The Where and When of Finding New Friends: Analysis of a Location-Based Social Discovery Network

Terence Chen
National ICT Australia
School of Electrical Engineering
& Telecommunications, UNSW
terence.chen@nicta.com.au

Mohamed Ali Kaafar
INRIA, France
National ICT Australia
mohamed-ali.kaafar@inria.fr

Roksana Boreli
National ICT Australia
School of Electrical Engineering
& Telecommunications, UNSW
roksana.boreli@nicta.com.au

Abstract

With more people accessing Online Social Networks (OSN) using their mobile devices, location-based features have become an important part of the social networking. In this paper, we present the first measurement study of a new category of location-based online social networking services, a location-based social discovery (LBSD) network, that enables users to discover and communicate with nearby people. Unlike popular check-in-based social networks, LBSD allows users to publicly reveal their locations without being associated to a specific "venue" and their usage is not influenced by the incentive mechanisms of the underlying virtual community. By analyzing over 8 million user profiles and around 150 million location updates collected from a popular new LBSD network, we first present the characteristics of spatial-temporal usage patterns of the observed users, showing that 40% of updates are from the user's primary location and 80% are from their top 10 locations. We identify events that trigger bursts of growth in subscriber numbers, showing the importance of social media marketing. Finally, we investigate how usage patterns may be utilized to re-identify individuals with e.g. different identifiers or from datasets belonging to different online services. We evaluate re-identification by usage, spatial and spatial-temporal patterns and using a number of metrics and show that the best results can be achieved using location data, with a high accuracy: our experiments demonstrate that we can re-identify up-to 85% of users with a precision of 77% using monitored spatial data. Overall, we find that although users exhibit strong periodic behavior in their usage pattern and movements, the success rate of re-identification is highly dependent on the level of activeness and the lifetime of the users in the network.

Introduction

The widespread availability of positioning technologies like GPS in smartphones and other mobile devices has promoted the use of real-time location updates in mobile apps and location based services. Location-Based Social Networks (LBSNs), like Foursquare, Gowalla and Facebook Places, provide a platform for updating one's location, that is viewable by friends, by checking into a set of venues in the geographical proximity of the user's current location. An emerging category of LBSNs are Location Based Social Discovery (LBSD) networks, that are specifically designed to enable establishment of (new) connections to nearby users. The popularity of LBSNs has also attracted research interest as they offer a new source of information that enables studies of Internet users' online and offline behaviors. While the check-in- and content sharing-based LBSN have been studied extensively with a focus on various topics, including users' behavior (Li and Chen 2009),(Scellato and Mascolo 2011), mobility prediction (Scellato, Noulas, and Mascolo 2011),(Cho, Myers, and Leskovec 2011) and social link recommendations (Crandall et al. 2010),(Cranshaw et al. 2010), the emerging LBSD apps are yet to be studied due to the young age of such services and unavailability of extensive datasets. On the other hand, location information has been used in re-identification studies, to evaluate potential for unique identification of users based on e.g. their home-work locations, or a set of unique locations they have visited in the past. Such studies are commonly based on mobile phone datasets (Wang et al. 2011).

In this paper, we investigate a new LBSD mobile network by studying the dataset collected from an increasingly popular social discovery application, "Momo"[1], launched in late 2011. Momo provides two OSN related functions: **social discovery**, which enables a user to discover surrounding people based on the geographical distance between them, and **instant messaging**, that allows users to (subsequently) communicate. By default the application updates the user's location to the server (unless users explicitly opt-out of status updates), hence a rich set of spatio-temporal information about the users is captured in our dataset.

Our motivation for this study is two-fold: first, we wish to characterize the evolution of a new LBSD network, both in regards to population growth and to the way the application is utilized by the newly signed-up users. Then, as the spatio-temporal information on how individuals utilize this application is available, we are interested in evaluating the potential for re-identification of users based on this data and using a representative set of similarity metrics. The re-identification scenarios logically follow from the current trend of user profiling by mobile carrier or providers, mobile analytic companies, location-based service providers and potentially by application developers. These entities possess pieces of information covering different aspects of the users' life and

[1] http://www.immomo.com

would be highly likely to have an interest in enriching their data by exchanging and aggregating the information from other parties. In most of the cases, mobile users present in different databases can be linked by one or multiple unique identifiers like Android ID, Apple ID and accounts that are associated with the mobile device. However these unique identifiers may not be available or applicable in some circumstances, for instance, if we try to link an individual who uses two or more devices, or the device identifiers have been changed during a system upgrade.

We have collected approximately 150 million location updates in a period of 38 days, from 19/5/2012 to 28/6/2012, and over 8 million user profiles. We analyze this dataset to provide insights about the new network comprising LBSD application users, growth in user numbers and the way it is used in regards to the user activities i.e. social discovery and messaging.

Our contributions are as follows. We **characterize user activity in a new LBSD mobile network** and show the differences between Momo and a check-in-based LBSN (Gowalla), with Momo having a higher user activity and a lower number of locations visited by users. We conclude that the latter is a likely consequence of the incentives given to users of check-in-based networks to accumulate a larger number of visited locations than what they would be normally inclined to do.

We **analyze the evolution** of the number of users and the way the LBSD application is utilized, including a study of the newly signed users during the monitoring period. We observe that the new users have a significantly higher activity in the first week after signing up for the service and that the vast majority of users, after that period, has a relatively low level of activity that relates to social discovery.

We **evaluate the potential for re-identification** of users, utilizing temporal, spatial, or spatial-temporal characteristics of their in-application activity and a number of selected metrics. We show that, overall, the spatial data provides the greatest accuracy in re-identifying users while the temporal data may be of limited value when used in isolation. On the other hand, using spatio-temporal data, particularly when a large data volume is available, also has good re-identification potential. We also show the relevance of varying level of available data, i.e. the user's level of activity and lifetime in the network, on the re-identification accuracy.

The rest of the paper is organized as follows. First, we outline relevant related work in Section 2. We describe the data collection and our dataset in Section 3, followed by the characteristics of user's activities in Section 4. In Section 5 we analyze the evolution of the user numbers and the way users utilize the application. Section 6 evaluates the potential for re-identification based on different data types and volume available. We conclude and outline future work in Section 7.

