

Received August 2, 2020, accepted August 22, 2020, date of publication September 3, 2020, date of current version September 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021462

Collaborative Cache Allocation and Transmission Scheduling for Multi-User in Edge Computing

BOCHENG YU¹, XINGJUN ZHANG¹, (Member, IEEE),
AND ILSUN YOU², (Senior Member, IEEE)

¹School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

²Department of Information Security Engineering, Soonchunhyang University, Asan-si 31538, South Korea

Corresponding author: Ilsun You (ilsunu@gmail.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0200902, and in part by the Soonchunhyang University Research Fund.

ABSTRACT Driven by the ubiquitous smart devices, wireless traffic data has increased significantly and brings a great burden on backhaul links. Edge caching and device-to-device are proved not only can overcome the above challenges, but also to reduce the energy consumption of devices. Meanwhile, the development of hardware makes it possible for terminal devices to cache popular content (referred to as “helpers”). In this paper, a device-to-device assisted edge caching network via helpers is designed to provide low-latency content access and alleviate the traffic load. Then, we formulate the problem of maximization of the successful delivery rate under the constraints of the Signal-to-Interference-plus-Noise Ratio and caching capacity as an integer programming and pose a novel efficient and effective scheme. Based on the algorithm, we first transform the problem into a tractable form. Then this paper presents a hybrid of heuristic approach and iterative selection method to limit the search scope. The simulations demonstrate the method can significantly reduce the system computation time. Moreover, the effects of different scenarios and parameter settings about the successful delivery rate of the Device-to-Device assisted edge caching networks are analyzed.

INDEX TERMS Cover inequality, D2D-assisted edge caching network, edge caching networks, NP-hard, SINR.


I. INTRODUCTION

With the rapid growth of new mobile applications such as live-stream, virtual reality (VR), and augmented reality (AR), video traffic has become the dominant mobile network worldwide. The Cisco Visual Networking Index predicts that mobile video will grow at a Compound Annual Growth Rate (CAGR) of 55 percent between 2017 and 2022 [1]. The increasing growth of traffic and ubiquitous applications present several challenges, such as network congestion occurring during the peak-time period, high-level power consumption and undesirable radio interference [2].

To alleviate the traffic congestion on the backhaul network, the concept of edge caching has been proposed and attracted lots of attention in both academia and industry [3]. Edge caching is an emerging paradigm that stores popular content at the edge like Base stations (BSs) and access points (APs) [4], [5] rather than the centralized cloud, thereby

reducing latency and the burden of the backhaul network [6]. Although edge caching is now widely considered to be an efficient approach for alleviating the heavy network load, it still faces limitations. Due to a surge of traffic loads, the cellular network neither can fulfill all the requests of caching content and lead to lower quality of experience (QoE), especially for high-quality video, nor reduce the spectrum requirement.

Device-to-device (D2D), which supports the direct communications between users bypassing the base stations (BS) [7], is envisaged as another communication paradigm to offload cellular traffic [8]. D2D communications using the same spectral resource with cellular users avoid significant uncontrolled interference which is in the unlicensed band like WiFi and Blue-tooth technologies [9], [10]. Although non-negligible interference is produced, D2D can achieve a high data transmission rate during the traffic peak periods and enhance cellular spectral utilization. Meanwhile, with smart devices growing in popularity, they can not only store content for themselves but also share the data with proximate devices

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

through D2D communications. Therefore, the D2D communication can work as complementary to the edge caching. The combination of D2D and edge caching is introduced as an initiative that efficiently handles the costs of infrastructure and transmission delay. The D2D-assisted edge caching network has the following advantages: i) alleviating the backhaul burden by avoiding download the same content from the core net. The research results show that the active cache can save backhaul overhead by up to 22% [11]; ii) reducing latency by shortening the transmission distance [12].

We model the D2D-assisted edge caching scenario, where we select edge devices, such as smartphones with large storage capacity, to cache popular content. These devices are connected via D2D communication and update cached content during the off-peak time. These edge devices are referred to as helpers to establish an edge caching infrastructure. Generally, the feasible D2D communication distance depends on the interference among D2D pairs and quality of service (QoS), which is defined as the Signal-to-Interference-plus-Noise Ratio (SINR). However, interference in D2D communications is usually more complicated by reusing the cellular resources. Many studies in this field (such as [13]) have considered simplified interference models. Due to the uncertain distribution of helpers and devices, the interference of D2D links is irreducible. We consider more realistic modeling to analyze the effect on interference and popular content distribution. At the same time, since one helper can just set up a D2D link to one device and provide the required content which may be requested by many devices in the D2D-assisted edge caching network, different D2D pairing scheme has a significant impact on content delivery resulting from the interference. Therefore, our main focus is to maximize the Successful Delivery Rate (SDR) under the constraints of channel resources, caching capacity, and SINR. The problem is formulated as Integer Linear Programming which is NP-hard with devices distributed in Euclidean [14]. To the best of our knowledge, the conventional approach of previous literature is hard to solve the global optimum of the Integer Linear problem under SINR constraints. Different from the prior works, we proposed a novel algorithm to computing the global optimal solution for the D2D-link scheme. The model is transformed via more effective inequalities to achieve better performance. The main contributions of this paper are summarized as follows.

