A Study of Multicast Message Allocation for Content Distribution with Device-to-Device Communications

Jianguo Xie, Wei Song, and Xi Tao
Faculty of Computer Science
University of New Brunswick, Fredericton, Canada
Email: {jxie1, wsong, xtao}@unb.ca

***Abstract*—As an enabling technology for the fifth-generation (5G) wireless networks, device-to-device (D2D) communications can provide many promising applications such as message dissemination and content distribution. In this paper, we study an important problem for D2D-assisted content distribution, which allocates the message requests to be served by the cache devices via D2D multicast. Aiming to minimize the total transmission cost or maximize the gain in cost saving for the base station (BS), this message allocation problem can be formulated from different perspectives, as a weighted set cover problem (WSCP), a hypergraph matching problem, or a multiple-choice knapsack problem (MCKP). Here, we evaluate three approaches for the formulated problems, including a greedy algorithm, a heuristic algorithm based on Lagrangian relaxation, and a fully polynomial-time approximation scheme (FPTAS), respectively. Simulations are conducted to compare the performance in the static and dynamic scenarios in terms of total cost, unit cost, D2D offload ratio, and service latency. The results show that the MCKP based approach outperforms the other two because the approximation guarantee of the FPTAS results in solutions closest to the optimum.**

***Index Terms*—Device-to-device (D2D) communications, content distribution, traffic offloading.**

I. Introduction and Related Work

Nowadays wireless networks are evolving rapidly to accommodate the explosive growth of mobile traffic, smart devices, and assorted emerging applications. In particular, device-to-device (D2D) communications offer an appealing paradigm for the fifth-generation (5G) wireless networks. There are many promising D2D applications such as proximity services, emergency communications and content distribution. With D2D communications, certain devices can serve as the content providers to fulfill the content requests of other devices in close proximity with their cached content. The short transmission range can potentially achieve high data rates and low latency with low energy consumption [1]. This is particularly favourable for video services that demand large bandwidth and real-time delivery. In addition, D2D-assisted content distribution can offload traffic from the cellular network, thus benefiting other non-D2D cellular users as well.

D2D-assisted content distribution can be performed by unicast [2,3] or multicast [4,5]. In [2], Wang *et al.* propose a network formation game in which the users decide their D2D sharing strategies based on their historical records and are guaranteed positive payoffs. In [3], we study the D2D pairing problem that appropriately assigns a device requesting a content file with a nearby device which caches the requested file. For this NP-hard integer linear program (ILP), we propose a channel-aware heuristic algorithm to solve it. In [4], Zhu *et al.* propose a randomized auction mechanism to offload traffic by fulfilling message requests via D2D multicast. A helper device chooses only one message to send to the requesting devices that fall within the transmission range. In [5], we also consider D2D multicast and propose a truthful moneyless mechanism that selects a limited number of messages to broadcast to socially connected users in the vicinity.

As seen above, a key problem in D2D-assisted content distribution is to allocate the transmission messages for cache devices to fulfill the demands of requesting devices. On one hand, it is preferable that as much traffic as possible is relieved from the base station (BS), so that the benefit of D2D communications is exploited to minimize the transmission cost. On the other hand, it is essential to mitigate the interference among D2D transmitters in order to simultaneously satisfy as many requests as possible. Moreover, it is important to limit the amount of resources that each user is willing to devote into D2D content sharing. Though some previous works consider limiting the number of allocated messages to one [4] or more [5], such constraints are quite simplified and neglect the difference of D2D transmission costs due to varying distance and interference. In this paper, we consider a more flexible cost constraint for D2D transmission and study the multicast message allocation problem using the multiple-choice knapsack problem (MCKP) formulation. As demonstrated in the simulation results, the MCKP solution achieves better performance in terms of total cost, unit cost, D2D offload ratio, and average latency.

The remainder of this paper is organized as follows. In Section II, we present the system model of multicast message allocation with D2D communications. Section III introduces different formulations of the research problem and the corresponding solutions. In Section IV, we present simulation

This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada and NSERC Strategic Network FloodNet.

results of different solutions. Section V concludes this paper.