Related Work

We provide an overview of research work related to various aspects of our LBSD app study.

A number of previous studies based on **measurement of user's activities in LBSNs** focus on characterizing check-in-based social networks like Gowalla, Brightkite and Foursquare. (Li and Chen 2009), (Scellato and Mascolo 2011), (Cheng et al. 2011) analyze the use of LBSNs in regards to the volume of check-ins and their spatio-temporal characteristics. Cho et al (Cho, Myers, and Leskovec 2011) analyze the social and spatial characteristics of Gowalla and Brightkite check-ins, aided by (related) cell phone traces. They first show the link between human movement and, respectively, social relationships and periodic user behavior (social relationships can account for between 10-30% of all human movement, while 50-70% of movement relates to periodic user behavior) and develop a mobility model that incorporates their findings. Similarly, (Scellato, Noulas, and Mascolo 2011) propose a new mobility model based on features of visited locations. Works including (Crandall et al. 2010), (Cranshaw et al. 2010) study predictions (recommendations) of social links. **LBSN evolution** was studied in (Allamanis, Scellato, and Mascolo 2012), with researchers proposing a new model of network growth based on a combination of social and spatial factors. **Instant messaging (IM) applications** were characterized in (Leskovec and Horvitz 2008), however this was done based on fixed user data.

Our paper studies the characteristics and the evolution of an emerging LBSD app that combines social discovery and IM. We demonstrate the differences between the check-in-based LBSN and the LBSD (Momo) and characterize evolution of both individual user's traffic and growth in overall population, highlighting the link with major promotional activities of Momo. We show how the mode of use shifts for most users, with time, from social discovery to IM use. We additionally study re-identification in LBSD apps.

The extensive **research work on re-identification** spans a number of fields of study. Related to spatial information, researchers in (Golle and Partridge 2009) have analyzed the U.S. Census data and have shown that on average, close to 20 individuals from the datatset share the same home or work locations, and that 5% of people in the dataset can be uniquely identified by home-work location pairs. A related work by Zang et al (Zang and Bolot 2011) generalized the use of (home-work) location pairs to an approach that uses top N locations to evaluate the uniqueness of US cellphone users. Both works strongly support the case for using location information to derive quasi-identifiers for re-identification of users. A number of research works e.g. (Mohammed, Fung, and Debbabi 2009), (Bonchi, Lakshmanan, and Wang 2011) and (Shokri et al. 2011) have raised the privacy issues in publishing location data and have focused on theoretical analysis of obfuscation algorithms. Further, (Li et al. 2008) exploit location history and evaluate the merits of different metrics to re-identify users from a dataset collected from a (small) number of cellphone and GPS device users. We note that the location related re-identification research works that are based on experimental data, have focused on mobility traces from cellphone or GPS devices, while the different nature of LBSN/LBSD data (human driven updates related to activities in the social network) necessitates a study that specifically addresses this environment.

We evaluate the potential of re-identifying users by their temporal or spatial patterns (or a combination of those) in the LBSD app. As users reveal their location in a non-continuous manner (as opposed to having a dataset that includes user's full trajectory in a measurement time period), the mobility trace and location sequence approaches were not applicable in our study. We examine a number of basic user similarity metrics that have been proposed in uniqueness measurement (Zang and Bolot 2011) and link prediction studies (Wang et al. 2011), (Crandall et al. 2010). We stress that our focus is on evaluating the potential for user re-identification in a LBSD application, rather than on proposing new re-identification algorithms.

Data Collection and Datasets

In this section we first briefly introduce the Momo LBSD application and the associated network, and then describe our data collection methodology. We also provide a comparative analysis of the basic characteristics of the LBSD network and outline key features of Momo, when compared to a popular check-in based LBSN, i.e. Gowalla. Specifically, we highlight the main differences between the two networks from the usage pattern and user behavior perspectives.

LBSD Application: Momo

Momo is a location-based social discovery application which allows users to discover people located in close geographical proximity and to connect with them using IM. The Chinese-based company launched its first iOS application in August 2011 and an Android version was rolled out in December 2011. Since then, the service has accumulated over 10 million registered users[2]. New users in Momo have to create a profile containing basic personal information (name, age, gender and icon photo) and some optional attributes (e.g. occupation, company, school, interest, etc.). The application is given permission to access the device's location information, which is extracted either directly from the cellphone GPS or via Google Mobile Map API[3].

When a user launches the application, a location update is sent to the Momo server. User can obtain a list of nearby users from the server (discovery function). These users are then displayed in the application according to their proximity to the user location, i.e. ranked in terms of distance to the device's current location. User status and location is simultaneously publicly revealed to other nearby users (and friends[4]), unless the "invisible" option is selected by the user. However, we have observed that only about 6% of the total number of users in our dataset have their location hidden from public access. Finally, individuals can establish a direct connection to any of the discovered users, who can be added to their Friends lists. This friendship relationship is never made public.

Data Collection The Momo mobile clients communicate with the server(s) via a set of network APIs. The *profile* API allows a client application to fetch a full profile of a user, identified by a numerical identifier. We crawled user profiles by selecting an exhaustive set of IDs, that were accessed sequentially during the data collection. We used the *profile* API to collect our first dataset, `Profiles`, comprising of user profiles.

Our second dataset, `Updates`, was collected using the *nearby* API, which upon request provides a list of nearby users. By varying the geographical coordinate parameters over time, we collected an extensive set of real-time location updates originating from different monitored areas.

Collecting user updates for such a large and dynamic system is a challenging task. Due to the high number of updates during different periods of the day, and the diversity of possible world-wide user locations, we have optimized our crawling strategy by considering the following trade-off: while we were interested in minimizing the number of requests to be sent to nearby APIs, we also had an aim to avoid missing any of the user updates within a monitored area.

The *nearby* API dictates that the server will only respond with up to a maximum of 30 users per update request, and a maximum of 40 requests can be made from any single set of coordinates, thus limiting the coverage area. To maximize the amount of collected data while working within these constrains, we have designed a dynamic and distributed crawling mechanism that operates in two steps.

