IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 31, 2020, accepted September 21, 2020, date of publication October 5, 2020, date of current version October 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028565

Smart Technique for Cache-Assisted Device to Device Communications

AHMED HASSAN ABDEL SALAM, HUSSEIN M. ELATTAR, (Associate Member, IEEE), AND MOHAMED A. ABOUL-DAHAB, (Life Senior Member, IEEE)

Department of Electronics and Communications Engineering, Arab Academy for Science, Technology and Maritime Transport, Cairo 11799, Egypt

Corresponding author: Ahmed Hassan Abdel Salam (ahmedhsn@student.aast.edu)

ABSTRACT The emergence of smart terminals with the consequence of higher network traffic growth rates, has increased the demand for new techniques that are needed to meet user requirements for high-quality communication and social media services. One of these techniques is Device-to-Device (D2D) communication, allowing direct data transmission of popular files between users. This is an advantage during rush hours where there is no access to the traffic via a base station (BS), thereby reducing network loads. In this article, we propose a new caching paradigm for D2D communication to achieve the goal of rising network offloading. A Smart Adaptive Algorithm (SAA) is proposed that smartly selects users with higher probabilities of sharing more data, then places popular files in their cache. This is done to optimize network offloading at minimal consumed energy by the user (user cost), taking into account the user speed, location, predicted zones, interests, and the size of the requested file. The data offloading ratio is taken as a performance metric in evaluating network offloading. It is defined as the proportion of requested data that can be delivered via D2D links. The offloading ratio is calculated in different scenarios in order to assess the influence of proactive caching on core network load. To minimize the complexity of the algorithm, and to respond in a shorter time period to the number of requests, user demand shall be handled with both the base station and the user's device application. For the simulation and implementation of the proposed algorithm, a human mobility model that predicts different human mobility behaviors is used. The simulation results reveal that the offloading ratio is significantly affected by the size of the file transmitted via D2D connection as well as the user speed. The amount of increase in the offloading ratio due to the increase in the user speed is based on the size of the files transmitted through D2D connection. The smaller files transmitted via D2D connection would result in a significant increase in the offloading ratio, whereas the larger file size would result in a less change in the offloading ratio for higher speed users. The simulation results also reveal that higher speed users share more data at a high bit rates compared to users with lower speeds. Simulation results also show that the proposed SAA has superior performance as compared to other algorithms available in literature.

INDEX TERMS Device-to-device communication, offloading, proactive caching, cache-assisted, D2D.

I. INTRODUCTION

The successful implementation of the 5G system will pose different challenges in achieving its anticipated goals, such as low latency, high data rates, enhanced stability, improved performance, and security. Mobile data exchange will increase seven folds in the next few years [1]. Therefore, network congestions can be expected, in particular when streaming videos. This is a major challenge, as the already existing networks are suffering from unregulated fluctuations in download speeds. The user's off-peak data rate is expected to be much greater than that of rush hours in the 4G wireless networks. These variable rates are due to network congestions when they are overloaded with connections that significantly lower download speeds [2]. These speed changes will become an issue as mobile services expand. One of the approaches implemented to minimize cellular network congestion during rush hours is proactive caching. Popular files are repeatedly requested and transmitted redundantly in core networks, resulting in an increase in data traffic [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Khaled Rabie.

Proactive caching can take place among mobile users, small base stations, mobile base stations, or even femtocell stations. However, the work in this article focuses on caching files among mobile devices only. The principal concept of proactive caching in cellular networks is that the popular contents are pre-fetched to certain mobile devices, and then exchanged between mobile users through a device to device (D2D) communication to decrease network loads and improve video streaming capabilities [4]–[9]. Several previous studies on cache-assisted D2D communication were carried out. In [10], the authors adopted a probabilistic approach in handling D2D networks to increase the cache hit probability and maximize the density of the successfully served requests by mobile users. Moreover, scheduling and optimizing the cache for maximizing offloading were introduced in [11]. In [12], a Q-learning strategy to replace the base station cache by user device cache was introduced by the authors. In [13], the authors investigated the energy transmission performance of adding a cache to the base station in comparison with the way in which the server content was downloaded. In [14], the authors introduced a method to distribute the cache of the base station in order to reduce network delay. In [15], the authors discussed the issue of caching allocation at base stations and suggested a low complexity algorithm for resolving the problem. An algorithm to minimize the user download delay was introduced in [16]. The authors in [17] introduced a new scheme of data offloading taking into account data heterogeneity. A content-encoding strategy was introduced in [18] in order to minimize the transmission cost. The outage-throughput tradeoff was discussed in [19], whereas a multicast-aware caching problem was investigated in [20]. In [21], The author introduced a random caching algorithm, through which every file was cached randomly by each user and the probability of caching was correlated to the probability of file requests. A popular caching technique was introduced in [22] where every user cached the most popular files. However, these algorithms suffered from high data redundancy, in addition to ignoring other factors such as user's interests, location and speed, with the consequence of affecting cache utilization as well as the network offloading. The data offloading ratio is always taken as a performance metric in evaluating network offloading. It is defined as the proportion of requested data that can be delivered via D2D links. Considering the effect of user mobility on the offloading ratio and the cost of service, the authors in [23] demonstrated that the mobility improved caching decisions, and had a good impact on service costs. Furthermore, a mobility-aware caching incentive scheme in D2D cellular networks was proposed in [24] to reduce base station costs. It assumed that each mobile user's preferences were known and remained unchanged. The authors also grouped the minimum-delay users who could serve the file requestor. Among this group, the user with the lowest delay would handle the requestor. In [25], the authors demonstrated that mobile users with low speeds were less likely to meet others and less likely to improve D2D interaction, as compared to higher speed users. The cost-optimal problem of caching was dealt with in [26] and [27], considering user mobility, cache size, and maximum encoded file segments to reduce the service cost. A model was introduced in [28] through which the energy consumption for the applications that were running under the Internet of Things (IoT) services could be monitored. With direct communication capabilities between mobile devices, the possibility of mobile communication applications could be expanded to D2D cellular networks. While cache storage for every single mobile device was not inherently huge, a large cache space that could cache a large volume of multimedia content might be built up using the cache space for multiple devices. Since the mobile user cache space was limited compared to large data requests, the avoidance of redundancies and waste of space was one of the cache placement targets for content diversity. With the growth of mobile user's cache capacities, they were not only able to store data for themselves but could also share with neighboring mobile users their cache contents via D2D communications [29]. This resulted in changing the value achieved by proactive content cache while creating new challenges in D2D communication. The researchers in [30] introduced the idea of a femto-caching aid, where the backhaul storage space was substituted by small cell access point storage space. In [31], the authors proposed an architecture that enhanced video transmission throughput in cellular caching of video files with D2D controlled by the base station. In a device with a multi-cache unit, a content caching placement for each caching unit was proposed in [32] by excluding repetition to maximize the hit ratio. In [33] the authors used the device cache and the D2D services as a bootstrapping system in the mobile network and also proposed an algorithm in which users would cache random file segments and share them with their nearby devices. A greedy and dynamic programming algorithm for sub-optimal and optimal solutions respectively was proposed in [34], where it was demonstrated the algorithm was efficient and fast for sub-optimal solutions. It assumed that the probabilities of the user's requests and behavior were close to being uniform. In large and complex scenarios, often greedy algorithms did not offer the best solutions, in addition to the longer computational time needed. Adaptive programming could provide an optimal solution but might be much slower than the greedy one. However, this had the drawback of consuming memory and increase complexity as the number of users increase. Network offloading was investigated while not considering the effect of other parameters such as bit rate and size of a popular file. In [35] the authors adopted a process to model the time intervals of both contact and inter-contact durations, and investigated the effect of mobility on the caching process. However, this approach along with those adopted in [25] and [34] were based upon the mechanism that when a user requested a file, he would continue to ask others within his contact range for the requested one. This mechanism had a drawback that the number of responses would increase proportionally with the number of users. A variety of approaches were adopted to maximize network offloading in case of mobile users. The authors in [36] investigated the impact of node mobility in a cluster-based wireless caching network. Each node generated a request for the file and received it from the closest nodes via a D2D link under a fast mobility scenario. In [37], the authors formulated a mobility-aware cache placement problem to reduce both the average transmission cost and the average cache leasing cost. A heuristic distance-based pairing scheme for collaborative content distribution via D2D communications was analyzed in [38]. The authors in [39] analyzed the delivery of coded packets in a cluster-based mobile caching network under Markov process and reshuffling mobility models, where they demonstrated the effectiveness of concurrent transmissions in improving network performance. In [40], the throughput scaling of the network was improved by exploiting the global caching gain and the spatial reuse gain simultaneously. The authors in [41] proposed a hybrid coded caching approach to enhance cellular network performance by applying decentralized coded caching. This approach provided spatial reuse gain in the first sub-phase, and centralized coded caching enabling multicasting transmissions in the second sub-phase. In [42], the authors introduced a user-centric proactive caching policy that could minimize the energy cost for a user to download a file (user cost), while maximizing the offloaded traffic. The authors in [43] investigated the trade-off between buffer and cache capacity through cache placement and joint bandwidth allocation in order to reduce the transmission delay. In [44], the authors developed a framework that allowed the implementation of Federated learning algorithms in the wireless networks. They also solved an optimization problem that considered user selection and resource allocation. In this article, a new file request mechanism is proposed to overcome the problem of the large number of replies. In addition, a new cashing content placement algorithm is proposed through which users are selected for efficient caching of popular files to achieve a high network offloading rate.

