Channel-Aware Device-to-Device Pairing for Collaborative Content Distribution

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

***Abstract*—With the increasing penetration of smart devices, device-to-device (D2D) communications offer a promising paradigm to accommodate the ever-growing mobile traffic and unremitting demands. The redundant storage and communication capacities of smart devices can be exploited for collaborative content caching and distribution. In this work, we study the D2D pairing problem, which appropriately pairs a device requesting a content file with a nearby device which caches the requested file. First, we formulate the D2D pairing problem as an integer linear program (ILP). Due to the complexity of the problem, we further develop a heuristic channel-aware D2D pairing algorithm. Computer simulations are conducted to compare the channel-aware algorithm with the optimal solution, as well as a minimum distance-based algorithm and a random algorithm. We consider both the static scenario and the dynamic scenario with arrivals and departures of requesting and caching devices. The simulation results demonstrate that the channel-aware algorithm outperforms the random algorithm in terms of the total number of served D2D pairs and the average latency of served pairs.**

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

I. Introduction and Related Work

Mobile video traffic is expected to account for over 75 percent of global data traffic in mobile networks by 2020 [1]. Many mobile videos feature short durations and may quickly spread with the aid of social media. Traditionally, content providers often deliver their content via a content delivery network (CDN), which is a globally distributed network of proxy servers deployed in multiple data centers. The pervasive smart devices introduce abundant resources at the network edge, which can be exploited to complement traditional CDNs with lower delivery cost and higher performance. In addition, device-to-device (D2D) communications can facilitate such collaborative content distribution, while offering various benefits such as expanding coverage, offloading traffic, and improving energy efficiency. Particularly, D2D communications can assist with message dissemination in a disaster scenario where the network infrastructure is damaged.

In the literature, there have been some studies on content distribution over D2D communications. In [2], content distribution over Wi-Fi cooperation is compared with content distribution over D2D-based multicast. The results show that D2D multicast excels in terms of delivery time and energy consumption. In [3], Chen *et al.* consider a cache-enabled D2D content distribution scenario, where a helper device can send a file to a requesting device within a collaboration distance. A probabilistic caching policy is investigated so that each device caches a file according to a probability distribution. In [4], the local caching problem is formulated for D2D content distribution as a Knapsack problem to address the limited storage capacities of devices. In addition, they explore the sender-receiver pairing problem as a maximum weighted bipartite matching problem. In [5], we propose a three-phase approach that leverages mobile social networks to share content via D2D communications. The results demonstrate the merits of the approach in terms of user utility and completion time.

In this paper, we study the D2D pairing problem that matches a requesting device with a caching device which can deliver the requested content via D2D communications. We consider both the static scenario and the dynamic scenario. An integer linear problem (ILP) is formulated for the static scenario. A channel-aware heuristic algorithm is proposed to address the complexity of the ILP. Particularly for the dynamic scenario, we consider a waiting queue to keep the content requests unfulfilled by D2D communications so as to maximize the offloading traffic from the cellular network. The simulation results demonstrate that the channel-aware algorithm achieves good performance regarding the number of successful pairs and the average latency.

The remainder of this paper is organized as follows. In Section II, we give the system model for content distribution over D2D communications and introduce the research problem. Section III we formulate the D2D pairing problem as an ILP and propose a channel-aware heuristic algorithm. In Section IV, simulation results are presented to compare different pairing algorithms. Section V concludes this paper.

II. System Model and Research Problem

A. Content Distribution Model

In this paper, we consider a content distribution scenario depicted in Fig. 1. A set of requesting devices, D, are requesting content items (referred to as a "file" henceforth for simplicity) from a "library" of size m, denoted by M. The circular disk $\mathcal{C}$ denotes the coverage region of a base station (BS) centered at the origin. The requesting devices

This research was supported by Natural Sciences and Engineering Research Council (NSERC) of Canada, NSERC Strategic Network FloodNet, and Harrison McCain Foundation.

