IDENTIFYING INFLUENTIAL USERS IN MOBILE DEVICE-TO-DEVICE SOCIAL NETWORKS TO PROMOTE OFFLINE MULTIMEDIA CONTENT PROPAGATION

Hao Fan, Xu Tong, Qing Zhang, Tianxiang Zhang, Chenyang Wang and Xiaofei Wang

Tianjin Key Laboratory of Advanced Networking,
School of Computer Science and Technology,
College of Intelligence and Computing, Tianjin University
xiaofeiwang@tju.edu.cn

ABSTRACT

In recent years, due to the rapid development of mobile multimedia services integrated with online social networks, how to select influential users (seed users) to promote multimedia content propagation has attracted more and more attention. However, little work has been done for large-scale offline face-to-face (Device-to-Device, D2D) content propagation. Previous studies have much limitations in this scenario due to their small-scale or synthetic data sets. In this paper, we propose the algorithm of Weighted LeaderRank with Neighbors (WLRN) to select seed users in D2D mobile social networks with accuracy to promote offline multimedia content propagation. We consider the importance of users' 2-hop neighbors. The evaluation of our algorithm is carried out on a realistic large-scale D2D data set based on the high performance computing platform of *Apache Spark*. The experiment results show the efficiency of the algorithm both in terms of content propagation coverage and time cost.

Index Terms— Offline Multimedia Content Propagation, Mobile Social Networks (MSNs), Device-to-Device (D2D), Influence Maximization

1. INTRODUCTION

With the rapid development of mobile communications, more people tend to use mobile devices such as mobile phones and tablets to meet their social and entertainment needs. They also tend to download multimedia content onto their mobile devices, which will generate massive amounts of traffic. The explosive growth of traffic load on mobile networks poses a huge challenge to the mobile communication infrastructure [1]. Therefore, how to reduce the traffic load has become the focus of mobile network operators.

Studies [2, 3] have shown that there are serious problems with repeated downloads of popular multimedia content. e.g., the top 10% of videos on YouTube generate 80% of the total traffic, which seriously waste the resources of communication networks. Therefore, how to effectively reduce the repeated downloads within cellular networks through local short-range sharing technologies have become a hot research topic. An effective solution is to use the D2D opportunistic sharing mechanism to encourage users to obtain the multimedia content they need from the surrounding devices [4].

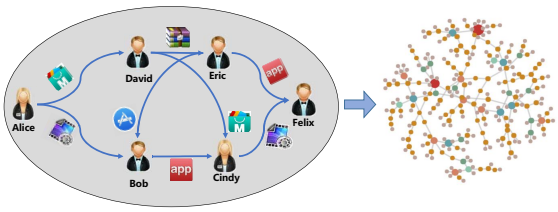


Fig. 1. Multimedia Content Propagation by D2D Sharing.

As shown in Fig. 1, the initial user Alice downloads a multimedia file on her mobile device, then shares this file with her friend Bob by D2D link, and Bob might share the file with other users. The people who carry on this offline short-distance sharing are likely to be friends, which can form D2D mobile social networks. Some studies [5, 6] have shown that there are opinion leaders in networks that have big effects on multimedia content propagation, so we can mine seed users according to the users' sharing behavior from social group to promote offline multimedia content propagation.

The contribution of this paper is modeling the process of maximizing D2D content propagation and proposing a new seed user selection algorithm to promote offline multimedia content propagation based on realistic large-scale D2D sharing data. We consider weights of D2D links and the importance of users' 2-hop neighbors in our algorithm, called Weighted LeaderRank with Neighbors (WLRN). The experiment results show the highest spreading propagation coverage of our algorithm by comparing total 6 different algorithms. In addition, the time cost of WLRN is almost the same as that of PageRank(PR) [7]. With 1 seed user, the coverage in 5 weeks

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00296

of WLRN is increased by 34.22% than PageRank.

The remainder of this paper is organized as follows. After reviewing related studies in Section 2, we describe the details of the D2D data set and users' social groups we use in Section 3. The new algorithm of seed user selection will be introduced in Section 4. A comparative analysis of the experimental results of all algorithms is shown in Section 5. The conclusion of this paper is given in Section 6.