- 1) We establish a stochastic edge caching networks model, in which randomly distributed helper device stores the popular content using their own storage resources, such as RAM memory, and sends the requested data via D2D communication to nearby devices. In our model, it also considers the popularity distribution of content, channel fading of D2D-link, network interference, and memory size.
- 2) We first formulate caching optimal problem about maximizing the successful delivery rate of D2D-assisted edge caching networks. End devices can get required contents using a D2D link from a nearby helper who

cached the file instead of getting it from the remote server. We analyzed the SINR and content popularity of the network and achieved the optimal value of SDR through link scheduling and content placement.

- 3) To the best of our knowledge, the previous methods are to solve the integer linear model to obtain global optimum. Due to the complexity of SINR constraints, a low-complex algorithm is designed in this paper. We first reformulate the complicated SINR constraints into more effective inequalities. Then, an iterative scheme is designed to select appropriate D2D-link for content sharing. Meanwhile, we make out the strict inequalities of constraints in each iteration to narrow down the search results. The new approach can obtain global optimum and reduce computation time.
- 4) With numerical results, the proposed solution is superior to conventional methods. Besides, the experiment validated the impact of key parameters and provided indispensable references for the future development of D2D-assisted edge caching networks.

The rest of this paper is structured as follows. In Section II, we discuss the related work; in Section III, the system model is presented and the performance metric is considered; then in Section IV, we formulate the system as an Integer Linear optimization problem; Section V details the proposed solution; simulation results are shown in Section VI. Finally, Section VII concludes the paper.

II. RELATED WORK

In the previous works, to deal with the congestion problem and handle the increase in video traffic. The small cell network, consisting of short-range base stations, such as picocells, microcells, or femtocells, is regarded as complementary to the existing macrocellular network to reduce network congestion [15]–[17]. However, this method faces many challenges: 1) To meet the data rate requirements, a high-speed backhaul is needed to connect the small cell access points (APs) to the core network [18]. In particular, the density of APs is equal to the spatial distribution of users in the current researches [19]. Therefore, it is necessary to deploy expensive backhaul cables such as optical fiber in this solution. 2) Densely deployed small cells and coexistence/interaction with existing macro base station (MBS) infrastructure bring about new problems, such as mutual interference and spectrum resource allocation, which requires additional efficient radio resource allocation and interference management technology [20] and is hard to manage the large scale network [21]. Meanwhile, based on the research, the most cause of network congestion is from repeated downloads of the popular data [4], which indicates that many users only view a small number of video contents that occupy most of the entire data traffic. Because of this, traditional distributed content delivery networks (CDN) transmits duplicate content multiple times, which results in the waste of network resources.

Mobile edge caching and D2D communication are two potential technologies to solve traffic overload problems. In this paper, we consider the integration of D2D into edge caching to overcome the above limitations. The most related literature to our research is categorized as follows.

A. CACHING ON THE EDGE

As edge computing is regarded as an effective approach to traffic offloading, several researchers proposed caching on the edge to reduce latency and backhaul load. In [22] Ahlegh *et al.* study the effect of the edge caching and prove it helps to improve the video capacity and minimize stalling. Kwark *et al.* [23] design a content control plan to obtain as much required content as possible. A performance-guaranteed algorithm is proposed and testified under three specific scenarios. In [24], Chen *et al.* analyze the data cognitive intelligence which is imported to provide personalized and smart service. They put forward an optimal caching strategy for both small-cell and macro-cell and results show its lower latency compared to conventional methods. Błaszczyszyn and Giovanidis [25] maximize the user's hit probability in wireless cellular networks by Poisson point processing. Their optimal policy increases the chances of the hit. Lin *et al.* [26] minimize the cost flow in wireless small cell networks and decrease the computation complexity using a learning approach to solve the problem in polynomial time. However, this solution consumes many spectrum resources and hard to cater to a sudden increase in traffic in peak-time.

B. D2D COMMUNICATIONS

Extensive works are conducted recently to offloading contents via D2D communication. Doppler *et al.* [7] prove that D2D is beneficial to enhance the total throughput and give the advantages of low delay and high transmission rate. Pei *et al.* [27] design a D2D resource allocation in non-orthogonal multiple access (NOMA) cellular network. They offer an iterative scheme to maximize energy efficiency under the Karush-Kuhn-Tucker conditions. Wang *et al.* [28] study resource allocation problem in UAV-assisted networks, in which UAV is regarded as an energy source for D2D devices. They aim to maximize the average throughput while satisfying the energy causality constraint and formulate the problem as mixed-integer nonlinear programming and present the Lagrangian relaxation method to resolve. In [29], Ma *et al.* propose a relay-assisted D2D communication in the millimeter-wave spectrum. They investigate a multi-objective combinatorial optimization problem about the trade-off of the transmit power and system throughput. Liang *et al.* [30] design a resource allocation solution under different QoS requirements for the D2D-enabled vehicular networks framework.

C. D2D ASSISTED EDGE CACHING

The combination of D2D communication and edge caching which is more complex than a single network model [31] is

expected to mitigate the drawbacks: repeated content transmission, high bandwidth consumption, and backhaul congestion. With the rise of the "Internet of Things" (IoT), they demand resources management and coordination [32]. In [33], to overcome the limited resources of servers and lots of services in a smart city, end IoT devices as caching helper is seemed to be a promising scheme for this challenge and enhance the cache-enabled network. Wu *et al.* [34] set up a collaborative content sharing framework based on D2D and caching and propose cache management about caching decision, update mode, and sharing mode selection. In [35] Li *et al.* propose a delay-aware algorithm for latency minimization in D2D assisted edge caching network. To address the drawbacks of the small-cell network, Zhao *et al.* [36] offer a caching scheme and develop three-link scheduling schemes to maximize the throughput. To solve the problem of content transmission, Waqas *et al.* [37] propose caching contents in a helper for content delivery using D2D communication. While the previous works studied the optimization in collaborative edge caching with D2D communication, however, they do not consider the constraints of SINR and the time-consuming optimization algorithm. Thus, we will formulate a caching delivery problem to maximize the SDR. Meanwhile, we develop a novel integer programming algorithm to meet delay-sensitive applications.

