Cache Placement Optimization in D2D-assisted Wireless Cognitive Networks

Cheng Zhan[1], Zhe Wen[1], Hong Lai[1], Zhiguo Qu[2]

[1] School of Computer and Information Science, Southwest University, China
[2] School of Computer and Software, Nanjing University of Information Science and Technology, China
zhanc@swu.edu.cn, wenzheswu@gmail.com, hlai@swu.edu.cn, qzghhh@126.com

Abstract

In this paper, we consider the cooperative caching in wireless cognitive device-to-device (D2D) network, where unlicensed users can use under-utilized licensed spectrum and cached contents can be shared by cooperating with each other. We formulate the cooperative caching problem to minimize the download time of mobile users. A heuristic caching strategy is then proposed to handle the content placement in the caches. Simulation results show that the proposed caching scheme can significantly reduce total download time compared with the traditional caching scheme.

Keywords: Cooperative caching, Content placement, Device-to-Device, Cognitive networks

1 Introduction

Device-to-device (D2D) communication enables nearby wireless devices to exploit their proximity and communicate directly with each other, bypassing their corresponding cellular base stations [1]. D2D communication offloads traffic from the cellular base stations and reduces congestion on radio resources. Bluetooth and WiFi-direct are examples of D2D technologies in unlicensed bands, however manual pairing and short range coverage have limited their functionality. Recent trends necessitate more sophisticated technologies for D2D communications possibly in cellular bands [2-3], incorporating the concept of cognitive networks [4-5] to deal with the spectrum shortage problem.

In wireless cognitive D2D network, unlicensed use of licensed spectrum is allowed. To avoid interference with licensed users, unlicensed users must vacate the spectrum when it is accessed by the primary users who are licensed to access the spectrum. It will take some time for the unlicensed users to switch to other available channels, therefore the transmission delay will be significantly increased [6]. As a result, it is hard to meet the delay constraints of many stream applications in cognitive networks. With an increasing popularity of smart mobile devices, mobile time-critical applications, such as audio and video streaming, cloud-based services, are more and more popular [7]. One way to reduce data access delay is through caching.

Today's smart mobile devices and networks possess a tremendous amount of storage capability. Cooperative caching offers an exciting new way to unleash the ultimate potential of wireless D2D networks. The concept of the cooperative caching is that the data contents can be cooperatively stored in the mobile terminals [8]. User can download the corresponding content from these cached devices instead of the base station (BS) as shown in Figure 1, which can reduce the average access latency, offload the network traffic and make BS more scalable.

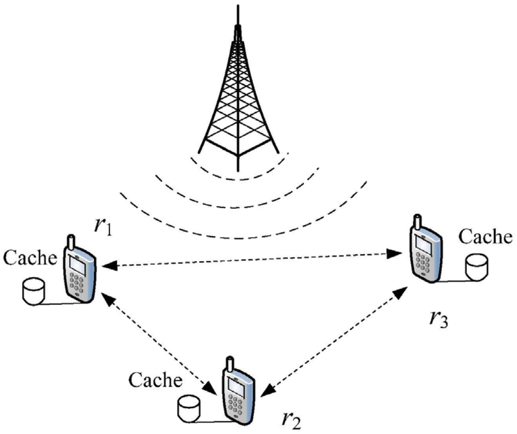


Figure1. Cooperative caching in wireless cognitive D2D networks

Although caching techniques have been well studied in traditional wireless networks, they cannot be directly applied to cognitive D2D networks. Traditional caching techniques assume the link transmission delay is known or just model the access delay by the number of hops [9-10]. The optimal caching problem to minimize the energy consumption in the wireless cooperative caching network was considered in [11].

*Corresponding Author: Cheng Zhan; E-mail: zhanc@swu.edu.cn

DOI: 10.3966/160792642018091905015

However, this work did not consider the cognitive network and they assumed there were some relay nodes which stored data. In cognitive D2D networks, the link transmission delay may vary significantly from time to time due to the primary user appearance. This makes it difficult to determine appropriate caching nodes to reduce data access delay. The primary user appearance makes it difficult to estimate various types of time cost, and further complicates the design of caching algorithms. To overcome the aforementioned difficulties, we use continuous time Markov chain similar as [12] to model the primary user appearance, and then derive the distribution of the data access delay.

