Serendipity of Sharing: Large-scale Measurement and Analytics for Device-to-Device (D2D) Content Sharing in Mobile Social Networks

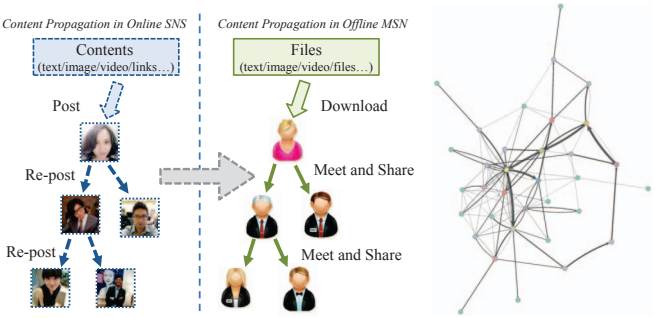
Xiaofei Wang[1], Hui Wang[1], Keqiu Li[1], Shusen Yang[2], Tianpeng Jiang[3]
[1]Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology,
Tianjin University, Tianjin, China
[2]National Engineering Laboratory for Big Data Algorithms and Analytics Technology, Institute of Information and System Science,
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
[3]Beijing Anqi Zhilian Technology Co. Ltd., Beijing, China
{xiaofeiwang, wanghui1994, keqiu}@tju.edu.cn, shusenyang@mail.xjtu.edu.cn, peter.jiang@shanchuan.cn

***Abstract*—The heavy multimedia traffic produced by mobile users poses great challenges for the mobile network operators, especially in the areas with large user densities but limited cellular network capacities (e.g. India). Recently, many studies demonstrate that exploiting the device-to-device(D2D) content sharing in offline Mobile Social Networks is a promising solution to cellular data offloading. However, such approaches are based on either unrealistic assumptions, or limited data analytics caused by small data size (e.g. hundreds of MSN users) or single-dimensional feature (e.g. human mobility only), which severely restricts their applications in practice. To address this issue, this paper performs the *first* large-scale data measurement and multi-feature analytics of D2D content sharing. Specifically, by using *Apache Spark* over a 20-server cluster, we analyze the behaviours of 30 million users (with 40 billion D2D transmissions and 16 million content files) of *Xender*, a leading global D2D sharing platform. Several important features are studied, including performance basics, content properties, location relations, meeting dynamics, and social characteristics. Furthermore, as a proof-of-concept study of our analytics, we also develop a multi-feature learning based framework, which demonstrates the large potentials of predicting and recommending D2D sharing activities using machine learning methods.**

I. Introduction

Recently, the demands for rich multimedia services over mobile networks have kept soaring at a tremendous pace [1]. The mobile traffic load explosion poses great challenges for the communication infrastructure (e.g. WiFi and Cellular Networks) of mobile networks operators (MNOs), and severely undergrades user experiences such as large access delay, slow downloading speed, and even frequent disconnections in peak hours. In fact, related studies [1]–[3] demonstrate that traffic explosion is mainly caused by the duplicate downloads of prevalent multimedia files, e.g., popular videos attract huge amount of downloads. For instance, Cha *et al* [2] reports that about top 10% of YouTube videos occupy nearly 80% views. An effective way to reduce such duplicate downloads is to cache and share the multimedia contents among geographically proximal mobile devices [4] through device-to-device (D2D) communications (e.g. WiFi Direct, Bluetooth, and LTE-direct [5]). In doing so, each user is likely to obtain popular contents from nearby devices, and only need to use the expensive cellular to download the contents that are unavailable in proximity, resulting in cellular traffic offloading.

It has been shown that exploiting the social behaviors of mobile users in Mobile Social Networks (MSNs) can significantly improve the efficiency of D2D sharing and cellular offloading, where socially-close users can exchange contents during frequent encounters, without relying on MNOs' communication infrastructures [6]–[8]. For instance, [7] *et al* reports that 86.5% cellular offloading can be achieved, by exploiting the social behaviors of mobile users. In fact, the D2D sharing approaches highly depend on various features of the complex mobile user behaviors, such as temporal and spatial dependency, underlying social network structure, delay tolerance, and content popularity and distributions, etc. However, most current approaches [9], [12], [14] are based on either unrealistic assumptions, or limited data analytics caused by small data size (e.g. hundreds of MSN users) or single-dimensional feature (e.g. large-scale human mobility measurement [15], [16]), which severely restricts their applications in the real-world D2D sharing scenarios.



(a) D2D sharing in *Xender* MSN *v.s.* content spreading in online SNSs (b) An example social graph in *Xender*

Fig. 1: *Xender*'s D2D sharing-based mobile social networks

Therefore, the collection and analytics of *large-scale* and *multi-dimensional* of real-world D2D sharing data trace are essential to deeply understanding the real-world user behaviors in D2D sharing, and to provide meaningful implications and

978-1-5090-6599-8/17/$31.00 ©2017 IEEE

guidelines for D2D sharing algorithms. However, there exists no such work yet. To bridge this gap, this paper presents comprehensive measurement and analysis based on *Xender*, one of the leading mobile applications (APP) for D2D contents sharing in the world (the largest one in India). As shown in Fig. 1(a), D2D sharing activities among *Xender* users form an offline MSN, which has similar content delivery pattern and social properties as online Social Network Services (SNSs) [11]. The contributions of this paper are two folds:

1) We collect one-month data from *Xender* server logs, which cover over 30 million active users, 40 billion D2D transmissions, and 16 million content files. We perform efficient and reliable analytics by running the distributed big data processing platform *Apache Spark* in a 20-server computing cluster. Five important features have been studied, including time series of performance basics, content properties, location relations, meeting dynamics and social characteristics. To the best of our knowledge, *this is the first work on both large-scale and multi-feature data analytics for real-world D2D sharing*.
2) Our analysis results demonstrate several meaningful implications and guidelines (e.g. The activities of offline D2D sharing are much more intense than other online services at weekend), which can be exploited to improve the quality of D2D sharing service. As a proof-of-concept study of our analytics, *we also develop an initial multi-feature learning based framework to predict the D2D sharing activities, using the Support Vector Machine (SVM) methodology*, which achieves reasonable prediction accuracy. The initial results demonstrate the great potentials of using machine learning methods for recommending accurate future D2D sharing services.