III. SYSTEM MODEL

An edge caching network based on Spatial Time Division Multiple Access (STDMA) is considered, in which the helpers are capable of caching popular content via D2D communication, as depicted in Fig. 1. In general, each device can be either a helper or a terminal device. We assume the network consists of $|H|$ helpers chosen from devices and $|U|$ users. The key notations are summarized in Table 1.

A. CACHING MODEL

In the network, we denote the set of cached files as F and the number of file is n , $F = \{f_1, f_2, \dots, f_n\}$. Each file with the same size is L bits. The caching capacities of helper is C bits. The preference content of each user refer to a kind of content $v \in V = \{v_1, v_2, \dots, v_n\}$. The probability that a user u prefers the content type v is $a_u^v \in [0, 1]$, and $\sum_{v=1}^V a_u^v = 1$. In each category of content v , the file f_i means the i -th most popular file according to Zipf distribution. The request probability of file f_i can be given as:

$$q_f^v = \frac{G(f_i, v)^{-\beta}}{\sum_{i=1}^f G(f_i, v)^{-\beta}}, \quad \forall f \in F, \forall v \in V$$

where $G(f, v)$ is the rank of file f for the content v . $\beta \geq 0$ defines the skewed popularity distribution, which means larger β reflects higher content reuse, indicating that few files are requested by most users.

B. CACHING COMMUNICATION MODEL

When a user u requests file f , it first checks its cache whether the content is stored locally. If the file is not in its device,

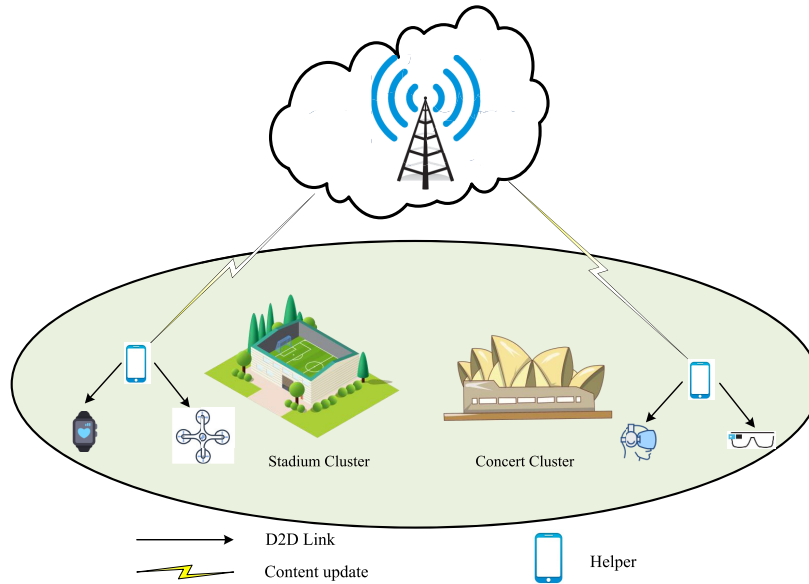


FIGURE 1. D2D-assisted edge caching networks.

TABLE 1. Summary of notation.

Symbol	Meaning
H	Set of helpers in the network
U	Set of end node
F	Set of the requested files
L	Length of the file
C	Helper caching capacity
V	Set of preference content
a_u^v	The probability of content v on user u
f_i	i -th most popular file
β	Zipf exponent
$G(f, v)$	Rank of file f for the content v
$s_{h,f}$	Helper h caches file f
$x_{h,u}$	Helper h send file to user u
q_f^v	The request probability of file f_i
η	Ambient noise
γ	Threshold of SINR
P_h	Transmit power of helper
g_{hu}	channel gain between helper h and user u

the user will send a request to MBS. MBS will allocate the request to the appropriate helper which stores the file. It's assumed that the MBS knows the caching condition of all helpers and the channel state information of the network. Meanwhile, we assume that:

- 1) A file cannot be cached at more than one helper. File stored at helper $h \in H$ is denoted as a binary vector $S = \{s_{h,1}, s_{h,2}, \dots, s_{h,f}\}$, s_{hf} represents whether helper h caches file f ;
- 2) Due to D2D communication between users and helper, a user can communicate with at most one helper at a time. The binary variable $x_{h,u}$ denotes whether helper h send the required content to user u ;
- 3) To ensure successful communication, every D2D-link has to satisfy the SINR threshold. u can receive the file from h only if $SINR \geq \gamma$, γ is SINR threshold.

