c)
- $\chi_i = 1 - c_i$, $\forall i \in \mathcal{C}$ (P1.d)
- $\Omega_{u,w,c_i}^y = \delta_{u,w} \phi_i \chi_i$, $\forall u \in \mathcal{U}$, $\forall w \in \mathcal{W}$ (P1.e)

Finally, if $\mathcal{I}_u = \emptyset$, $\mathbb{P}\left[ T_{u,f} = \emptyset \right]$ is written as

$$\mathbb{P}\left[ T_{u,f} = \emptyset \right] = \sum_{w=1}^{W} r_{u,w} \prod_{u' \in \mathcal{U} \setminus \{u\}} \left(1 - r_{u',w}\right).$$

(15)
Hence, problem (P2) is an ILP, which we model using the AMPL language [42], then we solve it using the IBM CPLEX Optimizer [43].

To study the solving complexity of problem (P2), we determine the total number of variables involved. For the variables in $\mathbf{x}$, the expected number of nonzero $x_{u,f}$ for each file $f$ can be defined by $N$. In (14), the number of variables is dominant in the last term. Indeed, we can rewrite

$$\Lambda_{x,y}^{u,v,w,F} = r_{u,w} r_{u',w'} \phi_x \phi_y$$

(15)

\[ \forall u \in U, u' \in U \setminus \{u\}, w \in W, f \in F, \text{ and } f' \in F \setminus \{f\}. \]

The maximum number of nonzero $r_{u,w}$ and $r_{u',w'}$ is limited by $R$, $\forall w \in W$, and that of $\phi_x$ (resp. $\phi_y$) is bounded by $1$, $\forall f \in F$ and $\forall x \in C \setminus \{u, u', x\}$, then, the total number of variables in the optimization problem (P2) is in the order of

$$\sum_{f=1}^{2N} \sum_{w=1}^{W} \sum_{u=1}^{U} \sum_{u' \neq u}^{U} \sum_{f'=1}^{F} \sum_{x \in C}^{x \in C} \sum_{y \in C}^{y \in C} 2^{2(R+1)} = O \left( WU(U-1)F^2(F-1)(|C|-2)(|C|-3)2^{N+2R+2} \right).$$

(16)

**B. HEURISTIC SOLUTION DESIGN**

Due to the high complexity to solve (P2) optimally, we propose the following heuristic channel allocation and caching algorithms. First, following a geometrical approach, channel allocation is realized to minimize the overall interference. Then, caching is performed to maximize the number of popularity-wise file-user pairs.

1) CHANNEL ALLOCATION

This process depends on the available number of channels and on the number of users. Indeed, if $W \geq U$, it is optimal to dedicate one channel per user. However, if $W < U$, channel allocation has to be adequately designed in order to minimize interference between simultaneous transmissions. To do so, we define the following parameters. Let $X = \text{div}(U, W)$ and $R' = \text{mod}(U, W)$ be the quotient and remainder of the Euclidean division of $U$ by $W$, respectively. For channel usage fairness, we impose that for $(W - R')$ channels, each channel has to be shared by $X$ users, while for the remaining $R'$ channels, each one is shared by $X + 1$ users. Finally, we set $R = X + 1 = \left( \frac{U}{W} \right)$.

Let $A_i = \{P_{1}^{i},...,P_{R'}^{i},\bar{P}_{1}^{i},...,\bar{P}_{R'}^{i}\} \cup \{P_{W-R'}^{j}\}$ be a set of $W$ polygons, where each polygon $P_{j}^{i}$ is a set of $X + 1$ vertices (users), $\forall j \in \{1,...,R'\}$, and each $\bar{P}_{j}^{i}$ is a set of $X$ vertices, $\forall j \in \{1,...,W - R'\}$, where

$$\bigcup_{j=1}^{R'} P_{j}^{i} \cup \bigcup_{j=1}^{W-R'} \bar{P}_{j}^{i} = \{1,...,U\}.$$


where in Algorithm 2, at each

Algorithm 2 Caching Policy Algorithm
1: Set $C = [c_{i,f}] = 0_{(|C|×|F|)}$ % Caching Matrix
2: $\forall f \in F$, calculate $Q_f$
3: Rank $Q_f$ from the largest to the smallest value in set $Q$
4: for $Q_f \in Q$ do
5: Set $C' = C$
6: Calculate SDP$_f'$ when $f'$ is placed in node $i$, $\forall i \in C'$
7: Find $i_0 = \arg\max_{i \in C'}$ SDP$_f'$ w.r.t. (P1.a)–(P1.c) and (P1.f)–(P1.g)
8: if $\sum_{f=1}^{F_0} c_{i_0,f} < C_i'$ then
9: Set $c_{i_0,f'} = 1$
10: else
11: $C' = C' \setminus \{i_0\}$
12: Return to step 6
13: end if
14: end for
15: Return $C$

