IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

SPECIAL SECTION ON D2D COMMUNICATIONS:
SECURITY ISSUES AND RESOURCE ALLOCATION

Received March 2, 2019, accepted April 1, 2019, date of publication May 3, 2019, date of current version May 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2914693

Socially-Aware Caching in Wireless Networks With Random D2D Communications

KHAI NGUYEN DOAN[1], (Student Member, IEEE), THANG VAN NGUYEN[2], (Member, IEEE), HYUNDONG SHIN[3], (Senior Member, IEEE), AND TONY Q. S. QUEK[1,3], (Fellow, IEEE)

[1]Singapore University of Technology and Design, Singapore 487372
[2]Department of Electrical and Electronic Engineering, University College Dublin, Belfield, Dublin 4, D04 V1W8 Ireland
[3]Department of Electronic Engineering, Kyung Hee University, Yongin 17104, South Korea

Corresponding author: Tony Q. S. Quek (tonyquek@sutd.edu.sg)

ABSTRACT Pushing contents to users with device-to-device (D2D) data sharing is considered as a promising solution to overcome backhaul congestion. However, this model is associated with critical issues which are user selfishness in terms of sharing personal resources and security concern when connecting with strange devices. Therefore, in this paper, we propose the notion of random D2D connection where the probabilities that the devices are connected for receiving or sharing data are manipulated by users themselves. Moreover, these probability values can also be treated as variables to be optimized. Based on that, we address the backhaul congestion issue and formulate it as a non-convex optimization problem. Two solving schemes are proposed based on primal decomposition and the alternating direction method of multipliers algorithm, respectively. In addition, these methods are designed in both centralized and distributed manners. Besides that, the congestion probability expression is derived in the special case giving an upper bound for the performance of our presented solution. The numerical results are provided to illustrate the effectiveness of the proposed method.

INDEX TERMS Caching, random connection, D2D caching, power allocation, backhaul congestion, selfishness, security.

I. INTRODUCTION

The growth in the number of smart electronic devices is pushing data traffic to the limits of our current cellular network infrastructure. To tackle this issue, caching has been considered as a promising solution, especially caching at users' devices [1]–[9]. This solution allows the caching resources to be extended with the number of users in the network and addresses the preference differences among users compared to caching at base station (BS) [10]–[12]. By caching the most popular content items at users' devices, there are higher chances to retrieve the desired data pieces right from the devices. This not only increases the user quality of experience, but also reduces the network data traffic, leading to the reduction of the system power consumption.

In [13], the authors considered a scenario where mobile devices that have caching capability were distributed randomly. Aiming to minimize the average caching failure rate, a searching algorithm was proposed to find an appropriate cache placement policy. In a small-cell-network context, a joint cache placement and D2D establishment scheme was investigated in [14] with the goal of maximizing an offloading probability. Besides that, in [15], two different D2D-based transmission modes called D2D enabled connected transmission and D2D opportunistic transmission were proposed to maximize the content downloading flows from all BSs to the content downloaders. Meanwhile, the authors of [16], [17] studied the power allocation of D2D communications in cellular networks to maximize the system sum-rate. In addition, separated D2D communication models were studied in [18]–[21] from an energy efficiency perspective where the approaches were based on energy harvesting, contract theory and user clustering with resource reuse policies. The common issue from these works is that they simply assume a full participation of users. However, for these models to be practically deployed, the problems from user selfishness and security concern need to be paid more attention. For this reason, some other works considering similar data sharing model try to address the user incentive problem.

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

For example, the authors of [22] proposed a memory-partition approach where a portion of device memory could be used to store user's favorite items and the remaining part was to serve the others via D2D connections. From another point of view, in [23], the authors exploited two different social relations called social trust and social reciprocity. Besides that, the rewarding mechanism based on Stackelberg game approach was widely investigated in many pieces of research such as [24]–[28]. In these works, rewards or payments were provided based on the contribution of users to the offloading target. This idea has also motivated several other works in which users were paid for either the amount of data they shared or the popularity of content items they cached [29]–[34]. Having a similar inspiration, the authors in [35] dealt with the user incentive by jointly addressing the resource management and reward design problems.

The recent aforementioned works mainly based on monetary benefits to embolden users. However, besides the earned profits, users also have other concerns such as how frequently they have to help the others, how to help someone that they prefer more than the others, how to stop sharing when their batteries are low or how to avoid connecting to devices that they think not secure. Therefore, from another viewpoint of incentive solutions, we propose in this work the notion of random D2D link activation where users are able to decide their own data sharing and receiving probabilities which instead, can also be treated as variables to be optimized. Then, users are able to address all the mentioned concerns by adjusting these probability values to desired levels. Based on that, we address the backhaul congestion problem in a cellular network. The contributions of our works are summarized as follows.

- The novel method of random D2D link activation is introduced in which the oftenness of connecting between devices can be either manipulated by users themselves or optimized by the central controller. The issue is formulated as a non-convex optimization problem.
- Two solving methods built on primal decomposition (PD) and alternating direction method of multipliers (ADMM), respectively, are proposed where both of their centralized and distributed versions are presented.
- To evaluate the performance of the proposed methods, we derive the congestion probability expression given any cache placement strategy in special cases. This is when the users' interaction time follows some specific distributions.

The rest of this work is organized as follows. Section II describes the network model under our consideration. Next, the problem formulation and solving methods are provided in Section III. Then, the special-scenario congestion probability is analyzed in Section IV. Subsequently, numerical results are given in Section V. Finally, Section VI concludes our work.

II. SYSTEM MODEL

We consider a D2D aided cache-enabled network as in Fig. 1, consisting of a user set $\mathcal{U}$ served by a BS in which user u has cache capacity B_u. The BS with a central controller is connected to a server containing a file library $\mathcal{F}$ via a limited-capacity backhaul link. Every file is encoded into several rateless Fountain coded segments [36]–[38], and users can reconstruct a file when they gather sufficient amount of coded segments. We denote S_f the total size of coded segments required for file f. Any pair of users u and v are able to exchange data based on their social interaction which is represented by the average contact time T_{uv} between them. We assume that $T_{uv}, \forall u, v$ are learnt and known by the BS. The quantity T_{uv} implies the interaction levels between users which is the first aspect of their social relationships. To avoid confusing, we note that the term ''contact time'' and ''interaction time'' will be used interchangeably.

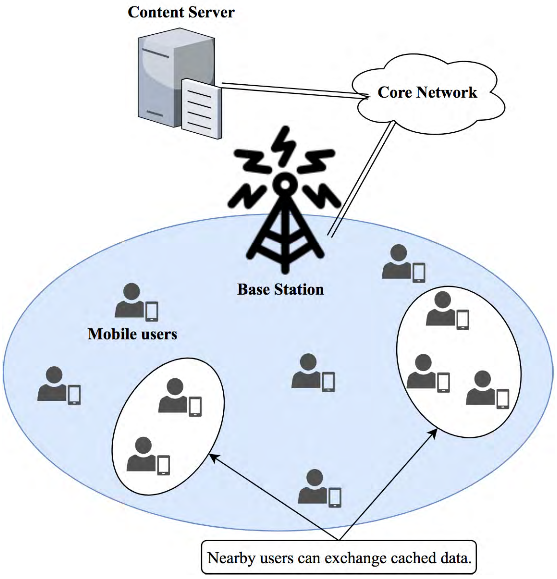


FIGURE 1. The system model under consideration with a BS and many users. Users' devices have cache memories for proactively storing data pieces. D2D connections have chances to be enabled for whoever in the vicinity of each other to exchange the cached information. As the exchanged data is not enough to reconstruct the desired files, the BS will assist to download and transmit the missing part.

The system operates in two phases called *caching phase* and *requesting phase*. The caching phase takes place during off-peak hours when the requests are sparse. In this one, a certain amount of data segments will be downloaded and stored into users' cache memories. In the requesting phase, each user requests a file in the library, and we assume this is a synchronous requesting scenario. Once the request is initiated, a user will connect to the others in the vicinity and receive the desired file from them via D2D links. Unless the data sent through D2D connections is sufficient to recover the requested file, the BS will download the remaining part from the server to satisfy the corresponding request, causing a certain backhaul load. Herein, the file transferring between users can only take place within the interaction time, and the connections will be established randomly. The random establishment mechanism should take into account users selfishness and security concerns. This is firstly because users may want to share data with only a specific group of people (e.g., friends), or care about how frequently they have to help the others due to the limits of personal resources. Secondly, they want to avoid interacting with suspiciously dangerous devices. To address these issues, our controlling scheme will allow users to manipulate their own connection probabilities.

Random D2D communications: Let q_{uv} be the probability that the connection is established for user v to share data with user u. Obviously, this probability depends on how likely user v want to share data with u and how likely u want to receive data from v. In order to define this quantity, firstly, users need to let the central controller know the probabilities that they are willing to share data with the others. This is generally letting the controller know how selfish they are. Secondly, they need to let the controller know the probabilities that they are willing to receive data from the others. This is generally letting the controller know how careful they are, because data received from other devices is not always secure. Then, the controller will define the connection probabilities such that the selfishness and carefulness of all users are satisfied. Note that the selfishness of user v is not satisfied if he has to share data to u with a probability higher than his desired probability. Similarly, the carefulness of u is not satisfied if he has to receive data from v with a probability higher than his expected one. The mentioned process is presented in the following two steps.

1) Each user u shares the set $\left\{\left(l_v^u, g_v^u\right), \forall v \neq u\right\}$ to the BS where $0 \leq l_v^u, g_v^u \leq 1, \forall v \neq u$. l_v^u represents how frequently u wants to receive data from v, and g_v^u represents how frequently u wants to send data to v.
2) The connection for v to send cached data to u is defined by the central controller as follows.

$$q_{uv} = \min\left(l_v^u, g_u^v\right) \tag{1}$$

Note that l_v^u is given by user u, and g_u^v is given by user v. The connection for v to share data to u is activated with probability q_{uv} following (1) will satisfy both the selfishness of v and the carefulness of u. The parameters g_v^u and l_v^u are reflections of the kindness and cautiousness of u in treating v, respectively. This is the second aspect of their social relationships which is exploited.