II. MOTIVATION AND CONTRIBUTION

In this article, we adopt an approach in which the time delay of the downloading process could be reduced. Unlike the previous work in [24], the proposed approach assumes a user preference that is not known beforehand. In addition, it is not the same all the time. We also study the offloading of network data by means of D2D communication using the human mobility model in such a way that each user can move on foot or use a bicycle, or vehicle. The model will calculate the contact duration (time duration at which users are within the communication range of each other), as well as the inter-contact duration (time span between contact durations) between users. Previous works investigated the impact of user speed on the network offloading [25], [34], [35]. We are continuing research on the impact of the user speed on the network offloading considering other factors that have a significant impact on the network offloading. These factors are the size of the file transmitted via D2D connection and the bit rate. A caching placement algorithm to fetch popular files into the user's cache is then implemented. The main contributions in this article are summarized as follows:

1. A caching content placement algorithm named the Smart Adaptive Algorithm (SAA) is proposed, in order to allow substantial improvements in the network offloading, taking into account the user's location, visited areas, speed, interest, bit rate, and the file size transmitted through the D2D connection.
2. A new approach is adopted to dramatically speed up the download process through D2D communication and reduce the energy consumed by the user (user cost).
3. The impact of the user's speed, the transferred file size through D2D, and the bit rate on the network offloading is investigated and analyzed.
4. A new factor called the "data exchange factor" is introduced to compare the amount of data transmitted by the user to data received from the D2D communication.

The rest of this article is organized as follows. The system model is defined in section III. Section IV analyzes the user's interests and discusses the estimation of the user's probability of sharing common interests. The new file request mechanism is given in section V. The offloading of the network is investigated in section VI. Section VII introduces the proposed smart adaptive algorithm. In section VIII, system analysis and results are illustrated and discussed. The paper ends with a conclusion in section IX.

III. SYSTEM MODEL

The system model utilized assumes a single cellular base station and a number of mobile users with index set $N = \{1, 2...., N\}$, where each mobile user assigns a cache (N_{cn}) for proactive caching. We assume that each mobile user assigns one gigabyte (1GB) from his cache for proactive caching. All mobile devices are assumed to have the same transmission power [34], and each is equipped with a global positioning system (GPS) capability to detect its location (N_{pn}) and speed (N_{sn}). Each user is capable of establishing a D2D connection under the control of the base station. The cellular base station controls the D2D connection with mobile users. By assuming orthogonal resource allocation [34], the inter-user interference will be neglected. This assumption is adopted in order to have the same parameters platform for a fair comparison between similar techniques. A fixed transmission rate is assumed to be within the coverage of one single base station, and there will be no handover. During the first contact, each newly registered mobile user in the base station must send an info message (*IMn*) to the base station in the form of:

$$IM_n = \{AD_n, Np_{n,x}, Np_{n,y}, Ns_n, Nc_n\} \tag{1}$$

where (AD_n) is the mobile user connection address, ($Np_{n,x}$) is the mobile user latitude, ($Np_{n,y}$) is the mobile user longitude, (Ns_n) is the mobile user speed, (Nc_n) is the mobile user cache available assigned for proactive caching. The user must update his data sent to the base station in the event of a major change in speed such as from walking state to a driving one. A list of all parameters used in the model is given in table 1.

TABLE 1. List of system parameter.

Symbol	Description	Symbol	Description
N	Users index set	U_w	pedestrian users set
N_{cn}	User cache	U_r	running users set
N_{pn}	User location	U_b	biking users set
N_{sn}	User speed	U_v	vehicular riding users set
IMn	Info message	$\frac{\sqrt{V}}{2}$	the maximum distance the user can move
AD_n	user connection address	M	connection matrix
Z_i	Zone number	l	connection status
Z_o	Initial zone	W	weight matrix
d_i	distance from zone Z_i,	$q_{n,m}$	weight of an interest category for a specific user n
α	power-law exponent	UI	User interest vector
υ	preference degree	Z_f	file size
ϱ	normalized constant	r_o	bit rate
V_{max}	Maximum speed	$T_{f,S,P}$	Download time
V_{min}	Minimum speed	F	popular files library
V_w	Pedestrian speed	γ	popularity factor
V_r	Running speed	$\mathcal{O}_t$	offloading ratio
V_b	Biking speed	G_n	Interest group
V_v	Vehicular speed	$x_{f,I,n}$	BS list for a file
$\mathcal{E}_u$	exchange factor	D_{tx}	sum of data transmitted
D_{rx}	the sum of data received	d	distance between user current zone to the user next zone

A. HUMAN MOBILITY MODEL

A Spatio-temporal parametric model given in [45] is adopted to simulate user mobility. The reason for selecting this model is that it describes a broad spectrum of patterns of human mobility in a finite space of locations. It uses the power law to predict human movement from one location to another. This model besides being simple, it describes the human geographic mobility. The model assumes that the area of interest is divided into a number of square zones, each is denoted by Z_i. This area is covered by one cellular base station. The distance between these zones is determined by a metric d. Users randomly move from one zone to another. Each user can remain for a certain period of time at each visited zone. Initially, each user is assigned an initial zone Zo, from which he moves a distance d to the next zone. The user moves to the next zone with his assigned speed, stays (rests) for a period of time, and then moves to a next zone and so on.

The distances moved are of random nature, and are assumed to follow the power-law distribution with a probability density function (pdf) given by:

$$P\left[L = d_i\right] = \frac{\omega}{(1+d)^{\alpha}} \tag{2}$$

where d_i represents the distance from a zone Z_i, α is the power-law exponent that represents the attractor power, d is the distance between user current zone to the next zone, and ω is a normalizing constant. The probability of the user moving from zone Z_o to zone Z_1 depends on the attractor power α as follows:

- $\alpha < 0$: the user moves longer to reach the next zone Z_i, and therefore Z_i becomes more repulsive than being an attractive one.
- $\alpha > 0$: the user is more localized to his current zone.
- $\alpha = 0$: the user moves from one zone to another randomly with a probability of uniform distribution.

Using a discrete-time Markov chain with N states that has been utilized in [45], where each state corresponds to one zone, then the probability of the user choosing his next zone Z_i is given by:

$$\boldsymbol{P[Z_i|d_{Z_iZ_o} = l] = \frac{1}{S \times l} P\left[L = l\right] = \frac{\omega}{S \times l(1+l)^{\alpha}}} \tag{3}$$

using the power law, the rest time at each zone can be expressed as:

$$\boldsymbol{P\left[T = t\right] = \frac{\varrho}{t^{\upsilon}}} \tag{4}$$

where υ is the user temporal preference degree, and ϱ is a normalized constant. The Spatio-temporal parametric model given in [45] does not consider the type of mobility of the user which was directly related to his speed of motion. Therefore, a modification is introduced to this model to establish a simulation close to the realistic scenario. The model allocates randomly to each user maximum (V_{max}), and minimum (V_{min}) speeds. A modification is introduced to take different user mobility scenarios into account. In this respect, user mobility is divided into four sets depending on the user speed, namely: pedestrian user (average speed $V_w = 1.4$ m/s (5 km/h)), running user (average speed $V_r = 3$ m/s (11 km/h)), biking user (average speed $V_b = 5.5$ m/s (20 km/h), and vehicular riding user (average speed $V_v = 16.6$ m/s (60 km/h))). Thus, the user set U is given by:

$$U = \{Uw, Ur, Ub, Uv\} \tag{5}$$

where U_w denotes pedestrian users set, U_r denotes running users set, U_b denotes biking users set, U_v denotes vehicular riding users set.

Algorithm 1 describes the modified human mobility model. It starts by allocating an initial zone to each user. Each user belongs to one of the four user sets, where each set represents a speed category (pedestrian speed, running speed, bike speed, and vehicle speed). Then distance d is randomly calculated from Z_o for each user, which is the distance between the user's current zone and the next zone.

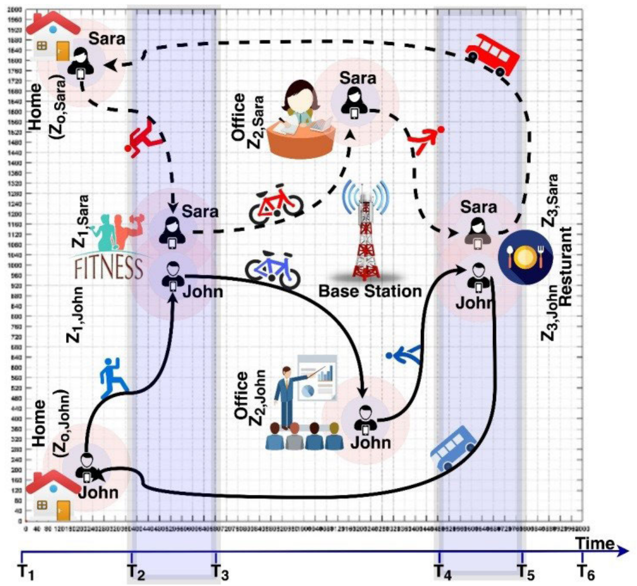


FIGURE 1. Daily user trajectory.

The user then moves to the next zone with his assigned speed, rests for a period of time and then moves to the next zone.