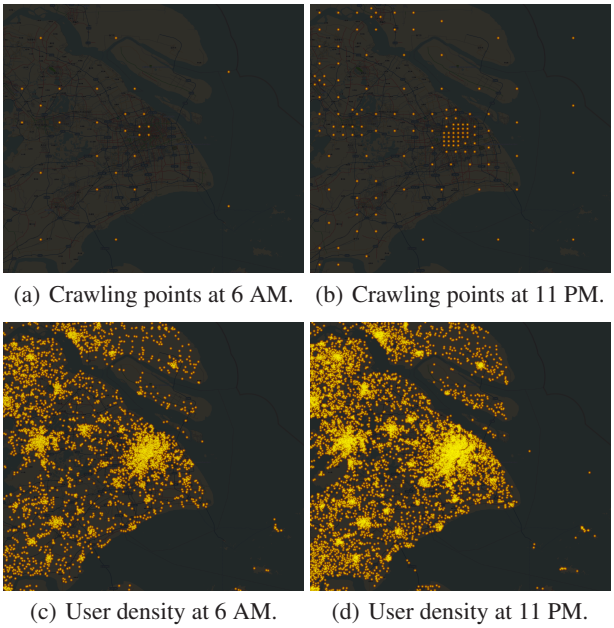
First, the crawler schedules a set of monitoring points and their corresponding coverage areas. This step is based on a modified version of the 2-dimensional closest point search algorithm in lattices (Agrell et al. 2000), which in our case aims to discover the minimum number of monitoring points by estimating the total number of active users in a given initial lattice area. This is achieved by recursively selecting potential monitoring points and verifying whether or not they are capable of discovering all other users activity within their currently defined monitoring area. Specifically, by simply requesting the last retrieved set of of nearby users of each selected monitoring point, we can decide whether the monitoring point can reach the edge of its assigned lattice. If the number of expected users exceeds the crawling capability of a single crawling (monitoring) point, i.e. more than $30 \times 40 = 1200$ users, the crawling area is further divided into four equal size lattices. Specifically, for each of the selected cities globally, we monitor a lattice area of 60×60 km^2 around the city center. The crawler scheduling process is performed every hour. The output of this step is a set of task parameters consisting of target coordinates, expiration time and the area to be covered.

Second, each of the scheduled crawling points requests the close users' locations from the server. Every 15 minutes, the crawling points selected at the first step (i.e. chosen hourly) update the location of the users observed in their vicinity. It should be noted that in case of multiple updates made between two different requests, the crawler can only

[2]http://siliconangle.com/blog/2012/08/28/chinas-dating-app-momo-proves-attractive-to-investors

[3]https://developers.google.com/maps/mobile-apps

[4]The notion of friends in Momo is similar to other OSNs. A friend's location is unconditionally shared with the user, i.e. users can always know where their friends have checked-in, irrespective of their current location.



(a) Crawling points at 6 AM. (b) Crawling points at 11 PM.

(c) User density at 6 AM. (d) User density at 11 PM.

Figure 1: The deployment of crawling points and example user density at 6 AM and 11 PM, in the Shanghai Area.

record the last update. Although this prevents our crawler from guaranteeing that all user updates are collected, we believe that the chosen request period of 15 minutes is already fine grained enough to allow us to capture the vast majority of updates. We have indeed tried different request periods, and empirically decided that such a choice is a good compromise between the number of recorded updates and the number of requests to the server (e.g. compared to a 2 minute requests period, less than 1% of the user updates are missed by having the period set to 15 minutes).

Figure 1 shows the example allocation of crawling points and user density at 6 AM and 11 PM, around Shanghai Area. The populated areas can be easily identified from the density map. During peak hours, the crawling range can be as small as 0.5 km around the populated area, which means that up to 1200 users can be discovered within 1 km^2 area.

The collected `Profiles` dataset consists of 8 million user profiles, each containing the user's most recent update time and associated GPS location. These profiles also include the publicly available user's personal information, namely: nickname, gender, age, interest, occupation and identities in other popular social networks. Every record in the `Updates` dataset consists of a user ID, timestamp and GPS coordinates (latitude and longitude). This dataset contains 150 million updates from 3.3 million active users in 48 cities (including cities from Australia, Canada, China, France, Germany, Italy, Japan, Korea, Singapore, Spain, US, and UK), over a period of 38 days during May-June, 2012. The `Updates` dataset contains approximately 65% [5] of all active users in the entire Momo network, present at the time of our data collection. Due to the rapid growth of the network, we have observed an increase in both the number of new users, which varied between 650K to 800K new users per day, and the total number of updates, for which we observed an increase from 3.5 million to 4.5 million updates per day, during the monitoring period.

Gowalla Dataset

Gowalla is a popular location-based social networking service that allows users to check-in their current location and share it with friends. The properties of the Gowalla LBSN have been studied in a number of research works e.g. (Cho, Myers, and Leskovec 2011).The dataset used in this work, consisting of more than 6 million check-ins from more than 196K users over the period of Feb. 2009 - Oct. 2010, was published in the SNAP website[6]. We use this dataset to evaluate the main differences between the usage and behavioral patterns of the two services: Momo as an LBSD network and Gowalla as a check-in-based LBSN.

Characterizing User Activity

We analyze user activity in the Momo network and provide a comparison of various related metrics to those derived from the Gowalla dataset.

Activity Distribution We start by examining the extent of difference in the user activity distribution in both networks. For the Momo service, we use the `Updates` dataset. In order to have a meaningful comparison of the two datasets in regards to the measurement duration, we use a 38-day portion of the Gowalla dataset (the same duration of the Momo dataset). In addition, as the location updates of Momo do not correspond to a specific "spots" as they do in Gowalla, we define a unique location as a $1 \times 1 km^2$ grid, which is then considered as a specific location for the `Updates` dataset.