In the next section, we present our methods which jointly perform caching and control the power allocation for all D2D connections while consider the activation probability of each link. Before closing this section, we introduce Table 1 which summarizes the meaning of notations used throughout this work.

III. JOINT CACHING AND CONTROLLING SCHEME

In this section, we present the problem formulations and solving methods for defining cache placement at users' devices and power allocation for each D2D link. In the presentation hereafter, $q_{uv}, \forall u, v$ can be considered as parameters fixed from user's manipulation or as variables to be optimized.

A. PROBLEM FORMULATION

In order to combine the connection probabilities and the D2D transmission power, we define the transmission power from user v to u as a random variable

$$P_{uv} = \begin{cases} p_{uv}, & \text{with probability } q_{uv}. \\ 0, & \text{with probability } 1 - q_{uv}. \end{cases} \tag{2}$$

For the ease of presentation, we refer (v, u) to the link where v transfer data to u. Then the total data transferred from v to u under the request of file f is

$$D_{uvf} = \min\left\{R_{uv}, C_{vf} S_f\right\} \tag{3}$$

where R_{uv} is the total data amount that v can send to u, defined as

$$R_{uv} = T_{uv} \log\left(1 + \frac{\mathsf{G}_{uv} P_{uv}}{\sum\limits_{(n,m)\neq(v,u)} \mathsf{G}_{un} P_{mn} + \sigma^2}\right) \tag{4}$$

with coefficients G_{uv} represents the channel including fast fading and the path loss between user v and u depending on the separated distance. Since D2D is a short-range communication, and the channel variation in small, $\mathsf{G}_{uv}, \forall u, v$ are assumed to be known average values. $\mathsf{G}_{uv} = \infty$ for $u = v$ implying that user u can get all data in his cache for himself. The term $\sum_{(n,m)\neq(v,u)} \mathsf{G}_{un} P_{mn}$ is the total interference from all other communication links on the (v, u) link. $0 \leq C_{vf} \leq 1$ is the portion of file f that v has cached. Note that, since the contact times of users are different, the interference term in the denominator of (4) is applicable only in dense networks. It is the sum of multiple random variables, hence, when the number of users is large, this sum will approach a constant even when the contact times of users are different. We aim to jointly minimize the average backhaul load and the total power consumption in the system which is expressed as follow.

$$\sum_{f,u} \zeta_{uf} \left\{ S_f - \mathbb{E}\left[\min\left(\sum_{v\neq u} D_{uvf} + C_{uf} S_f, S_f\right)\right]\right\} + \sum_{u,v} \epsilon_u p_{uv} \tag{5}$$

where ζ_{uf} is the probability that user u requests file f which is a known quantity. In fact, it can be learned from users' request history. The first term is the average backhaul load, and the second term is the total power consumption whose impact can be adjusted by weight set $\left\{\epsilon_1, \ldots, \epsilon_{|\mathcal{U}|}\right\}$. For example, user u can restrict the power consumption in the system by rising the corresponding weight ϵ_u. The term $\sum_{v\neq u} D_{uv}$ is the total data sent to a specific user u via D2D links. As shown in [39], that sum is upper bounded by a convergence bound when the number of random variables in the sum is sufficiently large. Therefore, in a dense network, the first term of (5), with the aid of (3), can be approximated by

$$\sum_{f,u} \zeta_{uf} \times \mathcal{J}\left(\mathbb{E}\left[R_{uv}\right]\right) \tag{6}$$

TABLE 1. Notations and meaning.

Notations	Meaning
$\mathcal{F}$	The set of files
$\mathcal{U}$	The set of users
ζ_{uf}	The probability that user u requests file f
f_u	The file requested by user u
$\mathbf{F}$	The set of requested files from every user
$\mathbf{F}'_i$	The i^{th} specific combination with repetition of $\lvert\mathcal{U}\rvert$ among $\lvert\mathcal{F}\rvert$ files
A_n^k	The total number of combinations with repetition of k elements among n elements
$\mathcal{L}_B$	The backhaul load
θ_{th}	The congestion threshold
S_f	Total size of data segments for reconstructing file f
C_{uf}	The portion of file f cached by user u
B_u	Size of user u's cache
T_{uv}	The total time duration that user v can send data to u (contact time)
p_{uv}	The transmission power from user v to u through D2D connection
$p_{\max}$	The maximum D2D transmission power
l_v^u	A parameter indicating the probability that user u wants to receive data from v
g_v^u	A parameter indicating the probability that user u wants to share data to v
q_{uv}	The actual probability (decided by the central controller based on l_v^u and g_v^u) that user v can connect to transfer data to user u
D_{uvf}	The data amount regarding file f transferred from user v to u
R_{uv}	The data amount that user v can possibly send to u for a given contact time and channel condition
G_{uv}	The channel gain between user v and u including fast fading and path loss coefficients
σ^2	The noise power
$(x)\,[t]$	The value of some quantity x at the t-th iteration in algorithms

where

$$\mathcal{J}(x) = \left\{ S_f - \min\left\{ \sum_{v\neq u} \min\left\{x, C_{vf}S_f\right\} + C_{uf}S_f, S_f \right\}\right\} \tag{7}$$

Applying Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}\left[R_{uv}\right] &= \mathbb{E}\left[\mathbb{E}\left[R_{uv}\right] | P_{mn} \forall\left(n,m\right) \neq \left(v,u\right)\right] \\
&\geq q_{uv}T_{uv}\log\left(1 + \frac{\mathsf{G}_{uv}p_{uv}}{\sum\limits_{(n,m)\neq(v,u)} \mathsf{G}_{un}\mathbb{E}\left[P_{mn}\right] + \sigma^2}\right) \\
&\geq q_{uv}T_{uv}\log\left(1 + \frac{\mathsf{G}_{uv}p_{uv}}{\sum\limits_{(n,m)\neq(v,u)} \mathsf{G}_{un}q_{mn}p_{mn} + \sigma^2}\right)
\end{aligned} \tag{8}$$

Using (8), we can derive an upper bound of (6).[1] Aiming to minimize this bound, we have the following problem formulation.

$$\min_{\mathbf{C},\mathbf{P},\mathbf{Q}} \; \mathcal{J}\left(\psi_{uv}\right) + \sum_{u,v} \epsilon_u p_{uv} \tag{9}$$

$$\text{s.t.} \quad 0 \leq q_{uv} \leq 1, \quad \forall u, v \tag{10}$$

$$\sum_{u} q_{uv} \leq q_{s,\max}, \quad \forall v \tag{11}$$

$$\sum_{v} q_{uv} \leq q_{r,\max}, \quad \forall u \tag{12}$$

$$0 \leq C_{uf} \leq 1, \forall u, f \tag{13}$$

$$\sum_{f} C_{uf}S_f \leq B_u, \quad \forall u \tag{14}$$

$$0 \leq p_{uv} \leq p_{\max}, \quad \forall u, v \tag{15}$$

where

$$\psi_{uv} = q_{uv}T_{uv}\log\left(1 + \frac{\mathsf{G}_{uv}p_{uv}}{\sum\limits_{(n,m)\neq(v,u)} \mathsf{G}_{un}q_{mn}p_{mn} + \sigma^2}\right) \tag{16}$$

C, **P** and **Q** are matrices of caching, power allocation and connection probability variables, respectively. Note that the above formulation is a general one. Later we will alternatively consider two subcases of when **Q** is fixed according to user's desire and when it is a variable matrix.

Remark 1: The first term in the objective function presents the cost from the backhaul load, while the second term is added to minimize the total power consumption in the system. The sets of constraints (11) and (12) are to control the total D2D links that can be established for each device to send and receive data, respectively. Note that we do not strictly constraint the number of links but the probability for it to exceed a certain level, i.e., $q_{s,\max}$ and $q_{r,\max}$, respectively. Therefore, in practice, there is a possibility that the number of should-be-activated links exceeds the maximum one that a device can bear. In those cases, redundant links can be randomly dropped. However, we can always adjust $q_{s,\max}$ and $q_{r,\max}$ to suppress the occurrence probability of those unexpected events to an insignificant level. Especially, when the number of users is large, it is even more reasonable to control the link activation probabilities.

[1] In dense networks, the interference term in the denominator of (8) approaches a constant, hence, the Jensen's equality holds (see Section V for the example). In addition, instead of minimizing the original objective function, we choose minimizing its upper bound. This is firstly because this bound is tight in the dense-network scenario. Secondly, minimizing the upper bound will also decrease our original objective function. Therefore, this is a reasonable approach to reach suboptimal solution and also to simplify the problem.

In terms of time scale, the cache placement and power allocation can be done at the beginning of a time interval, then, the results are applied within that interval. The time interval can be one or more days depending on how fast users' preference varies which can be learnt in practice. The original problem formulation can be transformed into a differentiable version by introducing additional variables k_{uf} and z_{uvf}. Then, we have a new objective

$$\min_{\mathbf{C},\mathbf{P},\mathbf{Q},\mathbf{k},\mathbf{z}} -\sum_{f,u} \zeta_{uf} k_{uf} + \sum_{u,v} \epsilon_u p_{uv} \tag{17}$$

with the following additional constraints (besides (10)-(15)),

$$\begin{aligned} k_{uf} &\leq \sum_{v} z_{uvf} + C_{uf} S_f, \quad \forall u, f && (18)\\ k_{uf} &\leq S_f, \quad \forall u, f && (19)\\ z_{uvf} &\leq \psi_{uv}, \quad \forall u, v, f && (20)\\ z_{uvf} &\leq C_{vf} S_f, \quad \forall u, v, f. && (21) \end{aligned}$$

It can be verified that $z_{uvf} = \min\left\{\psi_{uv}, C_{vf} S_f\right\}$ and $k_{uf} = \min\left\{\sum_{v} z_{uvf} + C_{uf} S_f, S_f\right\}, \forall u, v, f$ at the optimum. The proposed solving methods will be presented in the next subsection.