The user will keep moving from one zone to another until the end of the simulation. Figure 1 shows an example of a user daily trajectory of Sara and Peter. Each of them runs from her/his home (the initial zone) to the fitness center (zone 1) where Sara and Peter meet each other between time instance T_2 and T_3. They stay there for a while and then each moves with her/his bicycle to reach their jobs (zone 2). After finishing their works, Sara and Peter walk to the restaurant (zone 3) where they meet again between time instance T_4 and T_5, then each of them takes the bus to return back home. We can see from this example that different users share favorite places such as fitness centers, sports clubs, restaurants, etc.

Taking advantage of the time that users meet to share data, will increase network offloading and reduce network congestion. The modified human mobility model that is used for simulation will make it possible to simulate users on a daily basis, some users have short mobility patterns, which means that their favorite zones are close to each other, while others have a longer mobility pattern, which means their visited zones are far from each other. Each user can walk, run, ride a bicycle, or a vehicle from a zone to another.

B. CONTACT AND INTER-CONTACT DURATIONS

The two users, Sara and Peter, are said to be in contact if both are in the same zone, while the movement of each user is independent of the other. The probability that Sara and Peter will be in contact at a certain time instant is given in [45] by:

$$\begin{aligned} P_c &= P_S(Z_o)\,P_P(Z_o) + \cdots . + P_S(Z_{n-1})P_P(Z_{n-1}) \\ &= \sum_{j=0}^{n-1}[P(Z_i)]^2 = \sum_{j=0}^{\frac{\sqrt{V}}{2}} \frac{1}{S\times j}\left[\frac{\omega}{(j+1)^{\alpha}}\right]^2 \end{aligned} \tag{6}$$

Algorithm 1 Modified Human Mobility Algorithm

```
Inputs: - number of users ⟵ $N_n$
        - coverage area ⟵ $L \times L$ square meter
        - simulation interval ⟵ $T$ seconds
Dividing users into 4 sets $U = \{U_w, U_r, U_b, U_v\}$
Assign user speed to each user set $(V_w, V_r, V_b, V_v)$
Select initial zone for each user $(Z_o)$
repeat
    choose randomly a distance $(d)$ to select the next zone $Z_i$ from the probability distribution (2)
    choose randomly the next zone $Z_i$ that is d distance away from $Z_0$
    choose randomly a point (waypoint) in the next zone $Z_i$
    Go linearly to this point with a speed previously assigned to each user
    choose randomly a rest time t from the probability distribution (4)
    while $t < T$ do
        Run Random Waypoint movement in $Z_i$
    end
Until $t = T$ (end of simulation)
```

where $\frac{\sqrt{V}}{2}$ is the maximum distance the user can move. When the power-law exponent $\alpha = 0$, all users will move uniformly from one zone to another, so the inter-contact time distribution can be approximated with an exponential i.i.d (independent and identically distribution).

We define the inter-contact time (ICT) as the time duration between two successive contact durations, the pdf of the inter-contact time can be expressed as:

$$P_{(ICT=t)} = (1 - P_{c,s,p})^{t-1} P_{c,s,p} \tag{7}$$

When Sara and Peter are in contact, and Sara requests a file that is found in Peter's cache, then Peter can send the file to Sara using D2D connection. A connection matrix M with dimensions $N \times N$ will be generated during the simulation interval T at each instant t to track the user's contact duration length. The matrix M is expressed as:

$$M_{t=s} = \begin{bmatrix} l_{1,1} & \cdots & l_{N,1} \\ \vdots & \ddots & \vdots \\ l_{1,N} & \cdots & l_{N,N} \end{bmatrix} \tag{8}$$

The $N \times N$ (where N is the number of users) connection matrix $\boldsymbol{M}$ represents a contact status instant within the simulation interval between users, $\boldsymbol{l}$ represents the connection status such that if $\boldsymbol{l} = 1$, the two users are in contact whereas $\boldsymbol{l} = 0$, if they are not in contact. At the end of the simulation period $(t = T)$, we shall have a 3-dimensional matrix $\boldsymbol{N} \times \boldsymbol{N} \times \boldsymbol{T}$. Figure 2 shows the contact and the inter-contact time between two users, where the time period between $T2$ to $T3$, and $T4$ to $T5$ represent the contact duration between two users.

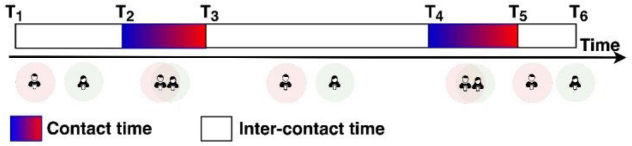


FIGURE 2. Sara & Peter contact and inter-contact durations.

IV. USER'S INTERESTS

According to [46], user interests are classified into fifteen categories, namely: reference, health, science, computers, business, society, adult, kids, teens, games, recreation, arts, news, shopping, and sports. According to [47], each interest is assigned a weight, the measurement of which is based on a weblog that is carried out for served users by the base station. For N users, the interest category weight matrix W can be expressed as:

$$W = \begin{bmatrix} q_{1,1} & \cdots & q_{1,15} \\ \vdots & \ddots & \vdots \\ q_{N,1} & \cdots & q_{N,15} \end{bmatrix} \tag{9}$$

where $q_{n,m}$ represents the weight of an interest category for a specific user, n is the user's identification number, ($n \in N$), and m is the interest category identification number. Each user's interest preference can be rearranged according to the weight category. We can rearrange the weight matrix so that each row is sorted in descending order according to the user's interest. This matrix is called the user‘s interest matrix given by:

$$W_{Ct} = \begin{bmatrix} Ct_{1,1} & \cdots & Ct_{1,15} \\ \vdots & \ddots & \vdots \\ Ct_{N,1} & \cdots & Ct_{N,15} \end{bmatrix} \tag{10}$$

In other words, each user will have an interest vector $\boldsymbol{UI}$ expressed as:

$$\boldsymbol{UI_n} = \{\boldsymbol{Ct_{n,1}}, \boldsymbol{Ct_{n,2}}, \ldots\ldots, \boldsymbol{Ct_{n,15}}\} \tag{11}$$

where $Ct_{n,m}$ is the interest category m of the specific user n. It is obvious that the first column of the matrix W_{ct} represents the highest interest for all users, whereas the last column represents the lowest interest of all users. Note that the interest category differs from one user to another, i.e., for the case of Sara and Peter, $Ct_{S,m} \neq Ct_{P,m}$ (where S stands for Sara and P for Peter). Assume that Sara and Peter exist in the same zone at a certain time instant. Let $Ct_{S,1}$ and $Ct_{P,1}$ be the highest weight interest category of Sara and Peter respectively, and let R to be the total number of interest categories ($R = 15$). Assume that the user's interests are uniformly distributed among each column of the interest category matrix W_{Ct}. The uniform distribution of user interest is a discrete one. The probability mass function for a uniform distribution taking one (Ct_m) of (R) possible value from the set $Ct = (Ct_1, Ct_2, \ldots, Ct_{15})$ is given by:

$$P\left(Ct_m\right) = \begin{cases} \frac{1}{N}, & Ct_m \in Ct \\ 0, & otherwise \end{cases} \tag{12}$$

The probability (P_i) of the two users (Sara and Peter) having common interest category is given by:

$$P_i = Pr\left\{Ct_{S,1}\right\} Pr\left\{Ct_{P,1}\right\} = \frac{1}{N^2} \tag{13}$$

Assuming that Peter has a number of files in his internal cache, Sara may or may not request one of these files. Let us consider that Sara requests one file at a certain time from Peter. Sara can either succeed in downloading her file from Peter or not. This could be described as a Bernoulli trial, where the probability mass function is denoted by $\boldsymbol{P_f}$, (this is the probability that Sara succeeds to download the file from Peter). If Sara fails to download the file from Peter, then the probability will be $(\boldsymbol{1 - P_f})$. The base station then calculates the total probability for user n in each interest category Ct_m, which is given by:

$$\boldsymbol{P_{t,S,P} = P_c \times P_i \times P_f} \tag{14}$$

The higher the value of $P_{t,S,P}$ of Sara, the more her chance to meet Peter and share the same interest. The total probability is then calculated for Sara with all the others. This is helpful for the base station in ranking mobile users. The base station shall be capable of sorting users in each interest category according to the total probability calculated for each in a descending order. The users with the highest total probability in each interest category will be chosen by the base station to cache the popular files. The probability that a number of users M ($M \epsilon N$) share the same interest P_I is given by:

$$\boldsymbol{P_I = \sum_{i=1}^{M} P_i} \tag{15}$$

V. A PROPOSED MECHANISM FOR FILE REQUESTS

The proposed file request mechanism flow chart is shown in Figure 3, and a relevant scenario is shown in Figure 4. The mechanism is described as follows: If Sara requests a file f, then, she first looks for it in her internal cache. If she finds it, then she will not have to download it. If the file f is not found in her cache, she will send a request to the base station. If the base station finds any user that stores file f, then it will send a reply message to Sara that contains the connection address, location, and speed of the users who store the file. If any of these users are found in Sara's communication range, then, she will send a direct request to that user having the file. If more than one user is detected, Sara will communicate with the user closest to her (Peter) to ensure a longer contact duration $\left(\tau_{S,P}\right)$. Sara begins downloading the file f from Peter for a period of time ($T_{f,S,P}$) given by:

$$T_{f,S,P} = \frac{Z_f}{r_o} \tag{16}$$

where Z_f if the file size, and r_o is the bit rate. At the end of this time period, Sara will check whether the download process is completed or not. If it is not completed, Sara will download the remaining part from the core network.