In this paper, we consider cooperative caching problem in the wireless cognitive D2D network that the communication channels dynamically vary due to the appearance of the primary user, the aim is to minimize the download time. With caching, a user can obtain the data contents through three ways: (1) Local caching: to find the content within a device's local cache; (2) D2D network caching: to find the content within other user's cache in the D2D network if the user cannot obtain the content from the local caching; (3) BS: to download the content from the BS when the user can find the content from neither the local caching nor the D2D network caching. The main contributions of this paper are summarized as follows:

- We formulate the optimal caching problem in the wireless cognitive D2D network to minimize the download time cost, considering communication channels dynamically vary due to the appearance of the primary user.
- By relaxing total download time to average download time, a heuristic caching scheme is proposed to get rid of the complex optimal caching scheme.
- The superiority of the heuristic caching scheme is verified with the simulation compared with the baseline. We can also obtain some valuable insights on the impact of different model parameters on the performance.

2 System Model and Problem Description

2.1 System Model

The architecture of the wireless cooperative caching network is illustrated in Figure 1. We consider a single cell, where there is one BS and n users. 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 has a certain cache capacity to store the frequently accessed content. We consider n users in a cognitive D2D network with equivalent cache capacity, and assume any two of them can establish a D2D communication.

There are m data contents in the BS. It is noticed that users usually access the useful and interesting contents from massive information, which leads to some parts of the contents are more popular than the rest. Further study found that the access frequency of contents conforms to Zipf distribution [11]. We assume that the statistical popularity of the data contents in the BS can be modeled as the Zipf distribution in this paper.

According to the Zipf distribution, the popularity of the j-th data content f_j is given by

$$f_j = \frac{1/j^{\theta}}{\sum_{i=1}^{m}(1/i^{\theta})}, \tag{1}$$

where $1 \le j \le m$. $\theta \ge 0$ is a constant, it reflects the skew of the popularity distribution. There are m data contents in the base station, and the popularity of j-th content depends on the index of j, $f_1 > f_j$, j>1, which means that the first content is the most popular content according to the Zipf distribution. The second content is the second popular content among the m data contents, and the m-th content is the most unpopular content. The distribution becomes increasingly skewed as θ increases from 0 (the uniform distribution) to 1 (highly skewed).

For simplicity, let all the contents have the same size. Let P_{ij}^{L} be the probability of finding j-th content in the local cache of user i, P_{ij}^{D} be the probability that the j-th content can be found by user i in the D2D network after its local search fails, and P_{ij}^{B} be the probability that the j-th content is found neither in the local cache nor in the D2D network, it can only be downloaded from the base station, $1 \le i \le n$, $1 \le j \le m$. Let $G = (g_{ij})$ be an n×m dimension content distribution matrix. If content j is cached at the i-th device, then g_{ij} = 1, otherwise, g_{ij}=0. The calculation of P_{ij}^{L}, P_{ij}^{D} and P_{ij}^{B} is based on the content distribution matrix G. If g_{ij} = 1 which means that content j is cached at the i-th device, then P_{ij}^{L} =1, otherwise P_{ij}^{L} = 0. If g_{ij} = 0, and there exist g_{ik} in G that g_{ik} = 1, $k \neq j$, which means that content j is not cached at the i-th device, but cached at other device k, then P_{ij}^{D} = 1, otherwise P_{ij}^{D} = 0. If g_{ij} = 0, and there do not exist any g_{ik} in G that g_{ik} = 1, k≠j, which means that content j is not cached at the i-th device, and not cached at any device k, then P_{ij}^{B} = 1, otherwise P_{ij}^{B} = 0. Based on the above analysis, we can formulate the calculation as follows,

$$P_{ij}^{L} = \begin{cases} 1, & if\ \ g_{ij} = 1 \\ 0, & otherwise \end{cases},$$

$$P_{ij}^{D} = \begin{cases} 1, & if\ \ g_{ij} = 0, and\ \exists k \neq j, g_{ij} = 1 \\ 0, & otherwise \end{cases},$$

$$P_{ij}^{B}=\begin{cases}1, & if\ g_{ij}=0, and\ \forall k\neq j, g_{ij}=0\\ 0, & otherwise\end{cases}$$

Therefore, we have

$$P_{ij}^{L}+P_{ij}^{D}+P_{ij}^{B}=1, \tag{2}$$

2.2 Channel 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. Each link can work on the channel which is not currently accessed by primary users. Similar as in [12], the primary user appearance can be modeled as the following continuous time Markov chain, based on which the link transmission delay is probabilistically estimated.

