Received: 12 August 2016 | Accepted: 7 September 2016
DOI: 10.1002/cpe.4021

WILEY

SPECIAL ISSUE PAPER

A measurement study of device-to-device sharing in mobile social networks based on *Spark*

Hui Wang[1] | **Xiaofei Wang**[1] | **Keqiu Li**[1] | **Jianji Ren**[2] | **Xiaohong Zhang**[2] | **Tianpeng Jiang**[3]

[1]Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin, China
[2]Institute of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China
[3]Beijing Anqi Zhilian Technology Co. Ltd., Beijing, China

Correspondence
Xiaofei Wang, Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin, China.
Email: dobbymmlab@gmail.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61372085

Summary

Because of the exponential growth of mobile users' demand for multimedia services in recent years, the increasing network traffic load gets a close attention of the mobile network operators. For the mobile traffic explosion issue to be solved, there are many efforts trying to offload the mobile traffic from infrastructure cellular links to direct local short-range communications among groups of users, which is called device-to-device sharing (D2D) in mobile social networks. Although there have been a number of studies for improving the exploitation of friends, contents, and sharing performance, there is no any large-scale measurement-based study to analyze the realistic D2D sharing service. We focus on the empirical trace from *Xender*, a popular mobile application for D2D sharing, and implement an effective big data processing platform based on *Spark* with customized algorithms. Extensive analysis and discussions are carried out from the perspectives of general time series statistics, content properties, and social graph basics. The trace-driven analysis exploits a number of implications regarding power law distribution for content popularity disparity, clustering effects of user relationships, and so on. We further discuss the potentials of improving *Xender*'s quality of service and optimizing its system resource, and hopefully, our study can offer useful guidelines for not only *Xender* but also those growing global social D2D sharing services.

KEYWORDS

device-to-device sharing, mobile social network, mobile traffic explosion, traffic offloading

1 | INTRODUCTION

Within recent years, the service demand for rich multimedia over mobile networks has kept being soaring at a tremendous pace.[1] More users choose to regard mobile devices as the main business communication and entertainment tools, for example, accessing games, videos, and folders through phones and tablets, which leads to the explosion of traffic load over mobile networks. The mobile traffic explosion problem poses great challenges for the communication infrastructure of mobile networks operators (MNOs) and severely downgrades user experiences such as large access delay, slow download speed, and even frequent disconnections in peak hours. Although wireless communication technologies such as Wi-Fi and LTE have been improved a lot in recent years, they are still difficult to meet users' daily needs on mobile networks traffic and bandwidth, which also limits mobile networks services development.

Recently, there have been many studies[2] pointing out that one of the essential reasons that cause traffic explosion is the duplicate download of the same prevalent files, especially when some videos and applications (APPs) attracted huge number of users over a period of time. For instance, top 10% of videos occupy nearly 80% of all the views in YouTube.[2] We discover that there is also much duplicate traffic load in mobile social networks (MSNs). An effective way to reduce such duplicate downloads is to cache and share the multimedia contents among geographically proximal mobile devices through device-to-device (D2D) communications (eg, Wi-Fi Direct, Bluetooth, and LTE-direct).[3] In doing so, each user is likely to obtain popular contents from nearby devices and only needs to use the expensive cellular to download the contents that are unavailable in proximity, resulting in cellular traffic offloading. What is more, the bandwidth of D2D connectivity is normally much fast and transmissions are totally free.

A number of studies explore the potential of D2D sharing for traffic offloading in MSNs, which can be considered as a special delay tolerant network associating with MSNs. For instance, Wang et al[4] report that, at the most, 86.5% cellular offloading can be achieved by analyzing social network services (SNSs) as well as mobile network services

trace and modeling the user-dependent access delay between contents created time and every user's access time. And hence, how to make full use of D2D sharing to significantly decrease the traffic load becomes a hot research topic in academic and industrial fields.

To exploit how to effectively offload duplicate traffic load and provide quality of experience (QoE) support for MSNs services, in this paper we take a trace-based measurement study of *Xender*, which is one of the world's leading APPs for D2D content delivery and sharing and the largest one in India, based on the big data processing platform *Spark*, to explore the characteristics of sharing activities, propagated files, social relationships, and so on. Our diverse measurement tests are able to describe dynamics of users, contents, sharing activities, and so on, allowing for an innovative analysis of the characteristics of the interactions among them. The discussions potentially reveal new research topics and offer implications and guidelines for optimizing the *Xender* system, improving the Quality of Service (QoS) of users, and helping most D2D sharing services to offload more mobile traffic load effectively.

As far as we know, this is the first realistic measurement study of social D2D sharing service over a very large-scale dataset of a huge number of users, contents, and activities. We summarize the major contributions of this paper as follows:

1. We consider the D2D file sharing tool as an enhanced content dissemination service and carry out realistic measurement. The trace is with quite large scale, and to the best of our knowledge, this is the first work on both large-scale and multifeature data analysis for real-world D2D sharing.
2. We analyze the trace from the aspects of content properties, user dynamics, and service performance over time series and obtain sufficient implications for improving the services.
3. Our measurement further studies the in-depth knowledge of user social networks (complex networks) over the users, including the features of social graphs and social clustering.
4. We also contribute a prototype test-bed framework to analyze *Xender*'s offline user-to-user file sharing service based on the big data processing platform *Spark*, along with necessary social network analysis and machine learning modules.

We organize this paper as follows. After reviewing related work in Section 2, we first present the measurement details in Section 3. Then we show some time-scale measurement results in Section 4. In Section 5, we analyze the characteristics of contents, while we next reveal some basic analysis results from the perspective of social graph in Section 6. Finally we discuss and unveil some potential future directions, and conclude the paper in Section 7.

2 | RELATED WORK

2.1 | D2D sharing in mobile social networks

The D2D sharing program[5] in MSNs is shown in Figure 1, which mainly depends on local short-range communication techniques and has similar content delivery patterns and social properties as online SNSs. For instance, Apple's Airdrop[6] provides conveniently local file delivery functionality via Wi-Fi techniques, while *Xender* is also offering D2D sharing capability for mobile devices. *Xender* users can send a number of files with high speed, which is almost hundreds of times of the rate of Bluetooth with several users, consuming zero cellular network traffic. By measuring the share activities among these users, we can analyze the social network generated by offline transmission. Then, we can exploit the social relationships as well as user's impacts to offload traffic validly and improve network performance.