2. RELATED WORK

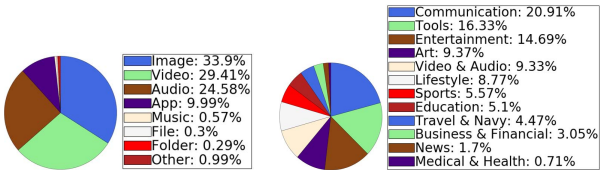
Mobile social networks are gaining more and more attention in many studies. Currently, a large part of the research work is to reduce the repetitive traffic within the network. Zhang et al. [8] and Li et al. [9] put forward a differentiation-based model to evaluate the performance of popular content sharing in MSNs. However, their experimental data only contains a small number of users.

Previous studies have shown that the influence of users is gradually accumulated hop by hop, and there is a significant extension of the re-sharing [5], which makes it possible to analyze and predict the sharing behavior of users and the propagation of popular mutimedia content in social networks. However, many of the existing researches are carried out for social networks formed by popular online websites such as Twitter [5]. Moreover, analyzing and predicting the sharing behavior of users in offline MSNs are much more difficult than online SNSs due to the temporal-spatial limitation. Even if some studies [10, 11] focus on the maximization of impact on offline social networks, the data set they use are based on either unrealistic assumptions, or limited data analytics caused by small data size, which do not fully consider the characteristics of offline users.

Identifying influential users is critical to promote multimedia content dissemination and reduce traffic load in cellular networks. PageRank [7] and HITS [12] are both well-known web page sorting algorithms and can be used to rank the user's influence. SeedRank(SR) algorithm is proposed to select seed users in social networks for the sake of improving PageRank [13]. In addition, there are some centrality metrics to rank nodes in networks, such as degree centrality [14], closeness centrality [15]. One of the most classical algorithms of influence maximization problem is greedy algorithm [16], but it is time-consuming for large-scale scenarios.

3. DATA SET

The big data set we use is based on realistic offline D2D sharing activities, derived from an APP called *Xender*. Users can directly share the multimedia content through the D2D communication mode without accessing the cellular network, thus promoting the spread of offline multimedia content. In order to improve the big data processing capability, we build an efficient big data processing cluster based on *Hadoop* and *Spark*, which contains 8 nodes, 304 cores, 680 GBytes memory, and 50TBytes storage in total.



(a) Distribution of Multimedia Content Types (b)Distribution of Mobile Applications(Type App in Fig. 2(a))

Fig. 2. Multimedia Content Types by D2D Sharing.

The size of the original data set is approximately 3.6TB, which is a 13 weeks of D2D communication record from August 1, 2016 to October 29, 2016. The data set contains about 10 million records per day. The number of active users per day is about 100 million. The effective data set after processing is about 96GB, of which the number of D2D sharing activities is about 900 million, the number of users is about 24 million, and the number of files transferred is about 6.22 million. The data dimensions selected include 6 properties: $<$ file type, file's MD5 code, sender, receiver, file size, timestamp $>$.

We classify and count the types of offline multimedia content. As shown in Fig. 2(a), multimedia files that include image, video, audio etc. occupy a large proportion of offline transmission. In addition, we also count the transmission of offline mobile applications. In Fig. 2(b), the top 3 types of offline multimedia mobile applications are Communication, Tools, and Entertainment, which are also closely related to the multimedia content dissemination.

As illustrated in Fig. 1, when users use D2D communication for multimedia content sharing, the relationship among users can form social graphs similar to online social networks. Hence, we can define a directed social graph of a D2D social group by $\overrightarrow{G} = \{V, \overrightarrow{E}, \overrightarrow{W}\}$ concerning sharing directions. A user social group consists of user pairs who have act a content sharing activity at least once. A vertex $v_i \in V$ represents a user in a social group, and an edge $\overrightarrow{e_{ij}} \in \overrightarrow{E}$ represents the transmissions from user v_i to user v_j, and $\overrightarrow{w_{ij}} \in \overrightarrow{W}$ is the sharing frequency.