Sara will notify the base station once she finishes downloading her requested file, to add her to users list who have

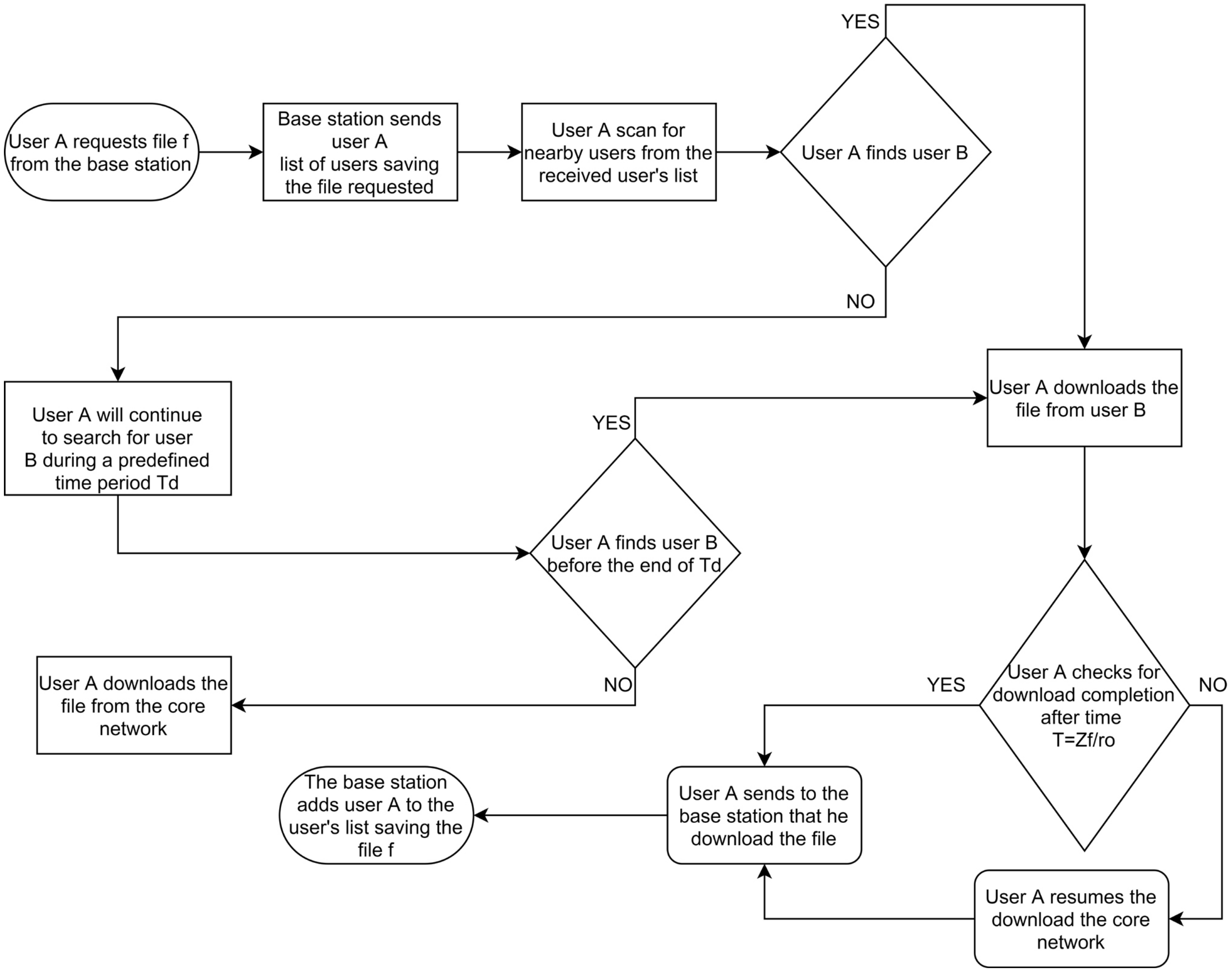

User A requests file f from the base station
Base station sends user A list of users saving the file requested
User A scan for nearby users from the received user's list
User A finds user B
YES
NO
User A will continue to search for user B during a predefined time period Td
User A downloads the file from user B
YES
User A finds user B before the end of Td
NO
User A downloads the file from the core network
User A checks for download completion after time T=Zf/ro
YES
NO
User A sends to the base station that he download the file
The base station adds user A to the user's list saving the file f
User A resumes the download the core network

FIGURE 3. The proposed file request Flow chart.

the file in their cache. If the connection link is broken before $T_{f,S,P}$ is elapsed, Sara will search for other users from the list. If she detects another user, then she will submit another request, and if not, then she will resume downloading the file from the core network. If Sara fails to find a user from the start of her request, then she will wait for a predefined time delay period T_d. Sara will keep scanning for the users until the end of T_d. If no users are detected, then she will download the file f directly from the core network. Thus, the proposed file request mechanism lessens the number of replies for each file, and thus reduces the user battery consumption. It is worth noting that previous relevant works in [25], [34], [35] were based upon requesting files directly from neighboring users existing within the radio coverage of the requesting user. This was done through continuous broadcasting process. The technique proposed in this article is based upon broadcasting requests to a selected group of neighboring users. On the other hand, the technique which was introduced in [24] assumed that each mobile user's preferences were known and remained unchanged. However, the technique proposed in this article assumes unknown mobile user's preferences. Thus, if Sara, for example, requests data file f, then she will send to users within her radio range requesting the availability of file f in their cache and wait for a period of time greater than $T_{f,S,P}$. This, in turn, will increase the number of replies for each file. Assume that Sara is requesting a file in the cache of one of these users within her radio range, then the number of replies will be directly proportional to the number of users that Sara is asking.

VI. CALCULATING THE OFFLOADING RATIO

We assume a popular files library $F = \{1, 2 \ldots, f\}$ with a heterogeneous file size content, where each file is of size z_f. The users' requests will be considered with the same distribution which follows the Zipf given in [48] with parameter γ, called the popularity factor (where the higher value of γ means that most users request the popular files). The file request probability is expressed as:

$$p_f = f^{-\gamma} / \sum_{j=1}^{f} j^{-\gamma} \tag{17}$$

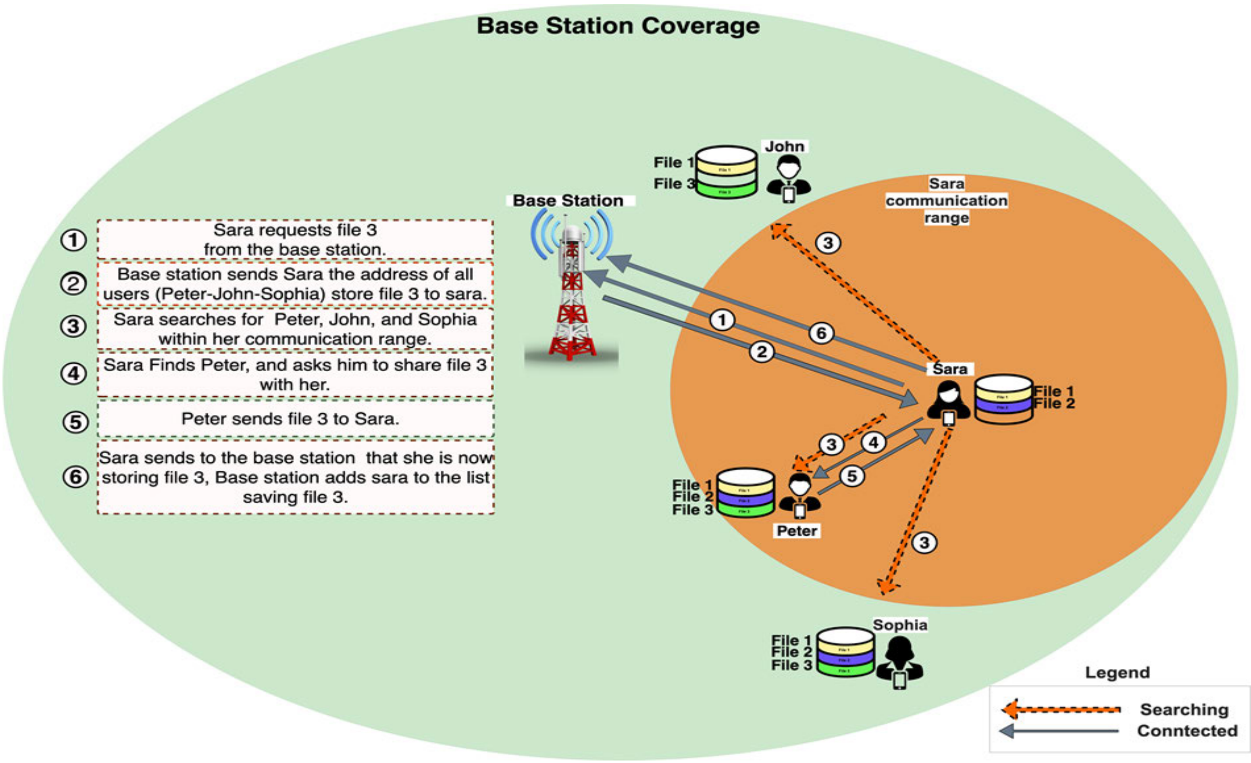

Base Station Coverage
① Sara requests file 3 from the base station.
② Base station sends Sara the address of all users (Peter-John-Sophia) store file 3 to sara.
③ Sara searches for Peter, John, and Sophia within her communication range.
④ Sara Finds Peter, and asks him to share file 3 with her.
⑤ Peter sends file 3 to Sara.
⑥ Sara sends to the base station that she is now storing file 3, Base station adds sara to the list saving file 3.
Base Station
File 1
File 3
John
Sara communication range
3
1
6
2
Sara
File 1
File 2
3
4
5
File 1
File 2
File 3
Peter
3
Sophia
File 1
File 2
File 3
Legend
Searching
Conntected

FIGURE 4. The proposed file request mechanism.