To model the channel state, we use 1 or 0 to represent channel is busy or idle. For link e, we denote the current primary user appearance by $M_e(t)=(M_{e,1}(t), M_{e,2}(t),\ldots, M_{e,c}(t))$, where $M_{e,c}(t)$ = 1 if channel c is accessed by some primary user, otherwise, $M_{e,c}(t)$ = 0. $M_{e,c}(t)$ follows a continuous-time Markov chain with two states, $M_{e,c}(t)$ = 1 and $M_{e,c}(t)$ = 0. Since $M_e(t)$ is a continuous-time Markov chain, the channel busy time and channel idle time obeys exponential distribution with parameters λ, and μ, respectively [13]. Since our focus is to deal with the disruption caused by primary user appearance, for the transmission delay of link e, we will focus on the time waiting for available licensed channels. According to the theory of Markovian stochastic process, the state transition probability matrix can be determined by Chapman-Kolmogorov equation and transition matrix. When the system tends to be stable, the average occupation time by primary user is $T=\mu/(\lambda+\mu)$ [13].

2.3 Problem Formulation

As described in Section 1, a content can be found in three scenarios with different time cost. We assume that there is no time cost when a user obtains the requested content from its local cache. We consider the time cost C_D as the download time when download a content from the D2D network, and similarly we have C_B as the download time when downloading a content from the base station. Assume every device has cached some contents at local storage in the cognitive D2D network.. There are k_l licensed channels and k_u unlicensed channels, and the D2D communication randomly chooses one of them. Communicate on licensed channels will generate time delay, the probability of choosing licensed channels can be represented as $k_0 = k_l/(k_l + k_u)$.

If device i wants to download the j-th content, there exits three cases.

- Case 1: the j-th content is cached at the device i. The download time of case 1 is 0 since the content is cached at the local cache, and the probability of case 1 is P_{ij}^{L}.
- Case 2: the j-th content is not cached at device i and cached at other device k, which $k\neq j$. The download time of case 2 is C_D+Tk_0 since the content is cached at the D2D network, and the probability of case 2 is P_{ij}^{D}.
- Case 3: the j-th content is neither cached at the device i nor cached at D2D network. The download time of case 3 is C_B since the content is cached at the base station, and the probability of case 3 is P_{ij}^{B}.

Since the popularity of the j-th content is f_j, the average of the total download time C can then be modeled as follows.

$$C=\sum_{i=1}^{n}\sum_{j=1}^{m}P_{ij}^{D}f_j(C_D+Tk_0)+P_{ij}^{B}f_jC_B \tag{3}$$

Since 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 K, where K is the uniform cache capacity size, we assume $K\leq m$. It is easy to obtain the following storage constraint.

$$\sum_{j=1}^{m}g_{ij}\leq K, 1\leq i\leq n \tag{4}$$

The goal of this paper is to minimize the download time of the wireless cooperative caching network through finding out the optimal content storage placement G in the D2D network and the local cache. Mathematically, we have the following optimization problem

$$\min_{G} C \tag{5}$$

$$s.t.\quad f_j=\frac{1/j^{\theta}}{\sum_{i=1}^{m}(1/i^{\theta})}, \tag{6}$$

$$P_{ij}^{L}+P_{ij}^{D}+P_{ij}^{B}=1, \tag{7}$$

$$\sum_{j=1}^{m}g_{ij}\leq K, 1\leq i\leq n \tag{8}$$

$$g_{ij}\in\{0,1\}, 1\leq i\leq n, 1\leq j\leq m \tag{9}$$

In the formulation, the objective function is the average of the total download time, and the goal of this paper is to minimize the average of the total download time. Since the objective function contains content popularity variable f_j, we give a constraint (6) to show the distribution of content popularity, the popularity of the data contents follows the Zipf distribution. Constraint (7) represents for the access constraints, and the calculation of P_{ij}^L, P_{ij}^D and P_{ij}^B is based on the content distribution matrix *G*, which is given in Section 2.1. Constraint (8) illustrates the cache storage constraints. Similar to [14], this optimization problem is an integer programming problem. It is difficult to find out the exact mathematical algorithm for this optimization problem. We thus present a heuristic solution in the following section.

