Automated User Profiling in Location-based Mobile Messaging Applications

Yao Cheng*, Chang Xu†, Yi Yang*, Linyun Ying*, Purui Su* and Dengguo Feng*
*Trusted Computing and Information Assurance Laboratory, Institute of Software,
Chinese Academy of Sciences, Beijing, China, {chengyao, yangyi, yly, supurui, feng}@tca.iscas.ac.cn
†School of Computing Engineering, Nanyang Technological University, Singapore, xuch0007@e.ntu.edu.sg

***Abstract*—Location-based messaging applications (LMAs), a kind of messaging applications for mobile devices which enable users to connect with people based on their geographical locations, have recently experienced a huge popularity growth. The killer feature in LMAs that embodies the concept of geo-based instant messaging, named *people nearby*, allows users at any place to search and communicate with other registered users nearby. In this paper, we discuss a common weakness in LMAs that relates to the abuse of the people nearby function. In this case, rich personal data of registered LMA users can be easily obtained, bringing a chance to perform automated user profiling in LMAs. Specifically, we build an automated and scalable system to construct "extended" profiles (or we call *life profile*) of LMA users, which contain not only personal information of LMA users but also the daily activities and social ties inferred from their leaked spatio-temporal privacy. The system is highly adaptable to various applications, requiring no modification of applications or trivial work on protocol reverse engineering. We conduct the evaluation on a large scale for the first time. In our experiment, we succeed to construct life profiles for more than 280,000 users from two popular LMAs. The results of empirical analysis not only validate the existence of the privacy issue in LMAs, but also demonstrate its severity.**

Keywords*—Mobile Security, Privacy Leakage, Location-based Messaging Applications.

I. Introduction

As the boom of location-based services (LBS) on the Internet in recent years, traditional mobile messaging applications also gain strength by making use of location data generated from users' mobile devices to achieve better communication experience, giving rise to the novel location-based messaging applications (LMAs). Users in LMAs not only can contact with people in the friend lists conventionally but can also communicate with strangers close by their current location. The dimension of location bridges the gap between the online chatting environment and the physical world, making LMAs become one of the most popular social applications for mobile users. For example, Skout (https://www.skout.com/), a popular LMA which is committed to facilitate instant nearby connections, has achieved over 350 million connections among users in 2013. Momo (http://www.immomo.com/), another popular LMA, has owned over 100 million users across 131 countries and regions by February, 2014.

The concept of geo-based instant messaging in LMAs is realized by the equipped killer feature - *people nearby*. With this feature, any user at any place can "search around" to find out who is nearby at the moment and subsequently send a greeting message if interested. However, the openness nature of LMAs inevitably brings about privacy issues. In LMAs, anyone, especially a stranger or even an attacker, can freely gain access to the personal information volunteered by nearby users, along with the relative location and time information about their presences. This case, which may bring new security challenges, differs from traditional LBS such as check-in services and geo-tagging where the access to the spatio-temporal information generated by users (e.g., users' check-in data, geo-tagged posts) is publicly and explicitly available in default privacy settings. In contrast, in LMAs, by default the spatio-temporal data of user presences can only be accessed by the ones nearby, and such data is not explicitly posted by users but implicitly measured by the server once the people nearby function has been activated, making LMA users more neglectful in taking necessary precautions to protect their location privacy, which should be considered more sensitive in this situation. Unfortunately, a LMA user is unaware of who has browsed his/her personal information via the people nearby function so that attackers can easily harvest the profiles without being noticed at all. Moreover, the presences of LMA users can be exposed at anywhere (e.g., home or workplace) at anytime (e.g., day or night) as long as they use LMAs from time to time in daily life. As such, the daily activities or even social connections of LMA users can be inferred from their spatio-temporal presences or co-presence histories, indicating significant privacy leakage of LMA users.

In this paper, we reveal privacy threats in LMAs caused by the killer function of discovering people nearby, based on which automated profiling can be achieved. Specifically, we have made the following contributions:

- *Understanding of privacy issues in LMAs.* We discuss and uncover three types of privacy threats in LMAs, i.e., the personal information privacy, the location privacy, and the social privacy, which will become reality if the people nearby function is abused in a systematic way. To highlight the severity of such privacy threats, we propose the notion of *life profile*, which covers 5 correlated dimensions of people's daily life, i.e., WHO is this, WHERE is (s)he and WHEN, WHAT (s)he is doing, and WHOM is that with him/her, to model the consequences of such privacy issues.
- *Concrete attack scenario and effective approaches for constructing user life profiles.* A concrete attack scenario is presented based on the abuse of the people nearby function. Effective approaches are proposed for the tasks of presence locating, activity induction, and social tie mining for LMA users, which are essential

for the construction of life profiles.

- *Implementation and evaluation.* A scalable and automated system is designed and implemented to construct life profiles of LMA users. The system design gives a promising adaptation for various applications, which requires no application modification or protocol reverse engineering. This work for the first time evaluates the severe consequences of such threat in LMAs in large scales. Extensive experimental results show our success in automated profiling a large number of users (over 280,000) from two popular LMAs, demonstrating that the privacy threats in LMAs are realistic and serious.

II. Methodology

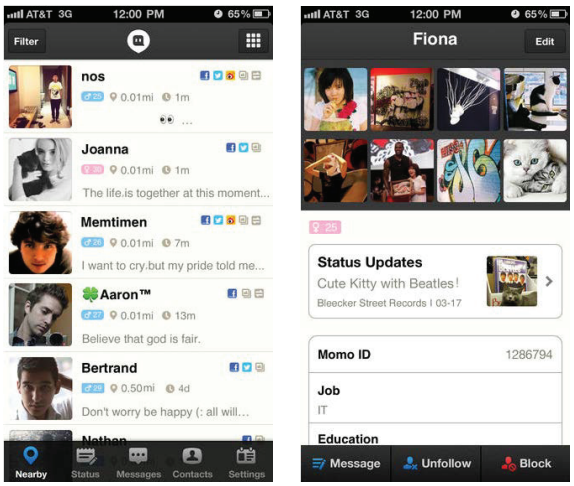
In this section, we first introduce the people nearby function in LMAs. Then an attack scenario is presented to show how this function can be abused by malicious users. Finally we discuss the raised privacy threats and the proposal of life profile to capture the consequence of such threats.