As the small group has less information, it is less significant for the process of seed users selection. We detect totally 3,965 large groups in our D2D sharing trace by k-means, and there are no less than 20 users in each group, and we randomly select 300 groups from them, which contain almost 8,000 users in total. Different algorithms are performed to select initial seed users in these 300 groups to maximize the number of users covered by offline multimedia content propagation.

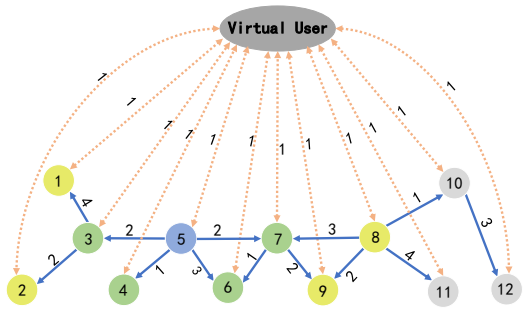


Fig. 3. Weighted LeaderRank with Neighbors.

Table 1. Scores of Weighted LeaderRank with Neighbors

Users	Neighbors		Scores		
	$1-hop$	$2-hop$	WLR	$WLRN^1$	$WLRN^2$
1	3	2,5	0.0612	0.0716	0.0737
2	3	1,5	0.0612	0.0716	0.0737
3	1,2,5	4,6,7	0.1040	0.1312	0.1334
4	5	3,6,7	0.0612	0.0762	0.0788
5	3,4,6,7	1,2,8,9	0.1500	0.1818	0.1851
6	5,7	3,4,8,9	0.0612	0.0853	0.0890
7	5,6,8,9	3,4,10,11	0.0904	0.1321	0.1352
8	7,9,10,11	5,6,12	0.1438	0.1734	0.1762
9	7,8	5,6,10,11	0.0612	0.0846	0.0882
10	8,12	7,9,11	0.0831	0.1036	0.1058
11	8	7,9,10	0.0612	0.0756	0.0779
12	10	8	0.0612	0.0695	0.0710

4. WEIGHTED LEADERRANK WITH NEIGHBORS

PageRank is a well-known link analysis algorithm generally used to rate the web pages [7]. However, the damping coefficient of PageRank is fixed in different situations and the social graphs of MSNs have a weak connectivity. Compared to PageRank, the original LeaderRank(LR) algorithm is parameter-free [17, 18], which can be modeled in Eq. (1). $R^*(u)$ is a set of users that have ever received content from u. $S^*(v)$ is a set of users that have ever sent content to v. $LR(t,v)$ is the score of user v in the t^{th} iteration. We propose and implement the Weighted LeaderRank with Neighbors(WLRN) algorithm to select the users with high scores as seed users.

$$LR(t+1,u) = \sum_{v \in R^*(u)} \frac{LR(t,v)}{|S^*(v)|} \tag{1}$$

For a social graph with n users and m edges, we add to it one virtual user u^*. There is a bidirectional edge with weight 1 between the virtual user and each of the n users. This virtual user plays a role as a mediation user to send files to and receive files from the other users in a social graph. The social graph now contains $n+1$ users and $m+2n$ edges in Fig. 3.

We consider the direction of the social graphs. In D2D social graphs, the users with more outlinks than inlinks are expected to be more active in sending files rather than receiving. Hence, users with high outdegrees in our algorithm will get higher scores. We consider the weight of edges in Weighted LeaderRank(WLR) in Eq. (2), where $W(u,v)$ is the weight of the edge of user pair (u,v). In our defined D2D social group, the weight of user pairs refers to the sharing frequency. And the weights between the virtual user and other users are all set to 1, which has no influence on the scores flowing in the social group. $WLR(t,v)$ is the score of v in the t^{th} iteration.

$$WLR(t+1,u) = \sum_{v \in R^*(u)} \frac{W(u,v)WLR(t,v)}{\sum_{s \in S^*(v)} W(s,v)}, \tag{2}$$

when the iteration in Eq. (2) converges, we divide the scores of the virtual user u^* equally to other n users by Eq. (3), where $WLR(u)$ is the final score of user u by Weighted LeaderRank, and t_c means when the iteration converges.