B. Q FIXED AS USERS' SATISFACTION

In this subsection, we first approach the problem under the condition of fixed $q_{uv}, \forall u, v$, i.e., this set of probabilities is manipulated by users as described in Section II. Thus, the constraints (10)-(12) can be removed. We are dealing with a non-convex optimization problem due to the constraint (20). For solving, our first introduced scheme is built on PD method.

1) PRIMAL-DECOMPOSITION-BASED ALGORITHM

Taking advantage of the problem structure, we observe that the two set of variables $\{\mathbf{C}, \mathbf{k}, \mathbf{z}\}$ and $\{\mathbf{P}\}$ are coupled by only the set of constraints (20). Therefore, they can be decoupled by transforming (20) with a parameter set $\omega_{uvf}, \forall u, v, f$ as follows

$$\begin{aligned} z_{uvf} &\leq \omega_{uvf} && (22)\\ \omega_{uvf} &\leq \psi_{uv} && (23) \end{aligned}$$

where the term ψ_{uv} contains variables $p_{uv}, \forall u, v$ as showed in (16). To this end, our problem can be split into two separated ones. The first problem ($\mathcal{P}_1$) associates with the variable set $\{\mathbf{C}, \mathbf{k}, \mathbf{z}\}$, the other ($\mathcal{P}_2$) associates with $\{\mathbf{P}\}$, and both are linear programming problems. $\mathcal{P}_1$ and $\mathcal{P}_2$ are not presented explicitly due to space limitation, but they can be trivially figured out from our original problem formulation.

Using PD method [40], it can be shown that we can find the optimal $\omega_{uvf}, \forall u, v, f$ such that solving two subproblems $\mathcal{P}_1$ and $\mathcal{P}_2$ is equivalent to solving the original problem. The centralized version of this is present in Algorithm III.1.

Remark 2: In this algorithm, the mentioned $\lambda^{(1)}_{uvf}$ and $\lambda^{(2)}_{uvf}$ are the dual variables associated with the constraint (22) and (23), respectively. For the presentation of all algorithms in this work, the operator $(x)[t]$ implies the value of x at the t-th iteration. Here, x can be a scalar, vector or matrix, and for the latter two, this operator is element wise. Note that α is a step size which, in general, can be different across ω_{uvf} and iterations, however, we do not show these dependences in the notation for simplicity.

Algorithm III.1 Centralized PD-based Algorithm.

```
1: Initialize $\mathbf{P}$.
2: repeat for each iteration $t$:
3:     Update $(\mathbf{C}, \mathbf{k}, \mathbf{z})[t]$ and $\left(\lambda^{(1)}_{uvf}\right)[t]$,
4:     Update $(\mathbf{P})[t]$ and $\left(\lambda^{(2)}_{uvf}\right)[t]$,
5:     $\left(\omega_{uvf}\right)[t+1] \leftarrow \left(\omega_{uvf} - \alpha\left(\lambda^{(2)}_{uvf} - \lambda^{(1)}_{uvf}\right)\right)[t]$,
6: until converge
```

Note that lines 3, 4, and 5 are executed sequentially $\forall u, v, f$. Line 3 is done by solving $\mathcal{P}_1$ with $\{\mathbf{P}\}$ is fixed, and line 4 is done by solving $\mathcal{P}_2$ with $\{\mathbf{C}, \mathbf{k}, \mathbf{z}\}$ is fixed. The most up-to-date values of all elements will be used in each line. In order to execute the algorithm in a distributed manner, each user u will need to solve a restricted version of $\mathcal{P}_1$ as described in Algorithm III.2 where $\{\mathbf{C}_u, \mathbf{k}_u, \mathbf{z}_u\}$ is the set of variables associated with index u.

Algorithm III.2 Distributed PD-based Algorithm.

```
Initialize $\mathbf{P}$.
repeat for each iteration $t$:
    User $u$ updates $(\mathbf{C}_u, \mathbf{k}_u, \mathbf{z}_u)[t]$ and $\left(\lambda^{(1)}_{uvf}\right)[t], \forall v, f$.
    Results from users are gathered by BS to update $(\mathbf{P})[t]$
    and $\left(\lambda^{(2)}_{uvf}\right)[t], \forall u, v, f$.
    Results from BS are broadcasted back to users where
    each user $u$ updates
    $\left(\omega_{uvf}\right)[t+1] \leftarrow \left(\omega_{uvf} - \alpha\left(\lambda^{(2)}_{uvf} - \lambda^{(1)}_{uvf}\right)\right)[t], \forall v, f$
until converge
```

2) ADMM-BASED ALGORITHM

The PD-based algorithm is applicable only when $\mathbf{Q}$ is fixed. We present here a second scheme based on ADMM [41] which can either treat $\mathbf{Q}$ as constants or optimize them as variables. To begin with, let us denote $\mathbf{x}$ the vector of all variables $\{\mathbf{C}, \mathbf{k}, \mathbf{z}, \mathbf{P}\}$, and form a set of linear constraints by treating (20) as a penalty term in the objective function. Specifically, the original problem is transformed to

$$\min_{\mathbf{C},\mathbf{k},\mathbf{z},\mathbf{P}} -\sum_{f,u} \zeta_{uf} k_{uf} + \sum_{u,v} \epsilon_u p_{uv} + \frac{\eta}{2} \sum_{u,v,f} \left(g_{uvf}\right)^2 \tag{24}$$

s.t. (13)-(15), (18), (19) and (21)

where

$$g_{uvf} = \exp\left(\frac{z_{uvf}}{q_{uv} T_{uv}}\right) - \frac{\mathsf{G}_{uv} p_{uv}}{\sum_{(n,m) \neq (v,u)} \mathsf{G}_{un} q_{mn} p_{mn} + \sigma^2} - 1 \tag{25}$$

and η is the penalty coefficient. The expression (25) can be linearized by Taylor approximation at $\tilde{\mathbf{x}} = \left\{\tilde{\mathbf{C}}, \tilde{\mathbf{k}}, \tilde{\mathbf{z}}, \tilde{\mathbf{P}}\right\}$ as

$$g_{uvf} \approx \tilde{g}_{uvf} + \frac{\partial \tilde{g}_{uvf}}{\partial z_{uvf}} \left(z_{uv}^{[f]} - \tilde{z}_{uv}^{[f]}\right) + \sum_{u,v} \frac{\partial \tilde{g}_{uvf}}{\partial p_{uv}} \left(p_{uv} - \tilde{p}_{uv}\right) \tag{26}$$

where $\tilde{g}_{uvf}$ is the function g_{uvf} evaluated at point $\tilde{\mathbf{x}}$. $\frac{\partial \tilde{g}_{uvf}}{\partial z_{uvf}}$ and $\frac{\partial \tilde{g}_{uvf}}{\partial p_{uv}}$ denote the partial derivative of g_{uvf} with respect to (w.r.t.) z_{uvf} and p_{uv} at point $\tilde{\mathbf{x}}$, respectively. From (24) and (26), the objective function can be approximated by a quadratic function at some point $\tilde{\mathbf{x}}$. Notably, the gradient direction of the quadratic function at $\tilde{\mathbf{x}}$ can be used to update our objective function. Therefore, the general idea of this method is to iteratively approximate the objective function by quadratic functions at points $\tilde{\mathbf{x}}\,[t]$. Finding and moving along the gradient direction with a step size to obtain a new point $\tilde{\mathbf{x}}\,[t+1]$ and repeat the process with this new point. Therefore, at each iteration our subproblem has the following form,

$$\min_{\mathbf{x}} \; \mathbf{x}^T \boldsymbol{\Upsilon} \mathbf{x} + \boldsymbol{\delta}^T \mathbf{x} \tag{27}$$

$$\text{s.t.} \quad \mathbf{A}_{\mathbf{x}} \mathbf{x} \leq \boldsymbol{\gamma} \tag{28}$$

where $\boldsymbol{\Upsilon}$ and $\boldsymbol{\delta}$ are matrix and vector, respectively, containing parameters in the objective function (24). Similarly, $\mathbf{A}_{\mathbf{x}}$ and $\boldsymbol{\gamma}$ are matrix and vector, respectively, containing parameters in the constraint set. To this end, we introduce a vector of auxiliary variables $\mathbf{l}$ with all nonnegative elements to transform the inequality constrained problem into the following equality constrained one

$$\min_{\mathbf{x},\mathbf{l}} \; \mathbf{x}^T \boldsymbol{\Upsilon} \mathbf{x} + \boldsymbol{\delta}^T \mathbf{x} \tag{29}$$

$$\text{s.t.} \quad \mathbf{A}_{\mathbf{x}} \mathbf{x} + \mathbf{A}_{\mathbf{l}} \mathbf{l} = \boldsymbol{\gamma} \tag{30}$$

where $\mathbf{A}_{\mathbf{l}}$ is a parameter matrix associated with auxiliary variable vector $\mathbf{l}$. Following ADMM, we solve

$$\begin{aligned} \min_{\mathbf{x},\mathbf{l}} \; & \mathbf{x}^T \boldsymbol{\Upsilon} \mathbf{x} + \boldsymbol{\delta}^T \mathbf{x} + \mathbf{y}^T \left(\mathbf{A}_{\mathbf{x}} \mathbf{x} + \mathbf{A}_{\mathbf{l}} \mathbf{l} - \boldsymbol{\gamma}\right) \\ & + \frac{\rho}{2} \|\mathbf{A}_{\mathbf{x}} \mathbf{x} + \mathbf{A}_{\mathbf{l}} \mathbf{l} - \boldsymbol{\gamma}\|^2 \end{aligned} \tag{31}$$

where $\mathbf{y}^T$ is a vector of dual variables associated with the constraint set and ρ is another penalty coefficient. The above is a quadratic function w.r.t. $\mathbf{x}$ and $\mathbf{l}$, respectively. Particularly, (31) w.r.t. $\mathbf{x}$ will take form

$$\min_{\mathbf{x}} \; \mathbf{x} \Psi_{\mathbf{x}} \mathbf{x} + \beta_{\mathbf{x}} \mathbf{x} + \gamma_{\mathbf{x}}. \tag{32}$$