A. "People Nearby" in LMAs

The people nearby function in LMAs provides a novel way for users to connect with people nearby. Once activated at a specific place by some user U, this function displays a list of users who have been around this place within certain ranges (e.g., 0.01 miles) and time periods (e.g., 1 minute ago) (Figure 1(a)). The presences of these displayed users are due to the fact that they have already used this feature a priori in the vicinity. Meanwhile, the presence of U may also appear in the people nearby lists of others nearby. In practice, some LMAs are found to automatically activate the people nearby function right after the launch, so as to guarantee the acquisition of the up-to-date presences of users at the server end that further determines the nearest ones for displaying. In the people nearby list, each record also contains a concise profile corresponding to each user, including summaries such as avatar, name, gender, age, and a link to the user's full profile page (Figure 1(b)) where detailed personal information can be found, such as photo, profession, education, and hobby. It can been seen that the people nearby function indeed provides an interface to the wealthy personal information, which gives us potential to profile LMA users.

B. Abusing "People Nearby" for Automated User Profiling

Assume that an attacker (e.g., Alice) who can arbitrarily fake the location information of her mobile device to disguise to be at any intended place on earth. The attack begins as follows (Figure 2): Alice first selects a list of targeted *disguised spots* $\mathbf{S} = \{S_i\}_{i=1}^{N}$, with each represented as a geo-coordinate $S_i = (\phi_i, \lambda_i)$, where ϕ_i and λ_i denote the latitude and longitude of S_i respectively. Next, for each S_i, at time T_i Alice changes the current location of her mobile device to S_i and activates the people nearby function. Then she obtains a people nearby list L_i, containing M unique records of users $\{u_k\}_{k=1}^{M}$ who have appeared around S_i within d_{ik} distance/range (geo-measure in LMAs) and at Δt_{ik} time ago (temporal-measure in LMAs). Each record also links to a profile page of u_k where his/her profile $profile_k$ can be obtained. As a result, Alice obtains a collection of LMA people nearby lists $\mathbf{L} = \{L_i\}_{i=1}^{N}$, corresponding to her disguised spot list $\mathbf{S}$. Particularly, what attracts Alice more is the case that one may appear multiple times in $\mathbf{L}$ (e.g., User 1 and 2 in Figure 2), which gives her the opportunity to explore the daily routines or activities of spotted users using such trajectories. For a specific LMA user, Alice stores his/her profile and the associated trajectory. Each user u_k is represented as a triple $(id_k, profile_k, \mathbf{r}_k)$, where id_k is the identifier of u_k, the spatio-temporal information $\mathbf{r}_k = \{(S_i, d_{ik}, t_{ik} = T_i - \Delta t_{ik})\}_{i=1}^{n_k}$ is a list of his/her range based LMA measures extracted from $\mathbf{L}$.



(a) People nearby list (b) Full profile page

Figure 1. People nearby function in Momo[1].

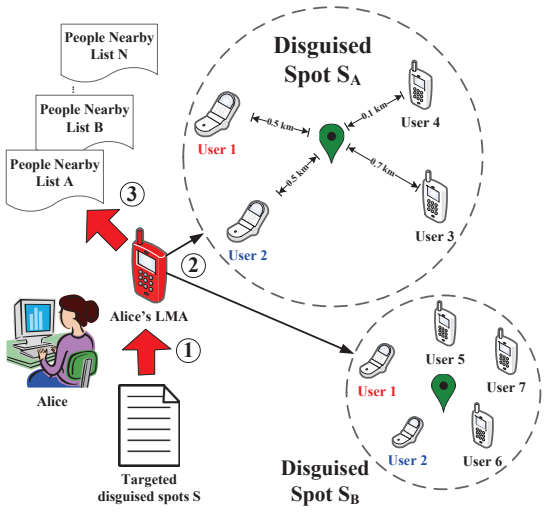


Figure 2. The attack scenario.

C. Privacy Concerns and Life Profile

Several privacy concerns are expected if the above strategy is exploited. The first type is about informative **personal details** of LMA users. This privacy data is leaked when attackers succeed in accessing users' profile pages via cheating the LMA servers to pretend to be in the vicinity of the targets.

[1]Figures come from Momo's official website.

For example, Figure 7 verifies this tactic by showing the personal details of over 280,000 users collected from two popular LMAs with good quality (the overall average profile completeness ratio is 74.42%). The same tactic can lead to the second privacy violation where massive **spatio-temporal information** of LMA users can be available. Location and presence of mobile users are two major sources of privacy leakage in mobile social networks [1]. In our data, over 23.9% users have more than one presence recorded, even 571 users each has left more than 20 presences. More importantly, the collected *multiple* presences of a user at different location and time can also help us to deduce his/her daily activities when combined with other geographical and temporal information. Note that this kind of knowledge can never be obtained by any user even the ones nearby. The final privacy concern constitutes when **social ties (weak or strong)** are inferred from spatio-temporal information of LMA users. This is not impossible if a group of socially related LMA users (e.g., friends, neighbours) activate the people nearby function "together" - at proximate locations and time periods, resulting in the formation of their *co-presence histories*. Certainly, just one co-presence may not be statistically significant to determine the existence of potential connections, however, if a group of users *frequently* show up at different location and time on the radar of attackers, the probability that they are socially related become much higher [2]. Moreover, other side information gathered in preceding steps, such as company, college, and interest group can be used to further determine the relationship.

In summary, the above privacy concerns can be reduced to three hierarchies. The first level exposes users' *personal information privacy* via the profile pages of LMAs, which benefits user identification and preference recognition, i.e., "Who is this?" and "What is (s)he like?". The second level discloses users' *location privacy* by gathering users' spatio-temporal trajectories. Such information can help attackers analyze daily routines of LMA users, i.e., "Where and when does (s)he do what?". Finally, the highest level involves the *social based privacy* of multiple users by deducing their social relations, i.e., "Whom is that with him/her?". This damages victims most as social relationship is considered one of the most sensitive privacy in human life. We thus propose the notion of *life profile* to properly capture the above mentioned privacy issues which consists of five dimensions, i.e., WHO, WHERE, WHEN, DOING WHAT, WITH WHOM, corresponding to the above three questions. Life profiles of LMA users are derived from the consequences of the abuse of the flawed people nearby function. From the perspective of attackers, the goal is to enrich the life profiles of LMA users as much as possible. In this work, we focus on the automated life profile construction to reveal the severity of such privacy leakage in LMAs.

