Cooperative Caching Placement in Cognitive Device-to-Device Networks

Bing Chen, Hong Shen, and Xiaoxiao Cao
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P. R. China
Email: {cb_china; sx1316012; chenjialiang}@nuaa.edu.cn

Abstract —Cognitive radio is regarded as a key technique in the fifth generation communication system. We consider that the popular contents can be cached in cognitive mobile users who compose a cognitive Device-to-Device (D2D) network. Cached Contents in this network can be shared by cooperating with each other. Consequently, contents can be downloaded from local cache, cognitive D2D network or cellular network with different access cost. It has communication cost when accessing cellular networks, and time delay and cooperative cost when accessing cognitive D2D networks. To minimize the total cost of all the mobile users, we formulate an optimal content placement problem. This problem is hard to solve, and we proposed a sub-optimal cache scheme, including cache placement and cache replacement. Our approach quantifies time delay cost by modeling the primary appearance as a continuous-time Markov chain. Numerical results show that our approach is low-complexity and implementable.

Index Terms—Cognitive radio, cache, cellular network, Device-to-Device (D2D) communication

I. Introduction

Cognitive radio is one of the key techniques in the fifth generation communication system. Both cellular networks and mobile users (we use user and device interchangeably in this paper) can be equipped with cognitive radio. Cognitive cellular networks have been studied widely [1]-[3] in previous work, while cognitive mobile users are rarely considered. Mobile users can communicate not only separately with native cellular base station, but also cooperatively with other mobile users. Moreover, the manner of Device-to-Device (D2D) communication can be decided by distributed or centralized protocol. Therefore, mobile users have great potential to improve the network performance. Specifically, cognitive mobile users can operate on wide spectrum bands which includes unlicensed bands (e.g., ISM bands) and licensed bands (e.g., 2G/3G/4G bands), and D2D communication provides more chances in future wireless communication system.

The spectrum scarcity problem can be alleviated by cognitive radio. However, the cost of downloading the contents is still high, because every requested content must go through the operation business network who will charge a cost to either the user or the Content Provider (CP). The operation business who maintains the cellular network determines the price of communication service at per multimedia data. This communication cost is not necessary if the contents are cached locally or in other cooperative devices. When fetching contents by D2D communication, a cooperative cost should also be paid to the provider device due to the selfishness of rational user. This cost is used for motivating the devices to cooperate. Practically, cooperative cost is much lower than communication cost for accessing the cellular network. There is another cost for D2D communication, i.e., delay cost. A cognitive mobile user, who is transmitting in a licensed channel, should wait when primary user appears. Users will determine to keep waiting or switch to download the content through cellular network.

Due to the limitation of cache capacity, it is impossible to cache all internet contents in mobile users, which leads to the problem of determining what contents should be cached. Recent research finds that abundant multimedia contents, e.g., videos, files, audios, occupy a large part in overall internet traffic. While a small portion of them (i.e., popular content) are frequently accessed by majority of users [4], [5]. These popular contents can be cached in local network, where D2D communication is available to all the users. On the other hand, recent devices (e.g., smartphones, laptops, access points) have large storage capacity and it is not expensive to increase the storage. They are able to cache more and more contents in the future.

Cognitive mobile users may gather physically, e.g., attending a meeting, working in the office. They compose a D2D network, where the communication channels dynamically vary due to the appearance of primary user. Many channel allocation and routing methods [6]-[8] can be implemented in such a network, so D2D communication can be always established by efficient channel allocation. In this paper, we focus on where and how to cache contents in every mobile device. With caching, a user will obtain desired contents from three placements: 1) local cache if the content is stored in the device, 2) cognitive D2D network if other users have cached the content, 3) CP's server through cellular network when and only when D2D network does not cache the content. We make a trade off in the communication cost, cooperative cost and delay cost to minimize the total cost of all the mobile users.

Manuscript received July 13, 2015; revised January 5, 2016.
Corresponding author email: cb_china@nuaa.edu.cn.
doi:10.12720/jcm.11.1.42-49

The contributions in this paper are as follows:

- We consider cognitive radio in mobile communication system, where cache is introduced to decrease the users' communication cost. This framework can be an alternative implementation in the fifth generation.
- We formulate the optimal caching problem to minimize the total cost of all the users. The cache capacity of every user is divided into duplicate and unique, and a distributed cache and replacement approach is proposed to maintain the content in the network.
- To quantify time delay cost, we model the primary appearance as a continuous-time Markov chain. This access delay indicates the performance of cognitive D2D network.
- We present the effect of the number of users, cache capacity, and time delay. Our approach can obtain the sub-optimal results in different scenarios

II. Related Work

In traditional wireless network, cache is used to reduce the transmission delay and improve the user experience. In [9], a cooperative caching method was proposed in social wireless networks, and they studied a case of Amazons Kindle electronic book delivery business to minimize electronic content provisioning cost. Energy consumption was considered in [10], and they formulated the optimal caching problem to minimize the energy consumption of base station and relay points. In [11], Zhao et al. designed and implemented a cooperative caching approach in wireless P2P networks. However, the available channels are time-varying due to the activities of primary user in cognitive radio networks. Existing caching techniques cannot be directly applied to cognitive radio networks.