3 Heuristic Solution for the Problem

The optimal solution becomes complicated and time consuming. In order to find out a general caching mechanism, we relax the model and the condition as follows. Let $U=C/n$ be the average download time of all users, as a result, minimize *U* is equal to minimize total download time.

We use P_L to represent the local hit rate, the average probability that a content can be found in a local cache. Let $s_j = \sum_{i=1}^{n} g_{ij}$, which indicates that there are s_j copies of the *j*-th content in the D2D network. Thus, P_L can be represented as

$$P_L = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij}^L f_j = \frac{1}{n}\sum_{j=1}^{m} s_j f_j \qquad \textbf{(10)}$$

Let *D* be the set of all the cached contents in the cognitive D2D network, therefore the elements in *D* are different from each other. Thus the probability that a content exists in this network is $\sum_{j\in D} f_j$, therefore $P_B = 1-\sum_{j\in D} f_j$, where P_B is the average probability that a content is neither in local cache nor in the cognitive D2D network. The average probability P_D that a content can be found in the cognitive D2D network can be expressed as

$$P_D = (\sum_{j\in D} f_j - \frac{1}{n}\sum_{j=1}^{m} s_j f_j) \qquad \textbf{(11)}$$

Since k_0 and *T* are constants, we define $\alpha = (C_D + Tk_0)/C_B$. From the above analysis, we can simply the objective as follows.

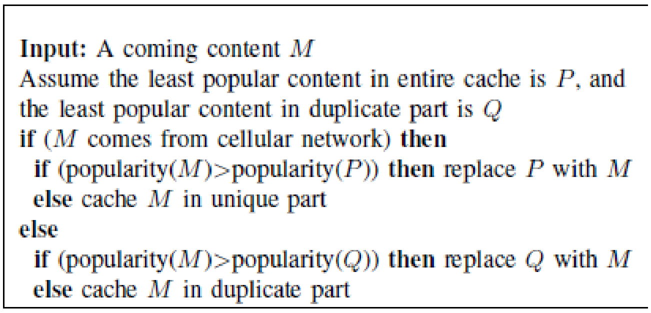
$$U = \left(1-(1-\alpha)\sum_{j\in D} f_j - \frac{\alpha}{n}\sum_{j=1}^{m} s_j f_j\right) C_B \qquad \textbf{(12)}$$

To solve this problem, we should determine the only variable s_j to minimize *U*. Since $s_j = \sum_{i=1}^{n} g_{ij}$, which means that there are s_j copies of the *j*-th content in the D2D network. Therefore, in order to minimize the total download time, we only need to optimize s_j. Given s_j, we can use a greedy choice to set g_{ij} to ensure every column of g_{ij} is equal to s_j.

We divide the cache space into duplicate and unique. Let β be the fraction of duplicate part, i.e., every user has βK duplicate contents. A duplicate part caches the contents that have more than one copy, while a unique part caches the contents that have only one copy in the whole network. Furthermore, the contents in these two parts have the following concerns.

If the content with higher popularity is missing, i.e., when the content j_1 is not in the network, the content j_2 ($f_{j_1} > f_{j_2}$) will never appear. Otherwise, when content j_2 is cached, we can always replace it with content j1 so that the average download time *U* will decrease. A content should not be duplicated unless all the content with higher popularity have been duplicated.