III. Design and Implementation

In this section, we describe the design and implementation of an automated and highly scalable system to construct life profiles of LMA users.

A. System Architecture

Figure 3 illustrates the system architecture, composed of two parts, i.e. information collection (LMA User Collector) and data analysis (Life Profiler). Preliminarily, the targeted disguised spot list should be specified to instruct the LMA User Collector, which is built on a modified mobile operating system (Android) that runs LMAs. The system can log down every visible element (e.g., texts and images) in any application once displayed on screen. This functionality (People Nearby List Logger) is implemented by monitoring display related APIs in Android framework layer such as android.widget.TextView and android.widget.ImageView. Upon receiving the targeted disguised spot list, the GPS Simulator sequentially sets the location information of the system to those in the list. At each location, the people nearby function is activated by simulating user actions with a sequence of Android Debug Bridge (adb) commands (e.g., touch, pop-down, etc.), triggering the update of the people nearby list. The entire contents displayed on screen are then logged down by People Nearby List Logger, including profiles and spatio-temporal data of all nearby users. After post-processing such as data extraction and formation, well-structured user profiles and spatio-temporal data are stored permanently in LMA User Database. Life Profiler is responsible for the construction of life profiles of LMA users. Its four sub-modules are committed respectively to generate each part of the life profile. One of them is Profile Identifier that manages all users' profile details which are anonymized for privacy protection. The implementation of the remaining sub-modules are presented in subsequent sections.

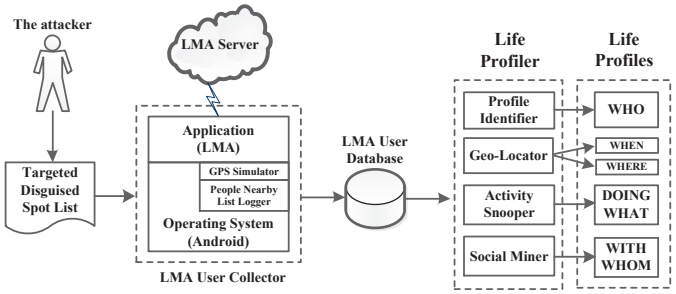


Figure 3. System Architecture.

B. Disguised Spot Selection

As the start of the whole abusing process, selecting suitable disguised spots is non-trivial, which will directly affect the subsequent LMA user collection. In order to facilitate the triangulation techniques adopted in the acquisition of precise presence location in Geo-locator, we split the intended region into grid-like cells, the side length of each cell spans d degrees of latitude and longitude, under which case the presence of a specific user at a specific time can be captured multiple times at different adjacent disguised spots. The challenge of the disguised spots selection not only lies in the population of the interested area, but also relies on the mobility of each user. Specifically, if the distance between two adjacent disguised spots is set too large, it will become harder to obtain multiple geo-measures of a specific user at a particular location for triangulation; otherwise, LMAs may not be able to detect the switching of the disguised spots, resulting in the people nearby lists not updated. Two distinct resolutions are adopted after various schemes are empirically tried, which will be presented in the experiment part (Section IV-B).

C. Geo-locator: Locating User Presences

WHEN and WHERE in life profile reveal location privacy of LMA users. To obtain such information, we try to locate each presence of LMA users based on the spatio-temporal data in LMA User Database through triangulation. Geometrically, the range-based LMA geo-measures can be represented as a circle with the numerical range as its radius. In theory, at most three different temporally proximate LMA geo-measures with respect to a specific presence are sufficient to calculate the exact WHEN and WHERE of that presence. However, we have to face two challenges in practice. First, in temporal dimension, it is full of uncertainty about the instant status of a specific user; we have no idea whether the geo-measures are collected during the moment the user appears, motionlessly. Second, in spatial dimension, the numerical precision of the geo-measures (the minimal recognizable ranges) in LMAs will cause another dimension of inaccuracy. Thus we have to resort to the approximation alternative. The approximation of a specific presence of LMA users can be achieved as follows. In terms of the temporal part, the entire geo-measures of a LMA user can be split into different sets. In each set the maximum time deviation is limited by a threshold reflecting the belief of the proximity degree in time so that this set of geo-measures can be considered corresponding to the same presence. The temporal part then can be obtained by integrating the timestamps associated with these geo-measures. In terms of the spatial part, an estimated region that covers all possible areas of the presence can be generated. There are three cases as shown in Figure 4: for each geo-measure set of a specific GSA user (i) if only one geo-measure is available, the area covered by this geo-measure is returned as the estimated region; (ii) if two geo-measures are available and their intersection (or tangency) exists, the overlapped area (or the tangent point) is returned as the estimated region; (iii) if three or more geo-measures are available, the area overlaid by all the geo-measures will be returned as the estimated region. In practice, we approximate the estimated region in either case 2 or 3 with the smallest circle covering that region shown as P in Figure 4.

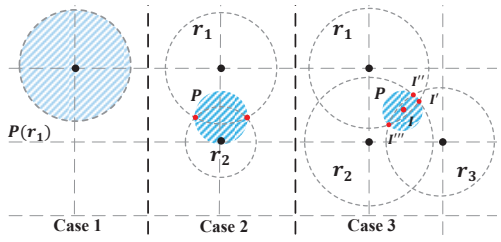


Figure 4. Geo-locating user presences in LMAs, where $\mathbf{r}$'s (dashed circles) denote the geo-measures displayed in the people nearby lists, $\mathbf{P}$'s denote the estimated regions (dashed blue area) for the presences.