Therefore, the optimal solution w.r.t. $\mathbf{x}$ ($\mathbf{l}$ is fixed) will take the following form

$$\mathbf{x} = -\frac{1}{2} \boldsymbol{\Psi}_{\mathbf{x}}^{-1} \boldsymbol{\beta}_{\mathbf{x}}. \tag{33}$$

where $\Psi_{\mathbf{x}}$ and $\beta_{\mathbf{x}}$ are formed from (31) by treating $\mathbf{x}$ as the only variable. Similarly, the optimal solution w.r.t. $\mathbf{l}$ ($\mathbf{x}$ is fixed) takes the following form

$$\mathbf{l} = \left[-\frac{1}{2} \boldsymbol{\Psi}_{\mathbf{l}}^{-1} \boldsymbol{\beta}_{\mathbf{l}}\right]^+ \tag{34}$$

where $\Psi_{\mathbf{l}}$ and $\beta_{\mathbf{l}}$ are associated with the second-order and first-order term of $\mathbf{l}$, respectively. They are formed by treating $\mathbf{l}$ as the only variable. The element-wise operator $[\mathbf{.}]^+$ is defined by

$$[l]^+ = \begin{cases} l, & l \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

This is due to the fact that $\mathbf{l}$ is a nonnegative vector. The Algorithm III.3 summarizes this method.

Remark 3: It is important to note that relaxing the nonnegative constraints then applying $[\mathbf{.}]^+$ to the final result is not equivalent to solving the constrained problem in general. However, since each element of $\mathbf{l}$ corresponds to a single constraint, the function in (31) w.r.t. $\mathbf{l}$ can be expressed as the sum of multiple quadratic terms, each associated with only one element in $\mathbf{l}$. In other words, optimizing (31) w.r.t. $\mathbf{l}$ is equivalent to optimizing multiple single-variable quadratic functions separately. Therefore, in this specific case, applying $[\mathbf{.}]^+$ to the relaxed problem solution provides the global optimal one for this subproblem.

Algorithm III.3 Centralized ADMM-based Algorithm.

```
1: Initialize x and l at iteration 0.
2: repeat for each iteration $t+1$:
3: $(\mathbf{x})\,[t+1] \leftarrow \left(-\frac{1}{2}\boldsymbol{\Psi}_{\mathbf{x}}^{-1}\boldsymbol{\beta}_{\mathbf{x}}\right)[t]$.
4: $(\mathbf{l})\,[t+1] \leftarrow \left(\left[-\frac{1}{2}\boldsymbol{\Psi}_{\mathbf{l}}^{-1}\boldsymbol{\beta}_{\mathbf{l}}\right]^+\right)[t+1]$.
5: $(\mathbf{y})\,[t+1] \leftarrow (\mathbf{y})\,[t] + \rho\,(\mathbf{A}_{\mathbf{x}}\mathbf{x} + \mathbf{A}_{\mathbf{l}}\mathbf{l} - \boldsymbol{\gamma})\,[t+1]$.
6: until converge
```

Recall that $(x)\,[t]$ implies the value of x at iteration t. Based on the above algorithm, a distributed version can be developed as presented in Algorithm III.4 where $\mathbf{x}_u = \{\mathbf{C}_u, \mathbf{k}_u, \mathbf{z}_u, \mathbf{P}_u\}$ is a set of variables associated with index u. Here, the matrix $\boldsymbol{\Psi}_{\mathbf{x}_u}$ and vector $\boldsymbol{\beta}_{\mathbf{x}_u}$ are formed from (31) by considering all the components other than $\mathbf{x}_u$ as constants. Note that from our constraint set, it can be verified that the objective function (24) is lower bounded by $-\sum_{f,u} \zeta_{uf} S_f$. In addition, according to the updating steps of Algorithm III.3 and III.4, the objective function (24) is monotonically decreased, therefore, converges [41]. Note that line 4 and 5 are done by BS after gathering $\mathbf{x}_u$ from all users.

Algorithm III.4 Distributed ADMM-based Algorithm.

```
1: Initialize $x_u$, $\forall u$ and l at iteration 0.
2: repeat for each iteration $t+1$:
3: Each user $u$ updates $(\mathbf{x}_u)\,[t+1] \leftarrow \left(-\frac{1}{2}\boldsymbol{\Psi}_{\mathbf{x}_u}^{-1}\boldsymbol{\beta}_{\mathbf{x}_u}^T\right)[t]$
4: BS updates $(\mathbf{l})\,[t+1] \leftarrow \left(\left[-\frac{1}{2}\boldsymbol{\Psi}_{\mathbf{l}}^{-1}\boldsymbol{\beta}_{\mathbf{l}}^T\right]^+\right)[t+1]$
5: and $(\mathbf{y})\,[t+1] \leftarrow (\mathbf{y})\,[t] + \rho\,(\mathbf{A}_{\mathbf{x}}\mathbf{x} + \mathbf{A}_{\mathbf{l}}\mathbf{l} - \boldsymbol{\gamma})\,[t+1]$
6: until converge
```

C. Q AS OPTIMIZATION VARIABLES

This is the ideal scenario when users are willing to dedicate all of their resources to achieve the maximum gain in terms of backhaul load alleviation. The optimal result is obtained by solving (17)-(21). Both the centralized and distributed version of the ADMM-based algorithm presented in Subsection III-B.2 can work effectively in this circumstance, which is another advantage of these schemes. Note that by applying the ADMM-based algorithm, the size of variable vector $\mathbf{x}$ is extended with some additional elements corresponding to the set of variables q_{uv}, $\forall u \neq v$, i.e., $\mathbf{x} = \{\mathbf{Q}, \mathbf{C}, \mathbf{k}, \mathbf{z}, \mathbf{P}\}$.

IV. CONGESTION PROBABILITY ANALYSIS IN THE SPECIAL CASE

To obtain an upper bound for the proposed caching strategy, we devote this section to analyze the congestion probability in an interference-free scenario given the cache placement of all users. However, for the notation simplicity, we do not indicate this condition in the probability expressions hereafter. The absence of interference can be achieved by allocating orthogonal subchannels to nearby links, and frequencies are reused for far-apart links, since unlicensed bandwidth can be used for D2D connections.

The authors of [42] suggest two models which are Power-Law and Weibull distributions to model the events that a group of people meet in person and when they interact with each other via mobile devices, respectively. In this work, we will alternatively assume that the user interaction time follows these two models and the corresponding analytical expressions regarding the congestion probability are derived. By assuming that the duration of the interaction time between users u and v follows Power-Law distribution, the probability distribution function (PDF) of T_{uv} is given by

$$f_{\mathrm{PL},T_{uv}}(t) = 2b\,(1+t)^{-2b-1} \tag{36}$$

where b is the shape parameter and the multiplication term $2b$ is the normalization factor. Then, by considering the Weibull distribution, we have the following PDF

$$f_{\mathrm{W},T_{uv}}(t) = \frac{k_{uv}}{\lambda_{uv}}\left(\frac{t}{\lambda_{uv}}\right)^{k_{uv}-1} \exp\left\{-\left(\frac{t}{\lambda_{uv}}\right)^{k_{uv}}\right\} \tag{37}$$

where $k_{uv} > 0$ and $\lambda_{uv} > 0$ are shape and scale parameters, respectively. These two parameters and b represent the strength of social interaction between users u and v. The load on the backhaul link is expressed as

$$\mathcal{L}_B = \sum_{u=1}^{|\mathcal{U}|} S_{f_u} - \sum_{u=1}^{|\mathcal{U}|} \min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\} \tag{38}$$

where D_{uv} is the amount of data transferred from user v to u, $f_u \in \mathcal{F}$ is the file requested by user u, and S_{f_u} is the size of file f_u. The expression of D_{uv} is as follows.

$$D_{uv} = \min\left\{T_{uv}\log\left(1 + \frac{\mathsf{G}_{uv} p_{\max}}{\sigma^2}\right), C_{vf_u} S_{f_u}\right\} \tag{39}$$

where $p_{\max}$ is the maximum D2D transmission power, and to achieve the above transferred data volume, the optimal power allocated to the $\{u, v\}$ link is given by

$$p_{uv} = \min\left\{\frac{\sigma^2}{\mathsf{G}_{uv}}\left(\exp\left(C_{vf_u} S_{f_u}/T_{uv}\right) - 1\right), p_{\max}\right\}. \tag{40}$$

with C_{vf_u} is the amount of file f_u cached by user v. To this end, we define the congestion as an event that the load on the backhaul exceeds a predefined threshold θ_{th}. Let $A_{|\mathcal{U}|}^{|\mathcal{F}|}$ be the total number of combination with repetition (cwr) of $|\mathcal{U}|$ files among $|\mathcal{F}|$ files, and denote $\mathbf{F}'_i$ to be the i-th cwr. Then the congestion probability is given by

$$\begin{aligned}\mathbb{P}\left[\mathcal{L}_B \geq \theta_{th}\right] &= \sum_{i=1}^{A_{|\mathcal{F}|}^{|\mathcal{U}|}} \mathbb{P}\left[\mathbf{F} = \mathbf{F}'_i\right] \\ &\times \mathbb{P}\left[\sum_{u=1}^{|\mathcal{U}|} \min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\} \leq \sum_{u=1}^{|\mathcal{U}|} S_{f_u} - \theta_{\mathrm{th}} \,\middle|\, \mathbf{F} = \mathbf{F}'_i\right]\end{aligned} \tag{41}$$

where $\mathbf{F}$ is the combination of requested files. Then, by leveraging the property of large user number in the cell, the approximation of (41) can be found with the *Central Limit Theorem* (CLT) as follows.