enter region $\mathcal{C}$ following a Poisson process of mean rate λ_r, and the requesting devices are uniformly distributed within $\mathcal{C}$. It is further assumed that the dwelling time of the requesting devices follows an exponential distribution with an average $1/\mu_r$. Each time a device $k \in D$ only requests one file from the library M independently according to a Zipf popularity distribution, which has been shown to be a good model that captures the popularity of video clips [6]. Specifically, device k requests file i with a probability

$$q_i = \frac{\frac{1}{i^{\gamma_r}}}{\sum_{j=1}^{m} \frac{1}{j^{\gamma_r}}}, \qquad 1 \le i \le m, \gamma_r > 0$$

where the exponent γ_r characterizes the relative popularity of files. A larger value of γ_r implies that more requests are concentrated on fewer files. Let m_k^{r} denote the file requested by device k. For the same requesting device, the idling time between a completed request and the next one follow an exponential distribution of mean $1/\lambda_q$. The set of requests from the devices in D is denoted by Q.

In addition, there are another set of devices, S, which enters the BS's coverage region $\mathcal{C}$ according to a Poisson process of mean rate λ_c and stays therein for an exponentially distributed dwell time with mean $1/\mu_c$. Considering the random caching policy studied in [7], we assume that each device $j \in S$ caches one file in the library M according to a Zipf distribution with exponent γ_c. That is, device j caches file i with a probability

$$f_i = \frac{\frac{1}{i^{\gamma_c}}}{\sum_{j=1}^{m} \frac{1}{j^{\gamma_c}}}, \qquad 1 \le i \le m, \gamma_c > 0.$$

Here, we denote the file cached at device j by m_j^{t}. The caching devices can be pooled together and form the fog nodes in fog computing. Then, instead of fulfilling each content request by the BS, it is potentially beneficial to serve some requesting devices in D via D2D communications with nearby caching devices in S. Considering that each device is equipped with a single omnidirectional antenna, each caching device can serve at most one requesting device. This D2D content distribution can not only offload traffic from the BS but also is more cost-effective in view of the close proximity.

B. 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 [8]. Assume that all potential D2D transmitters of the caching devices in S all share the same cellular uplink channel. This cellular channel is preferably unused, or it is allocated to a cellular user that is uniformly located within the coverage region of the cell, $\mathcal{C}$. Supposing that the request from device $k \in D$ is fulfilled by device $j \in S$, the received signal at D2D receiver k is written as

$$\begin{aligned} y_k =& \sqrt{P_d} d_{j,k}^{-\frac{\alpha}{2}} h_{j,k} x_k + \sqrt{P_c} d_{c,k}^{-\frac{\alpha}{2}} h_{c,k} x_c \\ &+ \sum_{j' \in S, j' \neq j} \theta_{j'} \sqrt{P_d} d_{j',k}^{-\frac{\alpha}{2}} h_{j',k} x_{\varphi(j')} + n_k. \end{aligned} \quad (1)$$

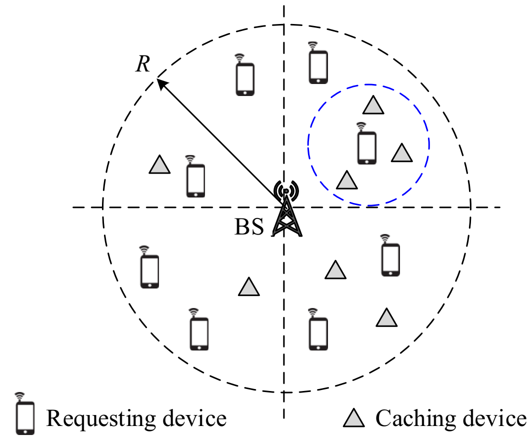

Fig. 1. D2D content distribution scenario..