Table 1 shows the user activity distribution function based on different parameters, with values shown for the 5, 25, 50, 75 and 95 percentiles. We define active days, as the number of days (out of the total number of observed days) on which users have checked-in at least once. Overall, the distribution statistics show that Momo users are more active than Gowalla users. While both Gowalla and Momo users do reveal almost the same number of unique locations per day, interestingly Gowalla users exhibit a higher number of total unique locations across the full observation period. Momo users, however, are more active, with almost two times higher median number of updates per day compared to Gowalla users. The latter also have a significantly lower number of location updates than the Momo users, with the median number of updates being only seven, while 50% of Momo users checked in at least 14 times. Notably, both networks have a similar distribution of active days for more than half of the user population. However, a considerable fraction of users in Momo publicly reveal a higher number

[5]The coverage of the dataset was estimated by the number of tracked users divided by the number of active users during the monitoring period (obtained from `Profiles` dataset)

[6]http://snap.stanford.edu/data/loc-gowalla.html

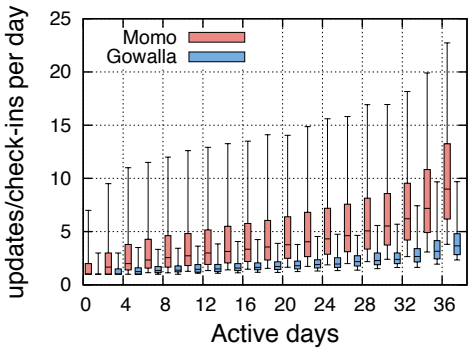


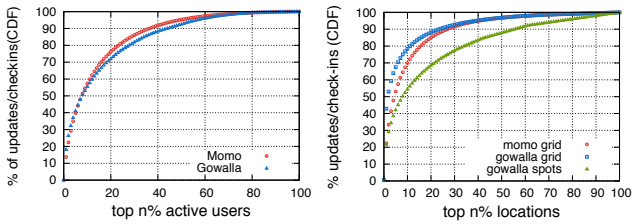
Figure 2: Updates/check-ins per day v.s. number of active days captured in the dataset

of unique locations compared to Gowalla users. While several reasons might explain the higher total number of visited locations for Gowalla users, we believe this is mainly due to the different nature of the two applications, where in Gowalla the users are incentivized to check-in at different locations that would accumulate in their history. Momo users, even though more active, seem unwilling to publicly share a number of locations they visit, and do not simply use the LBSD application in diverse locations. The two applications are designed for different purposes, and as such the user data collected from both services consequently reflects this difference in nature.

Correlation between daily activity and active days In Figure 2 we show the number of updates per day as function of the number of active days in the network, for both Momo and Gowalla. We can observe, for both services, that the increased loyalty to the applications results in a higher level of user activity. We then calculate a positive Spearmen coefficient of 0.625 in Momo and 0.552 in Gowalla, which indicates a high correlation between the number of active days and the number of updates per day. We again observe that the number of check-ins per day for Momo users is higher than for Gowalla, which is also illustrated by a sharper correlation slope.

Distribution of updates as a function of activity level Table 1 suggests a high skew in the distribution of updates, across users. To explore this further, Figure 3(a) depicts the cumulative distribution function of updates as a function of the top n% active users for both Momo and Gowalla datasets. Both networks show a similar distribution, with top 20% active users contributing 72% (resp. 75%) of total check-ins in Gowalla (resp. in Momo). This observation is in accordance with the Pareto principle ("a minor proportion of causes generate a major proportion of effects").

Distribution of updates/check-ins as a function of the number of locations Similarly to exploring the relation between users and updates, we examine the number of locations that attract the majority of check-ins. Figure 3(b)



(a) (b)

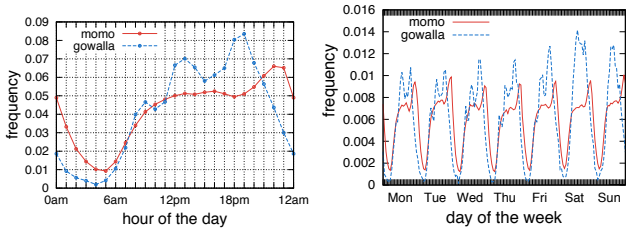
Figure 3: (a) % of total updates v.s. top n% active users; (b) % of total updates v.s. top n user locations

shows the cumulative distribution of updates as a function of the top n% locations shared by users, for both Momo and Gowalla datasets. In Gowalla, the check-in is associated with both geographical coordinates as well as with points of interests, called "spots", that could be e.g. a coffee shop or a library. As illustrated in the figure, Gowalla check-ins are distributed in a larger number of spots than grids, i.e. 20% spots attract less than 70% of check-ins while more than 85% of check-ins are done in 20% of grids. This can be explained by the fact that a number of spots can be aggregated in a fewer number of $1 \times 1km^2$ grids. We also note that Momo users' updates are distributed in more grids than Gowalla check-ins, which implies that Gowalla users are more likely to check-in within popular areas e.g. city centers.

Daily and weekly activity patterns Figure 4 shows the daily and weekly activity patterns of Momo and Gowalla users. For a meaningful comparison, we choose users from the same timezone. As previously observed in (Cho, Myers, and Leskovec 2011), during weekdays, the Gowalla curve shown in 4(a) exhibits two noticeable peaks around 12-2 pm and 6-8 pm, whereas the Momo activity curve shows a single peak around 10-11 pm. We also observe that there is less activity during late night and early mornings in Gowalla daily patterns, as opposed to the Momo users which exhibit a relatively high level of activity during these periods. Again, the different nature and the intended use of the two applications can explain such a different daily usage pattern. Gowalla users have limited opportunities to check-in to a place other than their home at late night and early morning hours, while Momo users can use the service in a meaningful way, by either discovering nearby people and/or chatting with friends from home; in both cases, their location is updated automatically when they refresh the user list or bring the application to the foreground. We also highlight the steady weekly pattern of user activity in Momo, with close to no variation between the weekdays and the weekends. Gowalla, on the other hand, shows a "typical" weekends/weekdays variation, with users being more active during weekends. This again reinforces our observation that the user activity level in Gowalla, and as such their location updates, seem to be closely linked with their visits to locations they have an incentive to check-in to.