$$WLR(u) = WLR(t_c,u) + \frac{WLR(t_c,u^*)}{n} \tag{3}$$

In addition, the influence of a user not only depends on its own influence, but also depends on its neighbors. The scoring schema by considering the neighbors of a user is more accurate in terms of selecting influential users [19]. Therefore, by considering the neighbors of users, we define the Weighted LeaderRank with Neighbors as follows.

$$WLRN(u) = WLR(u) + \sum_{l=1}^{L} a^l \sum_{v \in N^l(u)} WLR(u), \tag{4}$$

where $WLR(u)$ is the convergent Weighted LeaderRank score of user u. a is an adjustable parameter and $a \in [0,1]$. a^l represents the importance of $l-hop$ neighbors of u, indicating that the more hops the neighbors belong to, the less influence they have on u. $N^l(u)$ is the l-hop neighbors of user u. L is the number of hops we consider for users' neighbors, and it is proved that $L=2$ is a good choice to gain the highest performance [19]. Our objective is to select K of most influential users to form set T. The target optimization problem can be expressed as :

$$\begin{aligned} \max : & \sum_{k=1}^{K} WLRN(v_i) \\ & = \sum_{k=1}^{K} WLRN(WLR(v_i),a,L) \\ & = \sum_{k=1}^{K} [WLR(v_i) + \sum_{l=1}^{L} a^l \sum_{u \in N^l(v_i)} WLR(u)] \\ s.t : & v_i \in T; \\ & a \in [0,1]; \\ & k,K,L, \in N^*; k \le K; k \le L; \\ & u \in N^l(v_i). \end{aligned} \tag{5}$$

Algorithm 1 Calculate Scores of Users by WLRN

Require: $\overrightarrow{G} = (V, \overrightarrow{E}, \overrightarrow{W}), \varepsilon, L, a$

add a virtual user u^* and bidirectional links between u^* and each $u_i \in V$ to $\overrightarrow{G}$ and update $\overrightarrow{G}$ to $\overrightarrow{G^*} = (V^*, E^*, \overrightarrow{W^*})$;
find the direct receivers set $R(u_i)$ and senders set $S(u_i)$ of user $u_i \in V^*$;
forall $u_i \in V^*$ **do** $Scores[u_i] \leftarrow TempScores[u_i] \leftarrow 1$;
$e \leftarrow maxInt$;
while $e \geq \varepsilon$ **do**
    for $u_i \in V^*$ **do**
        $temp \leftarrow 0$;
        for $v_j \in R[u_i]$ **do**
            $temp \leftarrow temp + \frac{W^*[u_i,v_j] * Scores[v_j]}{\sum_{u_k \in S(v_j)} W^*[u_k,v_j]}$;
        end for
        $TempScores[u_i] \leftarrow temp$;
    end for
    $e \leftarrow ||TempScores - Scores||_\infty$;
    $Scores \leftarrow TempScores$;
end while
forall $u_i \in V$ **do** $Scores[u_i] + \frac{Scores[u^*]}{|V|}$;
for $u_i \in V$ **do**
    calculate 1 to L hop neighbors of user;
    get sum of L hop neighbors influence as $temp$;
    $Scores[u_i] \leftarrow Scores[u_i] + temp$;
end for
return $Scores$

The pseudocode of $WLRN$ is shown in Algorithm 1. First, we calculate the WLR scores of users in Fig. 3 and find out their 1-hop and 2-hop neighbors. Then the corresponding scores of each user can be obtained as shown in Table 1. In general, the higher total influence of users in initial seed user set T is, the more people will be covered by these seed users.

5. EVALUATION

We use different classical algorithms to select seed users from the 300 groups randomly selected in the previous eight weeks, including PageRank [7], HITS [12], Closeness Centrality [15], Greedy Algorithm [16] and Weighted SeedRank Algorithm [13]. Then we denote the number of users covered by seed users in the latter 5 weeks as the performance evaluation metric of different classical algorithms.

Fig. 4 refers to one group consisting of 24 users in the former 8 weeks, in which the red nodes denote the seed users, the dark gold and light gold nodes represent users that have ever received content from seed users respectively by 1-hop and 2-hop. The spot lines and gray nodes illustrate users who left the group in the latter 5 weeks or don't have any interaction with the seed users.