Here, α is the path-loss exponent, n_k is the additive noise at D2D receiver k distributed as $\mathcal{CN}(0, \sigma^2)$. Besides, $\theta_{j'}$ is a binary variable indicating whether transmitter $j' \in S$ is selected to serve a requesting device. For function $\varphi : S \mapsto D$, $\varphi(j')$ gives the receiver device in D that transmitter j' is matched to if there is one. For D2D transmitter j and the cellular user using the same uplink channel, x_k and x_c are their sent signals, respectively, P_d and P_c are their respective transmit power, $d_{j,k}$ and $d_{c,k}$ are their respective distance to D2D receiver k, and $h_{j,k}$ and $h_{c,k}$ are the corresponding distance-independent channel gain that captures the fading effect. Considering Rayleigh fading channels, $|h_{j,k}|^2$ and $|h_{c,k}|^2$ follow an exponential distribution of unit mean. As seen in (1), the received signal at D2D receiver k 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) at D2D receiver k is then given by

$$\xi_k = \frac{P_d d_{j,k}^{-\alpha} |h_{j,k}|^2}{P_c d_{c,k}^{-\alpha} |h_{c,k}|^2 + \sum_{j' \in S, j' \neq j} \theta_{j'} P_d d_{j',k}^{-\alpha} |h_{j',k}|^2 + \sigma^2}. \quad (2)$$

Then, the achievable data rate at receiver device k can be obtained as

$$r_k = B \cdot \log_2(1 + \xi_k) \quad (3)$$

where B is the carrier bandwidth. Given a minimum rate requirement $r_{\min}$, it implies that the received SINR cannot be less than a *decoding threshold* β, where $r_{\min} = B \cdot \log_2(1+\beta)$.

C. Research Problem

To fulfill the goal of offloading content distribution traffic from the cellular network via D2D links, we need to appropriately pair each requesting device with a caching device. On one hand, the caching device should satisfy the request with a transmission rate no less than $r_{\min}$ to avoid unacceptable latency. On the other hand, we must limit the D2D interference among the D2D links so that more D2D traffic can be accommodated. The pairing of requesting and caching devices is the research focus of this work.

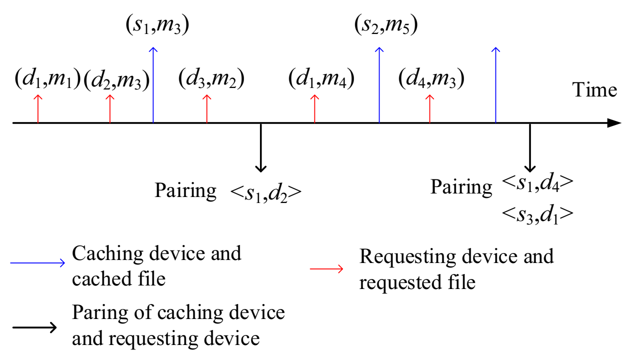

Fig. 2. Periodical processing of D2D pairing.

Specifically, we consider a time-slotted scenario where the content requests are processed periodically, as depicted in Fig. 2. At the beginning of each period, the requesting and caching information is collected to conduct a round of D2D pairing. It is possible that not all requests are fulfilled in the current round. Hence, the unpaired requesting devices will be placed in a waiting queue to be considered for the next round. As seen, the pairing result does not only affect the offloading efficiency in the current round but also influence the pairing subsequently. Hence, we need to consider the queueing mechanism that deals with the backlogged requests.

III. D2D Paring Problem Analysis and Solution

Given a particular pairing period in the content distribution scenario, we have set D for requesting devices with the corresponding content requests, and set S for caching devices with their cached files. Considering a requesting device is only paired with a feasible caching device, we model the feasibility relationship between S and D by a bipartite graph, for which Fig. 3(a) shows an example. An edge in Fig. 3(a) implies that a caching device contains the requested file of the corresponding receiver and satisfies the minimum data rate. In other words, the existence of edges only indicates the feasibility condition. If we want to relate a matching of the bipartite graph to the D2D pairing problem, we should further specify the edge weights to capture the D2D interference and SINR. When an edge is included in a matching for D2D pairing, the interference to be experienced by other potential pairs will be affected accordingly. As a result, the edge weights are not independent as in regular bipartite graphs but inter-dependent in our case. Hence, we cannot obtain an optimal D2D pairing by finding a maximum or minimum weighted bipartite matching, which is solvable in polynomial time.