Dataset	Momo					Gowalla				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95 %
Active days (out of 38 days)	1.0	2.0	5.0	14.0	31.0	1.0	2.0	5.0	12.0	26.0
Total updates/check-ins (in 38 days)	1.0	4.0	14.0	53.0	253.0	1.0	3.0	7.0	20.0	69.0
Updates/check-ins per day	1.0	1.444	2.48	4.8	12.889	1.0	1.0	1.333	1.9	3.8
Total unique locations (in 38 days)	1.0	2.0	3.0	8.0	24.0	1.0	2.0	5.0	11.0	28.0
Unique locations per day	1.0	1.0	1.272	1.807	2.2	1.0	1.0	1.222	1.588	2.53

Table 1: User activity distribution in Momo and Gowalla



(a) (b)

Figure 4: Comparison of (a) daily activity patterns (weekdays) and (b) weekly activity patterns between Momo and Gowalla

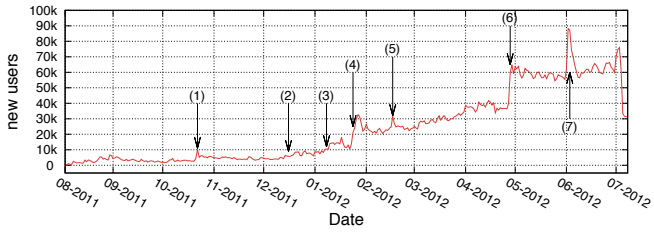


Figure 5: New users per day over time

Network and Usage Evolution

Next, we study the evolution of the Momo LBSD network, including the growth of user numbers and identifying events that have contributed to significant spikes in growth. We also analyze how the new users' activity evolves through time, by studying user retention rate, user behavior and the temporal evolution of the mobility patterns of Momo users.

We note that, due to limitations of the available information from the Momo APIs, we can only monitor in real-time the behavior of active users within the data collection period (38 days) and in the monitored geographical areas, as per the contents of the `Updates` dataset. We also have access to the last record of activity and the account age for all Momo users, as per the `Profiles` dataset.

User Volume Growth and Trigger Events

Since its launch date in August 2011, Momo has achieved significant growth, with a population size of more than 10 million users reached in less than one year[7]. Figure 5 shows the *daily* user growth between August 2011 and July 2012 (based on the data from the `Profiles` dataset, which includes the sign-up dates of users).

We observe that the application first experienced a five-month long slow start with, on the average, less than 5K new users per day. The pace of user growth became faster in early 2012 and received two major bursts around the end of April and the beginning of June 2012. To understand the external factors that drove the application growth, we have identified the events corresponding to the specific spikes of user growth, by associating them with various Momo company related information: (1) From 21/10/2011 to 23/10/2011, Momo launched an advertising campaign on Weibo (a Chinese popular micro-blog platform), promoting the product in 27 influential accounts that had millions of followers. (2) On 15/12/2011, the company released the first Android app version. Although we do not observe a sudden user growth after the Android version release, we can see that the user volume growth climbed steadily from the time of this event. (3) On 07/01/2012, Momo was awarded the "Best Social App 2011" by Geekpark.com[8]. (4) From 23/01/2012 to 29/01/2012, we observe a burst of user growth during the 7-day Chinese New Year public holiday. (5) On 17/02/2012, another popular advertising campaign was launched, receiving 15k re-posts soon after its release. (6) 27/04/2012, which corresponds to the most significant growth spike that occurred following a popular (funny) video mentioning the Momo application[9]. The video received one million views in 10 hours, and the volume of Momo's daily new users almost doubled on the following day. (7) 02/06/2012, Mainstream media "City Weekly" magazine (Southern Metropolis Daily) published a cover article on targeted social applications, with a section dedicated to introducing the Momo application[10].

Figure 5 suggests that although the two advertising campaigns in social media had immediate positive results, both online and traditional off-line media coverage (events (6) and (7)) attracted a steady and more significant user growth. Notably, we observe that online media coverage seemed to have a longer lasting period in which a high number of new users was attracted. This can be explained by the easy to access, long-tailed viral dissemination and reproducible characteristics of Internet content, as opposed to one-off conventional printed media content.

Likewise, the burst of user growth around the Chinese New Year indicates that special events like public holidays

[7]http://www.cnetnews.com.cn/2012/0803/2104590.shtml

[8]http://www.geekpark.net/event/view/details/152042

[9]http://v.youku.com/v_show/id_XMzg3MTk3ODQ4.html

[10]http://tech.sina.com.cn/i/2012-06-02/05217208517.shtml

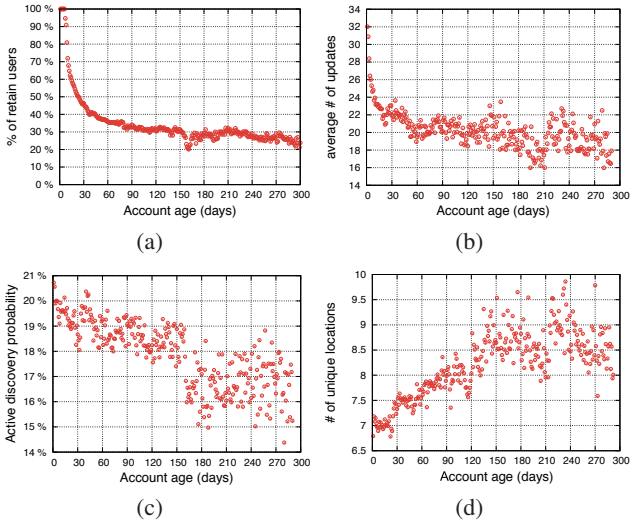


Figure 6: Evolution of user activities: (a) Percentage of retained users v.s. account age; (b) Average number of updates per week; (c) Active discovery probability over time; (d) Average number of unique locations v.s. account age

may massively attract new users and benefit the online social applications.

The Evolution of User Activity

We analyze a number of metrics related to the level of user activities in the Momo network.