The organization of this paper is as follows. Sec. II presents related work. *Xender* dataset and the *Spark* platform are described in Sec. III. Then, we present the measurement and analytics results, including time series in Sec. IV, content properties in Sec. V, location relations in Sec. VI, meetings dynamics in Sec. VII and social characteristics in Sec. VIII. Our machine learning based nowcasting framework is developed in Sec. IX, and we conclude this paper in Sec. X.

II. Related Work

Exploiting D2D communications for MSN content sharing not only attracts an increasing research interests from academia [9]–[11], [13], but also promotes a number of mobile applications such as Apple's Airdrop [17] and *Xender*. Many studies focus on the D2D epidemic content dissemination in MSNs over recent years for traffic offloading purpose. For example, Zhang et al. [18] and Li et al. [19] have utilized a differentiation-based model to analyze the performance of popular content sharing in MSNs. However, none of them tests over a large scale user base but a small group of people.

There exist several measurement-based studies for online SNSs and offline MSNs. The studies in [13], [20], [21] show the homophily and locality characteristics of users are observed in both MSNs and SNSs, which can be utilized to facilitate the D2D content dissemination [21]. In addition, Kwak *et al* [3] points out that there are obvious delays of re-sharing behaviors, while the spreading impact of each user is accumulated hop by hop. Such observation results enable the analysis, modeling, and prediction of the sharing activities and the content spreading of SNS users based on measurement traces [11], [22], [23]. There also exist large-scale measurements for human mobility [15], [16] that are related to our work. In reality, D2D sharing in offline MSNs is much more complex than information sharing in online SNS, due to temporal and spatial constraints [20], [23]. However none of aforementioned studies focuses on multi-dimensional feature analytics on large-scale D2D sharing data.

III. *Xender* Data and Processing Platform

Xender is a world-wide popular mobile application (APP) for D2D content deliveries, which provides users with the convenience of sharing various types of contents across a large diversity of mobile platforms(e.g., Android, iOS, and Windows), without using 3G/4G cellular network infrastructures. The D2D connection in *Xender* is mostly based on Wi-Fi tethering, while WiFi Direct and Bluetooth are also supported. 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.

We capture *Xender*'s trace for an one-month period from 01/02/2016 to 28/02/2016. According to our estimation, around 70% users are from India, which is probably due to their limited cellular network services. This motivates us to concentrate on analyzing Indian users in this paper. After cleaning unnecessary and invalid data (e.g., incomplete format, files with zero size, users with fake IMEI identity), there are in total 30,485,335 users with 4,434,440,043 times of transmissions conveying 16,785,175 content files.

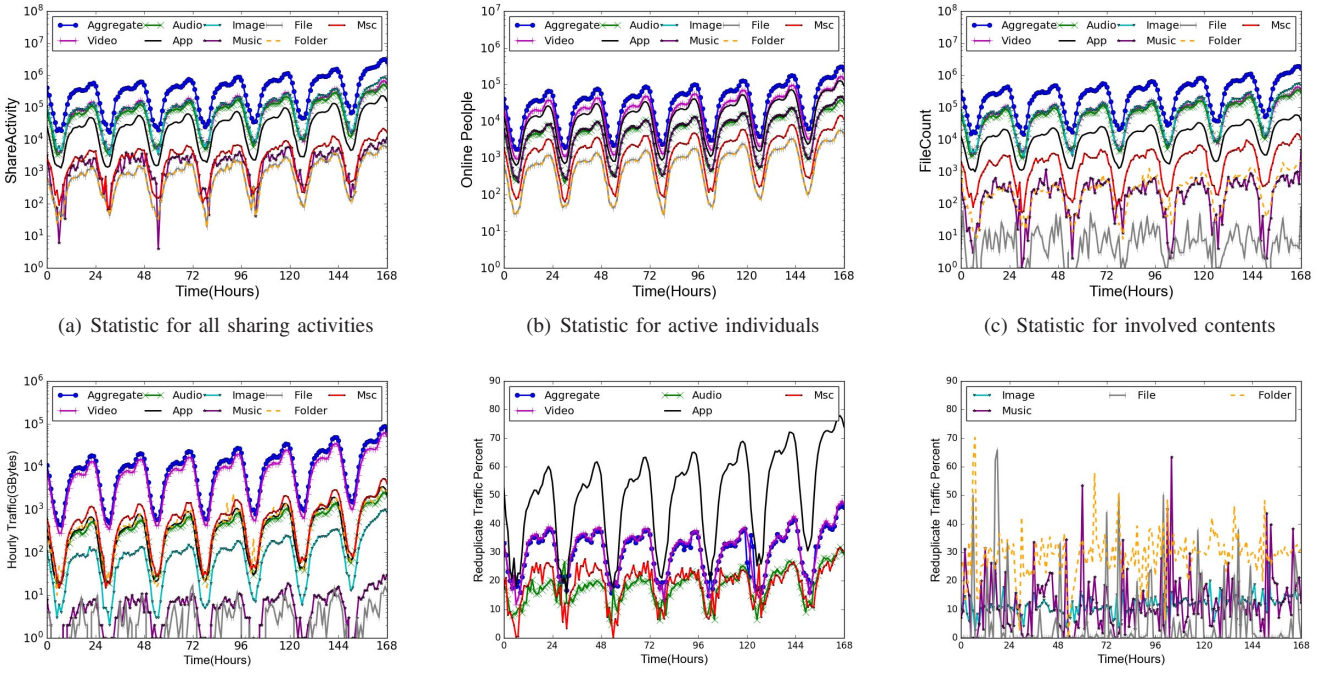
To perform efficient and reliable data processing of our large-scale data (with total size of 843 GB), we implemented the distributed data processing tool, *Apache Spark* 1.6, in a cluster with 20 servers interconnected by a Gigabit switch. Each server is with 4 CPUs, 32 GB memory, and 3 TB storage.

IV. Time Series of Performance Basics

A. Measurement Results of Time-Varying Performance Basics

We analyze the following five time-varying performance basics of the *Xender* data: Sharing Activities, Online Individuals, Involved Contents, Traffic Load and Duplicate Traffic Load. For a clear visualization, we only plot the date of 1st week of Feb. as the representative of the whole month in Fig. 2.