D. Activity Snooper: Inducting Activities

Having harvested a series of presences of a LMA user, the attacker can further induct his/her daily activities (DOING WHAT). The idea is that people's activities are to a certain extent related to the type of the locations and the time the activities take place. For example, if a user is spotted constantly near a residential area during the night, it is probable that (s)he lives there. Similarly, the Activity Snooper module works by integrating the location type information of users' presences with the corresponding temporal records, based on a set of rules. Each rule contains a Location Type field and a Time Period field together as the *precondition*, and an Activity field as the *consequent*. The ith rule can be expressed formally as:

$$r_i : (LT = V_{LT}) \wedge (TP = V_{TP}) \rightarrow Activity \qquad (1)$$

where V_{LT} is chosen from a set of location type names such as residential community, workplace, and restaurant. V_{TP} is chosen from a set of time intervals. The *Activity* includes typical daily events such as work, study, dining, and sleep. In practice, as the estimated locations of users' presences are represented as circles with error ranges, the type of a location can be determined by using the type of the smallest map-identified area that covers the circle. The segmentation of the time period can be decided following typical daily routines such as working hours (e.g., from 9am to 6pm) and sleeping hours (e.g., from 11pm to 6am).

E. Social Miner: Discovering Social Ties

The most profound part of life profiling is to infer social ties among LMA users based on their profiles and presence histories. Our hypothesis is that if a group of LMA users (approximately) co-appear with high frequency at different locations, the possibility of the existence of their social connections (strong or weak) will be much higher than that of the coincidence of strangers[2]. Based on this assumption, we deduce users' social ties by borrowing the idea of Co-location Pattern Mining[3] from the field of spatial data mining. In general, co-location pattern mining is to find spatial objects that are frequently located together. However, temporal information associated with the locations has not been particularly considered in this task. In our case, the co-presence of a group of LMA users involve both spatial and temporal proximity. In order to adapt this approach to our scenario, we propose a frequent co-presence mining algorithm (FCPM) for LMA users. The idea is that each presence of a particular user can be represented as a point in the spatio-temporal space (Figure 5). To consider both spatial and temporal dimension when mining frequent co-presences, the entire spatio-temporal space can be firstly partitioned, along the temporal dimension, into temporally consecutive subspaces (the planes stacked along the timeline in Figure 5). The height of each subspace along the timeline encodes the belief in the degree of simultaneity of the co-presence. Then we generate all co-locations (dashed shapes in each plane Figure 5) in each subspace, with each enclosing a set of nearby users in spatial dimension. Thus a *frequent* co-presence pattern of LMA users can be defined as a set of users showing up in at least θ_{min} co-locations.

The details of FCPM are shown in Algorithm 1. To generate co-locations in each subspace, density based clustering algorithms (e.g., DBSCAN [4]) can be utilized (Line 4), which take all presences and specific parameters as inputs, and output clusters of users as the found co-locations in each subspace. Finally, frequent co-presences can be found by using frequent itemset mining algorithm (e.g., Apriori [5]) (Line 8). In our context, a set of *items* is the set of all users collected. Each *transaction* is the set of users found in a particular co-location.

[2]This assumption may work empirically, as supported by the results of [2] that a relatively small number of co-occurrences between two people at distinct locations can rapidly increase the probability that they are socially connected.

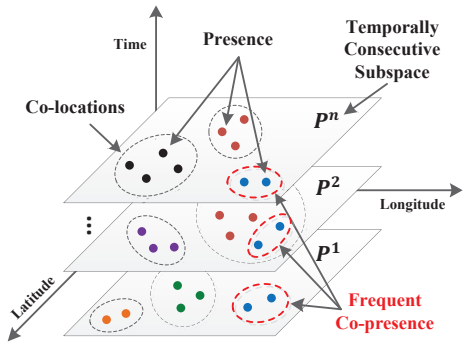


Figure 5. Illustration of the FCPM algorithm.

By mining frequent itemsets, we find sets of users who show up together in multiple co-locations (the pair of blue points in Figure 5).

Algorithm 1: The FCPM for LMA Users.

Input : Users' presence histories: $\{\mathbf{P}_k\}_{k=1}^{\mathbb{N}}$; The height of each subspace: δ; Parameters of density based clustering algorithm: $\mathbf{w}$; The minimum support of frequent co-presence: θ_{min};

Output: Sets of all Frequent co-presences F

// (1) Clustering

Along the timeline, split the collection of $\{\mathbf{P}_k\}_{k=1}^{\mathbb{N}}$ into consecutive subspaces $\mathbf{P}^1, \mathbf{P}^2, ..., \mathbf{P}^n$, such that the temporal differences among all presences in $\mathbf{P}^i$ are less than δ;

Initial $co\_location\_set$ as an empty set;

for $i = 1$ *to* n **do**

  $C^1_{\mathbf{P}^i}, C^2_{\mathbf{P}^i}, ..., C^m_{\mathbf{P}^i} = Clustering(\mathbf{P}^i, \mathbf{w})$;

  for $j = 1$ *to* m **do**

    merge all users in $C^j_{\mathbf{P}^i}$ to form a user set $U^j_{\mathbf{P}^i}$;

    add $U^j_{\mathbf{P}^i}$ to $co\_location\_set$;

// (2) Frequent Co-presence Mining

$\mathbf{F} = FIM(co\_location\_set, \theta_{min})$;

return $\mathbf{F}$;

IV. Experiment

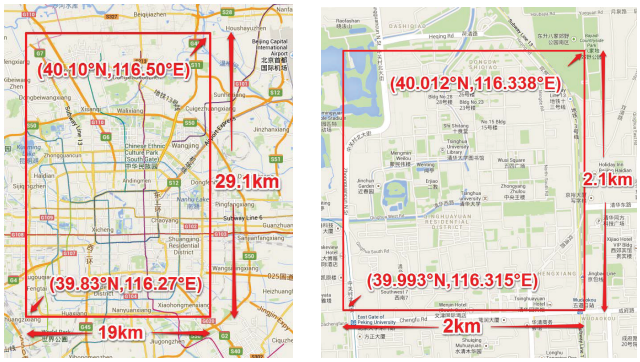
In this section, we conduct experiments to demonstrate the feasibility of automated profiling via abusing the people nearby function and evaluate the effectiveness of our system.