IV. PROBLEM FORMULATION

In this section, the joint caching placement and D2D link schedule optimization for maximizing SDR is formulated as follows. The conventional approach for global optimum is to solve the following Integer Linear problem.

$$[M1] \quad L^* = \frac{1}{u} \max \sum_{u=1}^U \sum_{f=1}^F \sum_{v=1}^V a_u^v q_f^v x_{h,u} \quad (1)$$

$$s.t. \quad \sum_h s_{h,f} \leq 1, \quad \forall f \in F \quad (2)$$

$$\sum_{f=1}^F s_{hf} \leq \frac{C}{L}, \quad \forall h \in H \quad (3)$$

$$\sum_{u \in U} x_{hu} \leq 1, \quad h \in H \quad (4)$$

$$P_h g_{hu} x_{hu} + M_{hu}(1 - x_{hu}) \geq \gamma \left(\sum_{k \neq h} P_k g_{ku} x_k + \eta \right) \quad (5)$$

$$x_{h,u} \in 0, 1, \quad \forall u \in U, \forall f \in F \quad (6)$$

$$s_{h,f} \in 0, 1, \quad \forall h \in H, \forall f \in F \quad (7)$$

where in (1) a_u^v, q_f^v denote the probability user u request content of type v and the request probability of content f in type v , respectively. Constraints (2) and (3) refers to each helper storage at most one content and the memory limits of helper h . Inequalities (4) state that at most one D2D-links can be activated. This corresponds to the first two constraints in Section 3. Inequalities (5) formulate the SINR requirement with the transmit power P_h . For x_{hu} , the constraint is always met by an adequately large number M_{hu} to satisfy all

x-variables setting equal to one. Solving the integer problem by the conventional approach relies heavily on the bound from the constant relaxation. The big number M , known as a big-M problem in integer programming, greatly weakens the continuous relaxation. Moreover, the values of propagation gain g in (5) vary remarkably in magnitude and cause numerical difficulties in problem solution.

V. PROPOSED SOLUTIONS

A. INEQUALITY TRANSFORM

If D2D-link (h, u) is activated, inequality (5) can be simplified to a knapsack constraint as $\sum_{k \neq h} \varpi_{ku} x_k \leq \varphi_{hu}$, where $\varpi_{ku} = P_k g_{ku}$, $\varphi_{hu} = \frac{P_h g_{hu}}{\gamma} - \eta$. Then we transform the inequality via cover inequality as the constraint of knapsack. In cover inequality, a set C is called a cover if it satisfies $\sum_{k \in C} \varpi_{ku} \leq \varphi_{hu}$. It is impossible to put all the elements of C in the knapsack, at most $|C| - 1$ elements can be selected. Based on cover inequality, we can transform equality (5) into $\sum_{k \in C} x_k \leq |C| - x_{hu}$. In the right-hand side of the reformulated inequality, at most $|C| - 1$ helpers can send content when D2D-link (h, u) is active between user u and helper h . The improved optimization algorithm via cover inequality is as follows:

$$\begin{aligned}
 [M2] \frac{1}{u} \max & \sum_{u=1}^U \sum_{f=1}^F \sum_{v=1}^V a_u^v q_f^v x_{h,u} \\
 \text{s.t.} & (2), (3), (4), (6), (7) \\
 & \sum_{k \in C} x_k \leq |C| - x_{ku}, \\
 & \sum_{k \in C} \varpi_{ku} \leq \varphi_{hu}, \quad h \in H, u \in U \quad (8)
 \end{aligned}$$

In the above algorithm, the reformuated inequality (8) does not contain big-M and gain values in (5) which lead to computational difficulties.

Theorem 5.1: The transformed formulation M2 is correct, that is, its optimal solution is a feasible solution of L^ of M1.*

Proof: Assume the solution violated SINR is infeasible for L^* and Z denotes the set of helpers. The solution has at least one D2D link (h, u) which $\sum_{k \neq h} P_k g_{ku} > \frac{P_h g_{hu}}{\gamma} - \eta$. The $Z \setminus \{i\}$ forms (8). Thus, the optimal solution satisfies L^* . Additionally, since (8) is derived from the SINR constraints, the inequality will include all feasible solutions. Therefore, the reformulation is correct. \square

B. OPTIMIZATION ALGORITHM

It is hard to solve the problem, since it has several constraints, such as equation (8), which increases exponentially with the number of users. Therefore, it is crucial to strengthen the restrictions on the range of D2D-links.

1) SIMPLIFYING THE CONSTRAINTS OF SINGLE OPERATIONS

The problem is solved by an iterative method with simple constraints such as inequality (9). In each cycle, a new constraint inequality will be introduced to test whether it

satisfies the requirement. If the solutions do not satisfy the inequalities, the model will be resolved.

$$\sum_{k \in C} x_k \leq |C| - 1 \quad (9)$$

2) A LINK-SELECTED HEURISTIC ALGORITHM

It can get the upper bound (UB) of the solution unless UB is an optimal solution. A link-selected heuristic algorithm is derived to generate a feasible set of D2D-links from helpers to suitable users. The interference caused by the links in UB is calculated and then removes D2D-link, causing the greatest interference in UB. The algorithm repeat the process until the SINR of link (h, u) , $\gamma_{h,v} \leq \gamma$. The detail of the algorithm is as follow:

Algorithm 1 Link-Select Algorithm

Input: Upper bound (UB) of solution D2D-link (h, u) ; SINR of link (h, u) is $\gamma_{h,u}$

- 1: **while** $\gamma_{h,u} \leq \gamma$ **do**
- 2: Sort the interference caused by the links in UB
- 3: Remove the largest interference element in UB
- 4: Calculate $\gamma_{h,u}$
- 5: **end while**