II. System Model

We consider a content distribution scenario within a cell of radius R as depicted in Fig. 1. Here, a set of requesting devices, D, are requesting content items (referred to as a "messages") from a "library" of size m, denoted by M. The requesting devices enter cell coverage following a Poisson process of mean rate λ_r, and the requesting devices are uniformly distributed within the cell. It is further assumed that the staying time of the requesting devices follows an exponential distribution with mean $1/\mu_r$. For the same requesting device, the interval between adjacent requests follow an exponential distribution of mean $1/\lambda_q$. In addition, there are another set of devices, S, $|S| = n$, which enters the BS's coverage according to a Poisson process of mean rate λ_c and stays therein for an exponentially distributed time with mean $1/\mu_c$. Each $s_i \in S$ already possesses a subset of messages $M_i \subseteq M$, $|M_i| = m_i$.

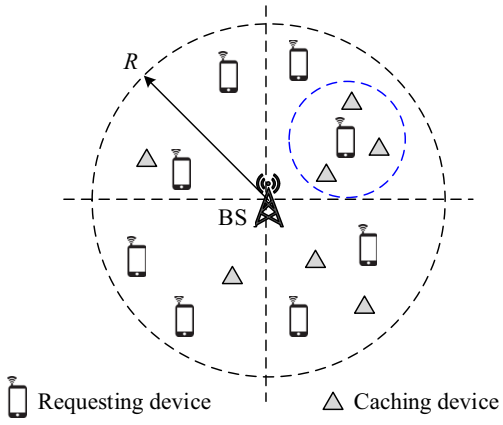


Fig. 1. Content distribution scenario.

If all requests are to be fulfilled by the BS, the BS has to unicast the requested messages and introduce a high cost. Hence, we allow the cache devices to multicast some of their possessed messages to multiple requesting devices that fall within the transmission range, denoted by L. Then, the goal of the multicast scheme is to satisfy as many requests as possible. As shown in Fig. 2, we consider a time-slotted scenario where the content requests are processed periodically. At the end of each period, the requesting and caching information is collected to allocate multicast messages for cache devices in the next period. In case that not all requests are fulfilled in the current round, the remaining requests are served by the BS.

A. Channel Model

Here, we consider a D2D underlaid cellular network, where the D2D links share the uplink spectrum of regular cellular users. There are several good reasons for favoring the use of uplink resources, such that the uplink resources are often less utilized, and the BS is more powerful in interference mitigation [6]. Assume that all potential D2D transmitters of the cache devices in S all share the same cellular uplink channel. This cellular channel is preferably unused or allocated to a cellular

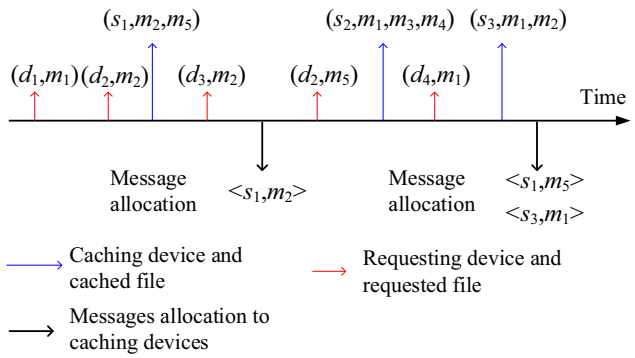


Fig. 2. Periodical message allocation.

user that is uniformly located within the cell. Supposing that the request from device $d_j \in D$ is fulfilled by device $s_i \in S$, the received signal at D2D receiver d_j is written as

$$y_j = \sqrt{P_d}\ell_{i,j}^{-\frac{\varphi}{2}} h_{i,j}x_i + \sqrt{P_c}\ell_{c,j}^{-\frac{\varphi}{2}} h_{c,j}x_c + \sum_{i' \in S, i' \neq i} \theta_{i'}\sqrt{P_d}\ell_{i',j}^{-\frac{\varphi}{2}} h_{i',j}x_{i'} + n_j. \quad (1)$$