The epidemic content dissemination[7] in MSNs has been widely researched over recent years for traffic offloading purpose. Zhang et al[8] and Li et al[9] have utilized a differentiation-based model to analyze the delay of popular content transfer and developed a high-performance content delivery framework. The work in Ioannidis[10] discusses the scalability and optimality of content transfer by exploring D2D contacts. Furthermore, the study in Watts et al[11] discovers that the "opinion leaders" plays a major part in information dissemination due to "word-of-mouth" effect,[12] and similarly, in social network formed by D2D file sharing, only a small number of users affect the sharing among most of other users. Therefore, leveraging seed users' social impacts to forecast content dissemination also becomes a study hotspot. An interesting study[13] analyzes the potential users for file transfer in MSNs by selecting the proper original users. Some researches have proposed to utilize probabilistic model to mine the activities of commenting and resharing information or contents to estimate the users' spreading influence.[14–16] In this paper, we also study the content dissemination by measuring the user behaviors, network dynamics, content properties, and social relationships based on the present significant research achievements in the literature.

2.2 | *Spark*-based social network analysis

Parallel processing techniques have been popularly used for processing huge amount of dataset effectively in recent years. *Spark* is an open source software framework to process large-scale data in a massively parallel way, having a developed directed acyclic graph execution engine, which supports cyclic data flow along with in-memory computing. Given that the social network dataset is fairly large, the number of records is up to a hundred million; it is necessary to introduce a distributed computing system with high performance. According to Cordeiro et al,[17] the MapReduce is suitable to be used to process SNS data, with the feature of high speed, reliability, and flexibility. The *Spark* runs machine learning like iterative programs much faster in either memory or disk. In addition, *Spark* provides a more abundant and convenient machine learning library and map and reduces Application Program Interface (API) that makes it possible to analyze social network data efficiently. There are also a number of SNS analysis studies based on *Spark*; for example, Yang[18] designed a social network analysis tools based on Hadoop-integrated environment, X-RIME, which realized prevalent social network analysis algorithms, including k-means, pagerank, and community detection. Xing[19] took a measurement on the feature of several social networks, verifying that the social network follows the power law and Rich-hub. According to the state-of-the-art framework, libraries, and algorithms, we then implement our *Spark*-based big data processing platform along with our customized and designed special functions.

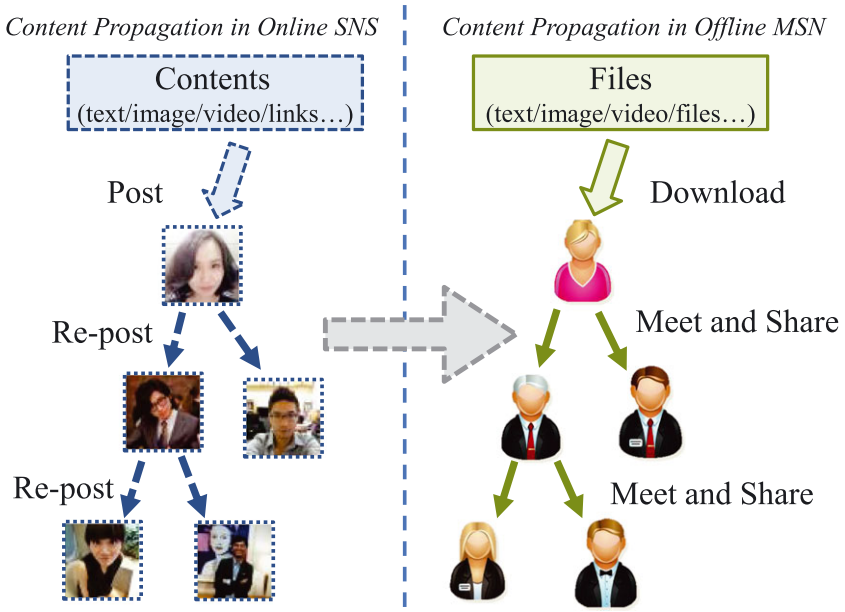


FIGURE 1 *Xender*'s device-to-device sharing-based mobile social networks

3 | DATASET AND PLATFORM

3.1 | Dataset details

Xender is a world-wide popular mobile APP for D2D content deliveries that provides users with the convenience of sharing various types of contents across a large diversity of mobile platforms (eg, Android, iOS, and Windows), without using 3G/4G cellular network infrastructures. The D2D connection in *Xender* is based on Wi-Fi tethering with a speed at around MBps, and transmissions are free as they utilize no mobile data. *Xender* has around 9 million daily and 100 million monthly active users, as well as about 110 million daily content deliveries.

Referring to the *Xender*'s related report, approximately 70% users are Indians for the relatively underdeveloped economic base and unavailable network. Therefore, in this paper, we concentrate on analyzing Indian users' measurement dataset. The basic descriptions of dataset are stated in the Table 1. We take 1 week to capture the trace from *Xender* for analysis, and there are 5 226 353 users sharing with each other for 90 715 241 times conveying 65 036 349 files. The data format is listed in Figure 2, and currently, we only pay close attention to *content type*, *content size*, *senderID*, *receiverID*, and so on to measure network states, content properties, social relationships, and so on.

TABLE 1 Details of dataset and platform

Dataset details	
Source	*Xender*
Periods	2016.02.01 - 2016.02.07
File size	153 GB
Platform details	
Nodes	20
CPU cores per node	4
Core frequency	1.4 Ghz
Memory per node	32 G
Storage per node	3 TB
Statistics	
User number	5 226 353
Activity number	90 715 241
File number	65 036 349

3.2 | Analysis platform

To analyze the very large-scale trace, we implement an effective parallel big data processing platform based on *Spark* along with customized algorithms in Python 2.7. Figure 3 shows our work flow for getting trace, obtaining general statistics, exploring advanced results (eg, social graph and content properties), and thus plotting/visualizing. The *Spark* big data platform includes 1 master node and 19 slave nodes. Each node is assembled with 4 CPUs, 32-GB memory spaces and 3-TB disk storage spaces. All those nodes are connected by a Gigabit Ethernet switch. Table 1 shows the hardware configuration of the platform.

We deployed *Spark* version 1.6 into the platform. According to the default configuration of *Spark*, we set the file block size of Hadoop Distributed File System to 64 MB, and the total copies of each file block to 3. To reduce the computing resource competition between concurrent running tasks, we put limitation on the total number of those tasks running on the same core. After considering the possible quantitative relation between map tasks and reduce tasks, we configured each node to simultaneously run 2 tasks.