To incorporate the interference effect, we modify the bipartite graph by replacing the vertices on the left side with the edges. For example, Fig. 3(a) is restructured to Fig. 3(b). Based on the modified bipartite graph, the D2D pairing problem can be reformulated in detail. Referring to the left-side vertices as set G, we use a binary variable $x_{\ell k}$ to indicate whether edge (ℓ, k), $\ell \in G$ and $k \in D$, is selected for the D2D pairing. For each edge (ℓ, k), we define two related vector variables, $p_\ell = \{p_{\ell,k'} : k' \in D\}$ and $w_\ell = \{w_{\ell,k'} : k' \in D\}$, to represent the desired signals and unwanted corresponding interferences with respect to each potential receiver in D, respectively. For instance, selecting edge $(3, a)$ in Fig. 3(b) indicates that transmitter t_2 is paired to receiver a. Accordingly, we have $p_1 = \{s_{t_2,a}, 0, 0\}$ and $w_1 = \{0, s_{t_2,b}, s_{t_2,c}\}$, where $s_{t_2,k} = P_d d_{t_2,k}^{-\alpha} |h_{t_2,k}|^2, k \in \{a, b, c\}$ gives the received power at device k due to the transmission by device t_2. In addition, let $w_{c,k}$ denote the interference from the cellular user to receiver k. Then, an optimal D2D pairing means a maximum one-to-one matching that satisfies certain additional constraints. Specifically, the D2D pairing problem can be written as

$$\max. \quad \sum_{k\in D}\sum_{l\in G} x_{\ell,k} \tag{4a}$$

$$\text{s.t.} \quad \sum_{k\in D} x_{\ell,k} \le 1, \forall \ell \in G \tag{4b}$$

$$\sum_{\ell\in G} x_{\ell,k} \le 1, \forall k \in D \tag{4c}$$

$$\frac{\sum_{\ell\in G} x_{\ell,k} p_{\ell,k}}{\sum_{\ell\in G}\left(\sum_{k'\in D} x_{\ell,k'}\right) w_{\ell,k} + w_{c,k} + \sigma^2} \ge \beta, \tag{4d}$$

$$\forall \sum_{\ell\in G} x_{\ell,k} = 1, k \in D.$$

Here, constraint (4d) ensures that a selected pairing for receiver k satisfies the decoding condition for successful transmission. Overall, problem (4) aims to maximize the total number of such successful pairs. The ratio of the objective value over the number of requesting devices (denoted by $N_r = |D|$) is actually the *satisfaction ratio* for the current period.

As constraint (4d) is defined only for the receivers that are successfully paired. This condition complicates the solution since it couples the problem with the potential solution. To remove this condition, we add N_r virtual transmitters, where each is dedicated to serve one distinct receiver in D. This extends the bipartite graph in Fig. 3(b) to Fig. 3(c), in which the set of vertices on the left side becomes $G' = G \cup H$ and H denotes N_r edges between the virtual transmitters and all receivers. For any edge $(\ell, k) \in H$ between a virtual transmitter j and receiver k, we have $w_\ell = \{0, ..., 0\}$ and $p_\ell = \{0, ..., s_{j,k}, ..., 0\}$, where $s_{j,k} = \beta \cdot (\sum_{\ell\in G} w_{\ell,k} + w_{c,k} + \sigma^2)$. This implies that a virtual transmitter causes zero interference to others, and the received power at its dedicated receiver is always high enough to achieve an SINR no less than the decoding threshold even if all potential real transmitters in S are selected for transmission. As such, the limiting condition in constraint (4d) can be relaxed so that (4d) is applicable to every potential receiver. This extension also ensures that there always exists a feasible solution that successfully pairs every receiver in D, which is to match every virtual transmitter to its dedicated receiver.

Correspondingly, the objective value for problem (4) becomes N_r. To prioritize the pairing of receivers with real feasible transmitters, we define a cost variable c_ℓ for each $(\ell, k) \in G'$ such that $c_\ell = \{c_{\ell,k'} : k' \in D\}$, where $c_{\ell,k'} = 0$ if $k' = k$ and $\ell \in G$, and $c_{\ell,k'} = \infty$ otherwise. Accordingly, problem (4) can be revised to (5). As seen in (5), the objective function (5a), constraint (5c), and constraint (5d) extend the corresponding objective and constraints in problem (4).