Here, φ is the path-loss exponent, n_j is the additive noise at D2D receiver d_j distributed as $\mathcal{CN}(0, \sigma^2)$. Besides, $\theta_{i'}$ is a binary variable indicating whether transmitter $s'_i \in S$ is selected to multicast to certain requesting devices. For D2D transmitter s_i and the cellular user using the same uplink channel, x_i and x_c are their sent signals, respectively, P_d and P_c are their respective transmit power, $\ell_{i,j}$ and $\ell_{c,j}$ are their respective distance to D2D receiver d_j, and $h_{i,j}$ and $h_{c,j}$ are the corresponding distance-independent channel gain that captures the fading effect. Considering Rayleigh fading channels, $|h_{i,j}|^2$ and $|h_{c,j}|^2$ follow an exponential distribution of unit mean. As seen in (1), the received signal at D2D receiver d_j includes the expected signal, the interference from the cellular user, the integrated interference from all other active D2D transmitters in S, and the additive noise. The signal-to-interference-plus-noise ratio (SINR) is given by

$$\xi_j = \frac{P_d\ell_{i,j}^{-\varphi}|h_{i,j}|^2}{P_c\ell_{c,j}^{-\varphi}|h_{c,j}|^2 + \sum_{i' \in S, i' \neq i} \theta_{i'} P_d\ell_{i',j}^{-\varphi}|h_{i',j}|^2 + \sigma^2}. \quad (2)$$

The achievable data rate at receiver device d_j is obtained as

$$r_j = B \cdot \log_2(1 + \xi_j) \quad (3)$$

where B is the carrier bandwidth.

B. Cost Model

Due to the periodical processing nature, we cannot use the instantaneous channel conditions in multicast message allocation. Instead, we use the average channel conditions to estimate resource costs. Accordingly, at the end of each period, the BS allocates the messages that the cache devices need to multicast for surrounding requesting devices.

Given cache device $s_i \in S$ that possesses message $m_k \in M_i$, let D_{ik} denote the set of destination devices that request message m_k and fall within the transmission range of s_i.

Assume that the co-channel interference between D2D and cellular users can be approximated by a Gaussian process similar to additive noise with mean I and N, respectively. As the cost of multicast transmission is constrained by that of the *worst* receiver, we estimate the SINR ξ_j at the worst receiver $d_j \in D_{ik}$ by an exponential distribution with mean

$$\overline{\xi}_j = \frac{P_d}{N+I} l_{i,j}^{-\varphi}. \tag{4}$$

As in [7], a receiver can successfully decode the received message only when the local SINR is not less than a threshold β. Then, the probability that message m_k is received successfully by all requesting devices in D_{ik} is given by

$$P_{ik} = \mathrm{P}\big[\xi_j \geq \beta\big] = e^{-\beta \frac{(N+I)}{P_d} \ell_{i,j}^{\varphi}}. \tag{5}$$

Assume that the resource cost of cache device s_i is proportional to the required SINR to achieve a minimum transmission success probability α, *i.e.*,

$$c_{ik} = z\left(-\frac{\beta \cdot l_{i,j}^{\varphi}}{\ln(\alpha)}\right) \tag{6}$$

where $z(\cdot)$ is a monotonic non-decreasing function which maps the required resource to a comparable cost.

III. Problem Formulations and Solutions

Given each cache device is subject to a multicast cost as described in Section II-B, the cache devices should be allocated multicast messages appropriately to achieve certain design goal. To minimize the total transmission cost, the multicast message allocation problem can be formulated as follows:

$$\text{minimize} \quad \sum_{i=1}^{n'} \sum_{m_k \in M_i} c_{ik} x_{ik} \tag{7a}$$

$$\sum_{i=1}^{n'} x_{ik} t_{ij} \geq 1, \forall r_{jk} = 1, m_k \in M_i, d_j \in D \tag{7b}$$

$$x_{ik} \in \{0,1\}, \forall s_i \in S \cup S_{BS}, \forall m_k \in M_i. \tag{7c}$$

Here, the binary parameter t_{ij} represents whether requesting device d_j falls within the transmission range of s_i, and r_{jk} is also binary indicating whether d_j requests message m_k. The set of requests is denoted by R, in which each element corresponds to a request (j,k) from d_j for message m_k. To ensure that there is always a feasible solution, we consider the BS as virtual cache devices and add a set of $|R|$ duplicated nodes, denoted by S_{BS}, to S. Each virtual device in S_{BS} only serves one request with a cost equal to the BS's unicast cost for the request. Then, the new set of cache devices is denoted by $S' = S \cup S_{BS}$, where $n' = |S'|$. Accordingly, x_{ik} is a binary variable indicating whether cache device $s_i \in S'$ sends message $m_k \in M_i$ to the reachable requesting devices.