4 | TIME SERIES ANALYSIS

In this section, we analyze *Xender* data on the basis of time to check daily user behaviors[20] and network traffic for every hour. We take this measurement from 5 main aspects including the statistics of Sharing Activities, Online Individuals, Sharing File, Traffic Load and Duplicated Traffic Load, for each hour in the week (168 hours in total). Then we collect the results and plot their time series figures as shown below.

Activity ID	Content Name	Content Type	Content Size	Sender ID	Receiver ID
500023	Instapic.apk	app	29502160 (Byte)	16b4467f03b10f	a3558b05f2386

Sender Package ID	Receiver Package ID	Receiver OS Language	Resource ID	Receiver Report Time
cn.xender	cn.xender	en-US	com.xender.instapic	1454903126367

Receiver Write Time	Receiver Report IP	Reciever Report MAC	3rd Party Account	3rd Party Source
1454845327520	27.154.121.237	BD:19:AD:3C:B6:44	1454903126367	weibo

Receiver OS Nation	Receiver GPS	Sender OS	Sender Phone Brand	Sender Phone Model
US	39.26247, 41.03254	Android	Samsung	GT-I9003

Receiver Phone Brand	Receiver Phone Model	Receiver Market Channel	Receiver Market Channel ID
Samsung	GT-I9003	andouya_google	245

Receiver Xender Version	Sender Market Channel	Sender Market Channel ID	Sender Xender Version
1.24.002	andouya_google	245	1.24.002

FIGURE 2 Data entry format from *Xender* trace

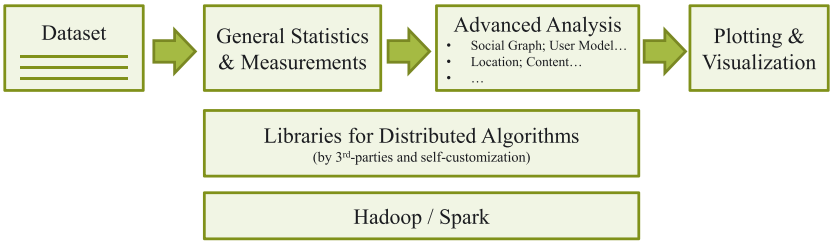


FIGURE 3 Illustration of our measurement and analysis working flow

4.1 | Sharing activities

Figure 4A is the statistics of the sharing activities over time, which shows the number of sharing events in each hour for the whole 24*7 period of the week. Note that the Y axis is in the log scale, and thus, it easily tells that, in the weekend, there are much more users to share contents than that in workdays. Also during each day, users become very active in the noon time as well as in the night. It is worth mentioning that, in India, families and friends always get together in the weekend, so groups of people may enjoy sharing interesting contents, and therefore, generally the number of sharing activities in weekend can be up to 2 to 3 times that in workdays, and the number of sharing activities in the evening can be even up to 10 times that in the morning time. Another interesting finding is that the sharing activities of multimedia files (ie, videos, audio, and images) take a large portion, which implies that the mobile users in India have a large potential to enjoy mobile multimedia services, but because of the not well-developed infrastructured cellular networks in most of the areas, people mostly use D2D sharing methods (eg, *Xender*) to share multimedia contents.

4.2 | Online individuals

We then analyze from the aspect of the number of online users, and users who share contents for many times in the time window are only counted for once. As shown in Figure 4B, still the user base for sharing videos takes a large portion. Also there are huge amount of users that share mobile APPs. Because of the inconvenient mobile APP market services via the mobile networks in India, many people rely on sharing interesting APPs to others via D2D sharing activities. And hence, sharing services (eg, *Xender*) becomes a more effective way for APP marketing strategies and video content distribution methods, which indicates another potential promotion methods for APP and video service providers to attract many potential users in the platform. There is no obvious difference among audio, images, and music. And we think that the audio files include music, which are just diverse types of music and organized, respectively, in *Xender*.

4.3 | Shared (involved) file

We then carry out measurement and analysis over the number of involved files with aggregate amount of all types and those of different types, respectively. If 1 file is shared by many people, it is only counted for once. As show in Figure 4C, multimedia types have nearly similar number of files, which is different from Figure 4B, many images are delivered with *Xender*, especially on weekends when users tend to hold get-togethers. And then the number of APPs shared via *Xender* is large as well, which reveals that APP has played a vital role in our daily life, even in India where cellular networks are not well equipped. There is very limited sharing of contents with other types that should be paid more attention.

4.4 | Traffic load

Figure 4D reveals the variant rules of traffic load, which is the total size of evolved content files at an hour window as time goes on. It is clear that videos take most of the traffic load; in other words, *Xender* (and similar D2D sharing platforms) helps the mobile networks offload many