$$\min. \quad \sum_{k\in D}\sum_{l\in G'} x_{\ell,k}c_{\ell,k} \tag{5a}$$

$$\text{s.t.} \quad \sum_{k\in D} x_{\ell,k} \le 1, \forall \ell \in G' \tag{5b}$$

$$\sum_{\ell\in G'} x_{\ell,k} = 1, \forall k \in D \tag{5c}$$

$$\frac{\sum_{\ell\in G'} x_{\ell,k}p_{\ell,k}}{\sum_{\ell\in G'}\left(\sum_{k'\in D} x_{\ell,k'}\right)w_{\ell,k} + w_{c,k} + \sigma^2} \ge \beta, \forall k \in D. \tag{5d}$$

Defining the inner summation in the denominator of (5d) as an additional variable y_ℓ, we can reformulate (5) as an integer linear program (ILP) in (6):

$$\min. \quad \sum_{k\in D}\sum_{l\in G'} x_{\ell,k}c_{\ell,k} \tag{6a}$$

$$\text{s.t.} \quad y_\ell \le 1, \forall \ell \in G' \tag{6b}$$

$$y_\ell = \sum_{k\in D} x_{\ell,k}, \forall \ell \in G' \tag{6c}$$

$$\sum_{\ell\in G'} x_{\ell,k} = 1, \forall k \in D \tag{6d}$$

$$\sum_{\ell\in G'}\left(y_\ell \beta w_{\ell,k} - x_{\ell,k}p_{\ell,k}\right) \le -\beta(w_{c,k} + \sigma^2), \forall k \in D. \tag{6e}$$

Even with the reformulation in problem (6), it is hard to obtain the optimal solution for each pairing period. Hence, we consider a heuristic channel-aware D2D pairing algorithm in Alg. 1, which is executed periodically as shown in Fig. 2 with the assistance of a central controller, *e.g.*, the BS. Intuitively, the distance and channel state of each potential pair of D2D devices are two important factors to be considered in the pairing. In addition, the arrival time of each content request also needs to be taken into account since the queueing order of requests affects the fulfilling latency. As seen in Alg. 1, we consider last-come first-served (LCFS) ordering. First, the requesting device with the most recently arrived request is chosen. Then, the algorithm identifies the set of candidate caching devices, which store the requested file and meet the minimum data rate requirement $r_{\min}$. After that, each candidate caching device is checked to see whether adding the new D2D pair will severely interfere with the existing D2D transmission such that some D2D pair may fail to satisfy the rate requirement $r_{\min}$. Finally, the current requesting device is paired to the candidate caching device which achieves the highest data rate and does not cause the above negative impact to existing D2D transmission.

IV. Simulation Results

In this section, we present simulation results to evaluate the performance of different algorithms for the D2D pairing problem. Here, we focus on three metrics, *i.e.*, total number of served D2D pairs and their average latency and distance. In

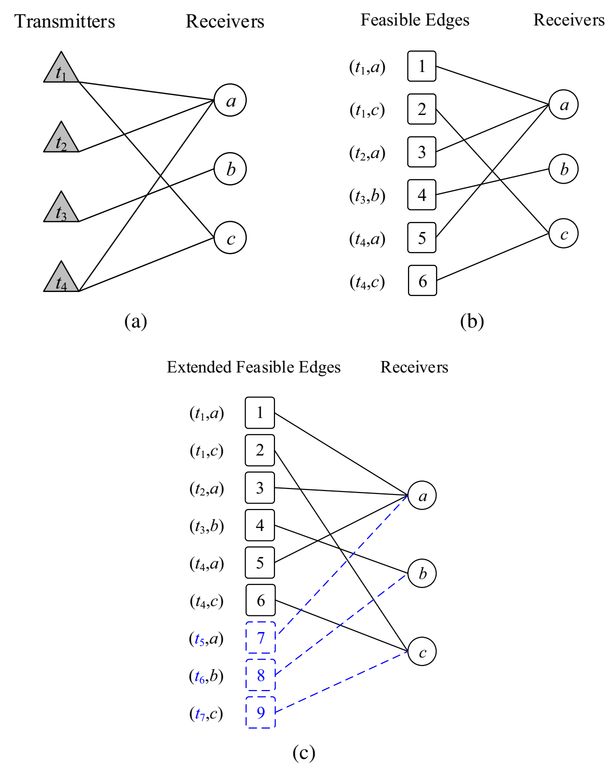

Fig. 3. Bipartite graph modeling for D2D pairing.