1) Sharing Activities: Fig. 2(a) shows the number of sharing activities over time. It can be seen that the sharing activities in weekend are around 2 to 3 times more than that in the working days. During each day, users are very more active (even up to 10 times) in the peak time (noon and night) than that through time (at dawn). Both the weekly and daily observations show typical life cycles with temporal regularity, and verify the fact that families and friends always get together in the weekend in India. This implies that there exists large



(a) Statistic for all sharing activities

(b) Statistic for active individuals

(c) Statistic for involved contents

(d) Statistic for traffic load

(e) Dup. traffic: Aggr., Video, APP, Audio, MSC

(f) Dup. traffic: Image, Music, File, Folder

Fig. 2: Measurement results of time series of performance basics

room of optimizing the time-varying system resources as well as effective event-based viral marketing strategies, particularly for India.

2) Online Individuals: We analyze the number of online individuals, where a user will be only counted for once during each time window (one hour), if this user shares at least a content file in this window. As shown in Fig. 2(b), the user who sharing videos takes the highest portion. Besides this, there are lots of users sharing APPs. This is mainly because of the inconvenient mobile APP market services via the MNOs' networks in India, where people often rely on exchanging interesting APPs via D2D sharing activities. Therefore, *Xender*-like services become more effective methods for video content dissemination and APP marketing.

3) Involved Contents: We turn to analyze the number of involved contents. Here, a content file is only counted for once according to its MD5 no matter how many times it has been shared by many people in the time window (1 hour). As shown in Fig. 2(c), multimedia types have nearly similar level of amount of involved contents. Different from Fig. 2(b), many images are delivered with *Xender*, especially on weekends when users tend to hold get-togethers. In addition, the number of mobile APPs shared via *Xender* is large as well, which implies that APPs have played a vital role in people's daily life, especially in India where cellular networks are not well equipped. There is very limited sharing of contents with other types. Another interesting finding is that the sharing activities of multimedia files (i.e., videos, audio and images) take a large portion, which indicates that the mobile users in India have great interests to enjoy multimedia services, due to the undeveloped cellular network infrastructure in most areas, people tend to use D2D sharing methods (e.g., *Xender*) to spread contents.

4) Traffic Load: Fig. 2(d) illustrates the time series of traffic load, i.e. the total size of all content files at each hour. It can be seen that that videos are still the majorities of the traffic load; in other word, *Xender*(and other similar D2D sharing platforms) helps the mobile networks to offload many videos into D2D sharing activities because the sharing is free and fast. And hence the service may emphasize more on improving the quality of video content sharing. Note that the traffic load on Sunday raises to 5 to 10 times than that in working days. This behavior could be utilized by *Xender* to initialize video-focused sharing events for improving the satisfaction of mobile users more conveniently. Moreover, the traffic load of Folder and MSCs (miscellaneous, all other types that cannot be classified) are also ignorable. This is because the fact that they normally contains large-size compressed files, although the number of them is relatively small.

5) Duplicate Traffic Load: As presented in Sec. I, there is a large portion of duplicate traffic in current mobile networks. Therefore it is very important to investigate the **duplicate traffic ratio** via the offline D2D sharing activities in MSNs, which is computed as one minus the ratio of total size of involved content files to the total traffic load regarding each type of contents. As shown in 2(e), approximately 10% to 40% aggregated traffic load shared via *Xender* is redundant, which means a certain large percentage of mobile users request and

obtain the same popular contents via *Xender*. Another interesting finding is that APP type has the largest the duplicated traffic ratio at about 60%, which reveals that many users are attracted by the same popular APPs. In addition, it is quite obvious that there are also a large quantity of duplicate video traffic. We can actively encourage seeding users to explore popular trending videos and hence to spread to more users effectively by utilizing methods in studies like [7] and [24].

As shown in Fig. 2(f), there exists no obviously temporal similarity for different days over the week. This is mainly caused by intensive sharing activities among small amount of users or sharing extremely large files. *Xender*-like services need to pay attention to improving the service for those sudden large-scale transmissions.

B. Quantifying Online Factor

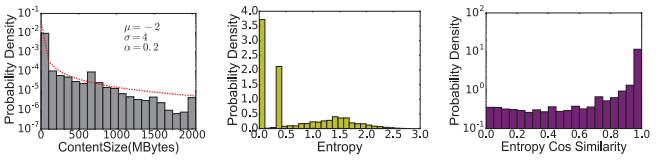
After analyzing performance basics over time, similar to [25], we obtain the online probability at different time zones for each user u_i, O_{iz}, where z is integers from 0 to 23 corresponding to total 24 hours, indicating the possibility that the user would like to use *Xender*. O_{iz} will be calculated for two types, averaged within working days (O^w_{iz}) and that within weekends (O^e_{iz}). We hence defined **Online Factor**, denoted by $\mathbb{O}_{ij}$, as the Cosine Similarity of two users's online probabilities as:

$$\mathbb{O}_{ij} = \frac{O_i \cdot O_j}{\|O_i\| \|O_j\|} = \sum_{z=1}^{24} O_{iz} O_{jz} \Big/ \left(\sqrt{\sum_{z=1}^{24} O_{iz}^2} \sqrt{\sum_{z=1}^{24} O_{jz}^2} \right) \quad (1)$$

Note that when carrying out prediction (nowcasting) in Sec. IX, online factor $\mathbb{Q}$ between requester u_i and potential sender u_i will be reflected as O_{jz} as the time when u_i's request is already given.

V. Content Properties

A. Content Size

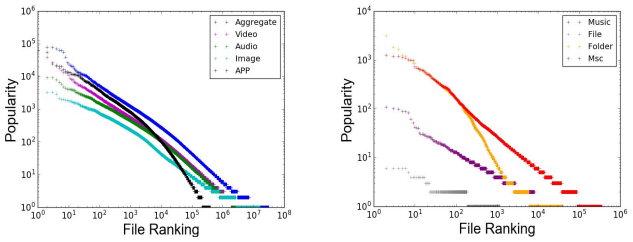


(a) Content size, with lognormal fitting (b) Content Entropy (c) Content Preference Similarity

Fig. 3: Measurement of content size, preference, entropy

Deep analysis of the contents (file objects) properties is essential to understanding the D2D sharing behavior in *Xender* data. Fig. 3(a) illustrates the probability density function (PDF) of file sizes. We can see that maximal content size is 2 GB, demonstrating the limitation of mobile APPs and mobile operation systems. The file size distribution follows the lognormal distribution with $\mu = -2$, $\sigma = 4$, $\alpha = 0.2$. This provides an useful implication for implementing techniques like content segmentation for efficient D2D transmission protocols. Also researchers can simply apply the empirical parameters for rapid modeling-based analytical researches.