$$\begin{aligned}&\mathbb{P}\left[\sum_{u=1}^{|\mathcal{U}|} \min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\} \leq \sum_{u=1}^{|\mathcal{U}|} S_{f_u} - \theta_{\mathrm{th}} \,\middle|\, \mathbf{F} = \mathbf{F}'_i\right] \\ &\quad = \Phi\left(\frac{\sum_{u=1}^{|\mathcal{U}|} S_{f_u} - \theta_{\mathrm{th}} - \sum_{u=1}^{|\mathcal{U}|} \mathbb{E}\left[\min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\}\right]}{\sqrt{\sum_{u=1}^{|\mathcal{U}|} \mathbb{V}\left[\min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\}\right]}}\right),\end{aligned} \tag{42}$$

Φ is the cumulative distribution function (CDF) of the standard normal distribution. Next, we compute

$$\begin{aligned}&\mathbb{E}\left[\min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\}\right] \\ &\quad = \mathbb{E}_{\leq}\mathbb{P}\left(\sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u}\right) + \mathbb{E}_{>}\mathbb{P}\left(\sum_{v=1}^{|\mathcal{U}|} D_{uv} > S_{f_u}\right)\end{aligned} \tag{43}$$

where

$$\mathbb{E}_{\leq} = \mathbb{E}\left[\min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\} \,\middle|\, \sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u}\right] \tag{44}$$

and

$$\mathbb{E}_{>} = \mathbb{E}\left[\min\left\{\sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u}\right\} \,\middle|\, \sum_{v=1}^{|\mathcal{U}|} D_{uv} > S_{f_u}\right]. \tag{45}$$

Using the CLT, i.e., considering $\sum_{v=1}^{|\mathcal{U}|} D_{uv}$ as a Gaussian random variable with mean μ_u and variance ϑ_u^2, we have

$$\begin{aligned}\mathbb{E}_{\leq} &= \frac{1}{\sqrt{2\vartheta_u^2\pi}} \int_0^{S_{f_u}} e^{-\frac{(x-\mu_u)^2}{2\vartheta_u^2}} x \, dx \\ &= \sqrt{\frac{\vartheta_u^2}{2\pi}} \left[e^{-\frac{{\mu_u}^2}{2{\vartheta_u}^2}} - e^{-\frac{\left(S_{f_u}-\mu_u\right)^2}{2\vartheta_u^2}} \right] \\ &\quad + \mu_u \left[\Phi\left(\frac{S_{f_u} - \mu_u}{\vartheta_u} \right) - \Phi\left(-\frac{\mu_u}{\vartheta_u} \right) \right] \end{aligned} \tag{46}$$

and $\mathbb{E}_{>} = S_{f_u}$. Due to the large number of users, we have the following approximations.

$$\mathbb{P}\left[\sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u} \right] \simeq \Phi\left(\frac{S_{f_u} - \mu_u}{\vartheta_u^2} \right) \tag{47}$$

and

$$\mathbb{P}\left[\sum_{v=1}^{|\mathcal{U}|} D_{uv} > S_{f_u} \right] \simeq 1 - \Phi\left(\frac{S_{f_u} - \mu_u}{\vartheta_u^2} \right). \tag{48}$$

Next, by the law of total variance, we have

$$\begin{aligned} &\mathbb{V}\left[\min\left\{ \sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u} \right\} \right] = \mathbb{P}\left(\sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u} \right) \\ &\times \mathbb{V}\left[\min\left\{ \sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u} \right\} \middle| \sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u} \right] \\ &+ \mathbb{P}\left(\sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u} \right) \mathbb{P}\left(\sum_{v=1}^{|\mathcal{U}|} D_{uv} > S_{f_u} \right) \left(\mathbb{E}_{\leq} - \mathbb{E}_{>} \right)^2. \end{aligned} \tag{49}$$

By definition, we have

$$\begin{aligned} &\mathbb{V}\left[\min\left\{ \sum_{v=1}^{|\mathcal{U}|} D_{uv}, S_{f_u} \right\} \middle| \sum_{v=1}^{|\mathcal{U}|} D_{uv} \leq S_{f_u} \right] \\ &= \frac{1}{\sqrt{2{\vartheta_u}^2\pi}} \int_0^{S_{f_u}} \left(x - \mathbb{E}_{\leq} \right) \exp\left\{ -\frac{(x-\mu_u)^2}{2\vartheta_u^2} \right\} dx \\ &= \left(\mu_u - \mathbb{E}_{\leq} \right)^2 \left[\Phi\left(\frac{S_{f_u} - \mu_u}{\vartheta_u} \right) - \Phi\left(-\frac{\mu_u}{\vartheta_u} \right) \right] \\ &\quad + \frac{\vartheta_u \left(\mu_u - \mathbb{E}_{\leq} \right)}{\sqrt{\pi/2}} \left[\exp\left\{ -\frac{{\mu_u}^2}{2\vartheta_u^2} \right\} - \exp\left\{ -\frac{\left(S_{f_u} - \mu_u \right)^2}{2\vartheta_u^2} \right\} \right] \\ &\quad + \frac{2{\vartheta_u}^2}{\sqrt{\pi}} \left[J\left(\frac{S_{f_u} - \mu_u}{\vartheta_u\sqrt{2}}, 3, -1, 2 \right) - J\left(-\frac{\mu_u}{\vartheta_u\sqrt{2}}, 3, -1, 2 \right) \right] \end{aligned} \tag{50}$$

The J function is given by

$$J(x, y, a, l) = -\frac{1}{(-a)^{\frac{y}{l}} l} \Gamma\left(\frac{y}{l}, -ax^l \right). \tag{51}$$

where $\Gamma(\bullet, \bullet)$ is a two-argument gamma function. Now, we compute the mean μ_u and the variance ϑ_u^2. From the CLT, we have

$$\mu_u = \sum_{v=1}^{|\mathcal{U}|} \mathbb{E}\left[D_{uv} \right] \tag{52}$$

and

$$\vartheta_u^2 = \sum_{v=1}^{|\mathcal{U}|} \mathbb{V}\left[D_{uv} \right] \tag{53}$$

where

$$\mathbb{V}\left[D_{uv} \right] = \mathbb{E}\left[D_{uv}^2 \right] - \left(\mathbb{E}\left[D_{uv} \right] \right)^2. \tag{54}$$

For the ease of presentation, let us denote

$$\mathcal{R}_{uv} = \frac{C_{vf_u} S_{f_u}}{\log\left(1 + \frac{\mathsf{G}_{uv} p_{max}}{\sigma^2} \right)} \tag{55}$$

By expectation definition, $\mathbb{E}\left[D_{uv} \right]$ and $\mathbb{E}\left[D_{uv}^2 \right]$ can be computed explicitly according to Power-Law and Weibull distribution, respectively, as follows.

1) Power-Law distribution

$$\begin{aligned} \mathbb{E}\left[D_{uv} \right] = \log&\left(1 + \frac{\mathsf{G}_{uv} p_{\max}}{\sigma^2} \right) (-2b + 1)^{-1} \\ &\times \left[\left(\mathcal{R}_{uv} + 1 \right)^{-2b} \left(2b\mathcal{R}_{uv} + 1 \right) - 1 \right] \\ &+ C_{vf_u} S_{f_u} \left(\mathcal{R}_{uv} + 1 \right)^{-2b} \end{aligned} \tag{56}$$

and

$$\begin{aligned} \mathbb{E}\left[D_{uv}^2 \right] = \log^2&\left(1 + \frac{\mathsf{G}_{uv} p_{\max}}{\sigma^2} \right) K\left(\mathcal{R}_{uv} + 1 \right) \\ &+ \left(C_{vf_u} S_{f_u} \right)^2 \left(\mathcal{R}_{uv} + 1 \right)^{-2b} \end{aligned} \tag{57}$$

where

$$K(x) = 2b \left(\frac{x^{-2b+2} - 1}{-2b + 2} - 2\frac{x^{-2b+1} - 1}{-2b+1} - \frac{x^{-2b} - 1}{2b} \right). \tag{58}$$

2) Weibull distribution

$$\begin{aligned} \mathbb{E}\left[D_{uv} \right] &= \log\left(1 + \frac{\mathsf{G}_{uv} p_{\max}}{\sigma^2} \right) \\ &\times \left[J\left(\mathcal{R}_{uv}, 1, -\frac{1}{\lambda_{uv}^{k_{uv}}}, k_{uv} \right) - J\left(0, 1, -\frac{1}{\lambda_{uv}^{k_{uv}}}, k_{uv} \right) \right] \end{aligned} \tag{59}$$

and

$$\begin{aligned} \mathbb{E}\left[D_{uv}^2 \right] &= 2\log^2\left(1 + \frac{\mathsf{G}_{uv} p_{\max}}{\sigma^2} \right) \\ &\times \left[J\left(\mathcal{R}_{uv}, 2, -\frac{1}{\lambda_{uv}^{k_{uv}}}, k_{uv} \right) - J\left(0, 2, -\frac{1}{\lambda_{uv}^{k_{uv}}}, k_{uv} \right) \right]. \end{aligned} \tag{60}$$

Finally, the first probability term in (41), due to the independence nature of users' requests, can be obtained as follows.

$$\mathbb{P}\left[\mathbf{F}=\mathbf{F}_i'\right]=\prod_{u=1}^{|\mathcal{U}|}\zeta_{uf_u},\ f_u\in\mathbf{F}_i' \tag{61}$$

where ζ_{uf_u} is the probability that user u will request file f.