A. Ethical Considerations

This paper presents a systematic method to derive LMA users' life profiles, which unavoidably brings some ethical and legal issues when conducting experiments involving diverse private information. However, nothing could be more reliable or convincible to verify our method than carrying out empirical experiments. In this work, to fully respect for user privacy, the authors make sure that any LMA users involved in the experiments are properly protected by anonymization. The data is used for research purpose only, to carry out statistic analysis. Individual identities of users are unknown to any participated personnels. We also claim that the conducted experiments are only for academic purpose and all privacy related data will never be used for further penetrations or provided to any irrelevant third party.

B. LMA User Collection

In our experiment, we select two areas (Figure 6) to generate the disguised spot lists: the larger one $\mathbb{C}$ covers a city-level region (about 550 km^2) which is discretized into a grid with the resolution of 0.01° in latitude and 0.01° in longitude, containing 672 disguised spots in total. The smaller area $\mathbb{U}$ targets a university-level region, covering a rectangle region about 4 km^2. 480 disguised spots are generated by a grid with the resolution of 0.001° in latitude and 0.001° in longitude.



(a) Area $\mathbb{C}$ (b) Area $\mathbb{U}$

Figure 6. The chosen targeted areas.

Momo

Profile Field	Area $\mathbb{C}$ (184,084 users) Completeness Ratio	Area $\mathbb{U}$ (12,320 users) Completeness Ratio
User Name	100.00%	100.00%
Age	100.00%	100.00%
Gender	100.00%	100.00%
Constellation	100.00%	100.00%
Photo & Avatar	100.00%	100.00%
Profession	68.79%	71.92%
Interests & Hobbies	62.88%	62.51%
Education	43.59%	51.44%
Company	37.72%	34.75%
About Me	76.38%	73.90%
Places often visit	65.11%	65.37%
Interest Groups	38.18%	33.16%

WeChat

Profile Field	Area $\mathbb{C}$ (85,722 users) Completeness Ratio
User Name	100.00%
Gender	30.60%
Photo & Avatar	66.57%
What's Up	75.76%
Region	99.65%

Figure 7. Collected LMA user profiles.

LMA User Collector is deployed on emulators and derived from Android 4.1.1. Two popular LMAs are chosen: Momo (4.3.2) and WeChat[3](5.0). In Momo, the cycle of scanning (line by line) through area $\mathbb{C}$ is about 6 days and 4 days for area $\mathbb{U}$. In WeChat, the cycle for area $\mathbb{C}$ is about 8 days[4]. Note that the scanning circle is measured on only one LMA User Collector, which can be reduced linearly by deploying multiple LMA User Collectors in parallel with little effort due to the system scalability. It means that if there are sufficient resources, e.g., to deploy one LMA User Collector for each disguised spot in area $\mathbb{C}$ separately, the overall time cost will be reduced to only

[3]WeChat (http://www.wechat.com) is also a popular LMA, who has owned over 400 million users by June, 2013.

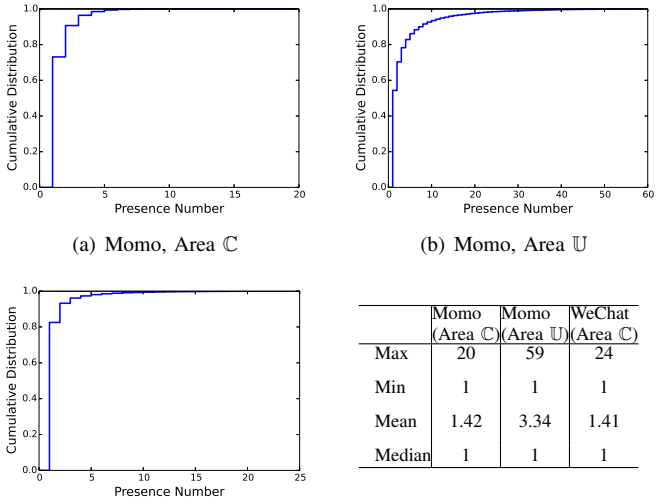
[4]We did not try WeChat on area $\mathbb{U}$ because the numerical distance precision in WeChat is 100 meters which is too large for the scale of the university.

several minutes. The statistics of the obtained user profiles are shown in Figure 7. During only one month period, we have collected 282,126 user profiles in total, with good quality of a high overall profile completeness ratio 74.42%. Specifically, 23.4% Momo users and 38.7% WeChat users mark themselves as female. These proportions do not agree with that in physical world where the proportion of female is much higher. It seems that males are keen on the way of making new friends based on locations more than females. In Momo, the average and median age of users are 27.8 and 25 respectively, indicating that LMAs are most popular among the youth. Our data also shows that 8.7% of users are students and 28.6% are attending or have attended universities/colleges. It is observed that the results rely much on the popularity of the LMA, and also on the different profile schemas in LMAs. Clearly, Momo provides much more fields than WeChat. In general, this result is encouraging for verifying our concern that it is viable to collect user profiles by abusing the people nearby function in LMAs.

C. Evaluation of Life Profiler

1) Geo-locator Evaluation: We first evaluate Geo-locator to see how well it can achieve in locating users' presences. Recall that the task of Geo-locator is to locate user presences as accurate as possible according to the range-based raw geo-measures displayed in LMAs. Among all the collected LMA users, 31.0% (87510/282126) have their presence accuracy improved. The average of the error ranges over the entire dataset has been reduced by 38.5%, from 0.631 km (raw geo-measures) to 0.388 km (approximate presences). The statistics of the presence number per user in Momo and WeChat are shown in Figure 8. First, we observe that the distribution of the number of presences per user conforms to the power law phenomenon [6] in all cases (Figure 8 (a-c)), that is, most of the users have only a few presences while this quantity for a minority of active users is relative larger. It can also be seen that the presences left by users in area $\mathbb{U}$ are in general richer than those in area $\mathbb{C}$. In Momo, the average number of presences per user in area $\mathbb{U}$ is over two times larger than that in area $\mathbb{C}$. One obvious reason is that the collection cycle of traversing through the university is shorter, resulting in more rounds of collection. A more subtle reason relates to the difference between the functionalities of these two regions. It is the general case that most students study and live in the university, akin to a local community. Their daily lives typically involve shuttling among different locations not too far away from each other within the university, such as dormitories, laboratories, and canteens. Therefore, it is not surprising that we can easily capture multiple presences of one user in this area.