Retention rate. This is one of the most important performance metrics that measures the percentage of retained users after a period of time. According to mobile analytic company Flurry's report[11], the average 90-day retention rate for social networking applications is around 34%. Although we cannot compute the absolute retention rate including all Momo users, based on the "join date" from user profiles, we are able to derive the retention rate of users for a specific account age, as shown in Figure 6(a). Similar to Flurry's approach, the "retained" users are defined as the users who login to the application at least once in the past 7 days. Figure 6(a) indicates that the percentage of retained users sharply drops in the first two months, and becomes stable after three months. The trend suggests that the users who stayed in the network for more than three months are likely to continue using the application for a longer time period. The 30-days retention rate is approximately 44.8% and is reduced to 33.1% after 90 days. Interestingly, we observed a reduction of retention rate around the account age of 160 days, which corresponds to users who joined during the Chinese New Year public holiday.

Engagement over time. We are interested in analyzing the activity level of users over time, demonstrating their involvement in using the application. Figure 6(b) shows the engagement of users with different account ages up-to 300 days, where the engagement is measured by the frequency of use per week. Figure 6(b) clearly shows that the newly joined users have a much higher activity level compared to more experienced users. The decline of activity slows down around 60 days and remains at the level of around 20 updates per week.

Probability of active discovery. In the Momo application, the mobile client updates its status and location when the user turns on the application, or when the user actively discovers nearby users, i.e. refreshes the user list. In the case of active discovery, multiple updates with short inter-arrival times can be observed from the server side. An active discovery session is defined as a set of consecutive updates, with each pair of updates occurring within a 30 minutes time interval. We measure users' discovery behavior by the *active discovery probability*, that is computed by the number of active discovery sessions over the total number of sessions. As shown in Figure 6(c), we observe that, for new users, as high as 20% of sessions are active discovery session, and that the average discovery probability decreases steadily as the account age of users increases, being reduced to an average of 15% for the account age of 300 days. This trend suggests that users spend less time on social discovery as they establish connections with (a sufficient number of) nearby users over time.

Unique location v.s. account age. Figure 6(d) shows the average number of unique locations visited by users with a specific account age. We can observe that more experienced users tend to use the application in more locations over the same period of time, compared to recently signed-up users.

User Re-identification

The extensive amount of information generated by users of Momo, including the locations they have visited when using the service, frequency and time of use, can also be utilised for service personalization, by profiling users in order to offer them e.g. recommendations on places to visit, or people to connect to. Considering such scenarios where specific users are targeted, in this section we evaluate the potential for re-identifying users, based on the pervasive spatial and temporal information that may be collected e.g. by the service providers.

As different types of information may be collected by different service providers, we evaluate their capability to re-identify users utilizing three different levels of background knowledge: using only the check-in time stamps, based on the subset of locations they have shared with the service and combining both temporal and spatial patterns of LBSD application use.

To better understand how the availability of data may affect the linkage performance, we also evaluate the performance of re-identification with different data collection period durations (e.g. comparing periods of 1 to 19 days of location information). We also vary the level of user activities and study whether it impacts the linkability of user patterns.

Next, we first introduce the methodology adopted to assess the re-identification capabilities, followed by the details of the similarity metrics used in our study.

[11] http://blog.flurry.com/bid/90743/App-Engagement-The-Matrix-Reloaded

Methodology

We first divide the `Updates` dataset records into two sets: a training set R and a test set $\hat{R}$. Each set comprises an equal number of observation days and depending on the selected strategy, the two sets contain either the user check-in time, the check-in location or both. Then, for each user i we extract all records r_i (resp. $\hat{r}_i$) from the training set R (resp. the test set $\hat{R}$) corresponding to the user activity during that period.

The goal of a re-identification classifier is to predict the linkage between user records in R and $\hat{R}$, assuming the identities in both sets are unknown or anonymized.

The classifier first computes the similarity score between r_i and all user records in $\hat{R}$. We have decided to adopt two different approaches: (i) a unique re-identification case where the classifier identifies the "best candidate" record $\hat{r}_j$ from $\hat{R}$, i.e. the record with highest similarity to r_i (amongst all records). (ii) a set of candidate records R_i^{pre}, predicted as being potentially linked to r_i as they have a similarity score higher than a predefined threshold. We define k as the size of the candidate records set, i.e. $k = |R_i^{pre}|$. The value of k will depend on the threshold value, as a more restrictive threshold will result in a smaller set of candidates.

In the following, we introduce the similarity metrics used in this study.

Temporal Cosine Similarity (TCS) : Considering the time-based vectors r_i and r_j, we compute their cosine similarity to capture the temporal similarity of the usage pattern.

Let $P(i,t) \equiv \frac{f(i,t)}{n(i)}$ be the probability that user i uses the application during a period of time t (in our case periods are defined as hours) and $f(i,t)$ is the cumulative frequency of updates in period t across the record duration. $n(i)$ is user i's total number of updates. Then $r_i = [P(i,0), P(i,1), ..., P(i,23)]$ and $r_j = [P(j,0), P(j,1), ..., P(j,23)]$. The TCS score between user i and j is computed as:

$$TCS(i,j) = cos(r_i, r_j) = \frac{r_i \cdot r_j}{\parallel r_i \parallel \times \parallel r_j \parallel}$$

Spatial co-location rate (SCR) : Considering the spatial usage patterns of users, we extract from r_i and r_j, the two sets L_i^r and L_i^r as the set of unique locations visited by user i and j. We then compute the Jaccard-index to measure the similarity of unique locations visited by both users.

$$SCR(i,j) = \frac{|L_i^r \cap L_j^r|}{|L_i^r \cup L_j^r|}$$

Spatial top co-location rate (STCR) : Considering users' top N locations, we aim to measure the similarity between popular co-locations. For each user i (resp. j), we only consider the locations from vector r_i (resp. r_j) and build a vectors L_i (resp. L_j), ranked according to the frequency of visited locations. The STCR score is then computed as:

$$STCR(i,j) = \frac{|\sum_{k=1}^{N} L_i \cap \sum_{k=1}^{N} L_j|}{N}$$

In the case where any of the two vectors contain less than N elements, we set $N = min(|L_i|, |L_j|)$. In our experiment we consider top 10 locations, i.e. $N = 10$.