2) CACHING PLACEMENT

For a given channel allocation strategy, problem (P1) can be decomposed into multiple cache placement problems described as

$$\max_{X,C} \text{SDP}_f = \frac{1}{U} \sum_{u=1}^{U} Q_{u,f} x_{u,f}$$

$$\text{s.t. (P1.a) – (P1.c), (P1.f) – (P1.g)}$$

where $\text{SDP}_f$ is the successful delivery probability of file $f$ and $Q_{u,f} = \sum_{i=1}^{K} p_k^u d_i^{f}$ is the averaged file $f$ popularity for user $u$, with respect to the $K$ user classes.

We define $Q_f = \sum_{u=1}^{U} Q_{u,f}$ the popularity metric of file $f$. By ranking $Q_f$ from the largest to the smallest value in a set $Q$, an iterative file placement approach can be implemented. The latter is described by the following steps:
1) Starting with the most popular file in $Q$, we find its best location within the network that achieves the maximal $\text{SDP}_f$, with respect to the constraints; 2) we update the caching memories; 3) we move to the next file in $Q$ and repeat the previous two steps; 4) we repeat the previous step until all files are processed. This approach is summarized in Algorithm 2.

The heuristic algorithm complexity can be seen as the dominant complexity issued from Algorithms 1 and 2. Complexity of Algorithm 1 is found straightforward as $O(UW)$, whereas in Algorithm 2, at each $Q_f$ loop, cache placement complexity is $O(U + S + 1)$. Hence, the overall complexity of Algorithm 2 is $O((U + S + 1)F)$. Finally, the complexity of the heuristic approach is $O(\max(UW, (U + S + 1)F))$, which is significantly lower than that of the optimal solution.

V. NUMERICAL RESULTS

In the following simulations, we assume a network of one MBS that occupies the center of a circular region, with radius equal to 100 meters, four uniformly located SBSs within $10\sqrt{50} \approx 71$ meters from the MBS, and $U = 8$ randomly located users within the MBS’s coverage area. Unless stated otherwise, the system’s key parameters are set as follows. We set $W = 3$ channels, using bandwidth $B = 10$ MHz [34]. Moreover, we assume that the path-loss exponent for the wireless channels follows from the urban micro street canyon model and is equal to $\alpha = 3.1$ [44]. The duration of a TS is fixed to 10 ms [45], the delay threshold is chosen stringent as $D_{th} = 5$ TS $= 50$ ms, which may correspond for instance to an infotainment service [46], while the average backhaul delay $\bar{T}_0 = 10$ TS $= 100$ ms [47]. The transmit signal-to-noise ratio (SNRs) are defined and set to $\gamma_m = \frac{P_u}{\sigma_0^2} = 20$ dB, $\gamma_S = \frac{P_u}{\sigma_0^2} = 17$ dB, and $\gamma_U = \frac{P_u}{\sigma_0^2} = 10$ dB. For caching, we assume that $K = 3$ user classes, for which we arbitrarily assign the probabilities of belonging to specific user classes by

$$\begin{bmatrix}
0.3, 0.5, 0.2, \text{ if mod}(u, K) = 0,
0.2, 0.3, 0.5, \text{ if mod}(u, K) = 1,
0.5, 0.2, 0.3, \text{ if mod}(u, K) = 2
\end{bmatrix}$$

while $\beta = 2$ reflects more centralized file requests [34]. For the sake of simplicity, we assume a small library $F = 100$ files, short files $L = 100$ kb, and small caching capacities $C_m = 500$ kb, $C_2 = 200$ kb, and $C_p = 100$ kb.