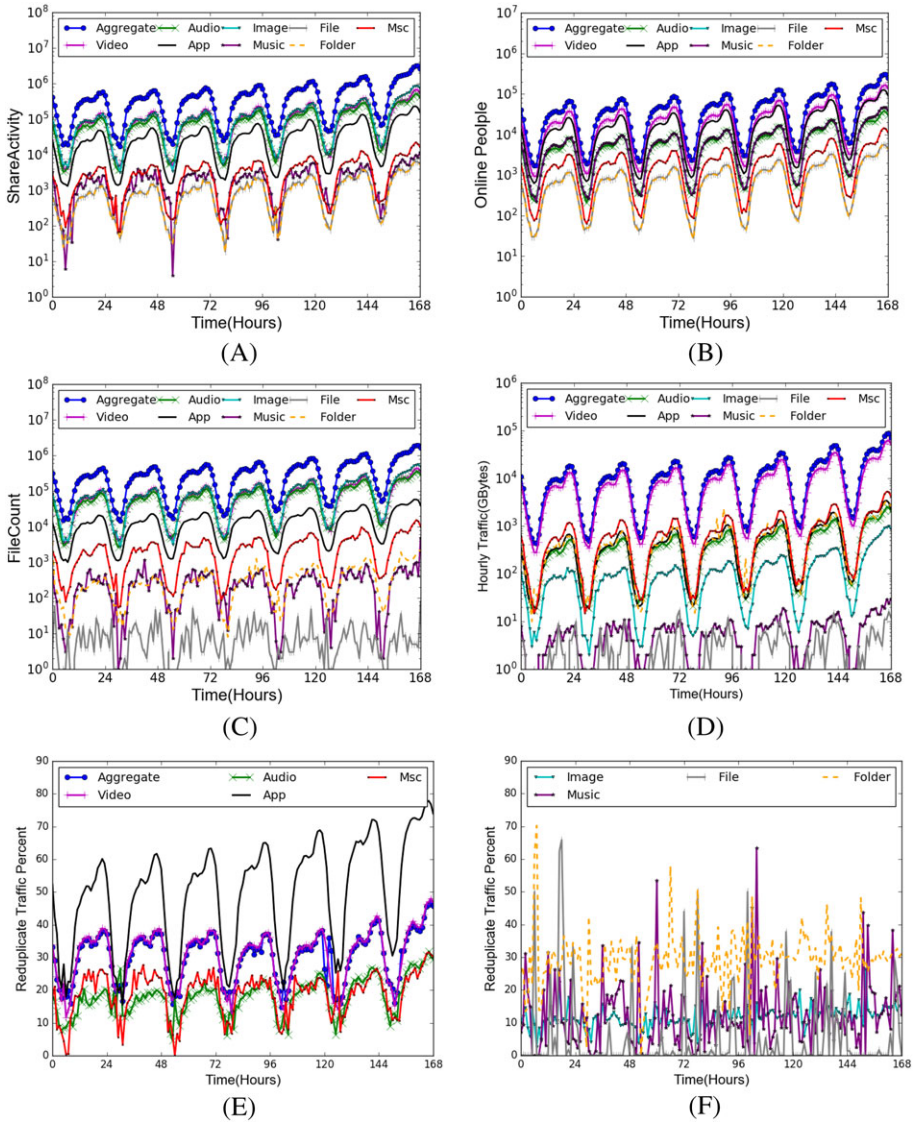


FIGURE 4 Time series statistics. A, Statistic for all sharing activities. B, Statistic for count of all involved people. C, Statistic for count of all involved files. D, Statistic for total traffic load. E, Percentage of redundant traffic load to total traffic for type: aggregate, video, application, audio, and miscellaneous. F, Percentage of redundant traffic load to total traffic for type: image, music, file, and folder

videos into D2D sharing activities because the sharing is free and fast. And hence, the service may emphasize more on improving the QoS and QoE of sharing video contents. Note that, on Sunday, the traffic load goes to up to 5 times more than that in work days. Therefore, *Xender* may effectively initialize video-focused events for improving the satisfaction of mobile users more conveniently. Moreover, the Msc. (miscellaneous: all other types that cannot be classified) also account for relatively much traffic load; hence, we will further analyze the part of files to verify if they are valuable to reduce traffic load. Comparing with other four results, folders show a clear feature in this figure that they occupy relatively much traffic due to the fact that they are very large, including many compressed files, although the number of them is small.

4.5 | Duplicated traffic load

As we have already presented in Section 1, there is a large portion of redundant traffic in current mobile networks. Therefore, it is of great importance of investigating redundant traffic load via the offline D2D sharing activities in MSNs. It is worth noting that the redundant traffic load is computed by comparing with its own type of files instead of all the files. As shown in Figure 4E, approximately 10% to 40% traffic load shared via *Xender* is redundant, which means a certain large percentage of mobile users' request and obtain similar popular contents via *Xender*. We further check the redundant traffic ratio of each type of contents. An interesting finding is that the percentage of duplicate traffic for APP is the largest, at 60%, which reveals there are many users being attracted by similar popular APPs and it is necessary to offload these redundant traffic. In addition, it is quite obvious that there are also a large quantity of redundant video content transmissions almost the same with total one, which means many users always watch the same video contents, and hence, we can actively push popular and trending videos to mobile users by related effective methods, eg, TOSS,[4,21] and let them share with each other for improving the user loyalty. Also from Figure 4F, it is shown that each type of contents may have the similar ratio of redundant traffic, but in many cases, there are peaks of the percentage and the pattern of the curves has limited time-scale similarity for the days in the week, which may be caused by some large scale of extensive sharing activities among small amount of users. *Xender*

needs to pay attention to improving the QoS and QoE of those sudden extensive transmissions. Video contents have the most stable values of duplicate traffic ratio and account for the most traffic during the week. Therefore, more investigations will be put to optimize video sharings via D2D in MSNs.

5 | CONTENT PROPERTIES

5.1 | Content size

As for *Xender*'s D2D sharing platform, it is challenging to carry out effective file storage and transmission techniques to optimize the performance regarding transmission bandwidth and security, and hence, the size of contents (file objects) should be analyzed in detail. We plot the probability density function of file sizes for each type of contents and aggregate contents in Figure 5, and thus, researchers can simply apply related fitting model with the empirical parameters to fast carry out modeling-based analytical research for content sharing in MSNs, which will be put to the enhancement of our upcoming study. It is obvious that maximal content size is 2 GB, demonstrating the limitation of mobile APPs and mobile operation systems; as we have guessed, the videos are the largest files. This provides very useful implication for future implementations to optimize necessary segmentation techniques for efficient D2D transmission protocols.