In summary, by giving the cache placement at the user side, the backhaul congestion probability can be approximated as in (42) with the expectation and variance of data obtained at a specific user are calculated by (43)-(48) and (49)-(51), respectively, where μ_u and ϑ_u^2 can be obtained from (52)-(60).

V. NUMERIAL EXAMPLES

In this section, we provide some numerical results to show the effectiveness of the proposed algorithms in a D2D caching network. The default values of our simulation parameters are summarized in Table 2 where the contact time T_{uv} is measured in second.

TABLE 2. Simulation parameters.

Parameters	Values	Parameters	Values
θ_{th}	1.5	k_{uv}	2
$\lvert\mathcal{F}\rvert$	5	ϵ_u	1
$q_{s,max}$	1	S_f	1
$q_{r,max}$	1	T_{uv}	$\sim Unif\left(0,10^3\right)$
α	0.01	l_v^u, g_v^u	$\sim Unif\left(0,1\right)$
η	100	ζ_{uf}	$\sim Unif\left(0,1\right)$
ρ	100	G_{uv}	$\sim Exp\left(0.5\right)$
λ_{uv}	2	σ^2	1

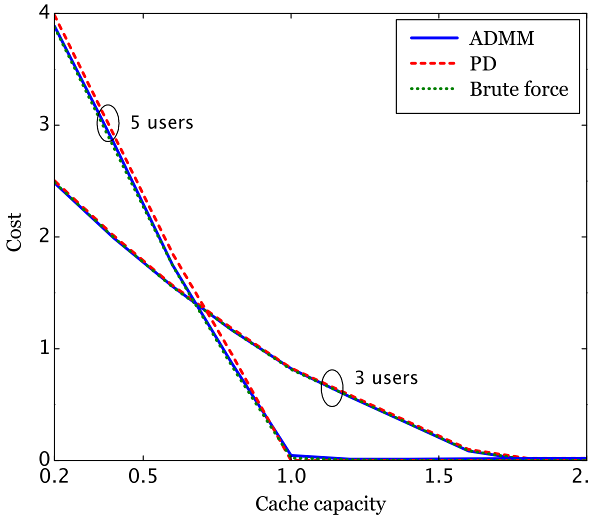


FIGURE 2. The effectiveness comparison between PD and ADMM algorithm in terms of cost minimization in (9) with the brute force serving as a benchmark. The simulation is conducted in two cases of 3 and 5 users.

Fig. 2 illustrates the performance of PD and ADMM algorithms relatively to the brute force result in terms of cost minimization in (9). We can see that both the PD and ADMM algorithms perform as well as the brute force method in this specific case, despite the non-convex objective function. This shows the effectiveness of our proposed algorithms to approach the suboptimal solution in (9).

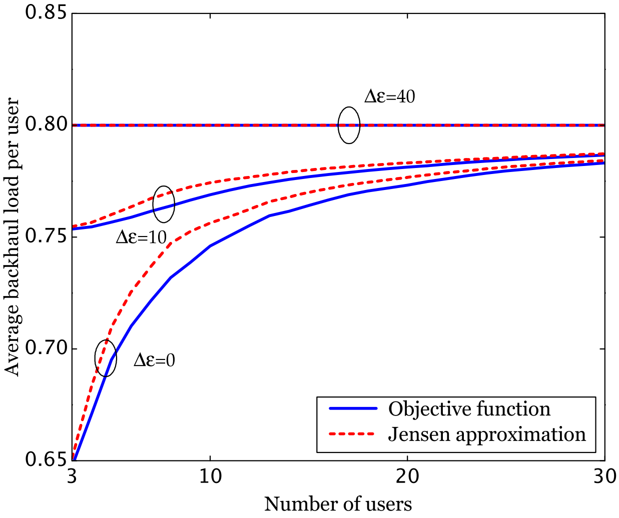


FIGURE 3. Average backhaul load per user as a function of the user number for two cases. The first case is when the average load is calculated by the original objective function (5). The second case is when the Jensen's approximation in (9) is used. Each case is investigated with three levels of power cost, $\Delta\epsilon = 0$ (no power cost), 10 (low power cost), and 40 (high power cost).

Fig. 3 shows the average backhaul load per user as a function of user number for the objective function in (5) and Jensen's approximation in (9). In this example, the average backhaul load per user is defined as the ratio between the expression in (5) and $|\mathcal{U}| \times |\mathcal{F}|$. Similarly for the case of Jensen's approximation which is the ratio between $\mathcal{J}\left(\psi_{uv}\right)$ and $|\mathcal{U}| \times |\mathcal{F}|$. In producing this figure, none of the files are significantly more popular than the others. The cost of power $\epsilon_i, \forall i = 1, \ldots, |\mathcal{U}|$, are initially given randomly small values between 0 and 10^{-2}, then added with $\Delta\epsilon = 0$ (low power cost), 10 (moderate power cost), and 40 (high power cost), respectively. As expected, the Jensen's approximation approaches to the original objective function as the network becomes denser. In particular, when the power is cheap, users can freely exchange data, and hence the best performance is associated with $\Delta\epsilon = 0$ followed by $\Delta\epsilon = 10$, and $\epsilon = 40$, respectively. When the power cost is too high ($\Delta\epsilon = 40$), D2D links are not activated, so users only rely on their own cached contents, and (6) is constant for a given cache placement. Throughout this illustration, it is reasonable to use Jensen's approximation instead of the objective function in (6) for designing suboptimal parameters in a D2D caching network, especially for dense networks.

Fig. 4 presents the backhaul congestion probability as a function of the cache capacity for two cases of 3 and 5 users using equal and optimal power allocations. In the equal-power case, the same power is given to all D2D links, while power is allocated by (40) for the optimal-power case.

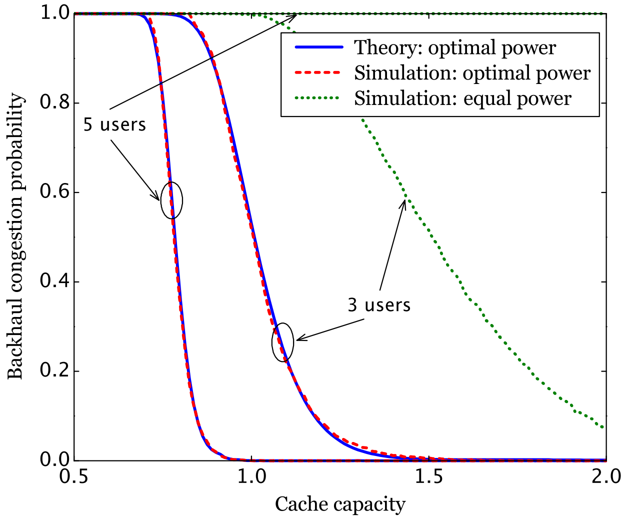


FIGURE 4. The effect from cache capacity enhancement on the backhaul congestion probability for two cases of 3 and 5 users using equal and optimal power allocation.

In this figure, we can observe that simulation results are almost coincide with our analysis. In addition, the equal-power scenario sees a more vulnerable system, since this policy does not exploit the channel conditions and cache states, which limits the amount of shared data within the user interaction time. This can conclude that a good power allocation is essential to enhance the performance of D2D-aided caching networks.

In Fig. 5, the congestion probability is depicted as a function of the cache capacity for (a) 3 users and (b) 5 users using optimal power allocation under low ($\lambda_{uv} = 0.4$), moderate ($\lambda_{uv} = 1$), and high ($\lambda_{uv} = 8$) interaction levels. The results tell us that the interaction growing with the number of users leads to the offloading improvement, evidenced by the outperformance of the denser network (5 users) due to more interactions among users. Obviously, for a given network size, higher interaction level also results in less congestion due to the fact that users have more chance to gather data from the others.

Fig. 6 embodies the backhaul congestion probability as a function of social interaction strength determined by λ_{uv} for two cases of 3 and 5 users using equal and optimal power allocations. In this example, higher λ_{uv} means a higher probability for the long interaction time between users u and v to occur. Generally, the congestion probability significantly decreases when λ_{uv} increases. In the optimal-power case, moreover, when the network becomes denser, the congestion is dramatically reduced due to the benefit from more interactions among users. Nevertheless, when the power is allocated equally, a denser network leads to more congestion due to large interference. Throughout this example, we conclude that it is essential to design an appropriate power allocation to exploit better the social user interaction aspect.

Fig. 7 shows the congestion probability as a function of the cache capacity and congestion threshold θ_{th} for (a) 3 users and (b) 5 users. The free and congestion regions indicate the congestion probabilities of zero and one, respectively. We observe that the congestion threshold θ_{th} significantly affects the congestion probability among networks. For example, when $\theta_{\text{th}} = 3$, the backhaul can handle the maximum load of the 3-user system, hence, no congestion even without caching. Meanwhile small cache capacity can lead to congestion for 5-user network size. However, for $\theta_{\text{th}} < 2$ the advantage of bigger network emerges in terms of facilitating more data exchanging, which allows the 5-user system to see less congestion. This implies that our considered model is suitable to reduce network congestion under limited backhaul capacity condition.