Spatial cosine similarity (SCS) : This metric captures the similarity between two users location frequency patterns. Let $P(i,l) \equiv \frac{f(i,l)}{n(i)}$ be the probability that user i utilizes the application in location l; $f(i,l)$ is the cumulative frequency of updates in location l across the recorded period and $n(i)$ is user i's total number of updates. For all locations in both sets, i.e., $L = L(i) \cup L(j)$, we then present location vectors of users i and j as $r_i = [P(i,0), P(i,1), ..., P(i,l)]$ and $r_j = [P(j,0), P(j,1), ..., P(j,l)]$. The SCS metric is measured as:

$$SCS(i,j) = cos(r_i, r_j) = \frac{r_i \cdot r_j}{\parallel r_i \parallel \times \parallel r_j \parallel}$$

Spatio-temporal co-location rate (StCR) : This metric is motivated by the observation that users may visit different locations on specific hours of the day. We first divide the location-based check-in vector r_i of each user i into T sub-vectors, corresponding to different periods of time duration T. We then compute the spatial co-location rate of each of the T sub-vectors between any two users. The StCR score of users i and j is defined as:

$$StCR(i,j) = \frac{\sum_{t=1}^{T} SCR(i,j)}{T}$$

which corresponds to the average SCR values across the different T periods of time. We choose $T = 24$ considering a daily based analysis of the location patterns.

Spatio-temporal top co-location rate (StTCR) : Similarly, we introduce the spatio-temporal version of $STCR$ as the similarity score between the top co-locations in time. Again, we divide the location-based check-in vectors into T sub-vectors and evaluate the spatial top co-location rate between different vectors, each corresponding to a specific period of time. The Spatio-temporal top co-location rate is computed as:

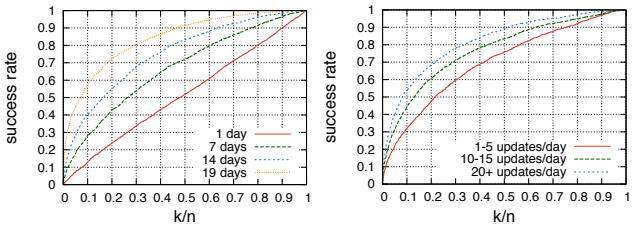
$$StTCR(i,j) = \frac{\sum_{t=1}^{T} STCR(i,j)}{T}$$

Spatio-temporal cosine similarity (StCS) : Another version of the STCR metric considers similarity between location check-ins on a periodic basis (daily if T =24). This metric is computed as follows:

$$StTCR(i,j) = \frac{\sum_{t=1}^{T} StTCR(i,j)}{T}$$

	matched	non-matched
linked	true positive (TP)	false positive (FP)
non-linked	false negative (FN)	true negative (TN)

Table 2: The possible outcomes of the re-identification process



(a) (b)

Figure 7: Success rate of re-identification using TCS, varying (a) the number of active days and (b) updates per day. k is the candidate set size and n the sample size, $n = 10,000$ in this experiment

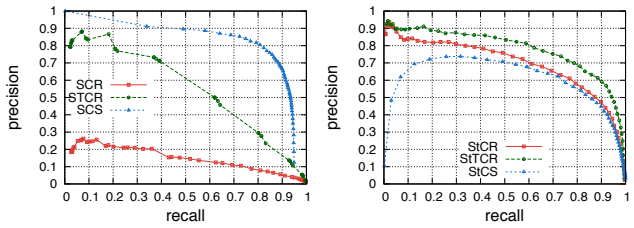
Performance Evaluation

Binary classifiers are commonly evaluated in terms of precision and recall. These are defined using the possible outcomes of a classifier, i.e. two records are deemed to be either **linked** or **non-linked**, and the resulting validity of the classification, i.e. two records will be **matched** if they belong to the same user, and **non-matched** otherwise. Table 2 shows the possible combinations, used to compute the Precision and Recall values as: $Recall = \frac{TP}{TP+FN}$ and $Precision = \frac{TP}{TP+FP}$.

For most cases, precision and recall provide a good indication of the performance for varying classifier parameters. However, when there is a large difference of e.g. a very small precision and a high (uncomparable) recall value, we use an alternative evaluation method by defining a successful prediction as the correct identification of a matched record that can be re-identified either uniquely or within a set of k candidate records. We then use the success rate as the % of successful predictions over the total number of linked records.

We now evaluate the performance of different similarity metrics.

Temporal approach Our initial experiments using TCS have shown a low accuracy of unique re-identification, therefore we consider the performance for a resulting candidate set. Figures 7(a) and 7(b) show the success rate of re-identification with a varying umber of active days and updates per day. We can observe that when using only a single day of user's data, the classifier provides almost random results. For the entire 19 days, the classifier can narrow down the candidate set to 10% of the sample size 60% of the time, which considering the sample size is not a promising result.



(a) (b)

Figure 8: Precision-recall curve for selected re-identification similarity metrics using (a) spatial information; (b) spatio-temporal information;

Spatial approach The precision-recall curves of SCR, STCR and SCS are shown in Figure 8(a). The SCR model has extremely low precision with all thresholds, which suggests that the probability of co-location is quite high amongst users and the low recall also suggests many users are inconsistent in location updates. By considering the visited location frequency, STCR and SCS metrics outperform SCR. STCR metric only considers the top 10 most relevant user's locations, ignoring the infrequent locations that are not likely to be revisited, and hence increases the linking probability. The SCS metric compares the direction of location vectors, i.e. cosine similarity. The frequently visited locations contribute more to the direction of the vectors, while the influence of infrequent or in-transit locations is suppressed. We can observe that SCS shows a solid performance, we can re-identify up-to 85% of users with a precision of 77%.

Spatio-temporal approach Figure 8(b) shows the precision-recall curves for StCR, StTCR and StCS re-identification results. We can observe that StCR and StTCR perform better than the original spatial-only approach, while the performance of StCS degrades compared to SCS. The improvement suggest that users are commonly co-located and may share the top locations, however within a time dependency. On the other hand, SCS computes the cosine similarity of the overall location frequency; segmenting the pattern into 24 sub-patterns, i.e. via StCS, results in a less identifiable pattern.