The total probability of a user to request the popular file f from the file library F is thus:

$$\sum_{f \in F} p_f = 1, \tag{18}$$

A better choice for users to proactively cache popular files is the one with a higher offloading ratio and with lower network congestion. Caching popular files to mobile users who have the chance to meet many other mobile users will increase the network offloading. The offloading ratio $\mathcal{O}_{S,f}$ is defined as the ratio between the percentage of file f downloaded through the D2D connection to the whole file size Z_{fn}, at a certain transmission rate $r_{S,P}$. Taking into account the file request probability p_f, the offloading ratio can be expressed as:

$$\mathcal{O}_{S,f} = p_f \sum_{S,P \epsilon N} r_{S,P} \tau_{S,P} / z_{fn} \tag{19}$$

where $\tau_{S,P}$ is the contact duration between Sara and Peter.

The overall offloading ratio for all users $\mathcal{O}_t$ is given as:

$$\mathcal{O}_t = \frac{1}{N} \sum_{A \in N} \sum_{f \in F} \left(\sum_{A,B \epsilon N} r_{A,B} \tau_{A,B} / z_{fn} \right) p_f \tag{20}$$

The optimum value of the offloading ratio is generally reached when the user with the highest total probability in each interest group is selected. This highest total probability is reached in the proposed algorithm through the searching, dividing, and sorting mechanism that is carried out at the base station. In other words, the proposed algorithm is considered a heuristic one, which depends on the prior knowledge by the base station, of the user info (speed, location, interest, cache size, and favorite zone).

VII. SMART ADAPTIVE ALGORITHM

An algorithm is proposed that is based on the base station as an active player in the caching process. Initially, the base station receives a message from each registered user containing his information (the connection address, location, speed, and cache capacity). Four probabilities have to be calculated, the first one is the probability of the user choosing his next zone Zi which relies on the power exponent α as described in section (III A). The second probability is the user's chance to encounter others in the same zone. Meeting more users will result in increasing the chance of downloading files from each other. The third probability is that if the two users (Sara and Peter) have the same interests. The last probability is when the user encounters another one with the same interest and finds his requested file in his user's cache. The total probability for each user is then determined. Higher total probability means that the user has a greater chance of meeting others with the same interest and a greater probability of accessing one of the files in their cache. The algorithm proceeds by dividing the users into groups, based on their interests, and sorts them in a descending order according to their total probabilities. Caching popular files in a user cache with a higher total probability is given more priority. Users in each interest group will cache popular files belonging to the same interest group, meaning that users interested in sports, for example, will cache popular files belonging to the sports category. The base station will start caching popular files into user's cache allocated for proactive caching until each user cache is full. The proposed algorithm is thus a smart and adaptive one, and will be denoted as smart adaptive algorithm (SAA). The pseudo-code for this algorithm is shown in algorithm 2. The base station receives the info message from each user which includes user connection address, location, speed, and cache capacity. The base station calculates the expected next zone for users according to the probability given in equation 3, from which it calculates the probability that users (e.g. Sara and Peter) are sharing the same zone to meet each other at the same time (equation 6). If Sara and Peter have a probability to meet each other at the same time, then the base station calculates the probability of both users (Sara and Peter) to have the same interests and the probability that Sara can find her file at Peter‘s cache (equation 13). The base station sorts the users into groups (G_n) according to their interests, then it sorts the users ($N_{I,n}$) of each group in a descending order according to the user total probability calculated in equation 14. The base station begins fetching popular files ($f_{i,I}$) for each interest group (F_I) into users' cache (C_n) starting from the user with the highest total probability. After finishing fetching a file into a user cache, the remaining cache ($U_{c,n}$) is calculated after subtracting the fetched file size (Z_f) from the user remaining cache and the base station adds the user to user list storing the file ($x_{f,I,n}$). The caching placement into the user‘s cache will continue until the all user‘s cache is full.

Algorithm 2 Smart Adaptive Algorithm

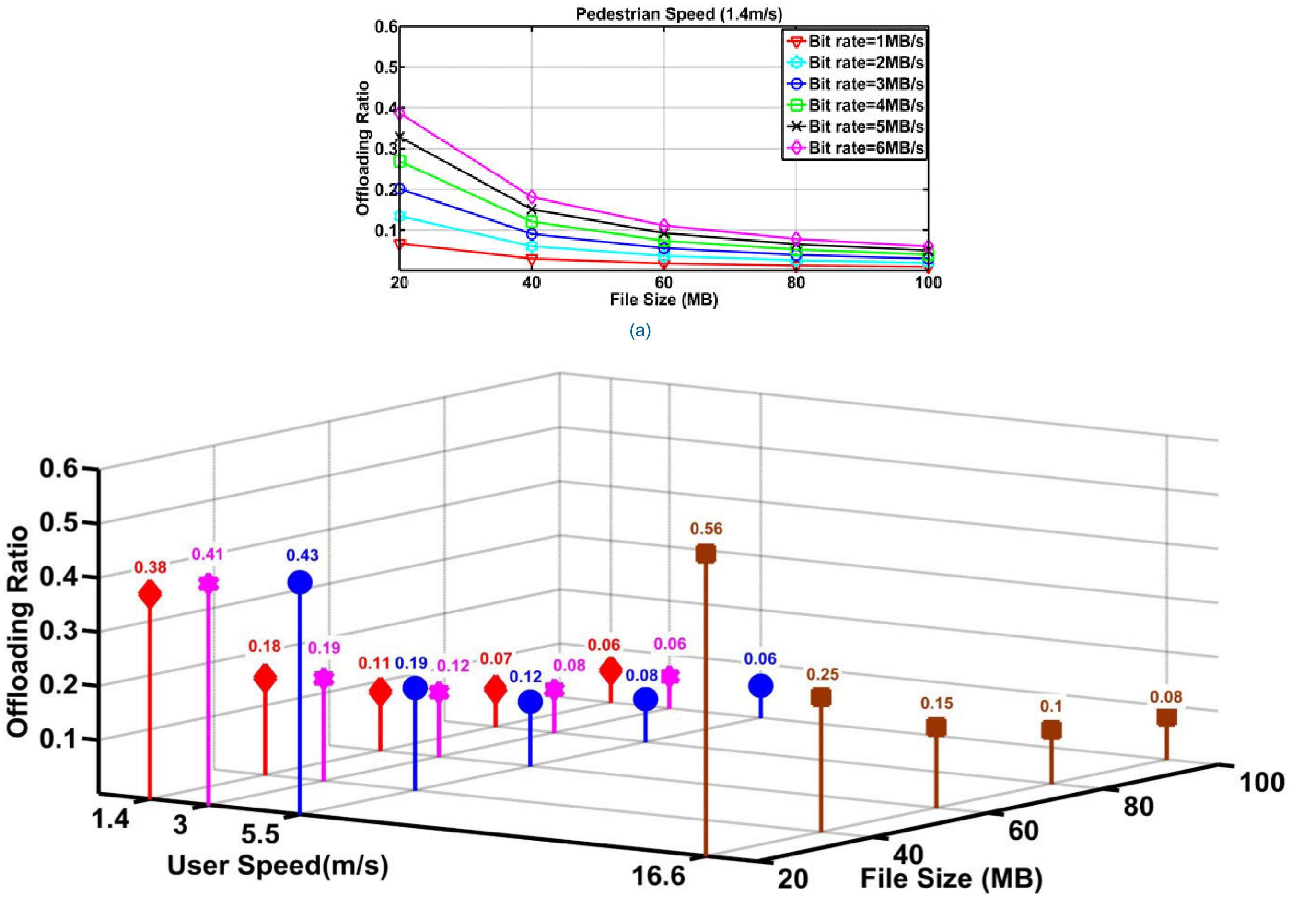
```
Inputs: - user connection address, location, speed, cache capacity eq (1).
Calculate the probability of user next zone eq (3).
Calculate the probability that two users (Sara, and Peter) are in the same zone at the same time instance eq (6).
Calculate the probability that the users (Sara, and Peter) have the same interests eq (13).
Calculate the probability that Sara will find her file at Peter's cache.
Calculate user total probability for each user eq (14).
Dividing users into groups according to their interests.
Sorting users of each group in descending order according to their total probability.
**Caching Files**
For $G_1 \leq G_I$
    For $N_{I,1} \leq N_{I,n}$
        For $U\_cn < C_n$
            For $f_{i,I} \leq F_I$
                $U_{cn} = U_{cn} + Z_f$.
                $x_{f,I,n} = 1$.
            end
        end
    end
end
```

VIII. SIMULATION ANALYSIS AND RESULTS

A. INVESTIGATIONS AND ANALYSIS

Several variables may affect the performance of the proposed (SAA), namely the user speed, the number of users, the bit rate, and the file size. The influence of these variables can be investigated using the human mobility model presented in Section II by utilizing the proposed (SAA). An area for the simulation scenario is assumed to be $L \times L$ meter where it is composed of a number of zones, each zone is $l \times l$ meter covered by one cellular base station. Each user moves randomly with an assigned speed which is either of the predetermined ones (pedestrian, running, biking, or vehicular speed) from one zone to another. A popular file library is assumed to contain F files belonging to the different categories of interest and different sizes between 20 MB to 100 MB. Each mobile device is assumed to have 1GB of his cache, reserved for the proactive caching activity.