Inspired by the above analysis, we have the following distributed caching replacement scheme. (1) If a content is downloaded from the cellular network, which means the content does not exist in D2D network, it will be compared with the least popular content of all the existing cache. The least popular content will be replaced when the popularity of coming content is higher, otherwise the coming content will be stored in the unique part. (2) If a content is downloaded from D2D network, which means there are at least two copies in the network after this content is cached, then the content with least popularity will be replaced when the popularity of coming content is higher, otherwise the coming content will be stored in the duplicate part. The process is described as shown in Figure 2.



```
Input: A coming content $M$
Assume the least popular content in entire cache is $P$, and
the least popular content in duplicate part is $Q$
if ($M$ comes from cellular network) then
  if (popularity($M$)>popularity($P$)) then replace $P$ with $M$
  else cache $M$ in unique part
else
  if (popularity($M$)>popularity($Q$)) then replace $Q$ with $M$
  else cache $M$ in duplicate part
```

Figure 2. The cache algorithm

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

$$f(x) \approx \int_1^x \frac{1/i^v}{\sum_{k=1}^{m} 1/k^v} = \frac{x^{1-v}-1}{m^{1-v}-1} \qquad \textbf{(13)}$$

As the system is in steady state, each device's duplicate part caches the same contents which include the most popular βK contents. While in the unique part, every device caches different contents. Therefore, P_L can be calculated as duplicate part hit rate plus unique part hit rate.

$$P_L = f(\beta K) + \frac{f(\beta K + (1-\beta)Kn) - f(\beta K)}{n} \quad \textbf{(14)}$$

where we assume that the contents in unique part are uniformly distributed. Similarly, P_D indicates the content is in other devices unique parts,

$$P_D = \frac{n-1}{n}(f(K\beta + (1-\beta)Kn) - f(K\beta)) \quad \textbf{(15)}$$

The average download time U can be rewritten with P_L and P_D. Given α, U is a function of the only variable β. Thus we can compute the optimal β for U by solving U' = 0, where U' is the first-order derivative of function U.

As described above, after determining the optimal β, which illustrates the cache space for duplicate and unique, the devices can cache and replace the coming contents according to the cache algorithm.

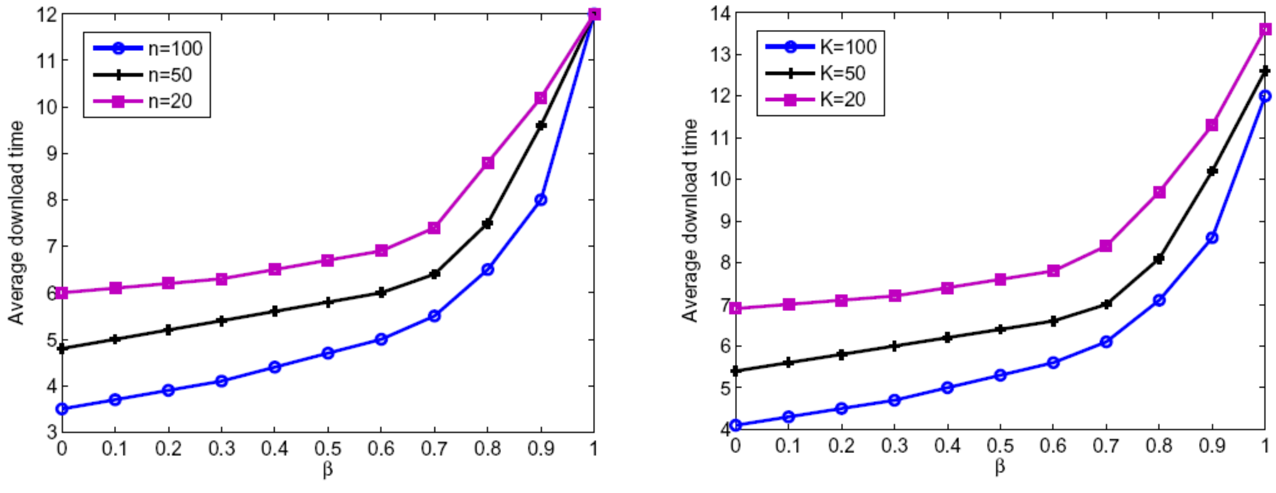
4 Simulations

In this section, we present simulation results to evaluate the performance of our proposed caching placement scheme. The non-cooperative caching approach and fully cooperative approach in [11, 15] are used for comparison. We also compare our heuristic solution with the globally optimal solution using integer programming considering small network size, since computing the integer programming will take more time when the network size is larger. The parameter setting is summarized in Table 1.