B. Content Popularity



(a) Content popularity (Aggr., Video, Audio, Image, APP) (b) Content popularity (Music, File, Folder, Msc.)

Fig. 4: Measurement results of contents size and popularity.

It has been reported that the contents in the Internet has very skewed popularity disparity [2], [3], [26]. This means that a very small number of contents would attract huge amount of individuals and therefore occupy the majority of traffic load. We illustrate the content popularity ranking results by log-log plots in 4(a) and 4(b). It is clear that contents shared via *Xender* have strong property of power law for both aggregate contents and any of the specific type. We further carry out fitting of the power law plots by Maximum Likelihood Estimation (MLE) method, and obtain the α (the slope) and *Xmin* of the curve, as shown in Table I.

TABLE I: 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	

Therefore this phenomenon may be exploited for concentrating the system resources 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 the the small amount of popular files in the servers of Content Delivery Networks to improve transmission quality and reduce the expensive cellular data usage.

C. 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 eight 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.

The probability density of content entropy values of all users are shown in Fig. 3(b). It can be seen that lots of users have 0

content entropy, which means they always exchange one 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 around 0.2 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 eight types.

D. Quantifying Content Preference Factor

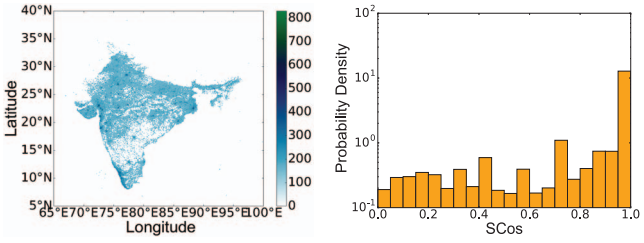
The content sharing among *Xender* users mostly relies on common interests of contents. Therefore, it is with significant importance to quantify how contents impact user sharing activities. To this end, we define the **content factor** for each pair of two users u_i and u_j by $\mathbb{C}_{ij}$, which is converted by calculating the Cosine Similarity of C vectors of u_i and u_j:

$$\mathbb{C}_{ij} = \frac{C_i \cdot C_j}{\|C_i\| \|C_j\|} = \sum_{y \in \Omega} C_{iy} C_{jy} \Big/ \left(\sqrt{\sum_{y \in \Omega} C_{iy}^2} \sqrt{\sum_{y \in \Omega} C_{jy}^2} \right) \quad (2)$$

Fig. 3(c) shows the content preference similarity of all user pairs that have sharing history from Eq. 2. It is shown clearly that a large number of users have exactly the same content preferences, while other user pairs have various similarity values from 0.1 to 0.8. Smaller values of content preference similarity means that two users have less common interests. Therefore, this is very essential factor for prediction of sharing activities and recommendation of users.

VI. Location Relations

According to the GPS records of all users, we first show a heatmap of *Xender*'s sharing activities in India in Fig. 5(a). Obviously, there are a great number of sharing activities in big cities with large populations (e.g. Mumbai, Delhi, Kolkata, Chennai, and Bangalore). It also shows that the *Xender*'s service has a very large coverage area in India.



(a) Heatmap of *Xender* users' activities in India (b) Homophily effect (location factor) of *Xender* users

Fig. 5: Measurement results of location relations

Because mobile users in offline MSNs mostly follow the phenomenon called "homophily" [20], which means each user has his/her own pattern of life cycle, spatially and temporally. Therefore, in order to find out how mobile users are tightly connected geographically, we collect the GPS records of each user u_i, denoted as l_i. We also carry out the analytics of location similarity (homogeneity) based on the the geographical Cosine Similarity [27], which is termed **Location Factor**:

$$\mathbb{L}_{ij} = \frac{L_i \cdot L_j}{\|L_i\| \|L_j\|} = \sum_{l \in l_i \cup l_j} L_{il} L_{jl} \Big/ \left(\sqrt{\sum_{l \in l_i \cup l_j} L_{jl}^2} \sqrt{\sum_{l \in l_i \cup l_j} L_{jl}^2} \right) \quad (3)$$

where L_{il} is the probability that the user u_i visits a particular GPS record $l \in l_i \cup l_j$, i.e. in the union of all visited places of u_i and u_j. The values of $\mathbb{L}_{ij}$ are within [0,1]. If $\mathbb{L}_{ij} = 1$, the two users u_i and u_j have exactly the same GPS records.

Fig. 5(b) illustrates the distribution of location factors of all active pairs of users. It is interesting that users who have been carrying out D2D sharing via *Xender* are mostly locally clustered, which verifies the homophily effect. Here, more than 80% pairs of users have score above 0.79, which implies that offline D2D sharing activities are mostly restricted by distances and locations (as users have to meet for exchanging contents). This is fundamentally different from online platforms that people can freely exchange contents via Internet, regardless their geographic locations. Therefore, the essential issue of facilitating *Xender*-like services for disseminating contents is to fully utilize or even encourage the user contact opportunities.

VII. Meeting (Contacts) Dynamics

The "homophily" phenomenon also applies to the temporal life cycles of user mobility. And thus we analyze the meeting dynamics of users. As illustrated in Fig. 6, for a typical pair of two users, they may meet (contact) and initiate several transmissions. We define the **inter-TX time** as the interval between each two continuous transmissions, and the average of inter-TX times between two users will be denoted by δ_{ij}. According to the time information, we consider if inter-TX times are larger than 3,600 seconds, the continuous transmissions end. The time of a round of continuous transmissions is defined as the the **contact time** and the interval between them as **inter-contact time (ICT)**. Averaged contact times is denoted by Θ_{ij} and averaged ICT is denoted by Δ_{ij}. Above three terms have been used in studies [7], [18], [19], [28], and can comprehensively describe the dynamics of opportunistic meetings, i.e., the mobility impact of exchanging contents between two users in the offline MSNs with respects to meeting frequency, meeting length, and sharing frequency.