Fig. 5 shows the propagation coverage of LeaderRank

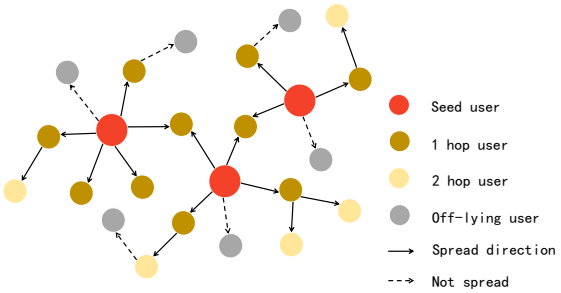


Fig. 4. Seed Users' Propagation Coverage.

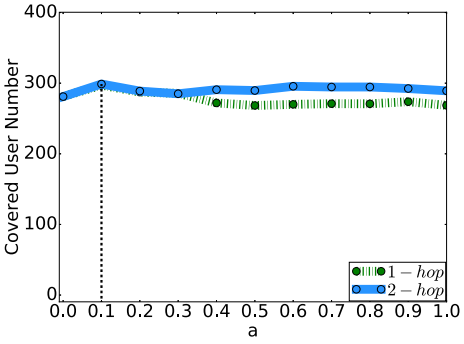


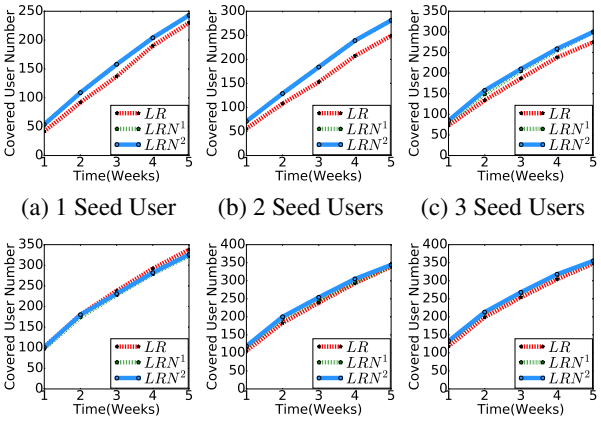
Fig. 5. Propagation Coverage Varies with a.

with 1-hop and 2-hop Neighbors with different a, and $a = 0$ represents the original LeaderRank. Obviously, the propagation coverage reaches the maximum when $a = 0.1$ for both 1-hop and 2-hop.

We compare LeaderRank with 1-hop and 2-hop neighbors shown in Fig. 6, which demonstrates the propagation coverage of LeaderRank under 3 different conditions. In Eq. (4), we set $a = 0.1$ and $L = 2$ as we have discussed above. LR is the original LeaderRank, and LRN^1 is LeaderRank with 1-hop neighbors, and LRN^2 is LeaderRank with 2-hop neighbors. All the 3 compared algorithms don't consider the weight of edges. As shown in Fig. 6, when the seed user number is 4, the propagation coverage of LRN^2 is not as good as original LR. But in other cases, LRN^2 is better than other 2 algorithms. Hence, LRN^2 achieves higher performance than LR and LRN^1 on the whole.

For Weighted LeaderRank, similar results are shown in Fig. 8. Fig. 6 and Fig. 8 prove that it is beneficial to improve propagation coverage by considering users' neighbors, and considering 2-hop is better than 1-hop in most cases. Taking both 1-hop and 2-hop neighbors into consideration, we compare Unweighted LeaderRank and Weighted LeaderRank in Fig. 7(a). The results prove that considering the weight gains higher propagation coverage.

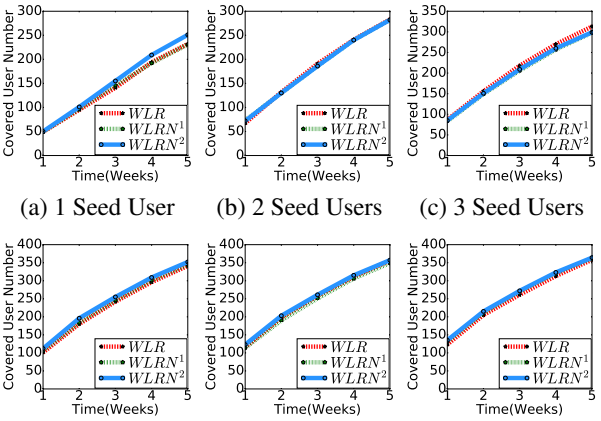
We show the propagation coverage of 6 different algo-