The relationships between the offloading ratio and the file size at different bit rates could be plotted for users at various speeds. An example is shown in Figure 5 (a) which illustrates such a relationship for the case of a pedestrian user with an average speed $Vw = 1.4$ m/s. It is observed that the offloading ratio declines when the file size increases at various values of bit rates. In addition, the higher the values of bit rates, the higher the offloading ratio. The same behavior could be observed at different user speeds as shown in Figure 5(b) which illustrates the relationship between the file size, and the offloading ratio at different user speeds. The figure also illustrates the shift in offloading ratio at various speeds and file sizes considering a bit rate of 6 Mbps. The percentage differences between the offloading ratios for different speeds at a fixed file size indicate higher offloading with higher user speeds. The offloading ratio is observed to depend on the file size as well. In the smaller file sizes, a better offload ratio is more noticeable than in a big one. For the 20 MB file, the rate of increase in the offloading ratio is greater than that for the 100 MB file. For the 20MB file, the offloading ratio of the running user is 3% higher than that of a pedestrian one, whereas the biking user is 2% higher than that at running one, and the vehicular user is 13% higher than that of a biking one. For the 100MB file, the offloading ratios of the pedestrian user, running user and the biking user are unchanged, while for vehicular speed users it is 2% higher. These results show that in small file sizes, the influence of the user speed is strongly observed, whereas this effect is less pronounced in cases of larger file sizes. The reason for this behavior is that fast moving users have higher probabilities to meet more people but with shorter communication time, enabling an effective exchange of small file sizes rather than larger ones. The proposed SAA algorithm can thus be practically implemented to allow caching the most popular small size files at the higher speed users.



(a)

(b)

FIGURE 5. **Relation between the offloading ratio and the file size for 80 users, 6 MB/s bit rate, area 1500m×1500m, 90 files at different user mobilities: (a) Pedestrian speed.
(b) All user speeds.**

The number of users is another variable that affects the offloading ratio. Figure 6 illustrates the relation between the number of users and the offloading ratio. It is shown that with an increasing number of users, the offloading ratio goes higher. This is because the amount of total cache reserved for proactive caching increases and more popular files are thus cached. Additional users will increase the chances of meeting each other and thus exchanging more files. As mentioned in section (III), if the user requesting a file does not identify users saving it, then he/she is told to wait for a predefined period of time.

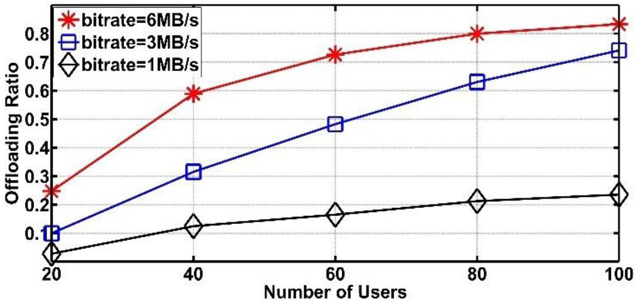


FIGURE 6. **Offloading Ratio vs Number of users in area 800 × 800m, 100MB file size, 90 files.**

The effect of waiting time is examined in Figure 7 for the case of 70 biking users moving in a zone of 1500 × 1500 m^2.

This figure illustrates the relationship between the waiting time and the offloading ratio at two different bit rates, namely 2 MB/s and 6 MB/s with a file size of 20 MB. It demonstrates that the more the ability of a user to wait, the higher the possibility of meeting users deposited the file size and this results in an increase in the offloading ratio. It is observed that the offloading ratio increases with the increase in the bit rate (by about 9% in the given case). It is also noted that the bit rate does not affect the rate of increase of the offloading ratio with the waiting time. Figure 8 illustrates the relation between the detection range of the wireless D2D connection and the offloading ratio at different bit rates. The figure illustrates the

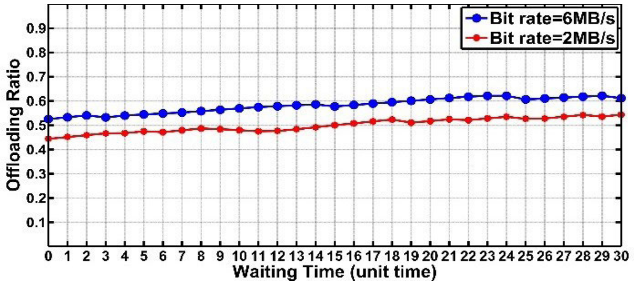


FIGURE 7. Waiting time for 70 biking users, area 1500m × 1500m, 20MB file size, 70 files.

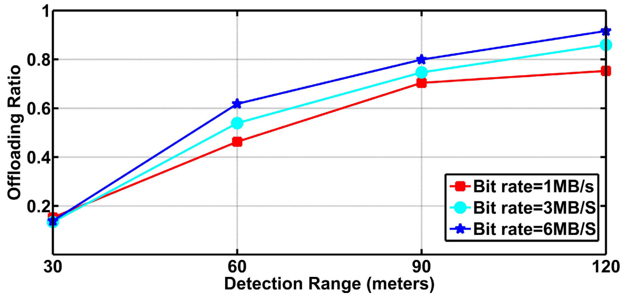


FIGURE 8. Detection range for 70 biking users, area 1000m × 1000m, 20MB file size, 70 files.

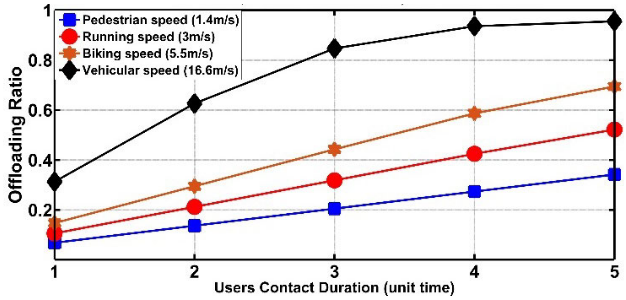


FIGURE 9. Users contact duration for 80 users, area 1000m × 1000m, 250MB file size, 6 MB/s bit rate.

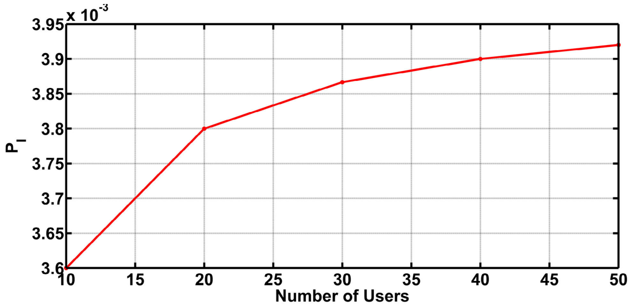


FIGURE 10. Sum of probability of interests (P_I) for all users versus the number of users for 1GB cache, 1500∗1500m area, 60 files, and 4 MB/s bit rate.

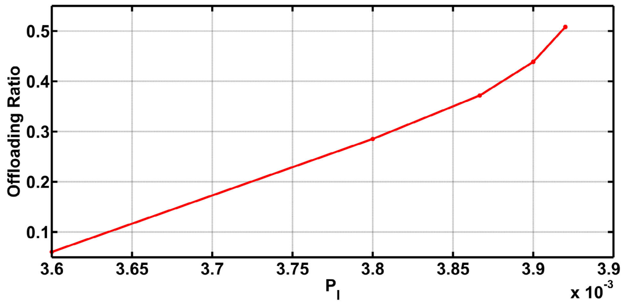


FIGURE 11. Offloading ratio versus the probability (P_I) of interests for 1GB cache size, 1500∗1500m area, 150 files, and 4 MB/s bit rate.

case of 70 bilking users cashing 70 files, each of size 20MB and moving in a zone of 1000 x 1000 m^2. It is clear from this figure that the wider the detection range, the more the offloading ratio.

The contact duration between two users is the duration in which they can establish a successful D2D link. Figure 9 shows the relationship between the users' contact duration and the offloading ratio for different user speeds.

The figure shows that the longer contact duration, the more data that can be shared, and thus higher offloading ratio could be achieved. It is also observed that for the users with vehicular speeds, the offloading ratio tends to saturate at higher contact durations indicating that the file is completely downloaded in the requestor cache. This ensures that the proposed SAA algorithm is more suitable for caching the most common small size files to higher speed users. From another perspective, higher speed users meet more users for a short duration, while lower speed users meet fewer ones for a longer contact period.

The relationship between the probability P_I and the number of users is shown in Figure 10. It is clear that for a number of users ranging from 10 to 50, the probability is slightly increasing. However, a significant increase in probability would be observed when the number of users is getting higher. Figure 11 shows the relationship between P_I and the network offloading for a number of users ranging from 10 to 50. It is clear from this figure that the more the users having the same interest, the higher the user's probability of finding his/her file in their cache, and hence higher network offloading.