We now consider how a varying amount of data available for re-identification would affect the accuracy of various similarity metrics.

Considering the available quantity of data Tables 3 and 4 show the success rate of **uniquely** re-identifying users, for the various similarity metrics and based on a selected number of active day and user activity levels (updates/day). With the exception of SCR, all other metrics have an improved performance with a longer monitored period. When only using a single day of records (i.e. one day each in the training and testing sets), approximately half of the users can be correctly re-identified. The success rate increases to 86 % when using a week of data records, and almost all users (97.7%)

	1 day	7 days	14 days	19 days
SCR	0.5634	0.6864	0.6789	0.6286
STCR	0.4358	0.7501	0.8031	0.8684
SCS	0.4082	0.8598	0.9281	0.9778
StCR	0.4396	0.7495	0.8305	0.8932
StTCR	0.5488	0.8333	0.8695	0.8926
StCS	0.5474	0.6587	0.6884	0.8801

Table 3: Success rate of the re-identification similarity metrics for a different number of active days. Results are generated as an average of 5 iterations.

	$\rho > 0$	$\rho > 5$	$\rho > 10$	$\rho > 15$	$\rho > 20$	$\rho > 25$	$\rho > 30$
SCR	0.5355	0.5055	0.494	0.4995	0.4874	0.4961	0.5142
STCR	0.8355	0.8663	0.8428	0.7813	0.7366	0.7585	0.7767
SCS	0.9057	0.9178	0.9418	0.9543	0.9594	0.963	0.9659
StCR	0.8462	0.8515	0.8666	0.8909	0.8934	0.8899	0.8922
StTCR	0.8448	0.8214	0.8464	0.8752	0.887	0.8933	0.8997
StCS	0.6596	0.7299	0.8209	0.8845	0.926	0.9408	0.946

Table 4: Success rate of re-identification similarity metrics for different activity levels (ρ = updates/day). Results are generated as an average of 5 iterations.

are re-identified with the best performing metric, SCS, using just under 3 weeks of data. The results from Table 4 indicate that users who have higher level of activity have a higher chance to be re-identified.

Conclusion

In this paper, we have characterized a new LBSD network, in terms of overall properties (when compared to more established LBSNs), and the way it is being used by it's subscribers. We have then analyzed the network evolution, identifying specific events contributing to extraordinary user growth, and the evolution of user's activity patterns. Finally, we have evaluated the potential for service providers who may monitor and retain data on user's activities, to re-identify LBSD users who may have e.g. multiple accounts or a number of mobile devices, by comparing the performance of a number of similarity metrics. Our results show that a significant quantify of data is required for re-identification. This indicates that re- identification may be applicable primarily to applications that are heavily used, preferably on a daily basis, e.g. from the social, gaming or business productivity categories.

References

Agrell, E.; Eriksson, T.; Vardy, A.; and Zeger, K. 2000. Closest point search in lattices. *IEEE TRANS. INFORM. THEORY* 48:2201–2214.

Allamanis, M.; Scellato, S.; and Mascolo, C. 2012. Evolution of a location-based online social network: analysis and models. In *Internet Measurement Conference*, 145–158.

Bonchi, F.; Lakshmanan, L. V.; and Wang, H. W. 2011. Trajectory anonymity in publishing personal mobility data. *SIGKDD Explor. Newsl.* 13(1):30–42.

Cheng, Z.; Caverlee, J.; Lee, K.; and Sui, D. Z. 2011. Exploring Millions of Footprints in Location Sharing Services. In *ICWSM*.

Cho, E.; Myers, S. A.; and Leskovec, J. 2011. Friendship and Mobility: User Movement in Location-based Social Networks. In *Proceedings of SIGKDD international conference on Knowledge discovery and data mining*, KDD '11.

Crandall, D. J.; Backstrom, L.; Cosley, D.; Suri, S.; Huttenlocher, D.; and Kleinberg, J. 2010. Inferring Social Ties from Geographic Coincidences. *Proceedings of the National Academy of Sciences* 107(52):22436–22441.

Cranshaw, J.; Toch, E.; Hong, J.; Kittur, A.; and Sadeh, N. 2010. Bridging the Gap Between Physical Location and Online Social Networks. In *Proceedings of the ACM international conference on Ubiquitous computing*, Ubicomp '10.

Golle, P., and Partridge, K. 2009. On the anonymity of home/work location pairs. In *Proceedings of the 7th International Conference on Pervasive Computing*, Pervasive '09, 390–397. Berlin, Heidelberg: Springer-Verlag.

Leskovec, J., and Horvitz, E. 2008. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 915–924. New York, NY, USA: ACM.

Li, N., and Chen, G. 2009. Analysis of a Location-Based Social Network. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, CSE '09, 263–270. IEEE Computer Society.

Li, Q.; Zheng, Y.; Xie, X.; Chen, Y.; Liu, W.; and Ma, W.-Y. 2008. Mining user similarity based on location history. In *Proc. of the 16th ACM SIGSPATIAL international conf. on Advances in geographic information systems*, GIS '08.

Mohammed, N.; Fung, B. C.; and Debbabi, M. 2009. Walking in the crowd: anonymizing trajectory data for pattern analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09.

Scellato, S., and Mascolo, C. 2011. Measuring User Activity on an Online Location-based Social Network. In *In Proceedings of IEEE 12th International Symposium on a World of Wireless, Mobile and Multimedia Networks*.

Scellato, S.; Noulas, A.; and Mascolo, C. 2011. Exploiting Place Features in Link Prediction on Location-based Social Networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, 1046–1054.

Shokri, R.; Theodorakopoulos, G.; Danezis, G.; Hubaux, J.-P.; and Boudec, J.-Y. L. 2011. Quantifying Location Privacy: The Case of Sporadic Location Exposure. In *PETS*, 57–76.

Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; and Barabasi, A.-L. 2011. Human Mobility, Social Ties, and Link Prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, 1100–1108. ACM.

Zang, H., and Bolot, J. 2011. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11, 145–156.