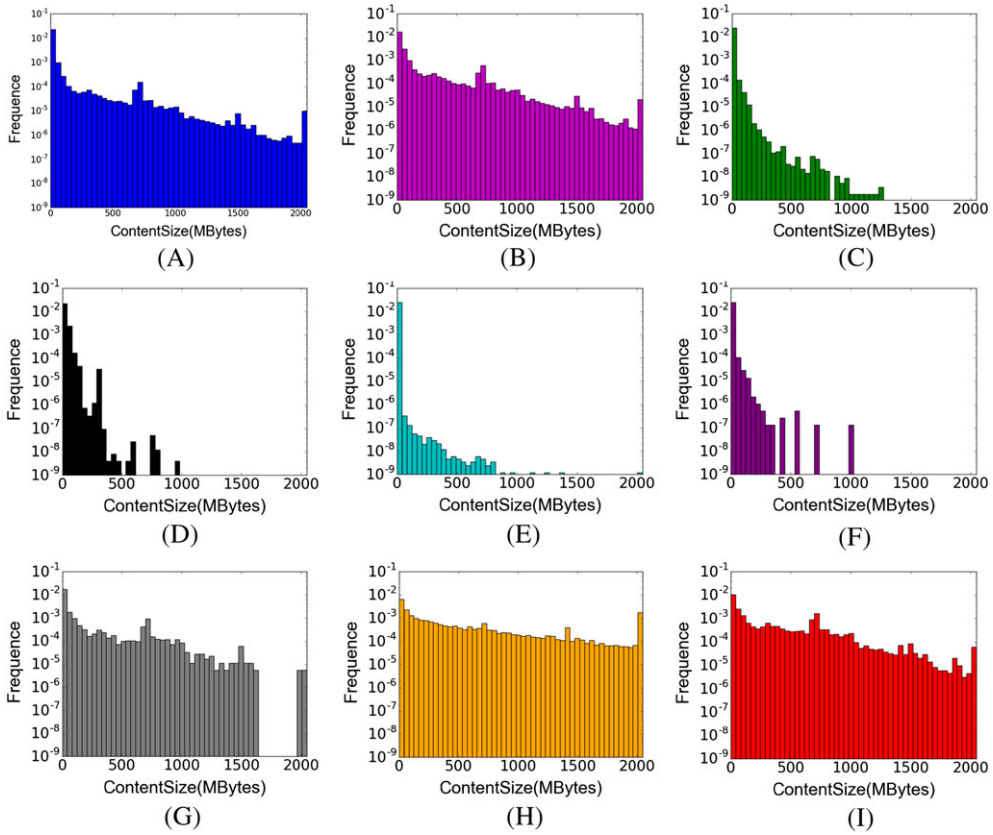
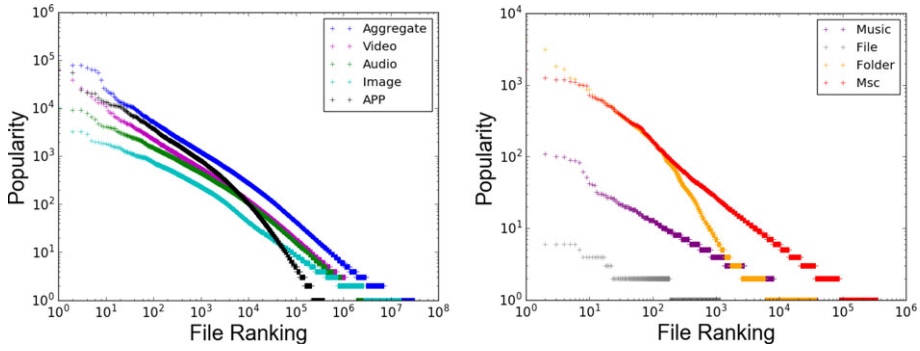


FIGURE 5 Probability density functions for different types of contents. A, Aggregate. B, Video. C, Audio. D, Application. E, Image. F, Music. G, File. H, Folder. I, Miscellaneous



(A) Content popularity power-law characteristics (aggregate, video, audio, image, APP)

(B) Content popularity power-law characteristics (music, file, folder, Msc.)

FIGURE 6 Content popularity. A, Content popularity power-law characteristics (aggregate, video, audio, image, and application). B, Content popularity power law characteristics (music, file, folder, and miscellaneous.)

5.2 | Content popularity

It is studied in Yang et al and Chaoji et al[22,23] that the contents in the Internet have very skewed popularity disparity, which means a very small number of contents always attract a huge number of individuals and occupy the majority of traffic load. We analyze the content popularity based on the trace and demonstrate results with log-log plots. As shown in Figure 6A,B, contents shared via *Xender* service have strong property of power law, if we analyze aggregate contents or any of specific type of contents.

We further carry out fitting of the power law plots by maximum likelihood estimation method and obtain the α factor (ie, the slope) and *Xmin* of the curve. The results are shown in Table 2. Therefore, this phenomenon may be exploited for concentrating the system resource optimization and user satisfactory improvement by focusing on a small amount of popular contents that can be discovered by our analysis. For examples, we can cache these files in the servers of content delivery networks to improve transmission quality and reduce the duplicate Internet traffic load.

5.3 | Content preference (entropy) of users

Different users have different tendencies of exchanging different types of contents via *Xender* service. And hence, we analyze users' content preferences (diversity index) by evaluating their *Shannon entropy* values, which describe how evenly sharing activities are distributed among file types for each user. Mathematically, the Shannon entropy H_i of each user u_i can be computed as $H_i = -\sum C_{iy} \log C_{iy}$, where Ω is the set of all 8 content types, $y \in \Omega$ is the type index, and C_{iy} is the probability of sharing contents with type y in the history of user u_i.

TABLE 2 Power law fitting of content popularity

Type	Aggregate	Video	Audio	APP	Image
α	2.619	2.656	2.88	2.518	2.56
Xmin	723	554	416	727	9
Type	Music	File	Folder	Msc.	
α	3.20	7.12	3.36	2.43	
Xmin	7	2	186	16	

APP, application; Msc, miscellaneous.

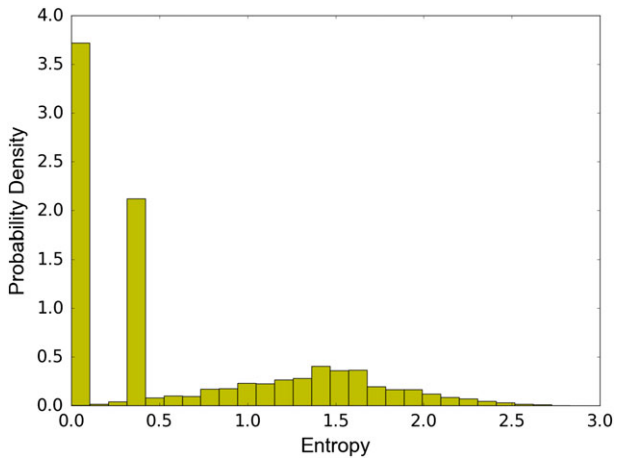


FIGURE 7 Content preference (entropy) of users

The probability density of content entropy values of all users are shown in Figure 7. It can be seen that lots of users have 0 content entropy, which means they always exchange 1 type of contents. This observation has a great potential for content recommendation and pushing strategies in the future.

Notably, there is also a peak of entropy values between 0.2 and 0.3, and as we estimated, lots of users have been involved of sharing contents with 2 types almost equally, which induces entropy values from 0.25 to 0.3. For other users, uniformly distributed content preferences are observed, while more users have entropy values around 1.5 at the most. This represents users with quite equal preference of all 8 types.

6 | SOCIAL-RELATED MEASUREMENT

As researched in many related studies,[4,24–26] offline D2D sharing of contents forms an opportunistic network, where the social links among users are represented by their sharing activities. Therefore, we carry out the analysis on user social characteristics in *Xender*'s social network. As shown in Figure 8, we illustrate the social graphs of 2 representative interconnected users (groups), where each node is for a user and each link represents sharing activities among a pair of 2 users. It is interesting that there are various types of social graphs (groups); in Figure 8A, 1 user sends to and receives from many users who instead never share with each other, but in Figure 8B, users share a lot with others tightly.