Output: Lower bound of solution

C. STRICT CONSTRAINTS OF THE D2D-LINK

In the left-side of inequality, it is easy to check whether (\bar{h}, \bar{u}) satisfies the constraint (9) when D2D-link $x_{hu} = 1$. If the SINR is under the threshold, the link x_{hu} will meet the requirement of (9). However, the limitation of (9) is too relaxation, and the solution range is too large. To get a more restrictive inequality, we need to make the minimum number of elements as the minimum cover. Assume $\sum_{k \neq h} \varpi_{ku} x_k > \varphi_{hu}$. The minimum number of elements is the sum of the number of interference D2D-link that just exceed φ_{hu} . To get the minimum number of elements, we first sort of the elements in set $(\bar{h}, \bar{u}) = 1$ in descending order. Second, adding the elements until the sum exceeds φ_{hu} . D denotes the index of these elements. The inequality is acquired as follows:

$$\sum_{k \in D} x_k \leq |D| - x_{hu}$$

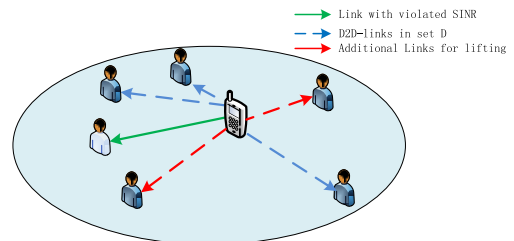


FIGURE 2. Lifting cover inequality.

Lifting in the right-side of inequality: Assume D2D-link (h, u) with $k \neq u$, as depicted in Fig. 2. i) $k \notin D$, If the

link (h, u) can maintain the maximum interference from other D2D-links in the D set which is b and $b \leq |D|$, the restrictive constraint is $\sum_{k \in C} x_k \leq |D| - x_{hu} - \alpha x_{hk}$, where $\alpha = |D| - b$. Specifically, D2D-link x_{hu} and x_{hk} cannot be activated at the same time. If $x_{hu} = 1$, the inequality is $\sum_{k \in C} x_k \leq |D| - x_{hu}$. Otherwise, if $x_{hk} = 1$, there are at most b links in set D which can be activated. To get b , we sort the links in D in ascending order according to the interference intensity of user k and calculate the links in the sorted list until the sum exceeds φ_{hu} .

ii) $K \in D$, the b is exclusive of set D and $\alpha = |D| - 1 - b$. The general form of lifted inequality is as follows:

$$\sum_{k \in D} x_k \leq |D| - x_{hu} - \sum_{k \neq h} \alpha x_{ku} \quad (10)$$

Algorithm 2 Lifting the Cover Inequality

Input: D2D-link (h, u) ; sum of interference; φ ;

- 1: Sort $\varpi_{ku}, k \neq h, k \in H$ in descend order
 - 2: **while** $sum \leq \varphi$ **do**
 - 3: $sum = sum + 1$
 - 4: $D = |sum|$
 - 5: **end while**
 - 6: **if** D2D-link (h, u) with $k \neq u$ **then**
 - 7: $\sum_{k \in D} x_k \leq |D| - x_{hu} - \sum_{k \neq h} \alpha x_{ku}$
 - 8: **if** $k \notin D$, the maximum interference of (h, k) can be suffered is b **then**
 - 9: $\alpha = |D| - b$
 - 10: **else if** $k \in D$ **then**
 - 11: $\alpha = |D| - 1 - b$
 - 12: **end if**
 - 13: **end if**
-

Algorithm 3 Summary of the Optimal Algorithm

- 1: Initialized $results, O$
 - 2: Calculate \bar{h}, \bar{u} under constraints of (1)(2)(3) (4)(6)(7)(9)
 - 3: Set $LB = Link - select(\bar{h}, \bar{u})$; Set $L = LB + 1$
 - 4: **while** result=has feasible solution **do**
 - 5: $O = O \cup Lifting(\bar{h}, \bar{u})$
 - 6: Set($results, \bar{L}, (\bar{h}, \bar{u})$)=Solve((1)(2)(3) (4)(6)(7)(9), O, L)
 - 7: **for** $results = Solutionexistand(\bar{h}, \bar{u})$ is feasible **do**
 - 8: Set $LB = LB + 1$; Set $L = LB + 1$ and Solve
 - 9: **end for**
 - 10: **if** results = solution exist **then**
 - 11: Set $LB = \max(LB, Link-select(\bar{h}, \bar{u}))$ and $UB = \min(UB, \bar{L})$
 - 12: **end if**
 - 13: **end while**
-

D. SUMMARY OF THE OPTIMAL ALGORITHM

The procedure of the optimal algorithm is summarized in Algorithm 3. In the first line, the problem is solved by Solver under the constraints (2), (3) and (9). The upper

bound (UB) of the feasible solution can be obtained since (9) just contains one interference link. The UB can meet the SINR threshold unless UB is an optimal solution. A heuristic named link-select algorithm is derived to generate a feasible link set. The link-select algorithm first calculates the interference caused by links in UB to link (h, u) . And then, the algorithm selects and deletes the link, which causes the largest interference. The heuristic algorithm repeats the process until the SINR of link (h, u) meets the SINR threshold. The Link-select algorithm calculates the lower bound (LB). Meanwhile, the set of lifting inequalities of (10) is initialized which denote by O

The algorithm constantly generates inequality (10) to update set O . The algorithm is completed when the solver finds link (h, u) with L feasible links and cannot find a further solution. Considering the computational efficiency, we do not require each iteration to be optimal. Once link (h, u) meets the SINR threshold, the algorithm update LB and repeats the process of line 8. If link (h, u) cannot satisfy the SINR, the link-select algorithm is employed to renew the LB and UB. And the process is followed by the next loop.