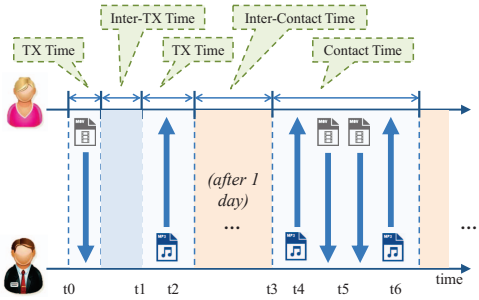
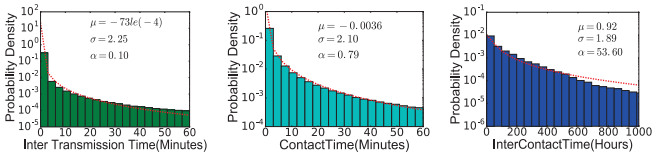


Fig. 6: Illustration of contact times, inter-TX times, and ICTs

As shown in Fig. 7, inter-TX times follow very perfect fitting of *lognormal distribution* with $\mu = -0.00073$, $\sigma = 2.25$, $\alpha = 0.10$. ICTs also match perfect lognormal distribution fitting with $\mu = 0.92$, $\sigma = 1.89$, $\alpha = 53.60$, while contact times mostly follow lognormal as well with $\mu = -0.0036$, $\sigma = 2.10$, $\alpha = 0.79$. Note that many studies like [18], [19], [28] carry out modeling of ICTs by exponential or power law distributions. In contrast, our large-scale measurement study provides an new evidence of lognormal distributions of all meeting dynamics. Hopefully this result can motivate new research on theoretical analysis and content sharing algorithm design.



(a) Inter-TX times, δ (b) Contact times, Θ (c) Inter-Contact Times, Δ

Fig. 7: Measurement results of meeting dyanmics

As δ, Θ and Δ values mostly reflect users' patterns of moving and potentials of sharing in real world, we further quantify the **Mobility Factor**, denoted by $\mathbb{M}$, by calculating (referring to *sigmoid* function, which maps all values into range of [0,1]):

$$\mathbb{M} = \alpha\left(2e^{-\delta}/\left(1+e^{-\delta}\right)\right) + \beta\left(2/\left(1+e^{-\Theta}\right) - 1\right) + \gamma\left(2e^{-\Delta}/\left(1+e^{-\delta}\Delta\right)\right) \tag{4}$$

where the coefficients for weighting the parameters should follow $\alpha+\beta+\gamma = 1$. $\mathbb{M}$ mobility factor will be quite important for profiling and predicting users mobile activities in *Xender*-like services.

VIII. Social Characteristics

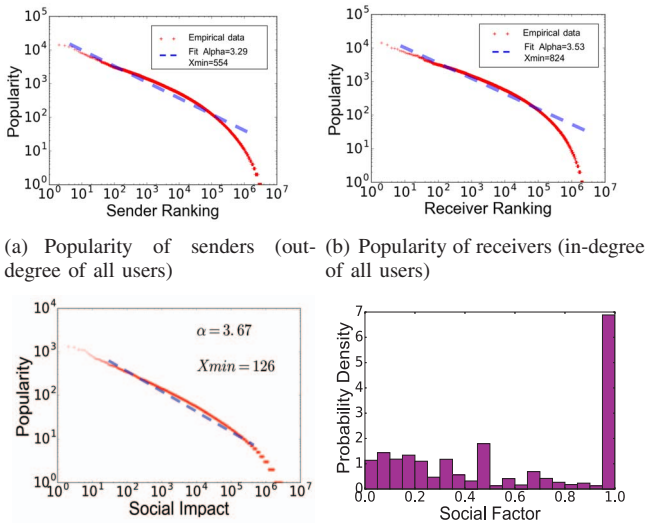
As shown in many recent studies [3], [29], [30], offline D2D content sharing can form opportunistic networks, where the social links among users are represented by their sharing activities. Therefore, we carry out the analysis on user social characteristics in *Xender*'s social network. According to the social graph definitions, we consider each user as the vertex, and each pair of users with sharing records as the edge, and the frequency of sharing as the weight of a given edge.

A. In-degree and Out-degree

We first investigate the in-degree and out-degree of each user to check the graph properties. For a given user u_i, the in-degree and out-degree of u_i are the numbers of other users who transmit files to, or receive files from u_i respectively. We plot the log-log figure of the ranking of the sender popularity (i.e., out-degree) and receiver popularity (i.e., in-degree) in Fig. 8(a) and Fig. 8(b), respectively. Here, 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 then implement fitting algorithm and obtain the slope (α factor) of the power law fitting with value of 3.29 and 3.54, which are quite skewed. This indicates that a very small number of users may have very strong capabilities of sharing contents with others. If we deploy related offloading schemes for propagating content, 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. By using the measurement results, services providers are able to mine these active users, gain their needs, and accelerate their activities to improve quality of services (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.



(a) Popularity of senders (out-degree of all users) (b) Popularity of receivers (in-degree of all users)

(c) Edge weights of all pairs (d) Social factors of all pairs

Fig. 8: Measurement results of social characteristics

So we define the **Social Factor** of each pair of sharing from users u_i to u_j as $\mathbb{S}_{ij}$ by considering the influence impact referring to studies [7], [11], [31]. We define U_i as the upstream neighbor set of user u_i, i.e. users in U_i can be considered as followees of u_i, who has shared (transmitted) contents to u_i. And likewise D_i as the downstream neighbor set of user u_i (followers of u_i), who have been shared (received) contents from u_i.