6.1 | In-degree and out-degree

We investigate the in-degree and out-degree of each user, while the degrees are actually the numbers of sharing to others and being shared by others. We plot the log-log figure of the ranking of the sender popularity (ie, out-degree) and receiver popularity (ie, in-degree) in Figure 9A,B, respectively, from which we can see clear trends of power law of the most parts of the curves but we also notice a sudden drop of the tail.

We implement fitting algorithm and obtain the slope (α factor) of the power law fitting with values of 3.29 and 3.54, which are quite skewed. This indicates that a very small number of users may have very strong capability of sharing content with others. If we deploy related offloading schemes for propagating contents, these users are considered as initial users with strong spreading impacts. It is also implied that a huge amount of users have infrequent requests of obtaining contents. According to the measurement results, services providers are able to mine these active users, gain their needs, and accelerate their activities to improve the QoS. Those social impact values are important for accelerating the sharing of contents and, along with fitting parameters, can be utilized for predicting user sharing probability, etc.

6.2 | Vertices and edges of social groups

We further obtain a set of nonoverlapping social groups (ie, 2 vertices in 2 social groups share no end-to-end path between them) and carry out analysis from the perspective of complex networks. Accord-

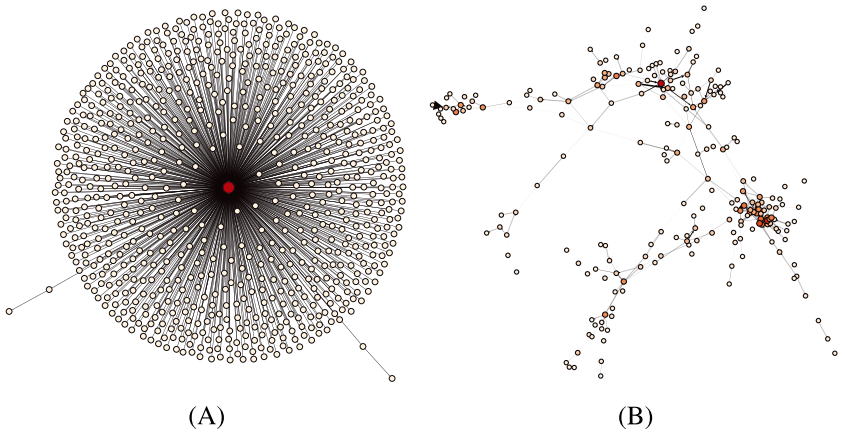


FIGURE 8 Two examples of the social graphs in *Xender*. A, Example 1 – social graph with 770 users. B, Example 2 – social graph with 443 users

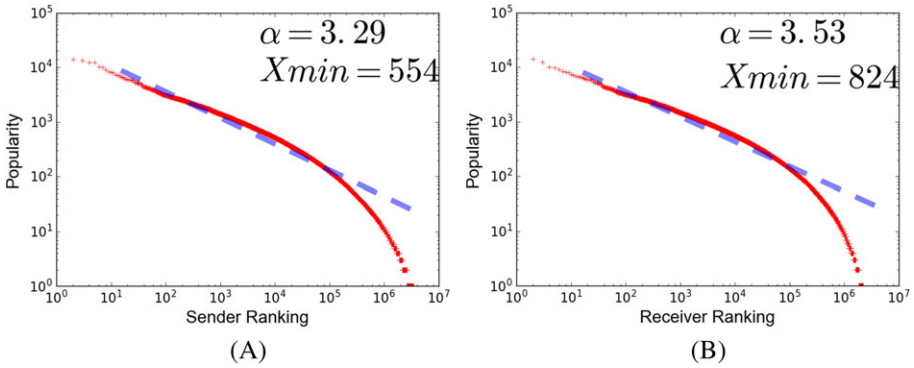


FIGURE 9 Content popularity. A, Popularity of sender (out-degree of all users). B, Popularity of receiver (in-degree of all users)

ing to the 2 example social groups in Figure 8, the maximal group sizes (number of vertex) are just 770 and 443, respectively, which are quite small compared with the total size of user base. This is a very important observation, which implies that MSNs for offline D2D sharing cannot support the extensive growth of social networks, because of time and space restrictions. From Figure 10A,B, we can see that the number of both vertex and edges perfectly fits power law pattern, which indicates there are a lot of small groups. Figure 10C shows that, as the vertex number increases, the edge number increases at approximate linear speeds rather than superlinear speeds. This implies that the connectivities of friends' friends are not very tight, and the group is not often closely clustered.

6.3 | Social clustering analysis

Because the measurement of user types and friendships can be used for recommendations and predictions, it is important to discover the social properties by further investigating social clustering[27] of user types and pair relationships. Specially, we define that a user can be any type among *sharing-focused*, *receiving-focused*, and *balanced*. And then we carry out k-means grouping method over the user base, where $K = 3$ at first. It is shown in Figure 11A that individuals who send and receive around less than 300 contents with others are grouped to *balanced* user. As for users who send more than 300 times and receive fewer files than sending are regarded as *sharing-focused* user, contributing to sharing. On the other hand, the number of receiving activities that *receiving-focused* users experience is over 300 and exceeds the number of sending activities. The coordinates of the cluster centers are (25,35) for balanced, (66,782) for receiving-focused, and (705,95) for sharing-focused.

We also define the following 3 types of relations for each pair of users: *close friends*, *normal friends*, and *unfamiliar friends*. Then, we are able to improve all pairs' satisfaction with certain promotion strategies.[28] We still carry out k-means grouping method over the base of pairs.[29,30] To ensure the correct clustering results and clear visualization, we follow the following rules: For each pair (*a*,*b*), where *a* is for the number of files shared from 1 user to the other and *b* is for the number of files shared reversely, if *a* is larger than *b*, we exchange *a* and *b*. As shown in Figure 11B, the X axis is *a* of all the pair and Y axis is *b* of all the pair. It can be seen that a small set of users are close friends, while the majority of users are unfamiliar friends. More specifically, when the unidirectional interactions between pairs are fewer than 200, we can judge they are unfamiliar friends with infrequent contacts. In contrast, close friends always share more than 1000 contents and the remaining pairs belong to normal friends group. The coordinates of the cluster centers are (20, 4) for unfamiliar friends, (342, 34) for normal friends, and (1564, 99) for close friends.