Cognitive radio is a promising technical to support increasing user demands for rich multimedia contents in the fifth generation communication system. Some key challenges were presented in [12]-[14]. Caching is also an efficient method to improve the throughput of wireless communication system. In [15], they cached the popular content in both core network and radio access network. The authors in [16] solved the cache placement problem in cognitive radio networks to minimize the total cost subject to delay constraint. Data replication techniques were proposed to improve data access performance in [17], where they attempted to determine the optimal replication location for maximizing the average data retrieval probability. While in [18], the authors were surprised to find that a very simple distributed caching policy, i.e., random cache, achieves the optimal scaling behavior in multi-hop wireless networks. However, few of previous work considered the cost which should be paid to the operation business for downloading contents through cellular networks. In this paper, we aim to minimize the total cost of all the users in D2D network.

III. Cognitive D2D Network and Problem Formulation

In this section, we first present a cognitive device-to-device (D2D) network framework, where mobile users, i.e., secondary users, communicate with each other through dynamic spectrum access. Transmission in licensed bands will cause access delay due to the activities of primary users. Mobile users also have cache capacity to store popular content locally. Then we model primary user appearance as a continuous-time Markov chain and the popular content is modeled as Zipf distribution. After that, the cost of downloading a content from content server is formulated, and we present the problem of finding an optimal caching placement approach to minimize the user cost.

A. System Model

Assume a group of users gather physically with cognitive devices, i.e., each device can sense vacant licensed spectrum and use the available spectrum to communicate with other mobile devices. Besides, every device is set a certain cache capacity to store the frequently accessed contents. These users actively decide whether to cooperative in content sharing. If some of them are agreed to cooperate, they form a cognitive D2D network, as shown in Fig. 1. Thus the cooperative users can fetch the requested contents either from cellular network or other cooperative users when local cache misses the contents. We consider D users in cognitive D2D network with equivalent cache capacity, and assume any two of them can establish D2D communication. To facilitate the cooperation, a user will obtain a rebate after it forwards a content to another user. The detail pricing model is presented in Subsection *D*, Section III.

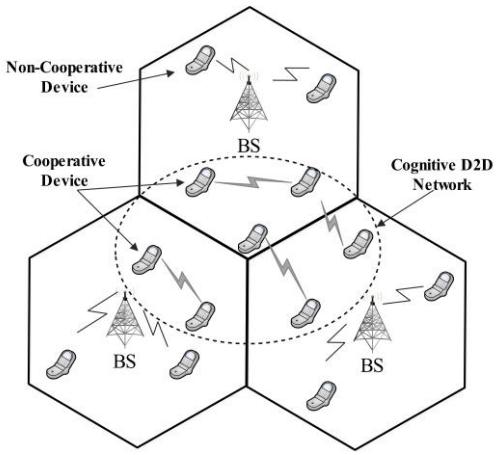


Fig. 1. The framework of cognitive D2D network

With caching popular contents in mobile devices (we consider the contents are time invariant so that content update is not a critical issue), the content access model becomes different from traditional communication system. When a user wants a content, it first looks up in local cache. If the content exists, the content will be directly used. Otherwise, this user will start and broadcast a

request in the cognitive D2D network. After the request is received by other mobile users, each of which will match the content in their caches and return a replay to tell the requested user that the content can be disseminated from this address. Note that a user may receive several replays after starting the content request, and it chooses the earliest replay device as content provider. If D2D search also fails, the requested content must be downloaded through cellular network, which will lead to higher communication cost than D2D network. Let P_L^{ij} be the probability that user i finds the content j in local cache, P_D^{ij} be the probability that user i finds the content j in cognitive D2D network, and P_C^{ij} be the probability that user i downloads the content j from cellular network. According to the above access protocol, we have the following equation