Figure 3 shows that the user average download time varies along with the fraction of duplicate parts. We assume that users have the same optimal fraction β. The download time of each user decreases if the number of users increases. The reason is that user has more chance to fetch the contents from D2D network with lower time when more users join in cooperative caching. The impact of cache capacity is presented in Figure 4. We observe that the cache capacity has the same influence trend with the number of uses. Larger cache size can cache more popular contents, which means that users will have more chance to obtain the contents in local cache. On the other hand, the D2D network will covers more popular contents, therefore less content need to be obtained from BS. Figure 4 only show partial result of the optimal solution, the reason is that computing the integer programming takes more time when the network size is larger, so we only obtain the results considering the small network size. From Figure 4 we can also see that the performance of our heuristic solution is very close to the optimal solution.

Table 1. Simulation parameters

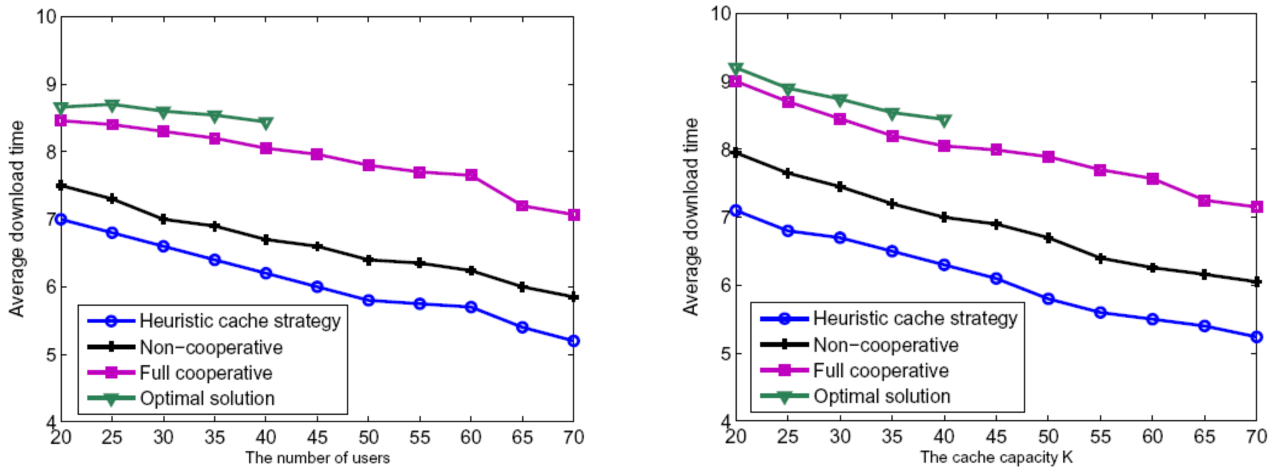
Parameters	Value range	Default
Number of data contents m	1000 ~ 2500	2000
Number of clients n	100 ~ 400	200
Cache size K	30 ~ 180	50
Skewness parameter of Zipf distribution (θ)	0 ~ 1	0.6
The time cost C_D	0 ~ 20	10
The time cost C_B	0 ~ 2	1
The average occupation time T	0 ~ 4	2
The proportion of licensed channel k_0	0 ~ 1	0.5



(a) U vs. n

(b) U vs. K

Figure 3. The impact of different duplicate part



(a) *U* vs. *n*

(b) *U* vs. *K*

Figure 4. The performance of our cache scheme

5 Conclusion

In this paper, we consider cooperative caching in the cognitive D2D network to minimize the total download time. Primary user appearance is modeled as a continuous time Markov chain. To solve the optimal placement problem, we proposed a heuristic approach to place and replace popular contents in the cognitive D2D network according to relax total download time to average download time. Simulation results show that the proposed caching scheme can significantly reduce total download time compared with the traditional caching scheme. We can also obtain some valuable insights on the impact of different model parameters on the performance.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61702426, No. 61702427), Fundamental Research Funds for the Central Universities (XDJK2018C048, No. SWU115091), and a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET).

References

[1] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, S. Li, G. Feng, Device-to-device Communications in Cellular Networks, *IEEE Communications Magazine*, Vol. 52, No. 4, pp. 49-55, April, 2014.

[2] F. Pantisano, M. Bennis, W. Saad, M. Debbah, Spectrum Leasing as an Incentive Towards Uplink Macrocell and Femtocell Cooperation, *IEEE Journal on Selected Areas in Communications*, Vol. 30, No. 3, pp. 617-630, April, 2012.