In Fig. 2 we present the SDP performance as a function of $\gamma_U$ for the “Optimal” solution, the proposed “Heuris-

These values are selected in a way to reflect the different communication capabilities of the MBS, SBSs, and D2D UEs [48].

Since we previously set stringent communication delays, the defined data characteristics correspond typically to low-rate networks with small-sized data (e.g., wireless sensor networks and Internet-of-things networks). Choosing different caching values that would correspond to high traffic-demanding multimedia services (i.e., data size in Mbits) would necessarily impact the communication assumptions and enforce setting either a long TS duration or very large delay thresholds [34]. Subsequently, the performance of these systems are expected to be similar.
FIGURE 2. Comparison of caching policies for different $P_U$ values ($U = 8$, $F = 10$).

(a) SDP vs. $\gamma_U$ ($W = 8$ and $D_{th} = 5$).
(b) SDP vs. $\gamma_U$ ($W = 8$ and $D_{th} = 11$).
(c) SDP vs. $\gamma_U$ ($W = 3$ and $D_{th} = 5$).
(d) SDP vs. $\gamma_U$ ($W = 3$ and $D_{th} = 11$).

TABLE 3. Execution times ($P_U = 0.1$ W).

| Algorithm   | $W = 8$, $D_{th} = 5$ | $W = 8$, $D_{th} = 11$ |
|-------------|-----------------------|------------------------|
| Optimal     | 483.59 sec            | 518.38 sec             |
| Heuristic   | 0.00629 sec           | 0.0063 sec             |
| No D2D      | 0.00266 sec           | 0.00255 sec            |
| $W = 3$, $D_{th} = 5$ | 74238 sec       | 1025.93 sec            |
| $W = 3$, $D_{th} = 11$ | 0.00562 sec       | 0.00664 sec            |
| No D2D      | 0.024 sec             | 0.0105 sec             |

QoS conditions, i.e., $D_{th} = 5$. Indeed, this case demonstrates the limits of the proposed heuristic approach, as it leads to sub-optimal performances. Nevertheless, the complexity is dramatically reduced compared to the optimal solution as illustrated in Table 3.

For the remaining of this section, unless stated otherwise, only the SDP performances of the proposed heuristic solution...
are shown, where $\gamma_U = 10$ dB, $L = 100$ kb, $U = 22$ users, and $W = 3$ channels.

Fig. 3 illustrates the SDP as a function of $C_U$, for different numbers of users. $C_U$ values have been chosen to correspond to the caching of a single or several contents at the devices. When $C_U$ increases, the SDP performance enhances. Indeed, higher $C_U$ favors more caching in users and leverages low interference D2D communications. Moreover, as $U$ grows, the SDP significantly improves due to a better caching policy, which places the files within better located users, compared to the scenario with a lower number of users. Hence, incentivizing users to participate by caching and delivering files to other users is advantageous in terms of SDP performance.

**FIGURE 3. Impact of number of users $U$ and caching capacity $C_U$ on SDP.**

Fig. 4 investigates the impact of the file length $L$ on the SDP for different numbers of SBSs. With larger files, caching within the network saturates rapidly and most files have to be delivered via the backhaul link. The latter, if very poor, significantly degrades the SDP performance. Hence, content segmentation may be beneficial when sufficient caching capacity is available within the devices. In addition, the SDP decreases with the number of SBSs. Indeed, a small $S$ means that less caching capacity and reliable wireless links are available to deliver files to users within the network. Consequently, network densification provides an SDP gain if the incurred interference can be efficiently mitigated.

**FIGURE 4. Impact of length of files $L$ and number of SBSs $S$ on SDP.**

In Fig. 5, the impact of the threshold $D_{th}$ on the SDP is studied for different $\bar{T}_0$ values. $D_{th}$ values are selected to reflect delay-tolerant, moderate, and critical services. When $D_{th}$ is small, the SDP performance decreases. This is expected since a stricter delay threshold would inevitably penalize the files delivery. In general, a better backhaul link (small $\bar{T}_0$) is expected to improve the SDP since more file deliveries can be conducted through the core network. However, we see that this does not apply for $D_{th} = 11$. Indeed, even with a better backhaul link, the caching policy prefers to rely on the devices within the network (i.e., MBS, SBSs, and UEs), enabled by the larger $D_{th}$ value, rather than on the core network that has a shorter backhaul delay.