$$P_L^{ij} + P_D^{ij} + P_C^{ij} = 1 \tag{1}$$

B. Channel Occupation Model

As all the mobile devices are equipped with cognitive radio, they can opportunistically utilize the licensed spectrum. However, primary user will take over the spectrum bands even though secondary users are transmitting in those bands, which will degrade the transition efficiency of secondary user. We express this effect as time cost, i.e., the secondary transmission will raise a delay when primary transmission appears. Note that the time delay when transmitting at unlicensed channel is limited so that we do not consider it in this paper, i.e., contend in wireless network is universal and inevitable. To model the channel state, we use 1 or 0 to represent channel is busy or idle. Assume the channel busy time and channel idle time obeys exponential distribution with parameters λ and μ, respectively. It indicates that before the busy state transits into idle, the residence time of busy state is a stochastic variable which conforms to exponential distribution with parameter λ, and before the idle state transits into busy, the residence time of busy state is an exponential distribution stochastic variable with parameter μ. From the above, we have that the process is a continuous-time Markov chain with two states.

Let $P_{uv}(t)$ be the transition probability from state u to state v in t time. Assume the transition probability satisfies the regularity condition, i.e., the state will not change in an infinitesimal time interval. Then we have

$$\begin{cases} P_{01}(h) = \lambda h + o(h) \\ P_{10}(h) = \mu h + o(h) \end{cases} \tag{2}$$

and let $\mathbf{Q}$ be the transition rate matrix, i.e.,

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix} \tag{3}$$

According to the theory of Markovian stochastic process, the state transition probability matrix can be determined by Chapman-Kolmogorov equation and transition rate matrix. The transition probability is associated with a period of time t. When $t \to 0$, the stationary distribution [19] is uniquely determined, i.e.,

$$\pi_0 = \mu_0, \pi_1 = \lambda_0 \tag{4}$$

where $\lambda_0 = \lambda/(\lambda+\mu)$ and $\mu_0 = \mu/(\lambda+\mu)$. This shows that when the system tends to be stable, the probability of channel busy is to μ_0, and the probability of channel idle is to λ_0 at any moment. Therefore, the average occupation time by primary user in a period t is $\mu_0 t$.

C. Content Distribution Model

It is noticed that users usually access the useful and interesting contents from massive information, which leads to some parts are more popular than the rest. Further study found that the access frequency of multimedia contents conforms to Zipf distribution [5, 20]. We divide these contents into M different ranks and the access frequency of i-th content is expressed as

$$y_i = \frac{1/i^v}{\sum_{k=1}^{M} 1/k^v} \tag{5}$$

where $v \geqslant 0$ is a constant. As shown in Fig. 2, the high ranked contents are accessed frequently, and when v approaches 1, the distribution is more uneven. For simplicity, let all the contents have the same size. It is worth mention that the number of contents is much greater than the number of users, i.e., $M >> D$.

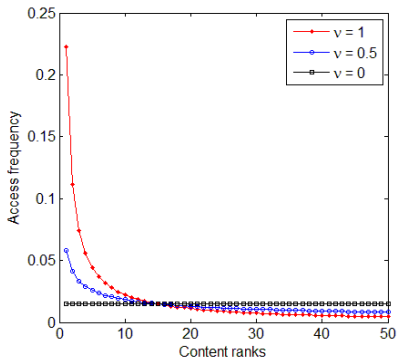


Fig. 2. Zipf distribution (M = 50)

D. Caching Placement Problem

As described above, a content can be found in three scenarios with different costs. 1) Local cache: it is totally free to get the content. 2) D2D network: it contains two costs, the one is time delay due to the activity of primary user, and the other is cooperative cost V_d paid to the user who provides the content for cooperation. 3) Cellular network: a higher cost V_c, we called communication cost, must be paid to the network operation business. Note that rational users are always selfish so that they would not cooperate without rebate. Instead, users would not fetch the contents from D2D network while the rebate is too high. Therefore, $V_d < V_c$ is always satisfied.

Assume every device has cached some contents at local storage in cognitive D2D network. Let **G** be the content distribution matrix, where dimension is defined as $D \times M$. If content j is cached at the i-th device, then $P_L^{ij} = G_{ij} = 1$. Otherwise, $P_L^{ij} = G_{ij} = 0$. There are m licensed channels and n unlicensed channels, and the D2D communication randomly chooses one of them. As mentioned above, only communicate on licensed channels will generate time delay. Thus we use ε to indicate the communication channel, i.e., if a licensed channel is chosen, ε = 1, otherwise ε = 0. Then, we can calculate the total cost for the cognitive D2D network

$$total = \sum_{i=1}^{D}\sum_{j=1}^{M}\left(P_D^{ij} y_j \left(V_d + w\varepsilon T\right) + P_C^{ij} y_j V_c\right) \quad (6)$$

where T is a stochastic variable that represents time delay when communicating at licensed channel. Since the time delay and traffic cost have different units, they cannot be added simply. We use ratio w to match the units and adjust the weight of time delay. Because each device has cache capacity limitation and contents have the same size, the number of contents that cached in the same device should not exceed C, where C is the uniform cache capacity size.