We define f_{ij} as the frequency of content sharing activities from u_i to u_j. Because the social factor should indicate the importance of u_j to u_i among all u_i's friends, as well as the importance of u_i to u_j among all u_j's friends, we need to consider: (1) the number of contents that u_i has shared to u_j, f_{ij}; (2) the number of contents shared from u_j's upstream neighbors to u_j; and (3) the number of contents shared from u_i's to all other downstream neighbors. Therefore, the Social Factor $\mathbb{S}_{ij}$ is calculated as follows:

$$\mathbb{S}_{ij} = \frac{f_{ij}}{\sum\limits_{u_r \in U_j} f_{rj}} \cdot \frac{f_{ij}}{\sum\limits_{u_k \in D_i} f_{ik}} \tag{5}$$

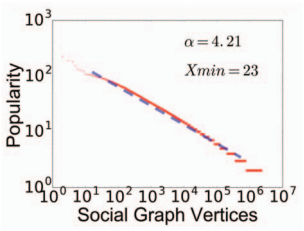
The value range of $\mathbb{S}_{ij}$ is within [0,1], and $\mathbb{S} = 1$ means that two users have strongest social impact that they only exchange

contents with each other. Fig. 8(d) shows the PDF of social factors for all pairs. It can be seen that the PDF curve of the social factors has a peak at values of 1, and nearly evenly distributed over other values. This implies that some user pairs have very strong social ties, but the majority of users are likely to have a uniformly distributed social impacts with others.

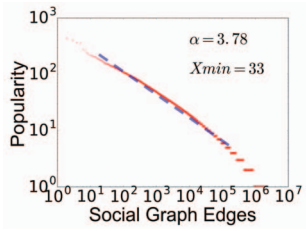
In all aforementioned sections, it is clearly shown that *Xender*-like D2D sharing services always have two distinct types of users: **bound users** who always have only 1 friend for exchanging contents; and **free users** who often exchange contents with many friends. *Xender*-like services should specialize the discovery and recommendation of friends or contents for different users based on their distinct types and impact factors.

B. Vertices and Edges of Social Groups

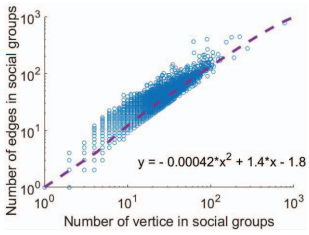
We further obtain a set of non-overlapping social groups (i.e. two vertices in two social groups shares no end-to-end path between them), and carry out analysis from the perspective of complex networks. One example social group is shown in Fig. 1(b), and the maximal group size (number of vertex) is just 771, which is quite small compared with the total size of user base. This is a a very important observation, which implies that MSNs for offline D2D sharing cannot support the extensive growth of social networks, due to time and space restrictions. From Fig. 9(a) and 9(b), we can see that the numbers of both vertex and edges perfectly fit lower law pattern, which indicates there are a lot of small groups. Fig. 9(c) shows that as the vertex number increases, the edge number increases at approximate linear speed rather than superlinear speed. This implies that the connectivities of friends are not very tight, and the group is not often closely clustered.



(a) Vertices of groups



(b) Edges of groups



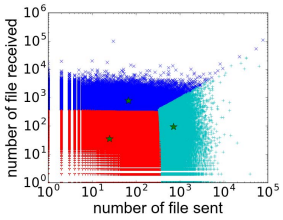
(c) Edges v.s. nodes

Fig. 9: Measurement results of social group properties

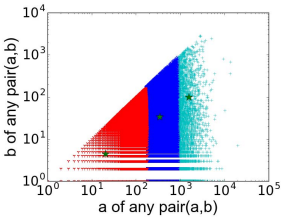
C. Social Clustering Analysis

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 [32] of user types and pair relationships. Specifically, we define that a user can be with any type among **sharing-focused**, **receiving-focused**, and **balanced**. And then we carry out k-means clustering method over the user base, where $K = 3$ at first. It is shown in Fig. 10(a) 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, respectively.

We also define the following three types of relations for each pair of users: **close friends**, **normal friends**, and **unfamiliar friends**. Then, we are able to improve all pairs' satisfactions with certain promotion strategies [33]. We still carry out k-means grouping method over the base of pairs [25], [34]. In order to ensure the correct clustering results and clear visualization, we follow the following rules: For each pair (a,b), where a represents the number of files shared from one user to the other and b represents the number of files shared reversely; if a is larger than b, we exchange a and b. As shown in Fig.10(b), 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 treat them as unfamiliar friends with infrequent contacts. In contrast, close friend always share more than 1000 contents. 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, respectively.



(a) Clusters of user types, k=3



(b) Clusters of user pairs, k=3

Fig. 10: Clustering (classifying) users and pairs by k-mean

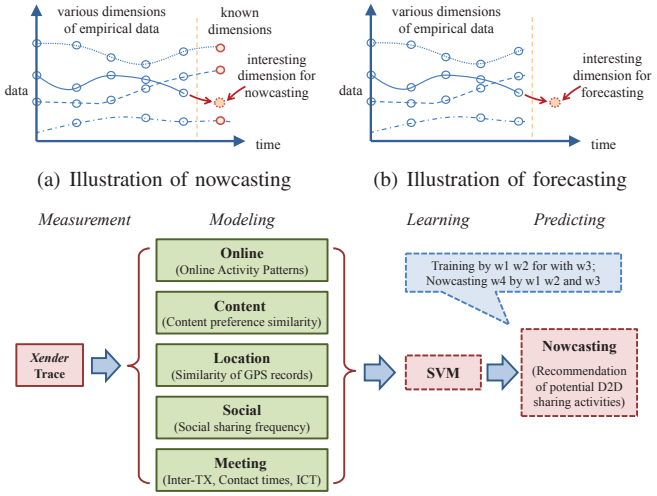
To simplify the calculation of social factors, it could be more efficient to utilize the clustered relationships among users. Therefore, we define a **simplified social factor**, $\mathbb{S}'$ based on above clustering results:

$$\mathbb{S}'_{ij} = \begin{cases} 0.1 & \text{if } u_i \text{ and } u_j \text{ are unfamiliar friends;} \\ 0.5 & \text{if } u_i \text{ and } u_j \text{ are normal friends;} \\ 0.9 & \text{if } u_i \text{ and } u_j \text{ are close friends;} \end{cases} \quad (6)$$

and we will test the performance in Sec. IX

IX. Nowcasting of D2D Sharing Service

An important application of data measurement and analytics is for the prediction of future D2D activities, while most of current prediction methods are based on one or two basic features, i.e. human mobility [12], [28]. In contrast, the multiple-feature analytics in our previous sections provide us much richer data of D2D sharing activities associated with users, contents, times and places, which motivates us to attempt to predict the future D2D sharing activities. To this end, this section will present an initial machine learning based "**nowcasting**" framework as the proof-of-concept case study of our data measurement and analytics.