**FIGURE 5. Impact of delay threshold $D_{th}$ and backhaul delay $\bar{T}_0$ on SDP.**

**VI. CONCLUSION**

In this paper, we studied the joint caching and channel allocation for D2D-assisted wireless HetNets. First, we derived the expression of the average content delivery delay under the fast varying channels condition. Then, the latter is used to formulate the SDP maximization problem with respect to the communication and caching constraints. The joint problem is identified as non-linear and non-convex. To solve it optimally, we proposed a linearization approach into an ILP. Since the optimal solution incurs a high computation complexity, we proposed a low-complex two-step heuristic method. In the first step, the channel allocation strategy is determined using a geometrical approach, while in the second step, an iterative caching policy is developed. Through numerical results, we showed that our proposed heuristic...
algorithm is able to achieve between 60% and 90% of the optimal solution’s SDP performance, in a very small execution time. Besides, it outperformed the conventional cache-enabled HetNet. Finally, the impact of key parameters is investigated. Obtained results provide the following design guidelines: 1) Incentivizing users to participate in the caching process is beneficial to the network; 2) network densifying may bring an additional performance gain; and 3) a better backhaul link is generally advantageous for stringent QoS requirements, however, loosened QoS favors more caching within the network devices and less use of the backhaul link.

**APPENDIX A**

**PROOF OF THEOREM 1**

By applying the Chernoff bound to (5), and assuming independence between random variables (RV) $Y_n$, $\forall n = 1, \ldots, T$, we obtain [49]

$$
P[Z_T < \frac{L}{\tau B}] \leq \min_{t>0} e^{\frac{t}{m}} \prod_{n=1}^{T} E[e^{-tY_n}].$$

(18)

Let $Z_n = e^{-tY_n}$ and $X_n = \text{SINR}_n(t)$, $\forall n = 1, \ldots, T$, then, the cumulative distribution function (cdf) of $Z_n$ is given by

$$
F_{Z_n}(z) = P[e^{-t \log_2(1 + X_n)} < z] = P[X_n \geq 2^{-\frac{\ln 2}{\tau} - 1}] = 1 - F_{X_n}\left(2^{-\frac{\ln 2}{\tau} - 1}\right) = e^{-\frac{z}{\theta_{ij}}} = 0 < z \leq 1
$$

(19)

where $F_{X_n}(x) = 1 - e^{-\frac{x}{\theta_{ij}}}$, $\forall x \geq 0$, is the cdf of RV $X_n$ in an interference-free environment. From (19), we derive the probability density function (pdf) of $Z_n$ as

$$
f_{Z_n}(z) = \frac{\partial F_{Z_n}(z)}{\partial z} = \frac{(\ln 2)e^{-\frac{z}{\theta_{ij}}}}{z \theta_{ij}^2}, \quad \forall 0 < z \leq 1.
$$

(20)

Consequently, the mean of $Z_n$ is calculated by

$$
E[Z_n] = \int_0^1 z f_{Z_n}(z) dz = \frac{1}{\theta_{ij}^2} \Gamma\left(1 - \frac{\ln 2}{\theta_{ij}}, \frac{1}{\theta_{ij}}\right).
$$

(21)

By combining (21) into (18), $\zeta_0(T, I_j)$ is obtained, given by (6).

In an interfered environment, the cdf of $X_n$ can be obtained from (Eq. A.4, [34]) as

$$
F_{X_n}(x) = 1 - \prod_{\ell' \in I_{j'}} \left(\theta_{ij} + \theta_{ij'} x\right), \quad \forall x \geq 0.
$$

(22)

Hence, the cdf of $Z_n$ is obtained similarly to (19), and its pdf is given by

$$
f_{Z_n}^0(z) = \left(1 + \theta_{ij} \sum_{\ell' \in I_{j'}} \left(\frac{\theta_{ij'}}{\theta_{ij} + \theta_{ij'}} \left(2^{-\frac{\ln 2}{\tau} - 1}\right)\right)\right) \prod_{\ell' \in I_{j'}} \left(\frac{\theta_{ij'} + \theta_{ij'}}{\theta_{ij} + \theta_{ij'}} \left(2^{-\frac{\ln 2}{\tau} - 1}\right)\right) \times \frac{(\ln 2)e^{-\frac{z}{\theta_{ij}}}}{z \theta_{ij}^2}, \quad \forall 0 < z \leq 1.
$$