In this problem formulation, the requests served by cache device s_i when sending message m_k is a subset of the requests in R. Problem (7) aims to find certain subsets whose union is the universe R and whose total cost is minimal. As seen, this is an NP-hard weighted set cover problem (WSCP). A greedy algorithm [8] can solve this problem with an approximation ratio $[1+\ln(d')]$, where $d' = \max\{|D_{ik}|, \forall s_i \in S', m_k \in M_i\}$ is the maximum cardinality of the subsets. Nonetheless, a main drawback of this solution is that the good-quality cache devices may be allocated multiple messages for multicast. This may quickly deplete the energy of these devices.

In [4], the number of messages that can be multicast by each cache device is simply limited by one. Here, we slightly modify their problem formulation as follows:

$$\text{minimize} \quad \sum_{i=1}^{n'} \sum_{m_k \in M_i} c_{ik} x_{ik} \tag{8a}$$

$$\text{subject to} \sum_{m_k \in M_i} x_{ik} \leq 1, \forall s_i \in S \cup S_{BS} \tag{8b}$$

$$\sum_{i=1}^{n'} x_{ik} t_{ij} \geq 1, \forall r_{jk} = 1, m_k \in M_i, d_j \in D \tag{8c}$$

$$x_{ik} \in \{0,1\}, \forall s_i \in S \cup S_{BS}, \forall m_k \in M_i. \tag{8d}$$

Similar to the formulation in (7), we add a number of duplicated nodes, S_{BS}, as virtual cache devices. As seen, (8b) is the extra constraint that limits the number of multicast messages allocated to each cache device. We note that this problem is a special hypergraph matching problem, *i.e.*, the three-dimensional matching problem, which is NP-hard. It is unknown whether constant-factor approximation algorithms exist. In [4], the authors first use the subgradient method for a Lagrangian dual to obtain a lower bound closest to the optimum and then apply a heuristics method to derive a feasible solution. Unfortunately, the approximation ratio of this approach is uncertain and not guaranteed.

From the perspective of the whole system, it is preferable that the total transmission cost is minimized, while the BS may also aim to maximize its gain over unicast by diverting the requests to cache devices. Hence, we consider a different formulation that focuses on the gain achieved by diverting the requests from the BS to the cache devices. Moreover, our formulation takes into account a more flexible constraint on the cost budget of the cache devices similar to [9]. Then, the message allocation problem can be formulated as

$$\text{maximize} \quad \sum_{i=1}^{n} \sum_{m_k \in M_i'} (c_{Bk_i} - c_{ik}) x_{ik} \tag{9a}$$

$$\text{subject to} \sum_{i=1}^{n} \sum_{m_k \in M_i'} c_{ik} x_{ik} \leq C \tag{9b}$$

$$\sum_{m_k \in M_i'} x_{ik} = 1, \forall s_i \in S \tag{9c}$$

$$x_{ik} \in \{0,1\}, \forall s_i \in S, m_k \in M_i'. \tag{9d}$$

Here, for each cache device $s_i \in S$, we add a dummy message m_d with a zero cost and zero gain to the subset of available messages at s_i, which gives $M_i' = M_i \cup \{m_d\}$ and extends c_{ik} and c_{Bk_i} correspondingly. For simplicity, we use the same notation for c_{ik} and c_{Bk_i} without confusion.

As seen in (9b), the BS limits the total cost for all cache devices allocated with one multicast message by a cost budget C. This is to further address the varying costs of the cache devices for message multicast. In the objective function (9a), c_{Bk_i} is the BS's total cost to satisfy the requests that cache device s_i is able to serve by multicasting message m_k, so $(c_{Bk_i} - c_{ik})$ is the gain by diverting requests from the BS to s_i. Assume that $c_{ik} \leq c_{Bk_i}$ and $0 < c_{ik} \leq C, \forall s_i \in S, m_k \in M_i'$. For reference convenience, we denote $(c_{Bk_i} - c_{ik})$ by a gain parameter g_{ik} and represent the objective function in (9a) by $g(x)$. Hence, this formulation targets at the maximum gain achieved by involving D2D-assisted content distribution. Note that we have equality in constraint (9c), because there always exists a feasible solution that allocates the zero-cost dummy message to a cache device. Then, problem (9) is translated into the multiple-choice knapsack problem (MCKP) [10].