Furthermore, if we set $K = 5$ for the grouping (classifying) users and pairs, the figures are shown in Figure 11C,D. The 2 cluster results reveal users' properties and social relationships in more detail, and when we are engaged in improving group-oriented service quality or predict user behavior, we can adjust the number of friendship clusters (K) to find out the most efficient clustering results for helping the user rec-

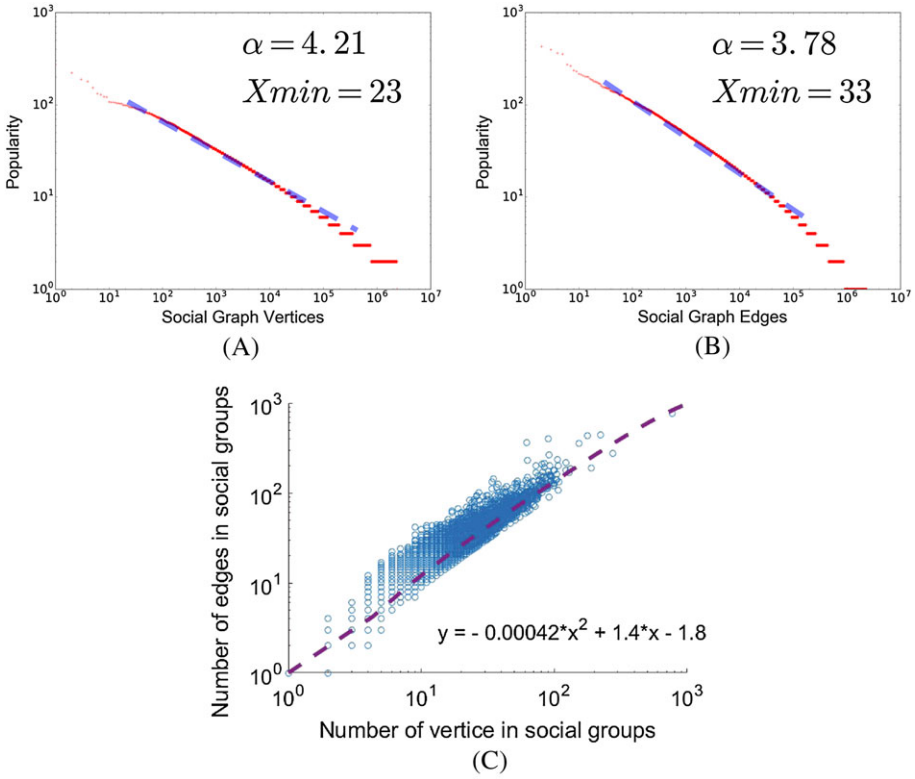


FIGURE 10 Measurement results of social group properties. A, Vertices of groups. B, Edges of groups. C, Edges vs nodes

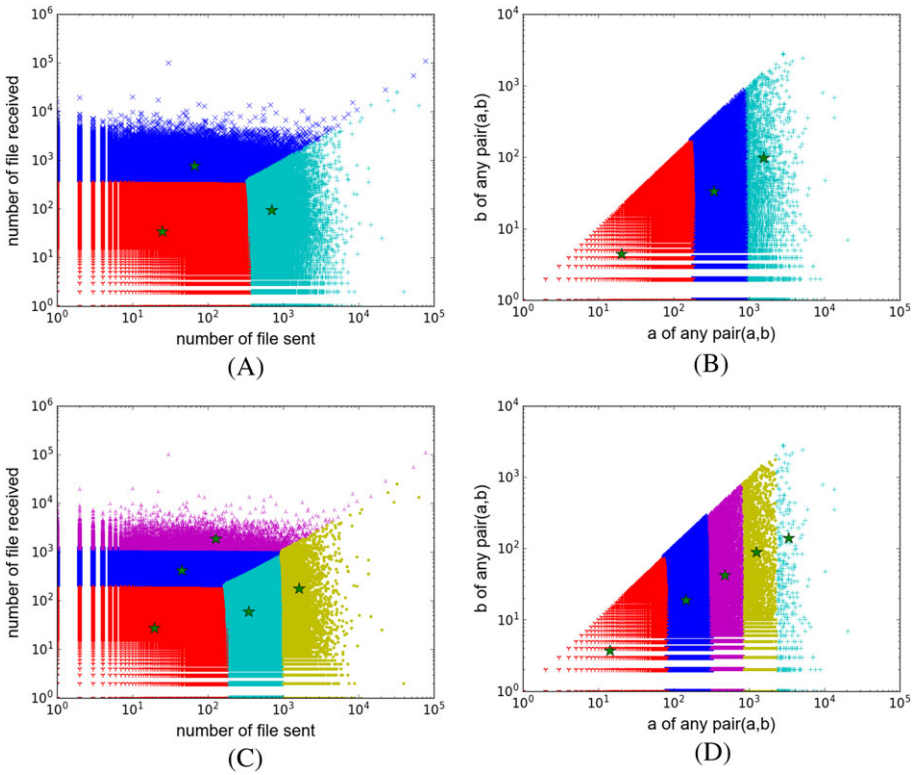


FIGURE 11 Grouping (classifying) users and pairs by k-mean clustering algorithm in *Spark*. A, Cluster results of user types, $k = 3$. B, Cluster results of user pair relations, $k = 3$. C, Cluster results of user types, $k = 5$. D, Cluster results of user pair relations, $k = 5$

ommendation and for improving users' loyalty. This kind of grouping (classification) methodologies further provides effective strategies for improving user QoS and QoE via differentiating users and applying effective marketing strategies not by arbitrary values of the sharing or receiving account but by reasonable ones for inducing all users and pairs to get into the most related group of relationships or user types.

7 | DISCUSSIONS AND CONCLUSIONS

In this paper, towards the concept of offloading the increasing mobile traffic into the MSNs among groups of mobile users, we carry out a practical study with trace-based measurement and analysis of a famous D2D sharing service, *Xender*, from the perspectives of general time series statistics, content properties, social graph basics, and so on. Our trace has a large user base with huge amounts of sharing activities, content files, and so on, and hence, we implement a parallel big data processing platform for carrying out the measurement and analysis based on *Spark* along with customized algorithms in Python. We further discuss the potential issues of improving *Xender*'s QoS and optimizing its platform resource, eg, promotion of popular video contents by the most influencing users, and eternal system optimization during work days but marketing acceleration during weekends, and so on. To the best of our knowledge, this is the first empirical study of a large-scale offline social D2D sharing service. In the future, we will focus on insightful research over the following topics: (1) tight integration with advanced D2D communication techniques[31]; (2) development of more efficient and scalable algorithms for analyzing the large-scale trace directly from *Xender* by streaming techniques; (3) analysis based on complex networks theories and exploitation of users' social properties; and (4) effective methodologies of modeling and predicting content popularity trends and user sharing activities[32].