addition to the heuristic channel-aware algorithm in Alg. 1, we also consider the optimal pairing solution, a random pairing algorithm, and a minimum distance-based pairing algorithm. The random algorithm chooses a requesting device randomly and pairs it with an arbitrary feasible caching device. The minimum distance-based pairing algorithm just picks a feasible caching device closest to the requesting device. In the following, we first show the results in the static scenario, where the caching devices, requesting devices and content requests are fixed. Then, we study the pairing performance in the dynamic scenario as defined in Section II, with arrivals and departures of requesting and caching devices and content requests.

A. Static Scenario

We fist evaluate the performance of the pairing algorithms in the static scenario. As it is time-consuming to obtain the optimal pairing result, we set the number of requesting devices as varying between 3 and 13 and the number of cache devices fixed at 10. The Zipf exponent for caching (γ_c) is set to 0.7, and the Zipf exponent for requesting (γ_r) is set to 0.9. Other important parameters are given in Table I. The results are based on the average of 10^4 runs.

Fig. 4 shows the satisfaction ratio of different D2D pairing algorithms with a varying number of caching devices. Here, we limit the number of caching devices by 14, as it is very time consuming to obtain the optimal solution to (6) when the problem size is large. As seen, the ratio of satisfied requesting devices increases fast with the number of caching devices. This is because more resources are provided by the caching devices.

Algorithm 1 A heuristic channel-aware pairing algorithm.

Input: S (requesting devices), D (caching devices), Q (requests), φ (current pairing), $r_{\min}$
Output: φ

$\mathbb{D} \leftarrow$ Reordered requesting devices D in a descending order of arrival time for requests in Q
for $d_k \in \mathbb{D}$ **do**
  $S_k \leftarrow \varnothing$ // Feasible cache devices as candidates
  $R_k \leftarrow \varnothing$ // Achievable data rates from feasible caches
  for $s_j \in S$ **do**
    if $r_{j,k} = B\log(1+\xi_k) \geq r_{\min}$ and $m_k^t = m_j^r$ **then**
      $S_k \leftarrow S_k \cup \{s_j\}$
      $R_k \leftarrow R_k \cup \{r_{j,k}\}$
    end if
  end for
  $\mathbb{S}_k \leftarrow$ Reordered cache devices in an ascending order of R_k
  for $s_j \in \mathbb{S}_k$ **do**
    $invalid \leftarrow false$
    for $\{s_{j'}, d_{k'}\} \in \varphi$ **do**
      // Adding the new pair will disable an existing
      // D2D pair in transmission
      if $r_{k'}(\varphi \cup \{s_j, d_k\}) < r_{\min}$ **then**
        $invalid \leftarrow true$
        break
      end if
    end for
    if $invalid = false$ **then**
      $\varphi \leftarrow \varphi \cup \{s_j, d_k\}$ // Add a new D2D pair
    end if
  end for
end for
return φ

In addition, it is found that the performance of the channel-aware algorithm is fairly close to that of the optimal solution, and slightly better than that of the min-distance algorithm when there are more caching devices available.

B. Dynamic Scenario

Next, we evaluate the performance of the D2D pairing algorithms in the dynamic scenario, where the devices move into and leave the cell dynamically. The D2D pairing is conducted in a periodical manner as illustrated in Fig. 2. Since it is not possible to predict the arrivals and departures of the devices, it is hard to determine an optimal pairing strategy that maximizes the overall pairing efficiency instead of focusing on the current period. Therefore, we only compare the channel-aware algorithm with the min-distance algorithm and the random algorithm for the dynamic scenario. Similar to the static scenario, the Zipf exponents for caching and requesting are set to 0.7 and 0.9, respectively. In addition, we fix the arrival rate of requesting devices at $\lambda_r = 1/s$ and vary the arrival rate of caching devices (λ_c) between 1 and 8 per second. The arrival rate of requests from each requesting device is set to $\lambda_q = 0.5/s$, which means that a requesting device waits for $2s$ on average after a completed transmission before sending a new request.