(a) 1 Seed User (b) 2 Seed Users (c) 3 Seed Users

(d) 4 Seed Users (e) 5 Seed Users (f) 6 Seed Users

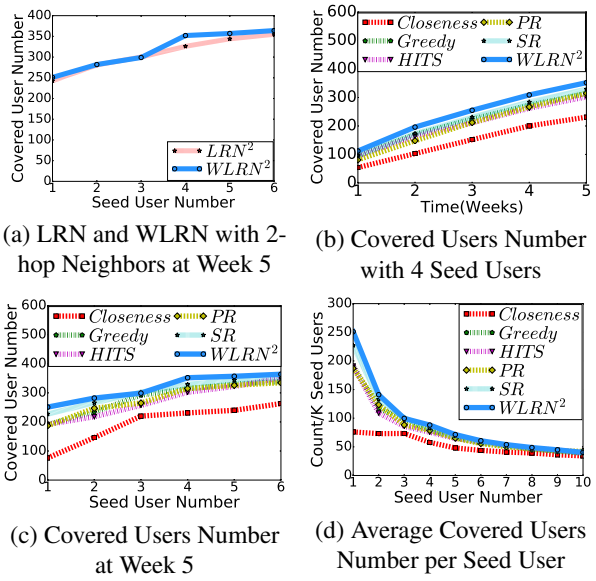
Fig. 6. LeaderRank with Neighbors.



(a) 1 Seed User (b) 2 Seed Users (c) 3 Seed Users

(d) 4 Seed Users (e) 5 Seed Users (f) 6 Seed Users

Fig. 8. Weighted LeaderRank with Neighbors.



(a) LRN and WLRN with 2-hop Neighbors at Week 5

(b) Covered Users Number with 4 Seed Users

(c) Covered Users Number at Week 5

(d) Average Covered Users Number per Seed User

Fig. 7. Propagation Coverage of Compared Algorithms.

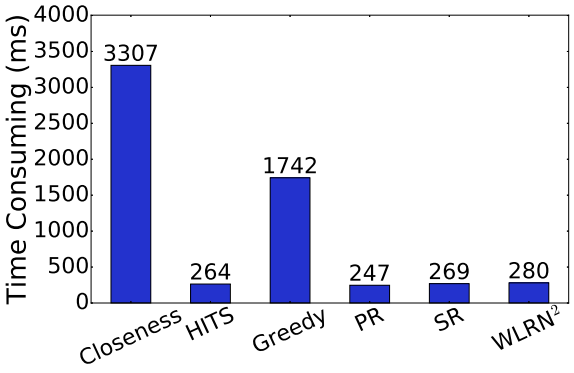


Fig. 9. Comparison of Algorithms' Running Time.

rithms with 4 seed users in Fig. 7(b). Furthermore, Fig. 7(c) shows the number of users covered in 5 weeks with 1 to 6 seed users. Obviously, the number of covered users increases with the accumulation of the number of seed users. When regarding to users' amounts, seed users selected by our algorithm cover more users than other 5 algorithms.

More specifically, $WLRN^2$ improves the propagation coverage by 31.41% compared with Greedy Algorithm and 34.22% compared with PageRank respectively with 1 seed user in 5 weeks. We show the relationship between average covered user number and seed user number in Fig. 7(d). With the number of seed users becoming larger, the users that one seed user can cover get fewer. It reveals that the influential users take up a small proportion in a social group but play an important role.

In Fig. 9, we shows the average running time consumed by the 6 compared algorithms through 30 repetitive experiments. Closeness Centrality takes the longest time because it calculates the shortest path length for each user pair. Greedy Algorithm needs to do several iterations, each of which takes much time to simulate information diffusion process in a social network. Among the rest 4 algorithms, $WLRN^2$ takes longest time. Nevertheless, the tiny time difference is negligible. Generally, our algorithm takes almost the shortest time and achieves the best performance in propagation coverage among all the compared algorithms.