We aim to minimize the total cost through finding out the optimal content storage placement in every cooperative device. Each device decides which contents should be cached to get the minimum total cost, i.e.,

$$\min_{G} total$$

Problem 1. Given a set of cognitive D2D devices with same cache capacity, and a certain number of ranked popular contents, find an optimal caching placement scheme by determining which contents that each device should store to minimize the total cost of the cognitive D2D net-work.

This problem is difficult to solve, because the time complexity of listing all possible placements grows exponentially with the size of the input. A simple greedy approach can be tried, i.e., each device caches as many different contents with higher access frequency as the cache capacity allows. After the network trends to steady, every device caches content from 1 to C. It is easy seen that there is no cooperation so that too many duplicates exist in the network. The other extreme is fully cooperative which tries best to cache distinctly contents in the whole networks which needs cooperative between devices. The result is that every device caches different contents which cover the first content to CD-th content. Simulations show that both greedy and fully cooperative cannot achieve optimal cache replacement. In the following, we present a sub-optimal solution which outperforms the above approaches.

IV. Heuristic Content Placement

To solve the problem, we relax some parameters as follows. Time delay T in objective function is a stochastic variable which can be replaced by its expectation. Assume the transmission time of a content is equal to t without primary user appearance. As describe in Subsection B, Section ш, the practical average transmission time is t/μ_o, therefore, the average delay time $\bar{t} = t(1-\mu_0)/\mu_0$. The channel chosen variable can be ε replaced by the probability of choosing licensed channels. Let $m_0 = m/(m+n)$, then the total cost function can be rewritten as

$$total = \sum_{i=1}^{D}\sum_{j=1}^{M}\left(P_D^{ij} y_j \left(V_d + wm_0\bar{t}\right) + P_C^{ij} y_j V_c\right) \quad (7)$$

We use P_L to represent the local hit rate, i.e., the probability that a content can be found in local cache. Let be the sum of k-th column of **G** which indicates the k-th content has g_k duplicates in the network. Obviously, $g_k < D$, k = 1, 2, … M, because one device caches two or more same contents is worthless. Then P_L can be represented as

$$P_L = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{M} P_L^{ij} y_j = \frac{1}{D}\sum_{j=1}^{M} g_j y_j \quad (8)$$

Let **Z** be the set of all the cached contents in cognitive D2D network. Note that **Z** does not contain the duplicates, i.e., the elements in **Z** are different with each other. Thus the probability that a content exists in this network is $\sum_{j\in Z} y_j$, i.e., $1 - P_C = \sum_{j\in Z} y_j$, where P_C is the average probability that a content is not in local cache but in cognitive D2D network. Combine (1) and (8), the average probability P_D that a content is not in cognitive network can be expressed as

$$P_D = (\sum_{j\in Z} y_j - \frac{1}{D}\sum_{j=1}^{M} g_j y_j) \quad (9)$$

Since m_0 and $\bar{t}$ is a constant, and w is fixed in advance, we define $\alpha = (V_d + wm_0\bar{t})/V_c$. If $V_d + wm_0\bar{t} > V_c$, i.e., $\alpha \geq 1$, then the user prefers to fetch the content from cellular network, i.e., each device caches the most popular contents is the best solution. While in our model, the cellular network is the last choice. Therefore, the following analysis is based on the assume that $\alpha < 1$.

As described above, the probability in (7) can be replaced by average probability. Let $U = total / D$ be the average cost, as a result, minimize U is equal to minimize $total$ cost. Then we will minimize U in the following. Using (1), (7), (8) and (9), U can be simplified as

$$U = \left(1 - (1-\alpha)\sum_{j\in Z} y_j - \frac{\alpha}{D}\sum_{j=1}^{M} g_j y_j\right) V_c \quad (10)$$

To solve this problem, we should determine the only variable g_i, i.e., every element in **G**, to obtain the minimum U. Similar to [9], we divide the storage space

into *duplicate* and *unique*. Let τ be the fraction of duplicate part, i.e., every device has τC capacity for duplicate contents. A duplicate part caches the contents that have more than one copies, while an unique part caches the contents that have only one copy in the whole network. For example, if there are five cognitive devices from A to E who form a cognitive D2D network. Assume $C = 4$ and $\tau = 0.5$, a possible placement is shown in Fig. 3. It is seen that twelve different contents are cached in the five devices, while content 1 and content 2 have a copy in all of them. Furthermore, the contents in these two parts have the following constraints.