TABLE I
SIMULATION PARAMETERS.

Symbol	**Definition**	**Values**
m	Library size	10
R	Radius of cell	300 m
P_d	Transmit power of D2D device	100 mW
B	Carrier bandwidth	10^6 MHz
α	Path loss exponent	4
β	SINR decoding threshold	-0.6 dB
σ^2	Noise variance	-147 dBm
γ_c	Zipf exponent for caching	[0.5, 1.2]
γ_r	Zipf exponent for requesting	[0.5, 1.2]
λ_r	Arrival rate of requesting devices	5
λ_c	Arrival rate of caching devices	[1, 8]
λ_q	Arrival rate of requests	0.5
μ_r	Departure rate of requesting devices	$\left[\frac{1}{120}, \frac{1}{20}\right]$
μ_c	Departure rate of caching devices	$\left[\frac{1}{120}, \frac{1}{20}\right]$

Fig. 5 shows the evolution of the number of served pairs in a single run consisting of 300 pairing periods. Fig. 6 further presents shows the average results from 200 runs in terms of the total number of served pairs and their average latency. As seen in Fig. 6(a), more D2D pairs are successfully matched when the arrival rate of caching devices increases. Obviously, this is because more resources become available at the caching devices to fulfill more requests. In addition, it is observed that the channel-aware algorithm outperforms both the random algorithm and the min-distance algorithm. On the other hand, Fig. 6(b) demonstrates that the channel-aware algorithm achieves a lower latency to serve the content requests. Here, the latency of each served pair includes the time that the requesting device spends in the waiting queue as well as the file transmission time with the paired caching device. Clearly, the distance between the D2D pair has an important impact on the data rate and consequently the transmission time. As seen in Fig. 6(c), the average distance between D2D pairs is generally decreasing when more caching devices become available. This shorter distance potentially reduces the transmission time. Meanwhile, since there are more candidate caching devices, the waiting time of the requesting devices is also shorter on average.

V. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the D2D pairing problem for collaborative content distribution, which can not only offload traffic from the cellular network but also serve end users with low delivery cost. The D2D pairing problem can be formulated as an ILP, which can serve as a foundation for future studies on this problem. On the other hand, due to the complexity in obtaining an optimal solution, a lightweight fast algorithm is also vital to deal with the dynamic scenario. The simulation results show that the channel-aware pairing

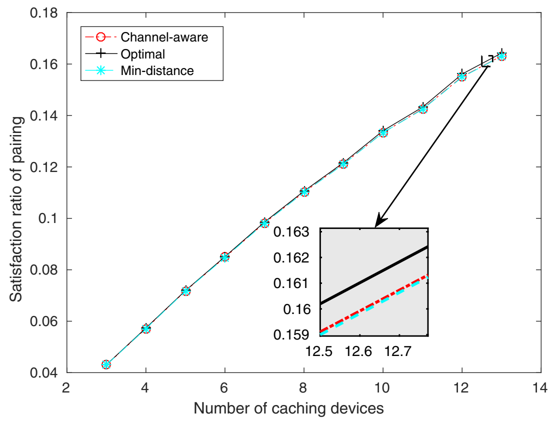

Fig. 4. Satisfaction ratio of pairing with a varying number of caching devices.

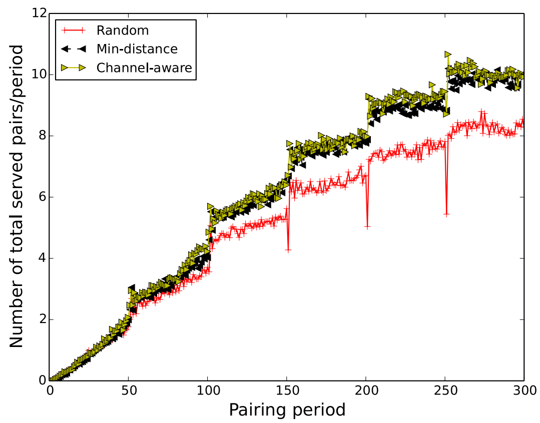

Fig. 5. Average number of served pairs per time slot.

algorithm can achieve good performance in terms of the number of successful D2D pairs and the average latency. In the future, it would be interesting to take into account multipath transmission [9,10] to accommodate service requests for large objects. Then, the one-to-one matching in the D2D pairing problem becomes the more complex one-to-many matching in which multiple source devices share the delivery.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," 2016.