B. DATA EXCHANGE FACTOR

As a matter of fact, users share the files with different percentages, some deliver more data than what they receive. In order to investigate the relation between the amount of transmitted and received data via a D2D communication link, a new parameter called the data exchange factor ($\mathcal{E}$) is proposed. It is defined as the data transmitted by a user over the data received via D2D communication. The exchange factor for each user is denoted by ($\mathcal{E}_u$), and is given by the sum of data transmitted (D_{tx}) over the sum of data received (D_{rx}). It is expressed as:

$$\mathcal{E}_u = \frac{\sum D_{tx}}{\sum D_{rx}} \tag{21}$$

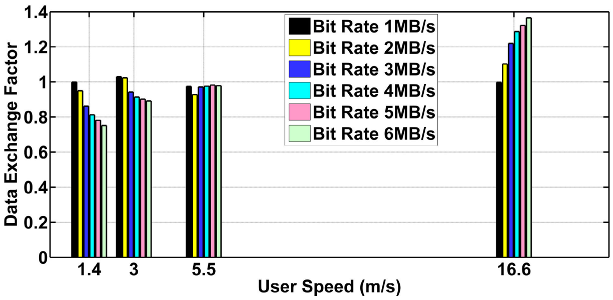


FIGURE 12. The relationship between the data exchange factor and the user speed with different bit rates.

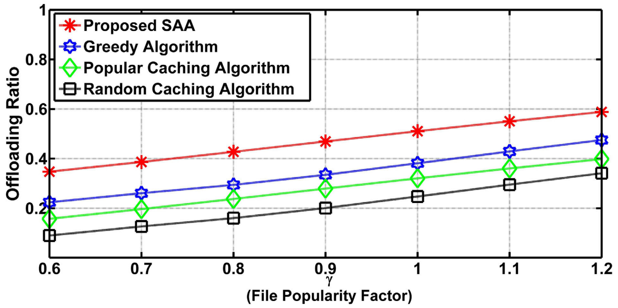


FIGURE 13. 120 users, 1GB cache, 600*600m area, 150 files, 6 MB/s bit rate.

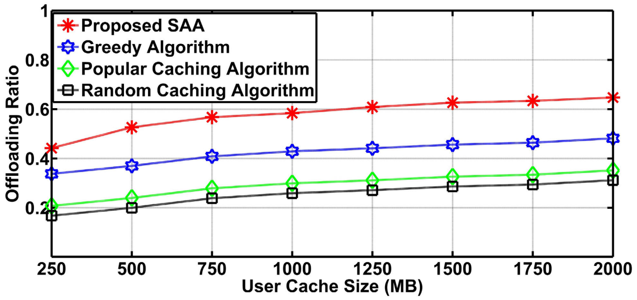


FIGURE 14. 120 users, 600*600m area, 150 files, 6 MB/s bit rate.

Since the users are divided into four sets according to their speeds, then the data exchange factor vector for all user sets is given as:

$$\mathcal{E} = \{\mathcal{E}_w, \mathcal{E}_r, \mathcal{E}_b, \mathcal{E}_v\} \tag{22}$$

where $\mathcal{E}_w, \mathcal{E}_r, \mathcal{E}_b, \mathcal{E}_v$ are the exchange factors for the four user speed sets (pedestrian, running, biking, and vehicular) respectively.

Figure 12 shows the relation between the data exchange factor and user speed at different bit rates. The figure shows that the variations in the exchange factor are based on the bit rate from one user's speed to another. For the pedestrian speed users (1.4m/s), the data exchange factor declines with the increase in the bit rate. This indicates that the data received by the user is higher than that transmitted when the bit rate is increased. The same behavior is experienced for running speed users (3m/s), noting that the data exchange factor declines at a lower rate than the users of walking speeds. For the biking speed users (5.5m/s), the data exchange factor almost does not change with the change in bit rate. For vehicular speed users (16.6m/s), the relationship is reversed from that of the walking and running users. With an increase in the bit rate, the data exchange factor improves, which means that the users' transmitted data is higher than the obtained information. This relationship is another proof of the rise in the offloading ratio when the user speed increases. The proposed data exchange factor would be a helpful merit for the base station to rank users who are more likely to utilize D2D communications to share their files.

C. PERFORMANCE COMPARISON

The performance of different caching placement algorithms is compared to the proposed smart adaptive algorithm (SAA) as illustrated in Figures 13 and 14. The utilized parameters are in accordance with those used in other algorithms that we compare with, in order to furnish a fair comparison between simulated results. Figure 13 shows the relationship between the offloading ratio versus file popularity factor Υ [48], illustrating that this ratio increases with the increase in this factor. It is quite notable that the proposed algorithm outperforms other algorithms with an increase in the offloading ratio by 11 %, 19 %, and 25 % compared to greedy [34], popular caching [20], and random caching [19] algorithms respectively.

Figure 14 shows the offloading ratio versus the user cache assigned to the proactive caching placement, indicating that the more cache assigned for caching popular files, the more satisfaction of the user requests, which results in higher offloading ratio. It is clear that the proposed algorithm outperforms the greedy, popular caching and random caching algorithms by an average increase in offloading ratio by percentages of 28%, 50%, and 52% respectively.

In comparison to these algorithms, the proposed SAA has two major advantages. The first one is the ability to identify users according to their preferred zones, then grouping them in each zone according to their interests. This is followed by caching popular files to the users with the same interest. This approach is considered a smart user selection one as compared to other algorithms. The second advantage is taking into consideration the requested file size, and investigating its effect on the relation between the offloading ratio and user's speed. The impact of the file size has not been taken into account in other algorithms.

As far as the proposed file request mechanism is concerned, its comparison to the conventional one is demonstrated in Figures 15, 16 and 17. Figure 15 illustrates the relation between the number of users received the file request and the number of replies. It is clear that in the conventional mechanism, the number of replies increases with the number of users receiving the requested file. However, in the proposed mechanism, the number of replies is independent of the number of users within the communication range. This is

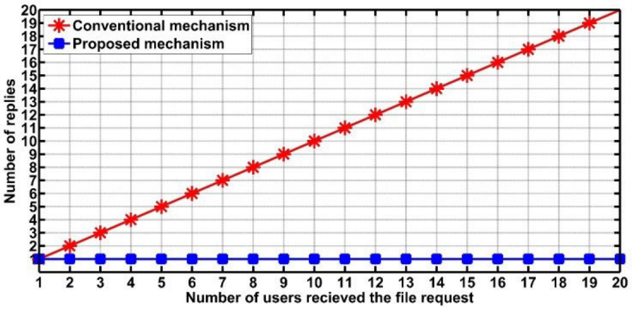


FIGURE 15. The number of replies versus the number of users.

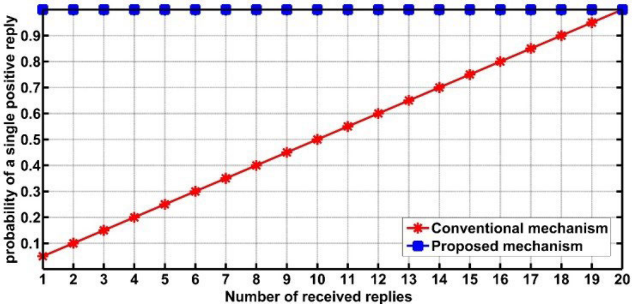


FIGURE 16. The number of received replies versus the probability of one positive reply, assuming 20 users, each of them sends one reply.

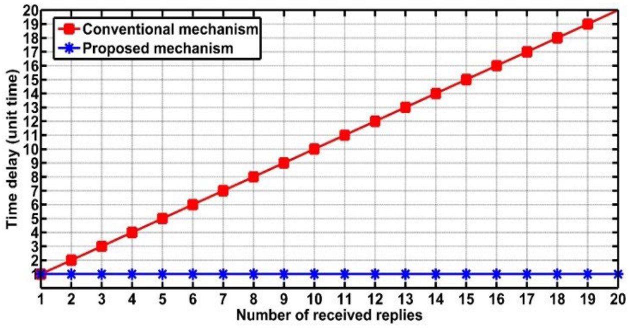


FIGURE 17. The number of replies versus the time delay.

obvious since the user requesting the file will search for the users' addresses sent to him instead of broadcasting a request message to the nearby users.

Figure 16 illustrates the relation between the number of received replies and the probability of a single positive reply It is shown that for the conventional mechanism, the probability to find the requested file is directly proportional to the number of received replies while in our proposed mechanism, the probability nearly approaches unity. If we assume that users' replies are received sequentially in the conventional mechanism (one reply per unit time), then the time delay until the download progress begins will be directly proportional to the number of received replies until receiving a positive one.

This is shown in Figure 17 which illustrates the relationship between the number of replies and the time delay (time delay until the requestor finds the user with the requested file in his cache). It is clear from this figure that with the proposed mechanism, there is no time delay if the user sorting the file exists within the communication range of the requesting one. This means that the proposed mechanism would result in less consumption of energy by the user. Lowering energy consumption would be directly reflected in the cost afforded by the user during the caching process. In other words, the user cost would be reduced.

However, a signaling link must be established between the BS and users to allow communicating the request, response and acknowledgment messages. This will add to the communication complexity of the system.