(23)

Due to the complexity of $Z_n$’s pdf, this expression cannot be used to calculate its mean. We propose to upper bound $f_{Z_n}^0$ given that $0 < z \leq 1$. Consequently,

$$
\frac{1}{\theta_{ij} + \theta_{ij'}} \left(2^{-\frac{\ln 2}{\tau} - 1} - 1\right) \leq 1, \quad \forall \ell' \in I_{j'}.
$$

(24)

and

$$
f_{Z_n}^0(z) \leq f_{Z_n}(z) \times \left(1 + \sum_{\ell' \in I_{j'}} \frac{\theta_{ij'}}{\theta_{ij} - 1}\right).
$$

(25)

By proceeding similarly as in (21), and combining the result with (18), $\zeta_0(T, I_j)$ is obtained as presented in (7). This completes the proof of Theorem 1.

**REFERENCES**

[1] C. Wang, Y. He, F. 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 Tuts.*, vol. 20, no. 1, pp. 7–38, 1st Quart., 2017.

[2] M. Agiwal, A. Roy, and N. Saxena, “Next generation 5G wireless networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[3] W. Cheng, X. Zhang, and H. Zhang, “Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3092–3107, Dec. 2016.

[4] R. I. Ansari, C. Chrysostomou, S. A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, and J. P. C. Rodrigues, “5G D2D networks: Techniques, challenges, and future prospects,” *IEEE Syst. J.*, vol. 12, no. 4, pp. 3970–3984, Dec. 2018.

[5] F. Malandrino, Z. Limani, C. Casetti, and C. F. Chiasserini, “Interference-aware downlink and uplink resource allocation in hetnets with D2D support,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2729–2741, May 2015.

[6] M. Ali, S. Qaisar, M. Naeem, and S. Mumtaz, “Energy efficient resource allocation in D2D-assisted heterogeneous networks with relays,” *IEEE Access*, vol. 4, pp. 4902–4911, 2016.

[7] 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. Commun.*, vol. 64, no. 8, pp. 3534–3547, Aug. 2016.

[8] W. Cao, G. Feng, S. Qin, and M. Yan, “Cellular offloading in heterogeneous mobile networks with D2D communication assistance,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4245–4255, May 2017.

[9] G. I. Tsiripoulos, A. Yadav, M. Zeng, and O. A. Dobre, “Cooperation in 5G HetNets: Advanced spectrum access and D2D assisted communications,” *IEEE Wireless Commun.*, vol. 24, no. 5, pp. 110–117, Oct. 2017.

[10] S. A. R. Naqvi, H. Pervaiz, S. A. Hassan, L. Musavian, Q. Ni, M. A. Imran, X. Ge, and R. Tafazolli, “Energy-aware radio resource management in D2D-enabled multi-tier HetNets,” *IEEE Access*, vol. 6, pp. 16610–16622, 2018.
[11] Y. Hao, Q. Ni, H. Li, and S. Hou, “Robust multi-objective optimization for EE-SE tradeoff in D2D communications underlying heterogeneous networks,” IEEE Trans. Commun., vol. 66, no. 10, pp. 4936–4949, Oct. 2018.

[12] 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. 82–88, Aug. 2014.

[13] 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 Tuts., vol. 20, no. 3, pp. 1710–1732, 3rd Quart., 2018.

[14] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, “Backhaul-aware caching placement for wireless networks,” in Proc. IEEE GLOBECOM, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[15] B. Blaszczyszyn and A. Giovanidis, “Optimal geographic caching in cellular networks,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2015, pp. 3358–3363.

[16] A. Khreishah and J. Chakareski, “Collaborative caching for multi-cell-coordinated systems,” in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), Apr. 2015, pp. 257–262.

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

[18] Y. Guan, Y. Xiao, H. Feng, C. C. Shen, and L. J. Cimini, “MobiCacher: Mobility-aware content caching in small-cell networks,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2014, pp. 4537–4542.

[19] M. Ji, G. Caire, and A. F. Molisch, “Wireless device-to-device caching networks: Basic principles and system performance,” IEEE J. Sel. Areas Commun., vol. 34, no. 1, pp. 176–189, Jan. 2016.