VI. PERFORMANCE EVALUATION

In this section, we assume that the devices are randomly distributed in a circular region with a radius equal to 100 meters as in [38]. Users' request probabilities follows the Zipf distribution with $\beta = [0, 1]$. The device transmission power $P_h = 0.1$ Watts. The gain parameter $g = d^{-\alpha}$, where d is the distance between helper and users, the path-loss exponent α is 3, and the SINR threshold of the channel is 3db. And there are 200 files, each of them with 50 bits in the networks. The caching capacities are 150 bits. We adopt CPLEX integer solver with its default options in implementing both M1 and M2 approaches.

TABLE 2. Time-cost (seconds) of computation.

User	Helper	ILP for M1	ILP for M2	User	Helper	ILP for M1	ILP for M2
40	10	2.6	0.97	60	10	15	9
40	20	5.3	1.3	60	20	29	13
40	30	8.4	2.4	60	30	36	11
50	10	12	5	70	10	42	17
50	20	23	6	70	20	34	10
50	30	17	5	70	30	36	12

In Table 2, we compare the time spending (seconds) for reaching the optimality of the optimal algorithm with the conventional integer-programming algorithm. We apply CPLEX to solve both optimization algorithms under different numbers of helpers. We compute the average time per case. From Table 2, we can see that the computation time based on our optimal algorithm is considerably reduced than the computation time of the conventional optimization algorithm. And as the scale of the instance increases, the computation time of the conventional algorithm increases significantly, while the time increase of the algorithm based on our algorithm is not obvious.

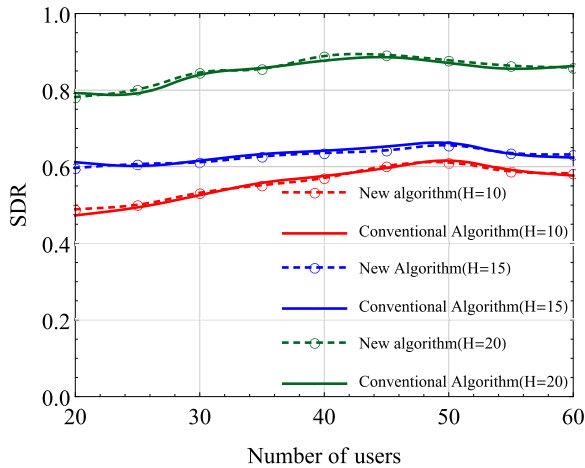


FIGURE 3. Accuracy of new algorithm and influence of number of users and helpers on SDR.

In Fig. 3, the curves are plotted using our algorithm and conventional algorithm (M1) for three cases with a different number of users and helpers. The results of our algorithm have similar accuracy with the conventional algorithm, which shows our approach does not affect the accuracy while reducing computation time (As presented in Table 2). Moreover, for the SDR, when the number of helpers increases, the SDR performance enhances. It means that more helpers are in favor of SDR. However, in the single case, the SDR climb steadily and then slip with the increase of users. This is due to larger interference to each D2D-links. Hence, it is crucial to choose the appropriate number of servable users in terms of helpers.

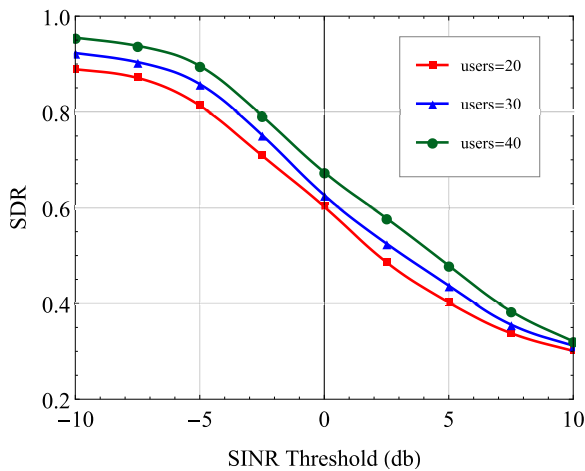


FIGURE 4. Influence of SINR threshold on SDR.

Fig. 4 illustrates the impact of SINR threshold on SDR for users = 20, 30 and 40 with helper =10 respectively. On one hand, with a larger SINR threshold, the SDR decline from about 0.9 to 0.3. The sensitivity of the device to interference becomes higher, due to the increased SINR threshold. The quality of communication from helpers to users become too stringent to establish. On the other hand, it shows more users cause higher SDR.

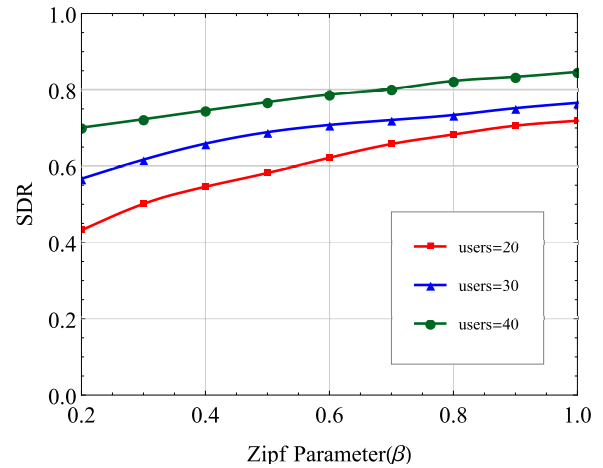


FIGURE 5. Influence of Zipf on SDR.