2) Daily Activity Induction: We exhibit the results of deducing daily activities of LMA users with the help of the Activity Snooper. In our experiment, two activity types with respect to home and workplace privacy are evaluated, i.e., work/study (daytime: assuming to be from 9am to 6pm on weekdays) and sleep (night: assuming to be from 11pm to 6am). Corresponding to these particular activity types, the selected location types contain "residential region" and "office/teaching region". The mappings from geo-coordinates to location types are built from the Web. We find that 26.5% (71369/269806) users in area $\mathbb{C}$ have been spotted at least once in areas covered by any



(a) Momo, Area $\mathbb{C}$

(b) Momo, Area $\mathbb{U}$

(c) WeChat, Area $\mathbb{C}$

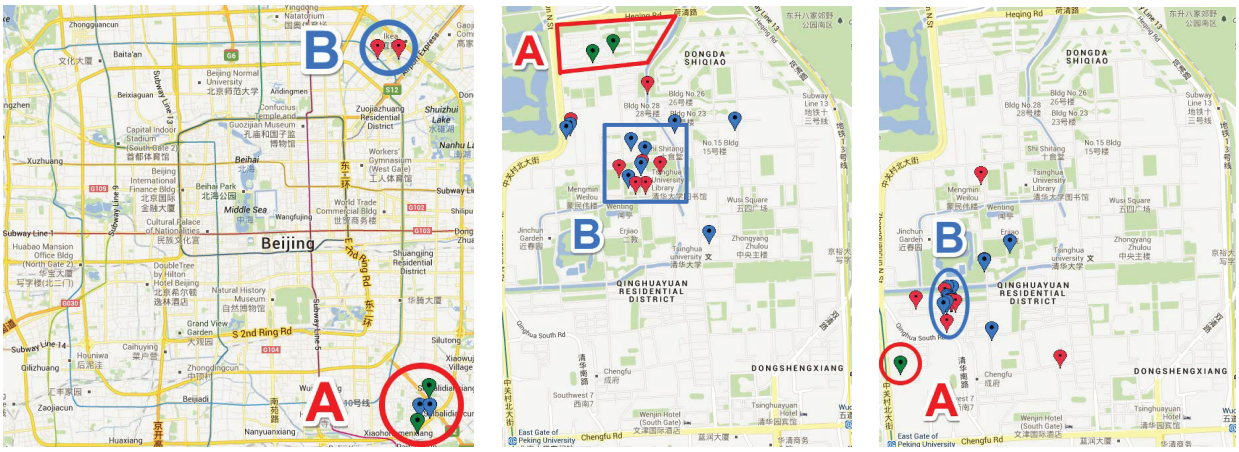
	Momo (Area $\mathbb{C}$)	Momo (Area $\mathbb{U}$)	WeChat (Area $\mathbb{C}$)
Max	20	59	24
Min	1	1	1
Mean	1.42	3.34	1.41
Median	1	1	1

(d) Statistics of the presence number per user

Figure 8. Statistics of presence number per user in Momo and WeChat. (a-c) show the cumulative distribution over the located presence number for each user in respective settings. (d) shows a statistical summary of such data.

residential region at night and 38.4% (103685/269806) in areas covered by any office/teaching region during the day. In area $\mathbb{U}$, such ratios are 40.5% (4984/12320) and 49.8% (6135/12320) respectively. We also show three representative cases found in our dataset (Figure 9). It is clear that their presences can be separated into two principle activity regions at different time periods: daytime near area B and night near area A. The user (male, engineer, according to his LMA profile) shown in Figure 9(a) appeared around area B for two times (the red spots) at working hours (17:14:07 and 17:44:59) on weekdays (Oct 17, 2013 and Oct 22, 2013), and showed up near area A for two times around 5 am (the green spots) on Oct 15, 2013 and Nov 1, 2013 (weekdays), separately. Based on these observations, it can be conjectured that this user might be a commuter who shuttles back and forth between his home near area A and the workplace near area B. If this is true, we then obtain more private information about the locations where he lives and works, which is not specified in his LMA profile. More interesting cases are found in the university area $\mathbb{U}$: a student (male) who studies at the university (Figure 9(b)) and a health care worker (male) who works for a health care center located in the university (Figure 9(c)). In Figure 9(b), area A is a part of residential districts of the university and correspondingly the student showed up two times in this area around midnight. In the daytime, the student was found mostly near area B which encompasses two libraries and several teaching buildings. Hence, we may carefully speculate that sleeping in area A and studying in different buildings in area B constitute two parts of his daily routines. In contrast, the health care center is located within area B as shown in Figure 9(c), so it is natural to find the health care worker many times in that area during the day on weekdays. Finally, the worker may live in area A, a residential region just outside the university, due to his presence at midnight over there.

3) Social Tie Inference: Finally, the performance of our FCPM algorithm is evaluated. In practice, the obtained results are hard to validate because the co-appeared users are



(a) An engineer (b) A student (c) A health care worker

Figure 9. Presence histories of three Momo users. One in area $\mathbb{C}$ (a) and two in area $\mathbb{U}$ ((b) and (c)). Only the presences with geo-location error range ≤ 200 meters are shown. The colors of map markers indicate specific time periods: red indicates working hours from 9am to 11am and 2pm to 6pm; green indicates sleeping hours from 11pm to 6am; remaining hours are expressed in blue.