1) According to the content access frequency, a content with lower access frequency will not appear in any parts, if the higher rank content is missing, i.e., when the content *i* is not in the network, the content *j* (*j* > *i*) will never appears. When content *j* is cached, we can always replace it with content *i* so that the average cost *U* of (10) will decrease.
2) Based the above constraint, a content should not be duplicated unless all the higher rank contents have duplicated. It can be inferred that if content *j* is a duplicate content, then all the devices must have cached content 1 to content *j* when the network is steady.

A	B	C	D	E
11	7	10	12	9
3	6	4	5	8
2	2	2	2	2
1	1	1	1	1

Fig. 3. A example of caching placement

If content *k* is the least duplicated content in cognitive D2D network, then content 1 to content *k* have a copy at all the devices, i.e., $g_i = D$, $1 \le i \le k$. Let *L* be the least popular content in the whole network, we have $L = (C-k)D+k$. Thus content *k* + 1 to content *L* have only one copy at a certain device, i.e., $g_i = 1$, $k+1 \le i \le L$. The rest content will not appear in this network, i.e., $g_i = 0$, $i > L$.

As a result, we have the following distributed caching and replacement scheme. For a device starting with empty caching, it will request contents for user requirement. The contents may come from local cache, D2D network or cellular network. 1) If a content comes from cellular network, which means this content does not exist in D2D network, it will be compared with the least popular content of all the existing caching. This least popular content will be replaced when the rank of coming content is higher than it; otherwise the coming content will be stored in the unique part. 2) If a content comes from D2D network, which means there are at least two copies in the network after this content is cached, then the least ranked content will be replaced by the coming content when the rank of coming content is higher than it, or store the coming content in duplicate part. 3) If a content is found in local unique part, it will be used to replace the lower content in the duplicate part, e.g., a device has cached content 1, 2 and 4 in duplicate part, and content 3 in unique part, when content 3 is requested, it will be transferred to duplicate part by replacing content 4. The process is described in Algorithm 1.

Algorithm 1 Content caching and replacement scheme

Input: An coming content *M*
Assume the least popular content in entire cache is *P* and the least popular content in duplicate part is *Q*
1: **if** (*M* comes from cellular network) **then**
2: **if** (rank(*M*) > rank(*P*)) **then** replace *P* with *M*
3: **else** cache M in unique part
4: **else**
5: **if** (rank(*M*) > rank(*Q*)) **then** replace *Q* with *M*
6: **else** cache M in duplicate part

To determine the fraction τ, we define function $f(x)$ as the probability of finding a content from caching which is filled with content 1 to content *x*, i.e., $f(x) = \sum_{i=1}^{x} y_i$. Replacing y_i with (5), this function can be expressed as

$$f(x) = \sum_{i=1}^{x} y_i \approx \int_1^x \frac{1/i^v}{\sum_{k=1}^{M} 1/k^v} d_i \tag{11}$$

where $\sum_{k=1}^{M} 1/k^v$ can be replaced by $\int_1^M \frac{1}{k^v} dk$. Therefore, $f(x)$ can be simplified as

$$f(x) \approx \frac{\int_1^x \frac{1}{i^v} d_i}{\int_1^M \frac{1}{k^v} d_k} = \frac{x^{(1-v)} - 1}{M^{(1-v)} - 1} \tag{12}$$

After the system becomes steady, each device's duplicate part caches the same contents which include the most popular τC contents. While in the unique part, every device caches different contents from content $\tau C + 1$ to content $\tau C + (1-\tau)CD$. Therefore, P_L can be calculated as duplicate part hit rate plus unique part hit rate, i.e.,

$$P_L = f(\tau C) + \frac{f(\tau C + (1-\tau)CD) - f(\tau C)}{D} \tag{13}$$

where we assume that the contents in unique part are uniformly distributed. Similarly, P_D indicates the content is in other device's unique parts, i.e.,

$$P_D = \frac{D-1}{D}\left(f(\tau C + (1-\tau)CD) - f(\tau C)\right) \tag{14}$$

Now, the average cost *U* can be rewritten with P_L and P_D, i.e., $U = P_D\left(V_d + w m_o \bar{t}\right) + P_C V_c$. Combined (1), we have

$$U = \left(1 - (1-\alpha)\sum_{j \in Z} y_j - \frac{\alpha}{D}\sum_{j=1}^{M} g_j y_j\right) V_c \tag{15}$$

where P_L and P_D can be expanded by (13) and (14). Given α, *U* is a function of the only independent variable

τ. Thus we can compute the optimal τ for U, i.e., minimize U, by solving $U' = 0$.