The classes for the MCKP is the cache devices in S. The messages in $M' = \sqcup M_i'$ are the items for allocation, where $\sqcup_i M_i'$ is the *disjoint* union of the subsets of messages of all cache devices, *i.e.*, the union of the elements in each M_i' by retaining the original set membership. Let $\eta = |M'|$ denote the problem size. As the MCKP is NP-hard, we can use a fully polynomial-time approximation scheme (FPTAS) to obtain a suboptimal solution. Here, we consider the FPTAS in [11], which gives a $(1+\epsilon)$-approximation for any arbitrary $\epsilon > 0$ in running time polynomial in both the problem size η and $1/\epsilon$.

The details of this approach are given in Alg. 1. First, the algorithm proposed by Dyer and Zemel in [10] is used to obtain a fractional optimal solution x^* for the linear relaxation of problem (9). Let P^* denote the objective value of x^*, which contains at most two fractional variables. Consider a simple rounding algorithm, which returns an integer solution x_0 that maximizes the gain among three alternative integer solutions [11]: 1) x_{ak_1} that only keeps one fractional variable; 2) x_{ak_2} that keeps the other fractional variable; and 3) $x_{\bar{a}}^*$ that discards both fractional variables from x^*. Here, the first two solutions are feasible since any $c_{ik} \leq C$. The objective value of x_0 is denoted by P_0, which is further used in the scale factor K. A new instance of the MCKP in (9) is generated by scaling down $(c_{Bk_i} - c_{ik})$ in (9a) to $\lfloor \frac{c_{Bk_i} - c_{ik}}{K} \rfloor$ and solved by dynamic programming. Let $F\{i, p\}$ denote the minimum cost of only allocating messages to the first i cache devices so that the total gain would be p. Then, the solution x_s for the scaled problem is obtained so that the corresponding objective value $\hat{P}_s = \max\{p | F(n, p) \leq C\}$. The gain of x_s in the original problem (9) is obtained as P_s. Comparing P_0 and P_s, Alg. 1 returns the more profitable solution between x_0 and x_s.

Algorithm 1: Multicast message allocation based on M-CKP formulation.

Input: S (set of cache devices), D (set of requesting devices), R (set of requests), C (limit of cost), ϵ, $\{c_{Bk_i}, c_{ik} : s_i \in S, m_k \in M_i'\}$
Output: $x = \{x_{ik} : s_i \in S, m_k \in M\}$

begin Obtain first solution x_0 and scale factor K
  Use Dyer-Zemel algorithm for (9) to obtain x^*
  if x^* *has no fractional variables* **then**
    Return $x \leftarrow x^*$
  `// Rounding algorithm`
  $\{x_{ak_1}, x_{ak_2}\} \leftarrow$ two fractional variables of x^*
  $p_1 \leftarrow g(x_{ak_1})$, $p_2 \leftarrow g(x_{ak_2})$, $p_3 \leftarrow g(x_{\bar{a}}^*)$
  $P_0 \leftarrow \max\{p_1, p_2, p_3\}$, $x_0 \leftarrow$ solution to achieve P_0
  $K \leftarrow \frac{\epsilon P_0}{n}$
begin Obtain second candidate solution x_s by solving new MCKP instance scaled by K
  for *each* $g_{ik}, \forall s_i \in S, m_k \in M_i'$ **do**
    $\hat{g}_{ik} = \lfloor \frac{c_{Bk_i} - c_{ik}}{K} \rfloor$
  Use Dyer-Zemel algorithm for scaled MCKP with $\hat{g}_{ik}$ and calculate $\hat{P}_0$ with the rounding algorithm
  for *each* $p \in \{1, ..., 3\hat{P}_0\}$ **do**
    $F(0, p) \leftarrow \infty$, $x_0^p \leftarrow \emptyset$
  for *each* $p \in \{1, ..., 3\hat{P}_0\}$ **do**
    for *each* $i \in \{0, ..., n-1\}$ **do**
      Update $F(i+1, p)$ and corresponding x_{i+1}^p
$\hat{P}_s \leftarrow \max\{p | F(n, p) \leq C\}$, $x_s \leftarrow x_n^{\hat{P}_s}$
$P_s \leftarrow \sum_{s_i \in S, m_k \in M_i'} (c_{Bk_i} - c_{ik}) x_s$
if $P_s > P_0$ **then**
  $x \leftarrow x_s$
else
  $x \leftarrow x_0$
Return x