neither volunteered nor hired, thus no ground truth about the real relationships is available. We can only try our best to speculate the relationships by referring jointly to other side information where we hope strong or weak clues could be found to support our inference. Specifically, we use the contents in LMA profiles as golden standard and try to find evidence from them. The success of this algorithm relies on two factors: *proximate in space* - a group of LMA users should be captured multiple times by our life profiling system at different locations; *proximate in time* - at each location, their presences should be within a short time interval. Thus the height of each temporally consecutive subspace $\delta = 30$ minutes. DBSCAN [4] is used as the density based clustering algorithm which requires two parameters: radius ϵ defining the notion of spatial proximity, and MinPts being the minimum number of points required to form a cluster. In our case, we set $\epsilon = 20$ meters and MinPts = 2. The frequent itemset mining algorithm is based on Apriori [5], the minimum support of frequent co-presence $\theta_{min} = 3$. This experiment is performed only on Momo data (196,404 users) since Momo's profile schema is much richer (Figure 7), which facilitates the process of social relationship verification. In the clustering phase, 1,534 subspaces are found, and 12,011 co-locations are generated at the end of the clustering. The final results consist of 1,818 frequent co-presences. By analyzing the Momo user profiles of involved users, it turns out that over half (942/1818=51.8%) of them join the same interest groups, 78.3% (1423/1818) attend the same universities or work in the same professions, and 39.2% (712/1818) often visit the same places. We argue that although these by no means necessarily indicate the existence of real social ties, such facts can somehow give us certain clues about the potential relationships of the frequent co-presence users in case of the lack of ground truth in our experiment. Finally, we present a real case of potential relationship found by the social tie inference with relative high confidence. Their co-presences are shown in Figure 10. As we can see, they are spotted to be very close in space and time in four different areas (A-D) on three separate days, where area B shows a clear companion pattern of these two users. In other three areas, their presences overlap with each other within the distance no more than 20 meters. The time intervals of their co-presences are shown in Table I. In addition, their LMA profiles also show some commonalities: 1) both of them are studying in the same university; 2) they both join the same interest group; 3) they are of the same age. Admittedly we have no knowledge about their real relationship, however, based on all evidence shown above, we can carefully draw the conclusion that the possibility of their being friends (or at least acquaintances) should be much higher than that of their being strangers.

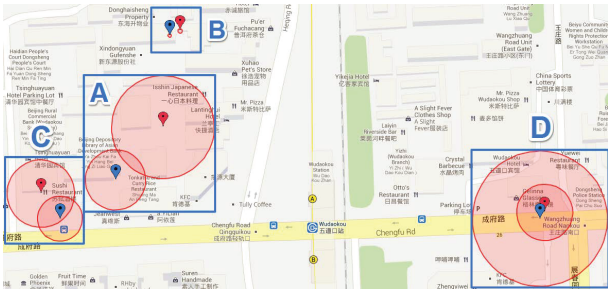


Figure 10. The presence histories (from Nov 5 to Nov 7, 2013) of two Momo users (red and blue) who are presumable to be friends with each other.

Co-presence	From	To	Time Interval
A	2013-11-05 17:53:36	2013-11-05 18:22:44	$\leq$ 29 min
B	2013-11-05 22:17:59	2013-11-05 22:40:20	$\leq$ 23 min
C	2013-11-06 23:41:03	2013-11-06 23:57:29	$\leq$ 27 min
D	2013-11-07 15:17:52	2013-11-07 15:45:52	$\leq$ 28 min

Table I. TIME INTERVALS OF THE 4 CO-PRESENCES.

V. DISCUSSION

In this section, we discuss some practical issues and considerations related to our method to automatically profile LMA users through abusing the people nearby function.

Limitations. First, the Geo-locator suffers from the inherent locating errors of the build-in location-acquisition technologies (e.g., GPS, Wi-Fi, and GSM networks) in mobile devices. Thus, geo-measures in LMAs may be inaccurate based on such location data, which will further affect the performance of Life Profiler. Fortunately, such errors can be considerably reduced if LMAs adopt more accurate location technologies which integrate data from multiple location information sources in mobile devices, e.g., utilizing the third-party location SDKs. Secondly, the construction of life profiles relies on the usage patterns of LMA users, specifically the usage frequency. If a targeted user seldom uses LMAs, the performance of Life Profiler will be limited by the amount of spatio-temporal data collected from that user. Finally, the completeness and truthfulness of profiles volunteered by users are hard to guarantee. The authenticity of user profiles have been discussed in [7] and [8]. Incorrect decisions caused by untruthful profiles can be mitigated with the help of authenticity judgements, which will be left for the future work.

Countermeasure. The fact that our method is based on the abuse of a normal function in LMAs gets us into a countermeasure dilemma. The people nearby function in LMAs is fundamental for connecting people in vicinity. This is also the root that leads to the rise of privacy issues presented in this work. Indeed, the most straightforward strategy on LMA server side that limits the frequency and quantity of issuing queries per user can weaken attackers who have limited resources. Nonetheless, it may fail to prevent such attack where multiple puppet accounts are used in a distributed way. On the other hand, from users' perspective, although they can hide their locations from disclosure, this however may affect the experience of location based socializing in LMAs, making it no advantage over traditional messaging applications. Having made such consideration, we hereby point out some pragmatic rules and designs to mitigate privacy risk in LMAs in practice. For example, in order to reduce the likelihood of potential privacy leakage, LMAs can strengthen the people nearby function by allowing users to make their LMA geo-measures invisible to nearby users at sensitive time or places, even during the use of this feature. Also, it is of great duty for LMAs to pop-up tips at appropriate time, so as to remind users of the potential exposure of their current locations, which may increase users' awareness about the concerned privacy risks effectively.

VI. RELATED WORK

Our current work relates to multiple lines of research. Li and Chen [9] find that people's attitudes towards location based privacy vary from their ages, genders, and geographic regions, nevertheless, the consensus is the concern about who can be aware of their temporal location privacy which is indeed sensitive and personal, not only for the reveal of people's personal itinerary but also for the associated implication and contextual meaning [10].

Location-related contents in online social networks have been studied for years [11]–[15]. Qu et al. [14] model trade areas and consumer-store interactions using User Generated Mobile Location Data (UGMLD) generated from users' check-in data. A probabilistic model is then applied to the task of location based mobile advertising by modeling users' preferences. Friedland et al. [12] discuss privacy implications in the context of geo-tagging and show how to compromise one's privacy by correlating geo-tagged data with corresponding publicly-available information, which typically comes from the postings on different online social networks, such as Twitter and YouTube. Location-related contents discussed in above literatures sometimes may not be accessible under strict user privacy settings. In this paper, we also face the similar problem as they do. However, the difference is that our focus is on privacy issues in a novel setting (i.e., location-based messaging applications) where the trade-off between the LMAs' usability and users' location privacy settings creates a protection dilemma. If LMA users choose to hide their locations, the LMAs would degenerate into "MAs", losing the strength of location based socializing. Otherwise, their location data can be accessed by our life profiling system once they connect with others via people nearby discovering.