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (61372085), National Key Research and Development Program of China No. 2016YFB1000205, the State Key Program of National Natural Science of China (Grant No. 61432002), NSFC Grant Nos. 61370199 and 61672379.

REFERENCES

1. CISCO Cisco Visual Networking Index. Global Mobile Data Traffic Forecast Update, 2014-2019, CISCO. Tech. Rep.; 2014.
2. Cha M, Kwak H, Rodriguez P, Ahn Y, Moon S. I Tube, You Tube, Everybody Tubes: analyzing the world's largest user generated content video system. *ACM IMC*, San Diego; 2007:1–14.
3. Device-to-device communications in 3GPP LTE standard, release 12. http://www.3gpp.org/specifications/releases/68-release-12. Accessed June 1, 2016.
4. Wang X, Chen M, Han Z, Wu D, Kwon T. TOSS: traffic offloading by social network service-based opportunistic sharing in mobile social networks. *IEEE INFOCOM*, Toronto; 2014: 2346–2354.
5. Steeg G, Galstyan A. Information transfer in social media. *WWW*, Lyon; 2012: 509–518.
6. AirDrop, Apple Inc. http://en.wikipedia.org/wiki/AirDrop. Accessed June 5, 2016.
7. Wang X, Chen M, Han Z, Kwon T, Choi Y. Content dissemination by push & share in mobile cellular networks. *MASS*, Las Vegas; 2012:353–361.
8. Zhang X, Neglia G, Kurose J, Towsley D. Performance modeling of epidemic routing. *Comput Networks*. 2007;51(10):2867–2891.
9. Li Y, Jiang Y, Jin D, Su L, Zeng L, Wu D. Energy-efficient optimal opportunistic forwarding for delay-tolerant networks. *IEEE TVT*. 2010;59(9):4500–4512.
10. Ioannidis S, Chaintreau A, Massoulie L. Optimal and scalable distribution of content updates over a mobile social network. *IEEE INFOCOM*, Rio de Janeiro; 2009: 1422–1430.
11. Watts D, Dodds P. Influentials, networks, and public opinion formation. *J Consum Res*. 2007;34(4):441–458.
12. Rodrigues T, Benvenuto F, Cha M, Gummadi K, Almeida V. On word-of-mouth based discovery of the web. *ACM IMC*, New York, NY, USA; 2011: 381–396.
13. Han B, Hui P, Kumar V, Marathe M, Shao J, Srinivasan A. Mobile data offloading through opportunistic communications and social participation. *IEEE TMC*. 2011;11(5):821–834.
14. Comarela G, Crovella M, Almeida V, benevenuto F. Understanding factors that affect response rates in Twitter. *ACM HT*, Milwaukee; 2012:123–132.
15. Yang J, Leskovec J. Patterns of temporal variation in online media. *ACM WSDM*, Hong Kong; 2011: 177–186.
16. Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks. *WWW*, Raleigh; 2010: 981–990.
17. Cordeiro R, Traina C, Traina A, Lopez J, Kang U, Faloutsos C. Clustering very large multi-dimensional datasets with MapReduce. *KDD*, San Diego;2011: 690–698.
18. Yang Y. X-rime: Hadoop-based large-scale social network analysis. *IC-BNMT*, Beijing; 2010: 901–906.
19. Xing L, Lv L. Analysis of the characteristics of social network based on Spark. *J Pingdingshan Univ*. 2014;29(5):80–83.
20. Gyarmati L, Trinh T. Measuring user behavior in online social networks. *IEEE Network Mag*. 2010;24(24):26–31.
21. Yu C, Doppler K, Ribeiro C, Tirkkonen O. Resource sharing optimization for device-to-device communication underlaying cellular networks. *IEEE TMC*. 2011;10(8):2752–2763.
22. Yang S, Adeel U, McCann J. Selfish mules: social profit maximization in sparse sensornets using rationally-selfish human relays. *IEEE J Sel Areas Commun*. June 2013;6:31.
23. Chaoji V, Ranu S, Rastogi R, Bhatt R. Recommendations to boost content spread in social networks. *WWW*, Lyon; 2012: 529–538.
24. Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? *WWW*, Raleigh; 2010: 591–600.
25. Toole J, Herrera-Yaque C, Schneider C, Gonzalez M. Coupling human mobility and social ties. *J R Soc Interface*. April 2015;12(105):266–271.
26. Wang X, Sheng Z, Yang S, Leung V. Tag-assisted social-aware opportunistic device-to-device sharing for traffic offloading in mobile social networks. *IEEE Wirel Commun*. August 2016;23(4):60–67.
27. Brown C, Lathia N, Mascolo C, Noulas A, Blondel V. Group colocation behavior in technological social networks. *PLoS ONE*. August 2014;8:9.
28. Li H, Chen Y, Cheng X, Li K, Chen D. Secure friend discovery based on encounter history in mobile social networks. *Pers Ubiquitous Comput*. 2014;19(7):999C1009.
29. Brown C, Nicosia V, Noulas A, Mascolo C. Social & place-focused communities in location-based online social networks. *Eur Phys J B*. June 2013;86(6):C10.
30. Wu C, Chen X, Zhou Y, Li N, Fu X, Zhang Y. Spice: socially-driven learning-based mobile media prefetching. *IEEE INFOCOM*, San Francisco; 2016: 2142–2150.
31. Li X, Ge X, Wang X, Cheng J, Leung V. Energy efficiency optimization: joint antenna-subcarrier-power allocation in OFDM-DASs. *IEEE Trans Wirel Commun*. 2016;PP(99):1–1.
32. Qiu T, Chen N, Li k, Qiao D, Fu Z. Heterogeneous ad hoc networks: architectures, advances and challenges. *Elsevier Ad Hoc Networks*. 2017;55:143–152.

How to cite this article: Wang H, Wang X, Li K, Ren J, Zhang X, Jiang T. A measurement study of device-to-device sharing in mobile social networks based on *Spark*. *Concurrency Computat: Pract Exper.* 2017;29:e4021. https://doi.org/10.1002/cpe.4021