[3] F.-H. Tseng, T.-T. Liang, L.-D. Chou, H.-C. Chao, Network Planning for Heterogeneous Cellular Network in Next Generation Mobile Communications, *Journal of Internet Technology*, Vol. 17, No. 6, pp. 1269-1277, November, 2016.

[4] S. Haykin, Cognitive Radio: Brain-empowered Wireless Communications, *IEEE Journal on Selected Areas in Communications*, Vol. 23, No. 2, pp. 201-220, February, 2005.

[5] Y.-W. Chen, P.-Y. Liao, B.-T. Huang, A Chip-Based Distributed Spectrum Adjustment for Fair Access in Cognitive Radio Network, *Journal of Internet Technology*, Vol. 17, No. 1, pp. 11-18, January, 2016.

[6] J. Zhao, G. Cao, Robust Topology Control in Multi-Hop Cognitive Radio Networks, *IEEE Transactions on Mobile Computing*, Vol. 13, No. 11, pp. 2634-2647, November, 2014.

[7] Z. Li, Z. Chen, J. Zhang, J. Zhu, N. N. Xiong, The Evolution of IoT Wireless Networks for Low-Rate and Real-Time Applications, *Journal of Internet Technology*, Vol. 18, No. 1, pp. 175-188, January, 2017.

[8] U. Niesen, D. Shah, G. W. Wornell, Caching in Wireless Networks, *IEEE Transactions on Information Theory*, Vol. 58, No. 10, pp. 6524-6540, October, 2012.

[9] Y. Fadlallah, A. M. Tulino, D. Barone, G. Vettigli, J. Llorca, J. M. Gorce, Coding for Caching in 5G Networks, *IEEE Communications Magazine*, Vol. 55, No. 2, pp. 106-113, February, 2017.

[10] Y. Zeng, M. Jin, J. Li, An Approach for Robust In-Network Caching in Information-Centric Networks, *Journal of Internet Technology*, Vol. 17, No. 3, pp. 503-513, May, 2016.

[11] C. Yang, Z. Chen, Y. Yao, B. Xia, H. Liu, Energy Efficiency in Wireless Cooperative Caching Networks, *2014 IEEE International Conference on Communications* (ICC), Sydney, NSW, 2014, pp. 4975-4980.

[12] S. Bayhan, F. Alagoz, A Markovian Approach for Best-fit Channel Selection in Cognitive Radio Networks, *Ad Hoc Networks*, Vol. 12, pp. 165-177, January, 2014.

[13] Q. Zhao, S. Geirhofer, L. Tong, B. M. Sadler, Opportunistic Spectrum Access via Periodic Channel Sensing, *IEEE Transactions on Signal Processing*, Vol. 56, No. 2, pp. 785-796, February, 2008.

[14] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, G. Caire, FemtoCaching: Wireless Content Delivery Through Distributed Caching Helpers, *IEEE Transactions on Information Theory*, Vol. 59, No. 12, pp. 8402-8413, December, 2013.

[15] J. Zhao, W. Gao, Y. Wang, G. Cao, Delay-Constrained Caching in Cognitive Radio Networks, *IEEE Transactions on Mobile Computing*, Vol. 15, No. 3, pp. 627-640, March, 2016.

Biographies

Cheng Zhan received the Ph.D. degree in computer science from the University of Science and Technology of China in 2011. He is currently in the School of Computer and Information Science, Southwest University, China. His research interests include network coding, wireless network optimization, multimedia transmission and distributed storage.

Zhe Wen was born in Anhui Province, China. He is an undergraduate student in the School of Computer and Information Science, Southwest University, Chongqing, China. His research interests include network coding, mobile multimedia transmission.

Hong Lai is currently an associate professor at School of Computer and Information Science, Southwest University, China. She received her PhD from Macquarie University and Beijing University of Posts and Telecommunications in 2015. Her current research interests lie in the areas of secret sharing, key agreement in wireless network.

Zhiguo Qu, received the Ph.D. degree in information security from Beijing University of Posts and Telecommunications, China, in 2011. He is currently a Lecturer in the College of Computer and Software, Nanjing University of Information and Technology in China. His research interests include secure communication, information hiding in wireless network.