In Fig. 8, we show the backhaul congestion probability as a function of the cache capacity in the ideal case without

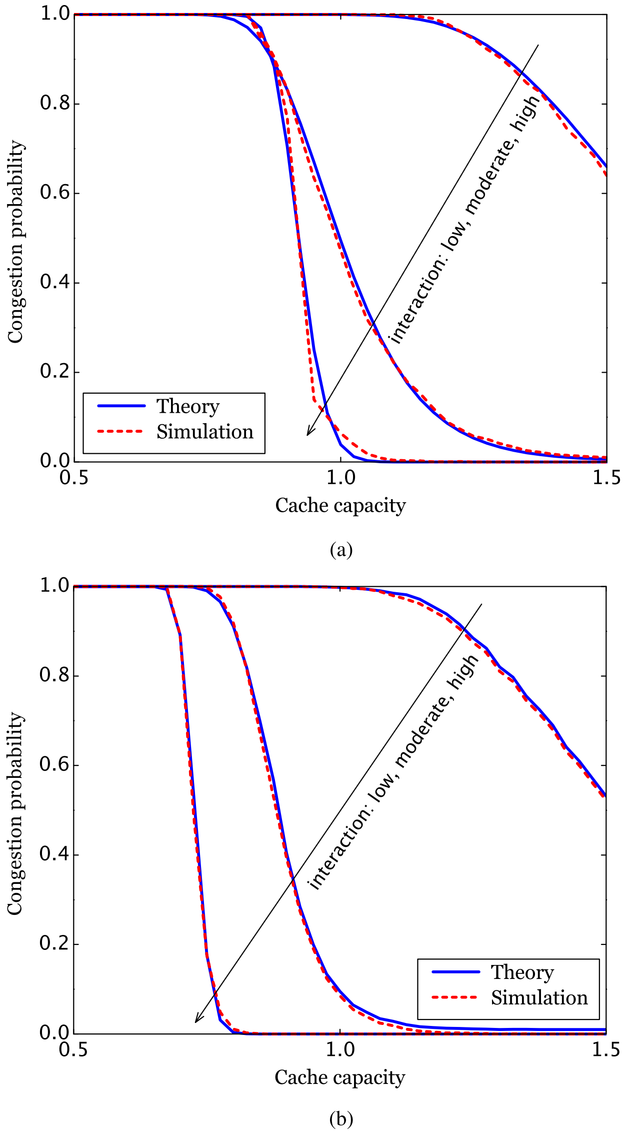


FIGURE 5. Congestion probability as a function of the cache capacity for (a) 3 users and (b) 5 users using optimal power allocation with low ($\lambda_{uv} = 0.4$), moderate ($\lambda_{uv} = 1$), and high ($\lambda_{uv} = 8$) interaction levels.

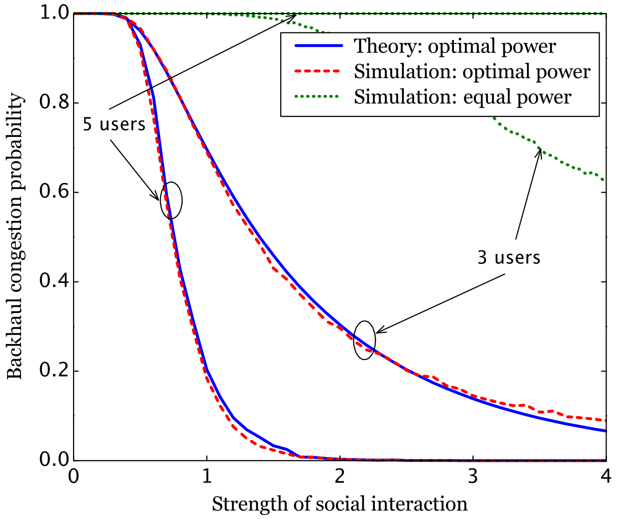


FIGURE 6. Backhaul congestion probability as a function of the social interaction strength λ_{uv} for two cases of 3 and 5 users using equal and optimal power allocation.

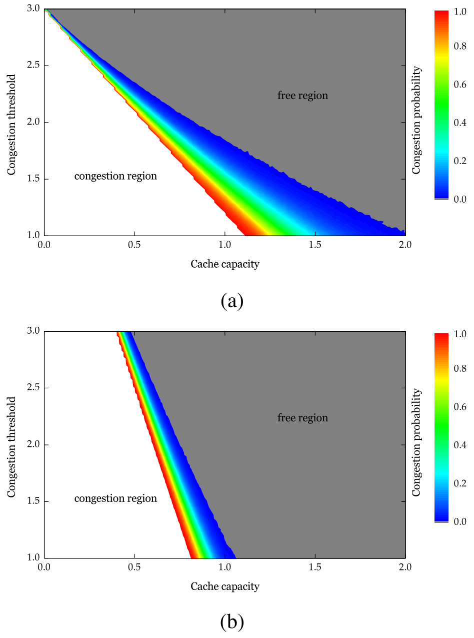


FIGURE 7. The illustration of backhaul load status under the joint influence of cache capacity and congestion threshold θ_{th} for (a) 3−user and (b) 5−user system. The free and congestion regions indicate the congestion probabilities of zero and one, respectively.

interference due to orthogonal channel allocation and the non-ideal case with interference due to resource sharing. High and low user preference levels are considered. In general, the network performance is restricted by the interference and the D2D connection limitation. Therefore, to emerge the effect of interference, we suppress the later negative effect by setting the connection probabilities all to 1 for the non-ideal case. For user preference, the low level corresponds to the requesting probability of 40% for the most favorite file, and the high level corresponds to 90%. Note that the total request probability of a user is 100%. We can see that the gap between cases with and without interference is narrower when users give high request probabilities to their most favorite files. The reason is that when we almost sure which file will be requested by each user, the optimal solution suggests users caching pretty much of those files. Therefore, data exchanging is nearly unnecessary, leading to a reduction in the influence of interference.

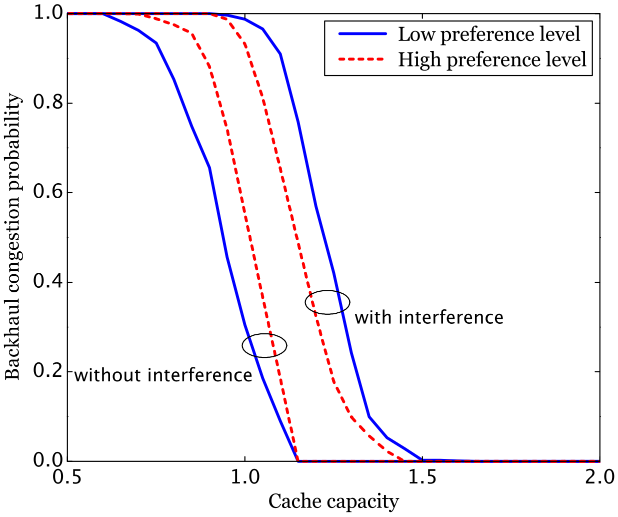


FIGURE 8. The variation of backhaul congestion probability w.r.t. the cache capacity under high and low user preference conditions.

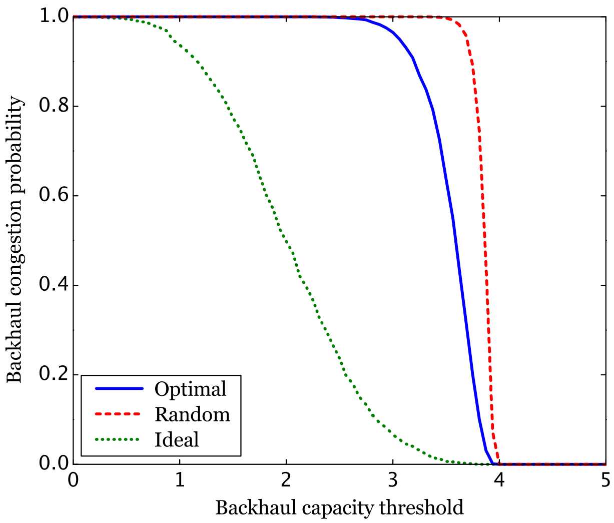


FIGURE 9. Backhaul congestion probability as a function of the backhaul capacity threshold θ_{th} for: Optimal (q_{uv} is optimized by ADMM), random (q_{uv} is randomly selected), and ideal (no interference) scenario. The random case is executed several times and the best result is chosen.

In Fig. 9, we define the congestion event as when the total load at the BS exceeds a threshold which takes values on the horizontal axis. In this experiment, we assume that q_{uv} are defined by the central controller without user involvement.

For ''Random'', q_{uv} are generated randomly for several times and the best result is picked. For ''Ideal'', there is no interference among users and it is added as a lower limit. Finally, for ''Optimal'' the values of q_{uv} are optimized by our method.

VI. CONCLUSIONS

In this work, a socially-aware framework is studied where there are two social aspects including user interaction levels and their willingness to participate in the data exchanging process. The latter factor is influenced by the user's concerns about their limited personal resources and risks of being harmed when connecting to suspiciously dangerous devices. To tackle the issue, we propose a concept of random D2D link activation allowing users to gain more control into the data disseminating process. This plays the role of addressing user incentive and encouraging the data flows between users.

Under this mechanism, we aim to alleviate the backhaul load by finding the joint cache placement and power allocation policy. We then, formulate the problem as a non-convex optimization and propose two optimization schemes which are designed in both centralized and distributed manners. Those methods are built on PD and ADMM algorithms, respectively. Finally, the congestion probability in the special case is derived to provide a performance upper bound to our proposed methods. The derived expression can also evaluate the effectiveness of any other caching schemes that match the considered scenario.

REFERENCES

[1] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, ''Wireless content caching for small cell and D2D networks,'' *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[2] J. Jiang, S. Zhang, B. Li, and B. Li, ''Maximized cellular traffic offloading via device-to-device content sharing,'' *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 82–91, Jan. 2016.

[3] M. Ji, G. Caire, and A. F. Molisch, ''Fundamental limits of distributed caching in D2D wireless networks,'' in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2013, pp. 1–5. doi: 10.1109/ITW.2013.6691247.

[4] 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. 849–869, Aug. 2017.

[5] L. Al-Kanj, H. V. Poor, and Z. Dawy, ''Optimal cellular offloading via device-to-device communication networks with fairness constraints,'' *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4628–4643, Aug. 2014.

[6] Y. Shen, C. Jiang, T. Q. S. Quek, and Y. Ren, ''Device-to-device-assisted communications in cellular networks: An energy efficient approach in downlink video sharing scenario,'' *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1575–1587, Feb. 2016.

[7] 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.

[8] B. Chen, C. Yang, and A. F. Molisch, ''Cache-enabled device-to-device communications: Offloading gain and energy cost,'' *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.