[20] L. Zhang, M. Xiao, G. Wu, and S. Li. “Efficient scheduling and power allocation for D2D-assisted wireless caching networks,” IEEE J. Sel. Areas Commun., vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[21] B. Chen, C. Yang, and Z. Xiong, “Optimal caching and scheduling for cache-enabled D2D communications,” IEEE Commun. Lett., vol. 21, no. 5, pp. 1155–1158, May 2017.

[22] C. Y. Tsai, C. Huang, and J. Cai. “An incentive mechanism integrating joint power, channel and link management for social-aware D2D content sharing and proactive caching,” IEEE Trans. Mobile Comput., vol. 17, no. 4, pp. 789–802, Apr. 2018.

[23] Y. Li, M. C. Gursoy, and S. Velipasalar, “A delay-aware caching algorithm for wireless D2D caching networks,” in Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS), May 2017, pp. 456–461.

[24] Y. Z. Chen, C. Yang, and S. Velipasalar, “Learning-based delay-aware wireless D2D caching networks,” IEEE Access, vol. 6, pp. 77250–77264, 2018.

[25] N. Gatsiougli, K. Ntontin, E. Kartasiki, A. Antonopoulos, and C. Verikoukis, “D2D-aware device caching in mmWave-cellular networks,” IEEE J. Sel. Areas Commun., vol. 35, no. 9, pp. 2025–2037, Sep. 2017.

[26] R. Awer, M. M. Butt, M. Bennis, and N. Marchetti, “Inter-cluster cooperation for wireless D2D caching networks,” IEEE Trans. Wireless Commun., vol. 17, no. 9, pp. 6108–6121, Sep. 2018.

[27] R. Sun, T. Yang, A. Wang, M. Qin, Z. Fei, and Y. Wang, “Cost-oriented mobility-aware caching strategies in D2D networks with delay constraint,” IEEE Access, vol. 7, pp. 177023–177034, 2019.

[28] T. Zhang, H. Fan, J. Loo, and D. Liu, “User preference aware caching deployment for device-to-device caching networks,” IEEE Syst. J., vol. 13, no. 4, pp. 226–237, Mar. 2019.

[29] T. Zhang, X. Fang, Y. Liu, G. Y. Li, and W. Xu, “D2D-enabled mobile user edge caching: A multi-winner auction approach,” IEEE Trans. Veh. Technol., vol. 68, no. 12, pp. 13124–13128, Dec. 2019.

[30] G. Kollias and A. Antonopoulos, “Joint consideration of content popularity and size in device-to-device caching scenarios,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2020, pp. 1–6.

[31] C. Yang, Z. Chen, Y. Yao, B. Xia, and H. Liu, “Energy efficiency in wireless cooperative caching networks,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2014, pp. 4975–4980.

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

[33] K. Li, C. Yang, Z. Chen, and M. Tao, “Optimization and analysis of probabilistic caching in N-tier heterogeneous networks,” IEEE Trans. Wireless Commun., vol. 17, no. 2, pp. 1283–1297, Feb. 2018.

[34] Z. Yang, C. Pan, Y. Pan, Y. Wu, W. Xu, M. Shikh-Bahaei, and M. Chen, “Cache placement in two-tier HetNets with limited storage capacity: Cache or buffer?” IEEE Trans. Commun., vol. 66, no. 11, pp. 5415–5429, Nov. 2018.

[35] G. Quer, I. Puppalardo, B. D. Rao, and M. Zorzi, “Proactive caching strategies in heterogeneous networks with device-to-device communications,” IEEE Trans. Wireless Commun., vol. 17, no. 8, pp. 5270–5281, Aug. 2018.

[36] W. Jaafar, W. Ajib, and H. Elbiaze, “Joint caching and resource allocation in D2D-assisted heterogeneous networks,” in Proc. 14th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WMob), Oct. 2018, pp. 1–8.

[37] W. Jaafar, W. Ajib, and H. Elbiaze, “Caching optimization for D2D-assisted heterogeneous wireless networks,” in Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), Sep. 2019, pp. 1–6.