As described above, every device in D2D network individually determines the optimal τ, i.e., the cache space for duplicate and unique is fixed. Then they cache and replace the coming contents according to Algorithm 1.

TABLE I: BASELINE PARAMETERS SETTINGS

Parameters	Values
The number of cooperative users(D)	30
Cache capacity of each device(C)	40
Zipf distribution constant(v)	0.8
The number of popular contents(M)	6000
The communication cost(V_c)	10
The cooperative cost(V_d)	1
The proportion of licensed channel(m_0)	0.5
The average transmission time($\bar{t}$)	2

V. NUMERICAL RESULTS

In this section, we present some numerical results to evaluate the performance of our proposed caching placement scheme. The non-cooperative caching approach and fully cooperative approach are used for comparison. We provide the baseline parameters in Table I. When evaluating some factors, the others are fixed as Table I gives.

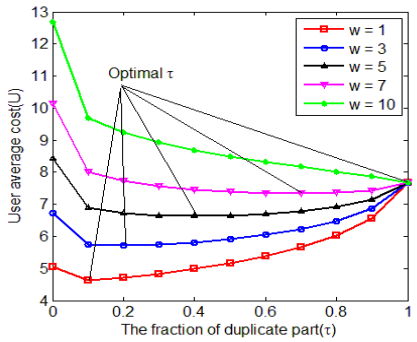


Fig. 4. The impact of time delay cost

Fig. 4 shows the impacts of time delay cost. As defined in (6), w indicates the importance of delay cost. The w = 1, i.e., α = 0.1, represents that the cost of fetching a content from D2D network is relatively small to communication cost. It is easy seen that the smaller w leads to the lower average cost of each user. When w = 10, α is 1, i.e., the cost of fetch a content from D2D network is equal to cellular network. Therefore, the optimal τ is 1 as the green line with star shows. In this case, every device caches the same contents when the network becomes steady, i.e., D2D communication is unnecessary. According to Fig. 4, the users can trade off communication cost and time delay. If they can endure long time wait, then they should choose a low w. Otherwise, a higher w is preferred.

Fig. 5 shows that the user average cost varies along with the fraction of duplicate part. Different numbers of users have the same optimal fraction τ, i.e., τ does not change when adding some users to the network. The cost of each user decreases if the number of users increases. It is because a user has more chance to fetch the contents from D2D network with lower cost when more users join in cooperative caching. The impact of cache capacity is presented in Fig. 6. We observe that the cache capacity has the same influence trend with the number of uses. The larger cache capacity is, the lower user cost is. Obviously, larger capacity can cache more popular contents, which means that users will have more chance to obtain the contents in local cache with free cost. On the other hand, the D2D network will covers more popular contents, i.e., less contents need to fetch from cellular network.

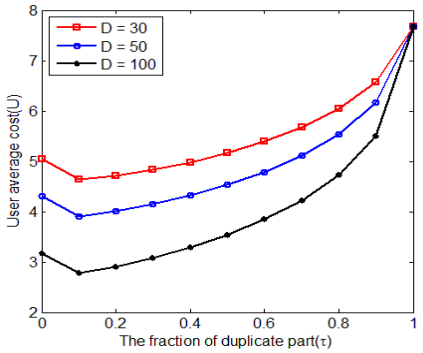


Fig. 5. The impact of the number of users (w=1)

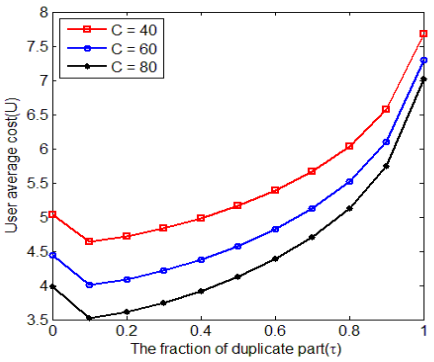


Fig. 6. The impact of cache capacity (w=1)

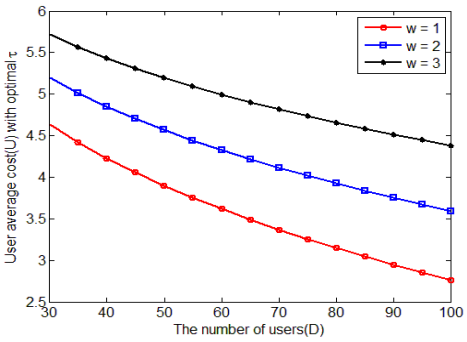


Fig. 7. The impact of the number of users and time delay

Fig. 7 depicts the combined impacts of the number of users and the weight of time delay cost, and Fig. 8 depicts the combined impacts of cache capacity and the weight of time delay cost. The results conform to the above analysis.