(a) Illustration of nowcasting (b) Illustration of forecasting

(c) Nowcasting methodology based on multiple dimensional features

Fig. 11: Nowcasting for D2D sharing activities.

Different from forecasting, nowcasting is defined as the prediction of an unknown feature value in the very near future, based on other factors known in the very recent past, mostly in real time [35], as illustrated in Fig. 11. The aim of prediction of nowcasting is that, by knowing current values of several features, we expect to accurately nowcast a D2D sharing activity of mobile users. A successful prediction is defined as that the user who may share the contents (at the highest probability) to the requesting user is just the actual one who sends the content from the log in the trace.

We use Python 2.7 with *Scikit Learn* library [36], and utilize SVM methodology with "rbf" and "linear" kernels (setting to use "norm L1" and "norm L2") along with Stochastic Gradient Descent (SGD) optimization method. In the framework of learning and nowcasting as illustrated in Fig. 11(c), we first transfer data of feature factors from previous sections, including **Online**, **Content**, **Location**, **Mobility**, and **Social**, by normalizing all values, and then design several nowcasting schemes with different combinations of the features to evaluate the tradeoff of running complexity and predicting accuracy. The data in the 1st and the 2nd weeks data are used as the base, and the 3rd week is used for training our models; finally based on three weeks, we predict the sharing activities in the 4th week. Note that the location factor $\mathbb{L}$ is the necessary one for prediction D2D sharing, as we use the calculated location similarity as the criteria of estimating, ranking, and justifying the prediction accuracy, and thus $\mathbb{L}$ is always included during training and predicting.

TABLE II: Nowcasting results

Used Features					Prediction Results (%)		
$\mathbb{O}$	$\mathbb{C}$	$\mathbb{M}$	$\mathbb{S}$	$\mathbb{S}'$	rbf	l. L1	l. L2
Y	Y	Y	Y	Y	57.86	49.69	47.90
Y	Y	Y	Y		48.00	49.06	49.67
Y	Y	Y		Y	52.62	42.78	45.52
Y	Y	Y			52.62	43.63	41.06
	Y	Y	Y		58.81	48.07	48.66
	Y	Y		Y	53.29	37.87	39.62
	Y	Y			53.26	40.44	39.15
Y		Y	Y		56.23	50.44	49.37
Y		Y		Y	42.25	43.89	39.67
Y		Y			42.25	42.70	42.42
		Y	Y		58.03	49.15	46.94
		Y		Y	42.25	28.34	28.14
		Y			42.25	28.70	28.73
Y	Y		Y		58.21	50.56	49.30
Y	Y			Y	52.65	42.56	43.00
Y	Y				52.65	46.09	41.35
	Y		Y		**58.95**	50.51	48.65
	Y			Y	53.26	39.80	39.93
	Y				53.40	41.40	44.04
Y			Y		56.30	52.58	48.45
Y				Y	42.25	45.30	42.73
Y					42.25	44.72	41.70

Table II shows the nowcasting results, each of which is averaged by 100 runs. We can see the prediction accuracy varies according to differences of the used features. By utilizing multiple features, the nowcasting has reasonable accuracy nearly up to 60%. Regarding learning performance, rbf kernel performs better than linear ones, as we can see significant accuracy downgrade of linear schemes. Also linear kernel using L1 is more adaptive than that using L2 over *Xender* trace. Generally, the accuracy decreases if we skip more features. $\mathbb{S}$ (social factor) plays the most important role on improving prediction performance, and hence verifies that offline D2D sharing service also highly depends on social relations of mobile users. However, although $\mathbb{S}'$ offers a engineering way with low-complexity to reflect user social relations, it cannot help to achieve satisfiable performance. Another major feature is $\mathbb{O}$ (online factor); when $\mathbb{S}$ is not applied, it can accurately target the potential user who will be able to share contents, because it reflects the probability of being online. Deploying learning with multiple features will induce better accuracy, but more engineering complexity and thus the significant factors can be chosen for real time prediction, e.g., $\mathbb{O}$ and $\mathbb{S}$.

It is worth noting that our nowcasting experiment is the first D2D sharing prediction with multi-dimensional features of a large scale data trace in the literature. By predicting the potentials users who have contents that requesting users might be interested in , it is capable of recommending contents for them and meet their needs as soon as possible. Therefore, offline contents transfer is able to be accelerated and the duplicate traffic can be largely offloaded by this method as

well. What's more, the D2D sharing technology will be widely applied to make full use of radio spectrum resources and the nowcasting strategies are beneficial to choose repeater stations as well as allocate spectrum.

X. Conclusion

In this paper, we present a large-scale and multi-feature data measurement and analytics for the offline device-to-device (D2D) content sharing in Mobile Social Networks (MSNs). Our study is based on the data collected from *Xender*, one of the largest D2D sharing platform in the world, which covers over 30 million users with 30 billion D2D transmissions and 16 million content files. Through efficient data analytics using *Apache Spark* over a 20-server cluster, we carry out comprehensively analytics over multiple important features, including content properties, location relations, meeting dynamics, and social characteristics. This results in a set of useful observations, such as the burst of D2D sharing at weekend and lognormal distributions of contact dynamics. Further more, we develop a learning-based nowcasting framework that can predict future D2D sharing activities with a reasonable accuracy. The initial results demonstrate the great potential of using multi-feature machine leaning techniques for the accurate recommendation of future D2D sharing services.

The implications and guidelines provided in this paper could be very useful to improve the quality of *Xender*-like D2D sharing services, especially in the areas with dense mobile users but limited cellular network capacities, such as India. Our future work will focus on the design of more efficient and accurate multi-feature nowcasting approaches.

References

[1] CISCO, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021, CISCO, Tech. Rep., 2017.

[2] M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in *Proc. ACM IMC*, pp.1–14, 2007.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?", pp.591–600, in *Proc. WWW*, 2010.

[4] Wang, Y., Wei, L., Vasilakos, A. V., Jin, Q., "Device-to-Device based Mobile Social Networking in Proximity on Smartphones: Framework, Challenges and Prototype," in *Future Generation Computer Systems*, 2016.