[9] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, ''Caching based socially-aware D2D communications in wireless content delivery networks: A hypergraph framework,'' *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 74–81, Aug. 2016.

[10] K. N. Doan, T. Van Nguyen, T. Q. S. Quek, and H. Shin, ''Content-aware proactive caching for backhaul offloading in cellular network,'' *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3128–3140, May 2018.

[11] A. Liu and V. K. N. Lau, ''Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming,'' *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57–69, Jan. 2015.

[12] 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.

[13] H. J. Kang and C. G. Kang, ''Mobile device-to-device (D2D) content delivery networking: A design and optimization framework,'' *J. Commun. Netw.*, vol. 16, no. 5, pp. 568–577, Oct. 2014.

[14] N. Zhao *et al.*, ''Caching D2D connections in small-cell networks,'' *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12326–12338, Dec. 2018.

[15] Y. Li, Z. Wang, D. Jin, and S. Chen, ''Optimal mobile content downloading in device-to-device communication underlaying cellular networks,'' *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3596–3608, Jul. 2014.

[16] W. Zhao and S. Wang, ''Resource allocation for device-to-device communication underlaying cellular networks: An alternating optimization method,'' *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1398–1401, Aug. 2015.

[17] C. Xu *et al.*, ''Efficiency resource allocation for device-to-device underlay communication systems: A reverse iterative combinatorial auction based approach,'' *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 348–358, Sep. 2013.

[18] H. H. Yang, J. Lee, and T. Q. S. Quek, ''Heterogeneous cellular network with energy harvesting-based D2D communication,'' *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1406–1419, Feb. 2016.

[19] A. H. Sakr and E. Hossain, ''Cognitive and energy harvesting-based D2D communication in cellular networks: Stochastic geometry modeling and analysis,'' *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1867–1880, May 2015.

[20] L. Xu, C. Jiang, Y. Shen, T. Q. S. Quek, Z. Han, and Y. Ren, ''Energy efficient D2D communications: A perspective of mechanism design,'' *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7272–7285, Nov. 2016.

[21] M.-C. Lee and A. F. Molisch, ''Caching policy and cooperation distance design for base station-assisted wireless D2D caching networks: Throughput and energy efficiency optimization and tradeoff,'' *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7500–7514, Nov. 2018.

[22] Y. Wang, J. Wu, and M. Xiao, ''Hierarchical cooperative caching in mobile opportunistic social networks,'' in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 411–416.

[23] K. Zhu, W. Zhi, L. Zhang, X. Chen, and X. Fu, ''Social-aware incentivized caching for D2D communications,'' *IEEE Access*, vol. 4, pp. 7585–7593, 2016.

[24] Z. Zheng, L. Song, Z. Han, G. Y. Li, and H. V. Poor, ''A Stackelberg game approach to proactive caching in large-scale mobile edge networks,'' *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5198–5211, Aug. 2018.

[25] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, ''Pricing and resource allocation via game theory for a small-cell video caching system,'' *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[26] L. Shi, L. Zhao, G. Zheng, Z. Han, and Y. Ye, ''Incentive design for cache-enabled D2D underlaid cellular networks using Stackelberg Game,'' *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 765–779, Jan. 2019.

[27] F. Shen, K. Hamidouche, E. Bastug, and M. Debbah, ''A Stackelberg Game for incentive proactive caching mechanisms in wireless networks,'' in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[28] Z. Chen, Y. Liu, B. Zhuo, and M. Tao, ''Caching incentive design in wireless D2D networks: A Stackelberg game approach,'' in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[29] S. Wang, X. Zhang, L. Wang, J. Yang, and W. Wang, ''Joint design of device to device caching strategy and incentive scheme in mobile edge networks,'' *IET Commun.*, vol. 12, no. 14, pp. 1728–1736, Aug. 2018.

[30] Y. Chen, S. He, F. Hou, Z. Shi, and X. Chen, ''Optimal user-centric relay assisted device-to-device communications: An auction approach,'' *IET Commun.*, vol. 9, no. 3, pp. 386–395, Feb. 2015.

[31] W. Wu, R. T. B. Ma, and J. C. S. Lui, ''Distributed Caching via Rewarding: An Incentive Scheme Design in P2P-VoD Systems,'' *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 3, pp. 612–621, Mar. 2014.

[32] T. Liu, J. Li, F. Shu, H. Guan, S. Yan, and D. N. K. Jayakody, ''On the incentive mechanisms for commercial edge caching in 5G wireless networks,'' *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 72–78, Jun. 2018.

[33] J. Wang, C. Jiang, Z. Bie, T. Q. S. Quek, and Y. Ren, ''Mobile data transactions in device-to-device communication networks: Pricing and auction,'' *IEEE Wireless Commun. Lett.*, vol. 5, no. 3, pp. 300–303, Jun. 2016.

[34] A. Ndikumana, N. H. Tran, T. M. Ho, D. Niyato, Z. Han, and C. S. Hong, ''Joint incentive mechanism for paid content caching and price based cache replacement policy in named data networking,'' *IEEE Access*, vol. 6, pp. 33702–33717, Jun. 2018.

[35] C. Yi, S. 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.

[36] D. J. C. MacKay, "Fountain codes," *IEE Proc. Commun.*, vol. 152, no. 6, pp. 1062–1068, Dec. 2005.

[37] 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.

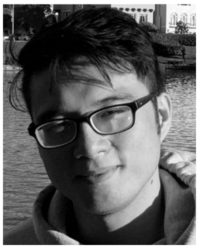
[38] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *Proc. IEEE ISIT*, Istanbul, Turkey, Jul. 2013, pp. 1017–1021.

[39] N. Chaidee and M. Tuntapthai, "Berry-Esseen bounds for random sums of non-IID random variables," *Int. Math. Forum*, vol. 4, nos. 25–28, pp. 1281–1288, 2009.

[40] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1439–1451, Aug. 2006.

[41] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[42] K. Zhao, M. Karsai, and G. Bianconi, "Models, entropy and information of temporal social networks," in *Temporal Networks* (Understanding Complex Systems), P. Holme and J. Saramki, Eds. Berlin, Germany: Springer, 2013, p. 95.



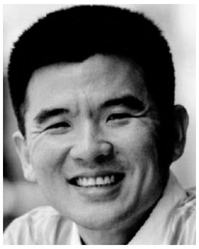
KHAI NGUYEN DOAN (S'18) received the B.E. degree (Hons.) in electronics and telecommunications engineering from the Ho Chi Minh City University of Technology, Vietnam, in 2014. He is currently pursuing the Ph.D. degree with the Singapore University of Technology and Design (SUTD), where he was a Research Assistant, in 2015. His major research interests include network optimization, the application of probability and statistical theories, resource allocation, machine learning, game theory, and signal processing.

THANG VAN NGUYEN (S'09–M'18) received the B.C. degree from the Ho Chi Minh City University of Technology, Vietnam, in 2008, and the M.S. and Ph.D. degrees from Kyung Hee University, Yongin, South Korea, in 2011 and 2014, respectively, all in electrical engineering. He is currently a Postdoctoral Research Fellow with the Singapore University of Technology and Design. His current research interests include wireless communications, compressive sensing, C-RAN, machine learning, physical-layer security, smart grid, localization, and cognitive radio networks.

HYUNDONG SHIN (S'01–M'04–SM'11) received the B.S. degree in electronics engineering from Kyung Hee University, South Korea, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 2001 and 2004, respectively.

During his Ph.D. research at the Massachusetts Institute of Technology (2004–2006), he was with the Wireless Communication and Network Sciences Laboratory, Laboratory for Information Decision Systems. In 2006, he joined Kyung Hee University, where he is currently a Professor with the Department of Electronic Engineering. His research interests include wireless communications and information theory, with the current emphasis on MIMO communication systems, cooperative communication networks, cognitive radio and networks, network interference, vehicular communication networks, physical-layer security, location-aware radios and networks, molecular communications, and quantum information science.

Dr. Shin received the Knowledge Creation Award in the Field of Computer Science from the Korean Ministry of Education, Science and Technology, in 2010, the IEEE Communications Society's Guglielmo Marconi Prize Paper Award, in 2008, and the William R. Bennett Prize Paper Award, in 2012. He has served as the Technical Program Co-Chair for the IEEE WCNC (PHY Track), in 2009, and the IEEE GLOBECOM (the Communication Theory Symposium, in 2012, and the Cognitive Radio and Networks Symposium, in 2016). He was an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2007 to 2012, and the IEEE COMMUNICATIONS LETTERS, from 2013 to 2015.



TONY Q. S. QUEK (S'98–M'08–SM'12–F'18) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology.

He is currently a tenured Associate Professor with the Singapore University of Technology and Design (SUTD). He also serves as the Associate Head of the ISTD Pillar and the Deputy Director of the SUTD-ZJU IDEA. His main research interests include the application of mathematical, optimization, and statistical theories to wireless communication, networking, signal processing, and resource allocation problems. His specific current research topics include network intelligence, wireless security, the Internet-of-Things, and big data processing. He is a coauthor of the book *Small Cell Networks: Deployment, PHY Techniques, and Resource Allocation* (Cambridge University Press, 2013), and the book *Cloud Radio Access Networks: Principles, Technologies, and Applications* (Cambridge University Press, 2017).

Dr. Quek has served as a member of the Technical Program Committee and the Symposium Chair for a number of international conferences. He is currently an Elected Member of the IEEE Signal Processing Society and the SPCOM Technical Committee. He has been actively involved in organizing and chairing sessions. He received the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the IEEE GLOBECOM 2010 Best Paper Award, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards–Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, and the 2017 IEEE ComSoc AP Outstanding Paper Award. He was an Executive Editorial Committee Member of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS, and an Editor of the IEEE WIRELESS COMMUNICATIONS LETTERS. He was the 2017 Clarivate Analytics Highly Cited Researcher. He is a Distinguished Lecturer of the IEEE Communications Society.