We compare our cache scheme with non-cooperative and fully cooperative approaches. As described in

Subsection *D*, Section III, the non-cooperative cache approach is greedy. Each device caches as many popular contents as cache capacity allows. The less popular contents will be replaced by higher ranked contents individually, and there is no cooperation between users in this approach. The fully cooperative cache approach is another extreme, i.e., each user only cache the contents that do not exist in D2D network. This approach neglects the local cache hit rate. From Fig. 9 and Fig. 10, we can see that our scheme performs better than both non-cooperative approach and fully cooperative approach. In Fig. 9, the non-cooperative approach has the same cost when the number of users changes, because every device caches the same contents. It only depends on the cache capacity, as shown in Fig. 10. The fully cooperative approach and our approach have similar cost reduction rate along with the increasing number of users. However, our approach reduces faster when the cache capacity increases.

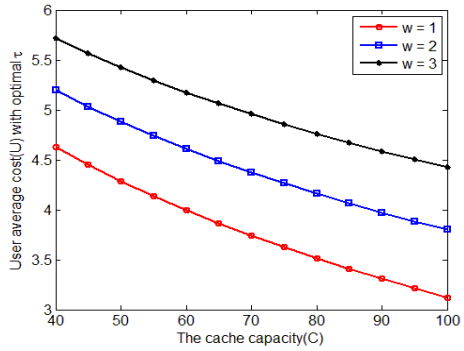


Fig. 8. The impact of cache capacity and time delay.

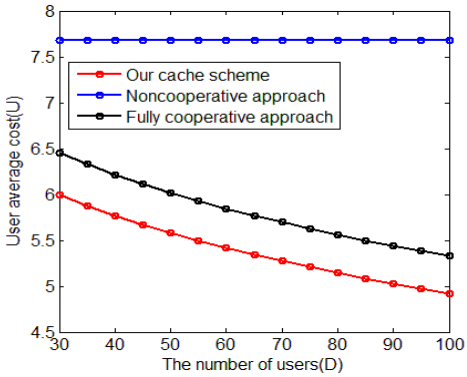


Fig. 9. The user cost of different cache approaches (w=4).

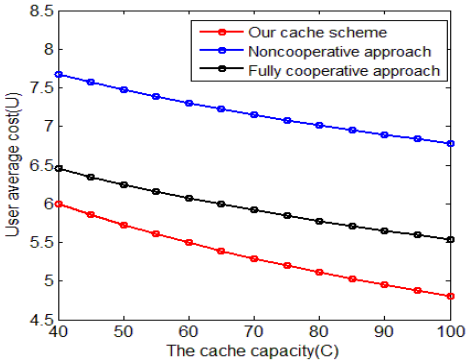


Fig. 10. The user cost of different cache approaches (w=4).

From the above analysis, our cache scheme has a better performance than cooperative and non-cooperative approaches. We obtain some valuable insights as follows.

The way to reduce the average cost of each user can be realized by increasing the number of users, increasing the cache capacity or choosing to wait for a long time, i.e., choosing a small w.

VI. Conclusion

In this paper, we consider caching placement problem in cognitive D2D network to minimize the total cost that involves communication cost, cooperative cost and delay cost. Primary user appearance is modeled as continuous-time Markov chain so that delay cost can be represented as expectation. Communication cost and cooperative cost are relaxed by average hit rate. To solve the optimal placement problem, we proposed a sub-optimal approach to place and replace popular contents in cognitive D2D network. We evaluate the performance in the aspect of the number of users, cache capacity, and time delay. The numerical results show our proposed scheme performs better in user average cost.

Acknowledgment

This work was supported in part by the Industry-University-Research Combination Innovation Foundation of Jiangsu Province (No. BY2013003-03) and the Industry-University-Research Combination Innovation Foundation of Jiangsu Province (No. BY2013095-2-10).

References

[1] C. X. Wang, *et al.*, "Cellular architecture and key technologies for 5G wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122-130, Feb. 2004.

[2] H. Bogucka, P. Kryszkiewicz, and A. Kliks, "Dynamic spectrum aggregation for future 5G communications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 35-43, May 2015.