[2] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Lera, and F. Fitzek, "Wi-Fi cooperation or D2D-based multicast content distribution in LTE-A: A comparative analysis," *IEEE WoWMoM*, pp. 1–6, 2015.

[3] B. Chen and C. Yang, "Energy costs for traffic offloading by cache-enabled D2D communications," in *Proc. IEEE WCNC*, 2016.

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

[5] Y. Zhao, W. Song, and Z. Han, "Social-aware data dissemination via device-to-device communications: Fusing social and mobile networks with incentive constraints," *IEEE Transactions on Services Computing*, 2016, to appear.

[6] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM Conference on Internet Measurement*, 2007, pp. 1–14.

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

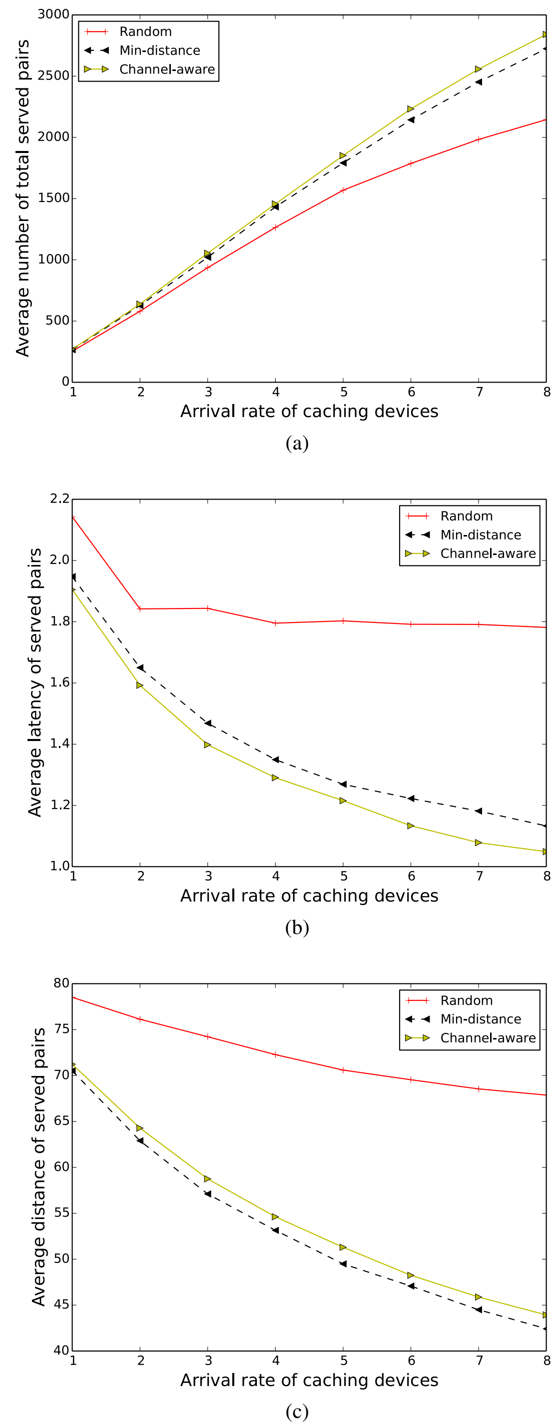

Fig. 6. Pairing performance with a varying arrival rate of caching devices.

[8] 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.

[9] D. Zhou, P. Ju, and W. Song, "Performance enhancement of multipath TCP with cooperative relays in a collaborative community," in *Proc. IEEE PIMRC*, 2012.

[10] D. Zhou, W. Song, and Y. Cheng, "A study of fair bandwidth sharing with AIMD-based multipath congestion control," *IEEE Wireless Commun. Lett.*, vol. 2, no. 3, pp. 299–302, 2013.