6. CONCLUSIONS

In this paper, we first analyze and measure the type of multimedia content based on a large volume of D2D data set, then

we propose a new algorithm of seed user selection, which takes users' L-hop neighbors' importance into account. We use it to select seed users in social groups and compare a total of 6 different algorithms both in terms of content propagation coverage and time cost. The results shows that our algorithm can effectively select seed users with high spreading capacity, thus accelerating offline multimedia content propagation.

7. ACKNOWLEDGMENT

This work is partially supported by the National Key R&D Program of China (2018YFC0809803), China NSFC (Youth) through grant 61702364, China NSFC GD Joint fund U1701263. We appreciate the valuable data provided by our partner Xender, and we also thank the anonymous reviewers for their valuable comments and suggestions that help improve the quality of this manuscript.

8. REFERENCES

[1] Cisco Visual Networking Index Cisco, "Global mobile data traffic forecast update, 2013–2018," *white paper*, 2014.

[2] Salvatore Scellato, Cecilia Mascolo, Mirco Musolesi, and Jon Crowcroft, "Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades," in *International Conference on World Wide Web*, 2011, pp. 457–466.

[3] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong Yeol Ahn, and Sue Moon, "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system," in *Acm Sigcomm Conference on Internet Measurement*, 2007.

[4] Xiaofei Wang, Min Chen, Zhu Han, Ted Taekyoung Kwon, and Yanghee Choi, "Content dissemination by pushing and sharing in mobile cellular networks: An analytical study," in *IEEE International Conference on Mobile Ad-Hoc and Sensor Systems*, 2012, pp. 353–361.

[5] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 591–600.

[6] Duncan J. Watts and Peter Sheridan Dodds, "Influentials, networks, and public opinion formation," *Journal of Consumer Research*, vol. 34, no. 4, pp. 441–458, 2007.

[7] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The pagerank citation ranking: Bringing order to the web.," Tech. Rep., Stanford InfoLab, 1999.

[8] Xiaolan Zhang, Giovanni Neglia, Jim Kurose, and Don Towsley, "Performance modeling of epidemic routing," *Computer Networks*, vol. 51, no. 10, pp. 2867–2891, 2007.

[9] Yong Li, Yurong Jiang, Depeng Jin, Li Su, Lieguang Zeng, and Dapeng Wu, "Energy-efficient optimal opportunistic forwarding for delay-tolerant networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 9, pp. 4500–4512, 2010.

[10] Ejder Bastug, Mehdi Bennis, and Merouane Debbah, "Social and spatial proactive caching for mobile data offloading," in *IEEE International Conference on Communications Workshops*, 2014, pp. 581–586.

[11] Anwar Said, Syed Shah, Hasan Farooq, Adnan Mian, Ali Imran, and Jon Crowcroft, "Proactive caching at the edge leveraging influential user detection in cellular d2d networks," *Future Internet*, vol. 10, no. 10, pp. 93, 2018.

[12] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, Jan. 1999.

[13] Yuhua Zhang, Zihan Huang, Shanjia Wang, Xiaofei Wang, and Tianpeng Jiang, "Spark-based measurement and analysis on offline mobile application market over device-to-device sharing in mobile social networks," in *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2017, pp. 545–552.

[14] Linton C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, Jan. 1978.

[15] Gert Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, no. 4, pp. 581–603, Dec. 1966.

[16] E. Tardos, D. Kempe, and J. Kleinberg, "Maximizing the spread of influence in a social network," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003, pp. 137–146.

[17] Linyuan L, Yi Cheng Zhang, Ho Yeung Chi, and Zhou Tao, "Leaders in social networks, the delicious case," *Plos One*, vol. 6, no. 6, pp. e21202, Jun. 2011.

[18] Shuang Xu and Pei Wang, "Identifying important nodes by adaptive leaderrank," *Physica A Statistical Mechanics and Its Applications*, vol. 469, pp. 654–664, Mar. 2017.

[19] Ying Liu, Ming Tang, Tao Zhou, and Younghae Do, "Identify influential spreaders in complex networks, the role of neighborhood," *Physica A Statistical Mechanics and Its Applications*, vol. 452, no. 3, pp. 289–298, Jun. 2016.