[3] S. Y. Lien, K. C. Chen, Y. C. Liang, and Y. Lin, "Cognitive radio resource management for future cellular networks," *IEEE Wireless Communications*, vol. 21, no. 1, pp. 70-79, Feb. 2014.

[4] K. Wang, Z. Chen, and H. Liu, "Push-based wireless converged networks for massive multimedia content delivery," *IEEE Trans. on Wireless Communications*, vol. 13, no. 5, pp. 2894–2905, May 2014.

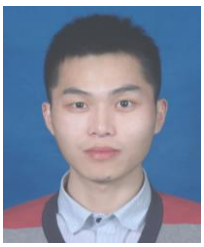
[5] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM SIGCOMM*, Kyoto, 2007, pp. 1-14.

[6] Y. Ding and L. Xiao, "Video on-demand streaming in cognitive wireless mesh networks," *IEEE Trans. on Mobile Computing*, vol. 12, no. 3, pp. 412-423, Mar. 2013.

[7] M. Alicherry, R. Bhatia, and L. E. Li, "Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks," in *Proc. ACM MobiCom*, Cologne, 2005, pp. 58-72.

[8] M. Pan, *et al*., "When spectrum meets clouds: Optimal session based spectrum trading under spectrum uncertainty," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 3, pp. 615-627, Mar. 2014.
[9] M. Taghizadeh, K. Micinski, C. Ofria, E. Torng, and S. Biswas, "Distributed cooperative caching in social wireless networks," *IEEE Trans. on Mobile Computing*, vol. 12, no. 6, pp. 1037-1053, June 2013.
[10] C. Yang, Z. Chen, Y. Yao, B. Xia, and H. Liu, "Energy efficiency in wireless cooperative caching networks," in *Proc. IEEE ICC*, Sydney, 2014, pp. 4975-4980.
[11] J. Zhao, P. Zhang, G. Cao, and C. Das, "Cooperative caching in wireless P2P networks: Design, implementation, and evaluation," *IEEE Trans. on Parallel and Distributed Systems*, vol. 21, no. 2, pp. 229-241, Feb. 2010.
[12] X. Hong, J. Wang, C. X. Wang, and J. Shi, "Cognitive radio in 5G: A perspective on energy-spectral efficiency trade-off," *IEEE Communications Magazine*, vol. 52, no. 7, pp. 46-53, July 2014.
[13] C. Mavromoustakis, *et al*., "Joint energy and delay-aware scheme for 5G mobile cognitive radio networks," in *Proc. of IEEE GLOBECOM*, San Diego, 2014, pp. 2624-2630.
[14] C. I. Badoi, N. Prasad, V. Croitoru, and R. Prasad, "5G based on cognitive radio," *Wireless Personal Communications*, vol. 57, no. 3, pp. 441-464, Apr. 2011.
[15] 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 Communications Magazine*, vol. 52, no. 2, pp. 131-139, Feb. 2014.
[16] J. Zhao, W. Gao, Y. Wang, and G. Cao, "Delay-constrained caching in cognitive radio networks," in *Proc. of IEEE INFOCOM*, Toronto, 2014, pp. 2094-2102.
[17] J. Zhao and G. Cao, "Spectrum-aware data replication in intermittently connected cognitive radio networks," in *Proc. IEEE INFOCOM*, Toronto, 2014, pp. 2238-2246.
[18] N. Golrezaei, A. Dimakis, and A. Molisch, "Wireless device-to-device communications with distributed caching," in *Proc. IEEE ISIT*, Cambridge, 2012, pp. 2781-2785.
[19] Q. Zhao, S. Geirhofer, L. Tong, and B. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Trans. on Signal Processing*, vol. 56, no. 2, pp. 785-796, Feb. 2008.
[20] M. Newman, "Power laws, Pareto distributions and zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323-351, Feb. 2007.

Bing Chen received B.S. and M.S. degree in computer engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1992 and 1995, respectively. He received Ph.D. degree in the College of Information Science and Technology from NUAA in 2008. He has worked for NUAA since 1998. His main research interests are computer network, embedded system and wireless communication.



Hong Shen received the B.S. in electrical engineer from China University of Mining and Technology, Xuzhou, China, in 2013. He is pursuing his M.S. degree at the Nanjing University of Aeronautics and Astronautics, Nanjing, China. His research interests are cognitive radio networks, and distributed computing.

Xiaoxiao Cao received the B.S in computer engineering from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2014. He is pursuing his M.S. degree at NUAA. His research interests are cognitive radio networks, and distributed computing.