IV. Simulation Results and Discussions

In this section, we compare the performance of the algorithms that solve the three problems formulated in Section III, including the WSCP in (7), hypergraph matching in (8), and MCKP in (9). We first consider the static scenario, in which all devices are distributed uniformly in the cell, the messages demanded by the requesting devices follow a Zipf distribution with exponent γ_r, and the messages stored at cache devices follow another Zipf distribution with exponent γ_c. Then, we evaluate the performance in a dynamic scenario where devices arrive at and leave the cell according to the system model described in Section II. The main simulation parameters are listed in Table I.

A. Static Scenario

In the static scenario, we assume that all requesting devices in the cell ask for two different messages from the library. Here, two important metrics are evaluated, *i.e.*, the total cost to fulfill all requests and the ratio of requests offloaded to D2D multicast. Fig. 3 presents the results with a varying number of cache devices. The result for each case is the average of 1000 random simulations. As seen in Fig. 3(a), the total costs of three approaches all decrease with more cache devices. This is because more requests can be offloaded to D2D multicast leading to lower total costs. The Lagrangian relaxation based

TABLE I
SIMULATION PARAMETERS.

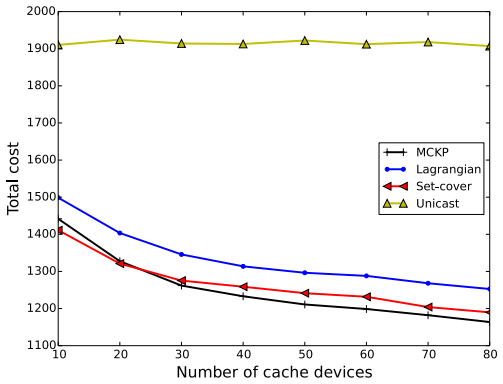
Symbol	Definition	Value
R	Radius of cell coverage (m)	500
L	Maximum D2D transmission range (m)	200
φ	Path-loss exponent	3
P_d	Power of D2D transmitters (mW)	100
P_c	Transmit power of cellular user (mW)	100
P_{BS}	Transmit power of BS (W)	10
B	Carrier bandwidth (MHz)	1
β	Decoding SINR threshold	2
α	Target transmission success possibility	0.8
m	Number of messages	10
m_i	Cache capacity at cache device s_i	2
γ_c	Zipf exponent for cached messages	0.7
γ_r	Zipf exponent for requested messages	0.9
λ_c	Arrival rate of cache devices (per unit time)	1
λ_r	Arrival rate of requesting devices (per unit time)	2
μ_c	Avg. staying time of cache devices (unit time)	30
μ_r	Avg. staying time of requesting devices (unit time)	30
λ_q	Average rate of requests (per unit time)	1
ϵ	Accuracy parameter for MCKP	0.1

approach results in higher costs than the other two approaches, which is partly due to the limit of one multicast message per cache device. Fig. 3(b) shows the ratio of requests offloaded to D2D multicast. As seen, the offload ratio increases with the number of cache devices since more candidates become available to multicast messages. The greedy algorithm for the WSCP achieves the highest offload ratio because there is no limit on the resource cost at each cache device. In contrast, the FPTAS for the MCKP and the Lagrangian relaxation based approach offload fewer requests to accommodate the resource constraints of cache devices.