[38] F. H. Panahi, F. H. Panahi, and T. Ohtsuki, “Energy efficiency analysis in cache-enabled D2D-aided heterogeneous cellular networks,” IEEE Access, vol. 8, pp. 19540–19554, 2020.

[39] Q. Fu, L. Yang, B. Yu, and Y. Wu, “Extensive cooperative content caching and delivery scheme based on multicast for D2D-enabled HetNets,” IEEE Access, vol. 9, pp. 40884–40902, 2021.

[40] A. Asadi, Q. Wang, and V. Mancuso, “A survey on device-to-device communication in cellular networks,” IEEE Commun. Surveys Tuts., vol. 16, no. 4, pp. 1801–1819, 4th quarter 2014.

[41] Y. Crama and P. L. Hammer, Boolean Functions: Theory, Algorithms, and Applications, Cambridge, U.K.: Cambridge Univ. Press, 2011.

[42] (2016). AMPL: Streamlined Modeling for Real Optimization. [Online]. Available: http://ampl.com

[43] (2016). IBM CPLEX Optimizer. [Online]. Available: http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer

[44] S. Sun, T. S. Rappaport, S. Rangan, T. A. Thomas, A. Ghosh, I. Z. Kovacs, I. Rodriguez, O. Koymen, A. Partlyka, and J. Jarvelainen, “Propagation path loss models for 5G urban micro- and macro-cellular scenarios,” in Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring), May 2016, pp. 1–6.

[45] S. Sciancalepore, M. Vaccini, G. Piro, G. Boggia, and T. Watteyne, “Layered security in TSCH networks: Effect on slot duration,” Trans. Emerg. Telecommun. Technol., vol. 28, no. 1, Jan. 2017, Art. no. e3089.

[46] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big data analytics in intelligent transportation systems: A survey,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 1, pp. 383–398, Jan. 2019.

[47] Microsoft. Azure Network Round-Trip Latency Statistics. Accessed: Jul. 10, 2021. [Online]. Available: https://docs.microsoft.com/en-us/azure/networking/azure-network-latency

[48] B. Li, Z. Fei, Z. Chu, and Y. Zhang, “Secure transmission for heterogeneous cellular networks with wireless information and power transfer,” IEEE Syst. J., vol. 12, no. 4, pp. 3755–3766, Dec. 2018.

[49] W. Mittennacher and E. Ufahl, Probability and Computing: Randomized Algorithms and Probabilistic Analysis, Cambridge, U.K.: Cambridge Univ. Press, 2005.
AMINA MSEDDI (Student Member, IEEE) received the degree (engineering) in networks and telecommunications from Institut National des Sciences Appliquées et de Technologie (INSAT), Tunisia, in 2013, and the M.Sc. degree in computer science from the University of Quebec at Montreal, in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include resource management in large-scale distributed systems, content delivery networks, cloud and fog computing.

WESSAM AJIB (Senior Member, IEEE) received the diploma degree (engineering) in physical instruments from INPG at Grenoble, France, in 1996, and the master’s and Ph.D. degrees in computer networks from École Nationale Supérieure des Télécommunications, Paris, in 1997 and 2000, respectively. He had been an Architect and Radio Network Designer at Nor- tel Networks, Ottawa, ON, Canada, from October 2000 to June 2004. He had conducted many projects on the third generation of wireless cellular networks. He received a Postdoctoral Fellowship from École Polytechnique de Montréal, Montreal, QC, Canada, from 2004 to 2005. Since June 2005, he has been with the Department of Computer Sciences, Université du Québec in Montréal, Montreal, where he is currently a Full Professor. His research interests include wireless communications and networks, multiple access design, and traffic scheduling. He is the author or coauthor of many journal articles and conferences papers in these areas.

HALIMA ELBIAZE (Senior Member, IEEE) received the M.Sc. degree in telecommunication systems from Université de Versailles, in 1998, and the Ph.D. degree in computer science from Institut National des Télécommunications, Paris, France, in 2002. Since 2003, she has been with the Department of Computer Science, Université du Quebec Montréal, QC, Canada, where she is currently an Associate Professor. She had been awarded many research grants from both public agencies and industry. Her research interests include performance evaluation, traffic engineering, cloud computing, wireless networks, and next generation IP networks. In 2005, she received the Canada Foundation for Innovation Award to build her IP over the DWDM Network Laboratory.