The SDR of the different number of D2D links is compared in terms of Zipf in Fig. 5. For 10 helpers, the SDR grows with the Zipf parameter increases. This is because that larger β , which means higher content reuse increased the probability of user request content from helpers. Additionally, under the same conditions, more users can bring higher SDR since more user requests are set up.

In the Fig. 6, the impact of the file length L on SDR with 50, 100 and 150 bits are compared. With the larger size of the file, there is a gradual decline in the SDR. Because of the increases in the length of a file, the number of files stored in a single helper gets smaller. it requires high reliable D2D-links to transmit the required content. Besides, the results show a larger number of helper generates higher SDR. It is useful to incentivize more user caching files and share with others to improve SDR.

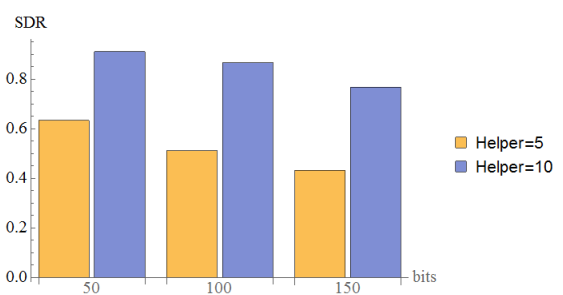


FIGURE 6. Influence of length of file L on SDR.

In the Fig. 7, the chart shows the comparison of the caching capacity of helpers. As the caching capacity of helper increases, the SDR grows steadily. Larger caching size is helpful to store more files, which can satisfy the request from users. Moreover, with the rise of users, the SDR is also improved under the same caching size. This is because of more users' requests are satisfied via the reasonable scheduling algorithm. Therefore, users should be encouraged to participate in enhancing SDR.

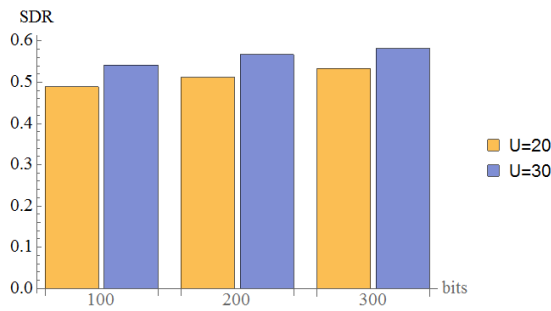


FIGURE 7. Influence of caching capacity of helper on SDR.

VII. CONCLUSION

Inspired by the D2D communication and edge caching network, in this paper we studied the successful data rate in a wireless network consisted of helper node and mobile devices. We aimed to maximize the SDR with considerations of energy, SINR, network interference, and caching capacity. And the problem is derived as an integer programming problem that is known as NP-hard. We designed a new algorithm to improve efficiency using more effective inequalities. Through the simulation, we have proved our proposed algorithm is superior to the conventional algorithm on reducing computation time without significantly affecting the accuracy. We also numerically compared and analyzed the impact of several parameters, like SINR threshold, length of files, Zipf distribution parameter, and caching capacity of helpers, on network performance. For the future work, we aim to extend the system to the multi-cell case and adjustable transmit power.