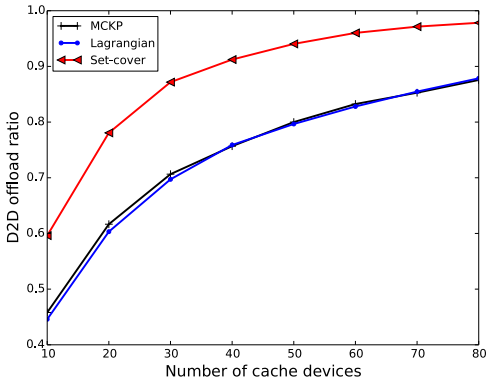
B. Dynamic Scenario

In the dynamic scenario, we consider arrivals and departures of devices and the dynamics of requests. Message allocation is conducted periodically as depicted in Fig. 2. In each simulation, we consider 100 time periods. For each allocation period, unfulfilled requests would be rolled back and satisfied by the BS. Four key performance indicators are considered in the dynamic scenario. In addition to the total cost and D2D offload ratio, we also evaluate the unit cost per message request and the service latency from the moment that a request arrives to the moment that it is successfully served.

Fig. 4 shows the results of the three approaches when the arrival rate of cache devices varies between 0.5 and 1.3. As seen in Fig. 4(a), the total costs of all three approaches decrease with a higher arrival rate of cache devices, which increases the density of cache devices. In addition, it can be observed that the MCKP solution achieves the lowest cost, while the greedy algorithm results in the highest cost. The different costs of the three approaches can be explained by the corresponding offload ratios shown in Fig. 4(b). As seen,



(a)

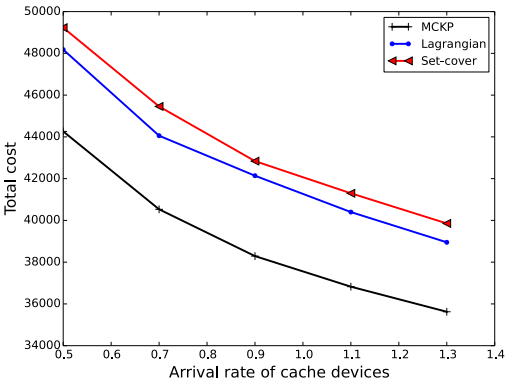


(b)

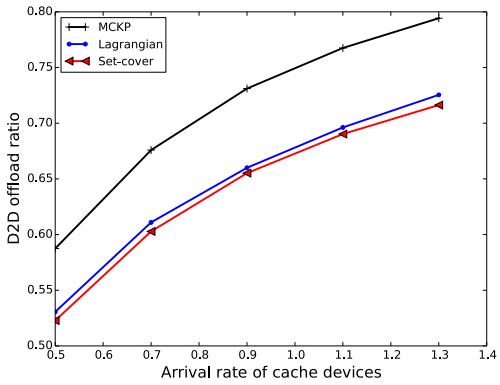
Fig. 3. Results when the number of cache devices varies.

the MCKP solution achieves the highest D2D offload ratio. Although the greedy algorithm allows to allocate multiple messages to a good-quality cache device, the number of allocated messages is still limited by the cache capacity. The Lagrangian relaxation based approach essentially solves the cost minimization problem in a heuristic manner without an approximation guarantee. As a consequence, the message allocation result may not be satisfactory in some cases. On the other hand, the MCKP based approach uses an FPTAS with an approximation guarantee that is made sufficiently close to the optimum by setting the accuracy parameter ϵ to 0.1. Hence, this approach achieves the highest offload ratio and lowest total cost. The consistency of the results in Fig. 4(a) and Fig. 4(b) also imply that almost all requests served by D2D multicast are successfully completed instead of rolling back to the BS. Consequently, Fig. 4(c) shows that the unit cost per served request exhibits a similar trend as the total cost in Fig. 4(a).

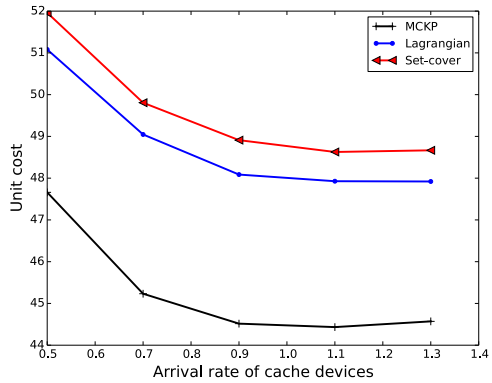
In terms of service latency, Fig. 4(d) shows that the FPTAS for the MCKP and the greedy algorithm for the WSCP achieve better performance than the Lagrangian relaxation based approach. As the resource cost is inversely proportional to mean SINR, a lower cost implies a higher transmission rate on average and thus leads to a lower latency. Due to the heuristic nature of the Lagrangian relaxation based approach, its latency performance is worse than the other two approaches.