[5] Device-to-Device Communications in 3GPP LTE standard, release 12, http://www.3gpp.org/specifications/releases/68-release-12

[6] Andreev, S., Pyattaev, A., Johnsson, K., Galinina, O., Koucheryavy, Y. "Cellular Traffic Offloading onto Network-Assisted Device-to-Device Connections." *IEEE Commun. Mag.*, vol.52, no.4, pp.20-31, 2014.

[7] X. Wang, M. Chen, Z. Han, D. Wu, T. Kwon, "TOSS: Traffic Offloading by Social Network Service-based Opportunistic Sharing in Mobile Social Networks," in *Proc. IEEE Infocom*, pp.2346–2354, 2014.

[8] X. Chen, B. Proulx, X. Gong, and J. Zhang, "Exploiting Social Ties for Cooperative D2D Communications: A Mobile Social Networking Case," *IEEE/ACM Trans. Netw.*, vol.23, no.5, pp.1471–1484, 2015.

[9] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile Data Offloading through Opportunistic Communications and Social Participation," *IEEE Trans. Mobi. Comput.*, vol.11, no.5, pp.821–834, 2011.

[10] R. Xiang, J. Neville, and M. Rogati,"Modeling Relationship Strengthin Online Social Networks," in Proc. *WWW*, pp. 981–990, 2010.

[11] G. Steeg and A. Galstyan,"Information Transfer in Social Media," in *Proc. WWW*, pp. 509–518, 2012.

[12] W. Gao and G. Cao, "User-Ccentric Data Dissemination in Disruption Tolerant Networks," in *Proc. IEEE Infocom*, pp.3119–3127, 2011.

[13] T. Rodrigues, F. Benvenuto, M. Cha, K. Gummadi, and V. Almeida, "On Word-of-Mouth Based Discovery of the Web," in *Proc. ACM IMC*, pp.381–396 2011.

[14] D. Hristova, M. J. Williams, M. Musolesi, P. Panzarasa, and C. Mascolo, "Measuring Uban Social Diversity Using Interconnected Geo-Social Networks," in *Proc. ACM WWW*, pp.21–30, 2016

[15] M. Gonzalez, H. Cesar, and B. Albert-Laszlo. "Understanding Individual Human Mobility Patterns," *Nature*, pp. 779-782, 2008.

[16] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He. "Exploring Human Mobility with Multi-source Data at Extremely Large Metropolitan Scales," in *Proc. ACM Mobicom*, pp. 201-212, 2014.

[17] AirDrop, Apple Inc., http://en.wikipedia.org/wiki/AirDrop

[18] X. Zhang, G. Neglia, J. Kurose, and D. Towsley, "Performance Modeling of Epidemic Routing," *Computer Networks*, vol.51, pp.2867–2891, 2007.

[19] Y. Li, Y. Jiang, D. Jin, L. Su, L. Zeng, and D. Wu,"Energy-efficient Optimal Opportunistic Forwarding for Delay-Tolerant Networks," *IEEE Trans. Veh. Technol.*, vol.59, no.9, pp.4500-4512, 2010

[20] K. Zhang and K. Pelechrinis. "Understanding Spatial Homophily: The Case of Peer Influence and Social Selection." in *WWW*, 2014.

[21] M. Wittie, V. Pejovic, L. Deek, K. Almeroth, and B. Zhao, "Exploiting Locality of Interest in Online SNS," in *Proc. ACM CoNEXT*, pp.2346–2354, 2010.

[22] E. Cho, S. Myers, and J. Leskovec." Friendship and Mobility: User Movement in Location-based Social Networks." in *Proc. ACM KDD'11*, pp. 1082–1090, 2011.

[23] L. Backstrom, E. Sun, and C. Marlow. "Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity." in *Proc. WWW*, pp.61–70, 2010

[24] C. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen,"Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks," *IEEE Trans. Mobi. Comput.*,vol.10, pp.2752–2763, 2011.

[25] C. Wu, X. Chen, Y. Zhou, N. Li, X. Fu, and Y. Zhang "Spice: Socially-Driven Learning-Based Mobile Media Prefetching", in *Proc. IEEE Infocom*,2016.

[26] V. Chaoji, S. Ranu, R. Rastogi and R. Bhatt, "Recommendations to Boost Content Spread in Social Networks," in *Proc. WWW*, 2012.

[27] D. Hu, S. Chen, L. Tu, B. Huang "Detecting Geographic Community in Mobile Social Network", in *Proc. IEEE GreenCom*, pp.343-347, 2012.

[28] S. Ioannidis, A. Chaintreau, and L. Massoulie, "Optimal and Scalable Distribution of Content Updates over A Mobile Social Network," in *Proc. IEEE INFOCOM*, pp.1422–1430, 2009.

[29] J. Toole, C. Herrera-Yaque, C. Schneider, M. Gonzalez, "Coupling Human Mobility and Social Ties." *Journal of The Royal Society Interface*, vol.12, no.8, 2015.

[30] J. Qiu, Y. Li, J. Tang, B. Chen, Q. Yang "The Lifecycle and Cascade of WeChat Social Messaging Groups", in *Proc. WWW*, pp. 311–320, 2016.

[31] J. Yang and J. Leskovec,"Patterns of Temporal Variation in Online Media," in *Proc. ACM WSDM*, pp.177–186, 2011.

[32] C. Brown, N. Lathia, C. Mascolo, A.s Noulas, and V. Blondel. "Group Colocation Behavior in Technological Social Networks," *PLoS ONE*, vol.9, no.8, pp.e105816, 2014

[33] H. Li, Y. Chen, X. Cheng, K. Li, D.g Chen, "Secure Friend Discovery based on Encounter History in Mobile Social Networks" *Personal and Ubiquitous Computing*, vol.19, no.7, pp.999–1009, 2015

[34] C. Brown, V. Nicosia, A. Noulas, and C. Mascolo. "Social & Place-focused Communities in Location-based Online Social Networks." *European Physical Journal B*, vol.86, no.6, pp.1–10, 2013.

[35] Y. Sun, N. Yuan, X. Xie, K. McDonald, R. Zhang, "Collaborative Nowcasting for Contextual Recommendation," in *Proc. ACM WWW*, pp.1407–1418, 2016.

[36] Scikit-learn, http://scikit-learn.org/