REFERENCES

- [1] Cisco. *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*. Accessed: Feb. 2019. [Online]. Available: <https://www.cisco.com>
- [2] R. Sun, J. Yuan, I. You, X. Shan, and Y. Ren, "Energy-aware weighted graph based dynamic topology control algorithm," *Simul. Model. Pract. Theory*, vol. 19, no. 8, pp. 1773–1781, Sep. 2011.
- [3] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.
- [4] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.
- [5] H. Kim, J. Park, M. Bennis, S.-L. Kim, and M. Debbah, "Mean-field game theoretic edge caching in ultra-dense networks," 2018, *arXiv:1801.07367*. [Online]. Available: <http://arxiv.org/abs/1801.07367>
- [6] 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.
- [7] K. Doppler, M. Rinne, C. Wijting, C. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.
- [8] N. Cheng, H. Zhou, L. Lei, N. Zhang, Y. Zhou, X. Shen, and F. Bai, "Performance analysis of vehicular device-to-device underlay communication," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5409–5421, Jun. 2017.
- [9] A. Pyattaev, O. Galinina, K. Johansson, A. Surak, R. Florea, S. Andreev, and Y. Koucheryavy, "Network-assisted D2D over WiFi direct," in *Smart Device to Smart Device Communication*. Cham, Switzerland: Springer, 2014, pp. 165–218.
- [10] A. Osseiran, K. Doppler, C. Ribeiro, M. Xiao, M. Skoglund, and J. Manssour, "Advances in device-to-device communications and network coding for IMT-advanced," in *Proc. ICT Mobile Summit*, 2009, pp. 1–8.
- [11] 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.
- [12] B. P. Rimal, D. Pham Van, and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2016, pp. 991–996.
- [13] 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.
- [14] O. Goussevskaia, Y. A. Oswald, and R. Wattenhofer, "Complexity in geometric SINR," in *Proc. 8th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, 2007, pp. 100–109.
- [15] Z. Qu, B. Ye, B. Tang, S. Guo, S. Lu, and W. Zhuang, "Cooperative caching for multiple bitrate videos in small cell edges," *IEEE Trans. Mobile Comput.*, vol. 19, no. 2, pp. 288–299, Feb. 2020.
- [16] X. Zhang, T. Lv, and S. Yang, "Near-optimal layer placement for scalable videos in cache-enabled small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 9047–9051, Sep. 2018.
- [17] M. Bennis, M. Simsek, A. Czyliwicz, W. Saad, S. Valentin, and M. Debbah, "When cellular meets WiFi in wireless small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 44–50, Jun. 2013.
- [18] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.
- [19] F. Rusek, D. Persson, B. Kiong Lau, E. G. Larsson, T. L. Marzetta, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
- [20] W. Cheng, X. Zhang, and H. Zhang, "Statistical-QoS driven energy-efficiency optimization over green 5G mobile wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3092–3107, Dec. 2016.
- [21] V. Sharma, F. Song, I. You, and H.-C. Chao, "Efficient management and fast handovers in software defined wireless networks using UAVs," *IEEE Netw.*, vol. 31, no. 6, pp. 78–85, Nov. 2017.
- [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] J. Kwak, Y. Kim, L. B. Le, and S. Chong, "Hybrid content caching in 5G wireless networks: Cloud versus edge caching," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3030–3045, May 2018.
- [24] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5G networks: Architecture and delay analysis," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 70–75, Feb. 2018.
- [25] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3358–3363.
- [26] P.-Y. Lin, H.-T. Chiu, and R.-H. Gau, "Machine learning-driven optimal proactive edge caching in wireless small cell networks," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, Apr. 2019, pp. 1–6.
- [27] L. Pei, Z. Yang, C. Pan, W. Huang, M. Chen, M. Elksashlan, and A. Nallanathan, "Energy-efficient D2D communications underlying NOMA-based networks with energy harvesting," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 914–917, May 2018.
- [28] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered D2D communication underlying UAV-assisted networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 14–24, Mar. 2018.
- [29] B. Ma, H. Shah-Mansouri, and V. W. S. Wong, "Full-duplex relaying for D2D communication in millimeter wave-based 5G networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4417–4431, Jul. 2018.
- [30] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186–3197, Jul. 2017.
- [31] V. Sharma, I. You, and R. Kumar, "Energy efficient data dissemination in multi-UAV coordinated wireless sensor networks," *Mobile Inf. Syst.*, vol. 2016, pp. 1–13, May 2016.
- [32] V. Sharma, I. You, D. N. K. Jayakody, and M. Atiquzzaman, "Cooperative trust relaying and privacy preservation via edge-crowdsourcing in social Internet of Things," *Future Gener. Comput. Syst.*, vol. 92, pp. 758–776, Mar. 2019.

- [33] L. Zhao, J. Wang, J. Liu, and N. Kato, "Optimal edge resource allocation in IoT-based smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 30–35, Mar. 2019.
- [34] D. Wu, L. Zhou, Y. Cai, and Y. Qian, "Collaborative caching and matching for D2D content sharing," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 43–49, Jun. 2018.
- [35] Y. Li, M. C. Gursoy, and S. Velipasalar, "A delay-aware caching algorithm for wireless D2D caching networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 456–461.
- [36] N. Zhao, X. Liu, Y. Chen, S. Zhang, Z. Li, B. Chen, and M.-S. Alouini, "Caching D2D connections in small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12326–12338, Dec. 2018.
- [37] M. Waqas, M. Zeng, Y. Li, D. Jin, and Z. Han, "Mobility assisted content transmission for Device-to-Device communication underlying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6410–6423, Jul. 2018.
- [38] M. Zhao, X. Gu, D. Wu, and L. Ren, "A two-stages relay selection and resource allocation joint method for d2d communication system," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Apr. 2016, pp. 1–6.



XINGJUN ZHANG (Member, IEEE) received the Ph.D. degree in computer architecture from Xi'an Jiaotong University, China, in 2003. From 1999 to 2005, he was a Lecturer and an Associate Professor with the Department of Computer Science and Technology, Xi'an Jiaotong University. From February 2006 to January 2009, he was a Research Fellow with the Department of Electronic Engineering, Aston University, U.K. From 2009 to 2013, he was an Associate Professor with the Department of Computer Science and Engineering, Xi'an Jiaotong University, where he has been a Full Professor, since 2014. His research interests include high-performance computer architecture, high-performance computing, big data storage systems, and computer networks.



ILSUN YOU (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Dankook University, Seoul, South Korea, in 1997 and 2002, respectively, and the Ph.D. degree from Kyushu University, Japan, in 2012. From 1997 to 2004, he was a Research Engineer with Thin Multimedia Inc., Internet Security Company Ltd., and Hanjo Engineering Company Ltd. He is currently an Associate Professor with the Department of Information Security Engineering, Soonchunhyang University. His current research interests include the Internet security, authentication, access control, and formal security analysis. He is also a Fellow of the IET (based on document published on September 2019). He has been serving as a Main Organizer for international conferences and workshops, such as MobiWorld, MIST, SeCIHD, AsiaARES, and IMIS. He is also the Editor-in-Chief of the *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications* (JoWUA). He is on the Editorial Board of *Intelligent Automation and Soft Computing* (AutoSoft), the *Journal of Network and Computer Applications* (JNCA), the *International Journal of Ad Hoc and Ubiquitous Computing* (IJAHUC), *Computing and Informatics* (CAI), the *Journal of High Speed Networks* (JHSN), and *Security and Communication Networks* (SCN).

...



BOCHENG YU received the M.S. degree in software engineering from the University of Southampton, U.K., in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Xi'an Jiaotong University. His current research interests include mobile edge computing, network optimization, and machine learning.