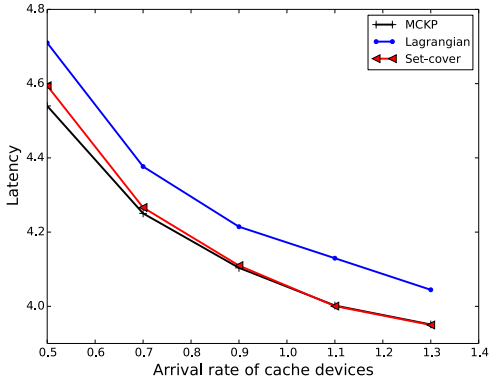
(a)



(b)



(c)



(d)

Fig. 4. Results when the arrival rate of cache devices varies.

V. Conclusion and Future Work

In this paper, we investigate the message allocation problem, which determines the messages multicast by each cache device to serve the content requests via D2D communications. The problem can be solved from different perspectives, aiming to minimize the total transmission cost or maximize the gain in cost saving for the BS. We evaluate three different approaches that formulate the message allocation problem as a weighted set cover problem (WSCP), a hypergraph matching problem, or a multiple-choice knapsack problem (MCKP). The three problems can be solved by a greedy algorithm, a heuristic algorithm based on Lagrangian relaxation, and an FPTAS, respectively. The approximation ratio of the greedy algorithm is limited by the maximum cardinality of the subsets, while the FPTAS for the MCKP can be sufficiently close to the optimum by choosing the accuracy parameter. In contrast, the performance of the heuristic algorithm based on Lagrangian relaxation can vary with the datasets.

We conduct simulations to compare the three approaches in the static and dynamic scenarios. The results show that the MCKP solution outperforms the other two in terms of total cost, unit cost per request, and service latency. This is because the MCKP solution can find better message allocation that diverts more requests to be successfully fulfilled by D2D multicast. In other words, the MCKP solution can achieve a higher offload ratio. In the future, it would be interesting to explore better approximation algorithms for the hypergraph matching problem to improve the heuristic algorithm based on Lagrangian relaxation. It would be favourable if the algorithm can provide approximation guarantee.

References

[1] K. Yang, S. Martin, C. Xing, J. Wu, and R. Fan, "Energy-efficient power control for device-to-device communications," *IEEE J. Select. Areas Commun.*, vol. 34, no. 12, pp. 3208–3220, 2016.

[2] T. Wang, Y. Sun, L. Song, and Z. Han, "Social data offloading in D2D-enhanced cellular networks by network formation games," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 7004–7015, 2015.

[3] J. Xie and W. Song, "Channel-aware device-to-device pairing for collaborative content distribution," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, 2017.

[4] Y. Zhu, J. Jiang, B. Li, and B. Li, "Rado: A randomized auction approach for data offloading via D2D communication," in *Proc. IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2015.

[5] Y. Zhao and W. Song, "Social-aware data dissemination via opportunistic device-to-device communications," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Fall, 2016.

[6] X. Lin, J. Andrews, A. Ghosh, and R. Ratasuk, "An overview of 3GPP device-to-device proximity services," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 40–48, 2014.

[7] W. Song, P. Ju, A. Jin, and Y. Cheng, "Distributed opportunistic two-hop relaying with backoff-based contention among spatially random relays," *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 2023–2036, 2015.

[8] N. E. Young, "Greedy set-cover algorithms," in *Encyclopedia of Algorithms*. Springer, 2008, pp. 379–381.

[9] W. Song and Y. Zhao, "A randomized reverse auction for cost-constrained D2D content distribution," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, 2016.

[10] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack Problems*. Springer, 2004, ch. The multiple-choice knapsack problem, pp. 317–347.

[11] M. Bansal and V. Venkaiah, "Improved fully polynomial time approximation scheme for the 0-1 multiple-choice knapsack problem," International Institute of Information Technology, Tech. Rep., 2004.