IX. CONCLUSION

In this article, we address network offloading through the D2D communication by proposing a smart adaptive algorithm (SAA). The algorithm makes use of the data relevant to user location, favorite zones, and interests to fetch popular files into user cache taking into account user speed, size of files transmitted via the D2D connection. To reduce the energy consumed by the user (user cost), a file request mechanism was proposed. A human mobility model was used for simulation to investigate the effect of human mobility on the network offloading. The simulation results showed that the amount of increase in the offloading ratio due to the increase in the speed of the user was based on the size of the files transmitted through D2D connection. Transmitting files with smaller sizes had resulted in a significant increase in the offloading ratio. On the other hand, transmitting files with larger sizes had resulted in a less change in the offloading ratio especially at higher user speeds. The results also showed that higher speed users could transmit more data at high bit rates as compared to users with lower speeds or lower bit rates. The results of the proposed algorithm were compared to those of other caching placement algorithms available in literature. The results showed that as far as the popularity factor is concerned, the proposed algorithm outperforms its counterparts with a significant increase in the network offloading by 11 %, 19 %, and 25 % compared to greedy, popular caching and random caching algorithms respectively. As for the amount of cache assigned for proactive caching, the proposed algorithm outperforms the above-mentioned algorithms by an average increase of 28%, 50% and 52% respectively. In addition, the proposed SAA had two major advantages over the other algorithms. The first one was the ability to identify users according to their preferred zones, then grouping them in each zone according to their interests. The second advantage was taking into consideration the requested file size and investigating their impact on the offloading ratio. Our proposed algorithm was validated only for low, and medium speed movements only and not for high and very high-speed movements. This is because the algorithm assumed concerned users to move within the coverage area of one base station, and hence no handover processes were considered. On the other hand, the user device had to be equipped with GPS capability, otherwise, the proposed mechanism would not be applicable. Also, each user had to assign a portion of his cache to be used for proactive caching.

However, a signaling link ought to be established between the cellular base station and users to allow communicating the request, response, and acknowledgment messages, which might add to the communication complexity of the system.

REFERENCES

[1] C. Shan, X.-P. Wu, Y. Liu, J. Cai, and J.-Z. Luo, "IBP based caching strategy in D2D," *Appl. Sci.*, vol. 9, no. 12, p. 2416, Jun. 2019.

[2] P. Skocir, D. Katusic, I. Novotni, I. Bojic, and G. Jezic, "Data rate fluctuations from user perspective in 4G mobile networks," in *Proc. 22nd Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2014, pp. 180–185.

[3] 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.

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

[5] 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.

[6] A. Altieri, P. Piantanida, L. R. Vega, and C. G. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 4958–4971, Sep. 2015.

[7] G. Chandrasekaran, N. Wang, and R. Tafazolli, "Caching on the move: Towards D2D-based information centric networking for mobile content distribution," in *Proc. IEEE 40th Conf. Local Comput. Netw. (LCN)*, Oct. 2015, pp. 312–320.

[8] R. Lan, W. Wang, A. Huang, and H. Shan, "Device-to-device offloading with proactive caching in mobile cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.

[9] A. Afzal, S. A. R. Zaidi, D. McLernon, and M. Ghogho, "On the analysis of cellular networks with caching and coordinated device-to-device communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[10] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[11] 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.

[12] W. Wang, R. Lan, J. Gu, A. Huang, H. Shan, and Z. Zhang, "Edge caching at base stations with Device-to-Device offloading," *IEEE Access*, vol. 5, pp. 6399–6410, 2017.

[13] Y. Gao, Y. Li, H. Yu, X. Wang, and S. Gao, "Energy efficient content aware cache and forward operation in 3GPP LTE-advanced base stations," in *Proc. 3rd Int. Conf. Comput. Sci. Netw. Technol.*, Oct. 2013, pp. 816–820.

[14] Z. Wang, M. Peng, D. Chen, Y. Li, and J. Zhou, "Delay-optimized small world model for base station caching," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Aug. 2015, pp. 1447–1452.

[15] J. Gu, W. Wang, A. Huang, and H. Shan, "Proactive storage at caching-enable base stations in cellular networks," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 1543–1547.

[16] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 1–6.

[17] Y. Li, M. Qian, D. Jin, P. Hui, Z. Wang, and S. Chen, "Multiple mobile data offloading through disruption tolerant networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 7, pp. 1579–1596, Jul. 2014.

[18] L. Wang, H. Wu, Y. Ding, W. Chen, and H. V. Poor, "Hypergraph-based wireless distributed storage optimization for cellular D2D underlays," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2650–2666, Oct. 2016.

[19] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 1461–1465.

[20] S. Krishnan and H. S. Dhillon, "Effect of user mobility on the performance of device-to-device networks with distributed caching," *IEEE Wireless Commun. Lett.*, vol. 6, no. 2, pp. 194–197, Apr. 2017.

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

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

[23] S. Hosny, A. Eryilmaz, and H. El Gamal, "Impact of user mobility on D2D caching networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[24] D. Wang, H. Li, S. Wang, J. Yang, X. Liu, and X. Zhang, "Mobility aware caching incentive scheme for D2D cellular networks," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.

[25] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[26] T. Deng, G. Ahani, P. Fan, and D. Yuan, "Cost-optimal caching for D2D networks with presence of user mobility," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[27] T. Deng, G. Ahani, P. Fan, and D. Yuan, "Cost-optimal caching for D2D networks with user mobility: Modeling, analysis, and computational approaches," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3082–3094, May 2018.

[28] C. X. Mavromoustakis, G. Mastorakis, and J. Mongay Batalla, "A mobile edge computing model enabling efficient computation offload-aware energy conservation," *IEEE Access*, vol. 7, pp. 102295–102303, 2019.

[29] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1–15.

[30] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, 2014, pp. 649–653, doi: 10.1109/ISWCS.2014.6933434.

[31] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

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

[33] N. Anjum, Z. Yang, H. Saki, M. Kiran, and M. Shikh-Bahaei, "Device-to-device (D2D) communication as a bootstrapping system in a wireless cellular network," *IEEE Access*, vol. 7, pp. 6661–6678, 2019.

[34] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Exploiting mobility in cache-assisted D2D networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5592–5605, Aug. 2018.

[35] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility increases the data offloading ratio in D2D caching networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[36] A. Ahmed, H. Shan, and A. Huang, "Modeling the delivery of coded packets in D2D mobile caching networks," *IEEE Access*, vol. 7, pp. 20091–20105, 2019.

[37] 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.

[38] W. Song, "Analysis of a distance-based pairing scheme for collaborative content distribution via Device-to-Device communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9245–9256, Sep. 2019.

[39] A. Ahmed, H. Shan, and A. Huang, "Modeling impact of concurrent transmissions in cluster based mobile caching networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4601–4605, Apr. 2020.

[40] A. Shabani, S. P. Shariatpanahi, V. Shah-Mansouri, and A. Khonsari, "Mobility increases throughput of wireless device-to-device networks with coded caching," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[41] S. B. Hassanpour, A. Khonsari, S. P. Shariatpanahi, and A. Dadlani, "Hybrid coded caching in cellular networks with D2D-enabled mobile users," in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–6.

[42] B. Chen and C. Yang, "Energy costs for traffic offloading by cache-enabled D2D communications," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2016, pp. 1–6.

[43] 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.

[44] M. Chen, Z. Yang, W. Saad, C. Yin, H. Vincent Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," 2019, *arXiv:1909.07972*. [Online]. Available: http://arxiv.org/abs/1909.07972

[45] A. D. Nguyen, P. Sénac, V. Ramiro, and M. Diaz, "STEPS—An approach for human mobility modeling," in *Proc. Int. Conf. Res. Netw.*, in Lecture Notes in Computer Science, Valencia, Spain, 2011, pp. 254–265.

[46] J. Y. Kim, K. Collins-Thompson, P. N. Bennett, and S. T. Dumais, "Characterizing Web content, user interests, and search behavior by reading level and topic," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 213–222.

[47] Y. Wang, "An algorithm about the measurement of user interest based on Web log," in *Proc. 3rd Pacific-Asia Conf. Circuits, Commun. Syst. (PACCS)*, Jul. 2011, pp. 1–3.

[48] 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.

AHMED HASSAN ABDEL SALAM received the bachelor's degree in electrical engineering from the Military Technical College, in 2001. He is currently pursuing the master's degree in electronics and communications engineering with the Arab Academy for Science, Technology and Maritime Transport. He is currently a Senior Engineer of Transportation Aircraft and the Chief Engineer of the Avionics Workshop for the Cargo Aircraft at the Egyptian Air Force. His research interests include proactive caching and cellular network offloading.

HUSSEIN M. ELATTAR (Associate Member, IEEE) received the B.Sc. degree from the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo, Egypt, and the M.Sc. and Ph.D. degrees from Ain Shams University, Egypt, all in electrical engineering (electronics and communications). He is currently an Associate Professor with the Department of Electronics and Communications Engineering, AASTMT. His current research interests include radio resource management, cognitive networks, optimization techniques, the Internet of Things (IoT), and 5G applications.

MOHAMED A. ABOUL-DAHAB (Life Senior Member, IEEE) received the B.Sc. degree in communication engineering from the Faculty of Engineering, Cairo University, in 1973, and the M.Sc. and Ph.D. degrees in communication engineering from Alexandria University, Egypt, in 1980 and 1986, respectively. He has been a Professor of Communications Engineering with the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Egypt, since 1999. During his long service at AASTMT, he had occupied the post of Head of the Department of Electronics and Computer Engineering, from 1995 to 2002, and the Dean of the College of Engineering and Technology (Cairo Campus), from 2002 to 2008. His publications are in the areas of antennas, channel coding, and wireless networks, in addition to some publications in the field of education. He was a member of the National Radio Science Committee, from 1993 to 2015, the Chairman of the National Committee of ITU, from 2015 to 2019, and has been the Chairman of the National Committee for Communications and Information Technology, since 2019, all of which are belonging to the Academy of Scientific Research and Technology, Egypt.