Location based mining has also drawn a lot of attention [2], [16]–[19]. Cranshaw et al. [19] collect traces of mobile users in the physical world to explore their relationships on online social networks. They collect data from volunteers who update their location every 10 minutes based on specific equipment. Their work proves the fact that users' online location information indeed can reflect the real-world relationships and inspires us to take full advantage of the user presences gathered from LMAs to explore richer social privacy of users. Besides, Crandall et al. [2] find that even a small number of co-occurrences can result in a high empirical likelihood of a social tie, which supports our experiment results under current condition of no ground truth. Meanwhile, the uncovered social ties from our experiments can be seen as a testimony to such conclusion.

On the defense side, a growing privacy preserving techniques have been proposed to protect LBS from privacy violation in various aspects [20]–[26]. In online social networks, user privacy access control is the only way provided to adjust the exposure degree of users' privacy data. Ho et al. [24] point out the insufficiency in the privacy protection mechanisms in most existing online social networks, and propose a more robust privacy management framework. Camilli et al. [22] address co-location privacy threat in geo-aware social networks which concerns the availability of information about the presence of multiple users in a same locations at given times. This threat is elevated by the sharing of geo-temporal tagged contents involving multiple users, and mitigated by generalizing the tags of resources via temporal cloaking or user erasure. Puttaswamy and Zhao [25] transform all users' locations shared with the server and encrypt all location data stored on the server with inexpensive symmetric keys so as to protect users' location data from being accessed by unauthorized users. However, these protections may be insufficient against our attack scenario. The reason is obvious that a standard function in LMAs is in effect utilized to harvest users' location and profile data, which makes our behaviours no difference with normal users if necessary operations are adopted (e.g., in a distributed manner).

VII. CONCLUSIONS

The location-based mobile messaging applications have won people's favor and enjoyed a large user population. However, the developers' failure in the lack of full control of resolving the trade-off between functionality provision and user privacy preservation results in some unnecessary privacy threat. In this work, we give a comprehensive understanding of the privacy threat in location-based mobile messaging applications. By abusing the build-in people nearby function, an automated user profiling system is built for the attack scenario. Particularly, the system has a very high level of scalability and adaptability, which is easy to deploy in a distributed way and requires no application modification or trivial protocol reverse engineering. Besides, the approaches of user presence locating, activity induction, and frequent co-presence mining succeed to profile LMA users from various aspects: Who is this, Where is (s)he and When, What is (s)he doing, Whom is that with him/her. For the first time, we conduct the experiment and evaluate the threat on a large scale. The experimental results containing more than 280,000 user profiles not only validate the viability of automated profiling through abusing LMAs' normal functions, but also show real cases of daily routine patterns of both individual and correlated users. All of these - real and severe ongoing privacy threats as we have seen in the evaluation - shed new light on the privacy issues in LMAs.

VIII. ACKNOWLEDGMENTS

This work was supported by National Program on Key Basic Research Project (Grant No. 2012CB315804), National Natural Science Foundation of China (Grant No. 61073179 and No. 91118006), and Beijing Municipal Natural Science Foundation (Grant No. 4122086).

REFERENCES

[1] Balachander Krishnamurthy and Craig E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the 3rd Conference on Online Social Networks*, pages 4–4, 2010.

[2] David J. Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[3] Yasuhiko Morimoto. Mining frequent neighboring class sets in spatial databases. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 353–358, 2001.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[5] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216, 1993.

[6] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.

[7] Marco Balduzzi, Christian Platzer, Thorsten Holz, Engin Kirda, Davide Balzarotti, and Christopher Kruegel. Abusing social networks for automated user profiling. In Somesh Jha, Robin Sommer, and Christian Kreibich, editors, *Recent Advances in Intrusion Detection*, volume 6307, pages 422–441. 2010.

[8] Yao Cheng, Lingyun Ying, Sibei Jiao, Purui Su, and Dengguo Feng. Bind your phone number with caution: Automated user profiling through address book matching on smartphone. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security*, pages 335–340, 2013.

[9] Nan Li and Guanling Chen. Sharing location in online social networks. *IEEE Network*, 24(5):20–25, 2010.

[10] Aristea-Maria Zafeiropoulou, David Millard, Craig Webber, and Kieron O'Hara. Privacy implications of location and contextual data on the social web. In *ACM Web Science Conference*, 2011.

[11] Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, and Christian S Jensen. Location-related privacy in geo-social networks. *IEEE Internet Computing*, 15(3):20–27, 2011.

[12] Gerald Friedland and Robin Sommer. Cybercasing the joint: on the privacy implications of geo-tagging. In *Proceedings of the 5th USENIX conference on Hot topics in security*, pages 1–8, 2010.

[13] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 34:1–34:10, 2008.

[14] Yan Qu and Jun Zhang. Trade area analysis using user generated mobile location data. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 1053–1064, 2013.

[15] Anthony Stefanidis, Andrew Crooks, and Jacek Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013.

[16] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, 25(3):12, 2007.

[17] Jin Soung Yoo, Shashi Shekhar, John Smith, and Julius P Kumquat. A partial join approach for mining co-location patterns. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages 241–249. ACM, 2004.

[18] Jin Soung Yoo, Shashi Shekhar, and Mete Celik. A join-less approach for co-location pattern mining: A summary of results. In *Data Mining, Fifth IEEE International Conference on*, pages 4–pp. IEEE, 2005.

[19] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pages 119–128, 2010.

[20] Sergio Mascetti, Dario Freni, Claudio Bettini, X Sean Wang, and Sushil Jajodia. Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies. *The International Journal on Very Large Data Bases*, 20(4):541–566, 2011.

[21] Greg Bigwood, Fehmi Ben Abdesslem, and Tristan Henderson. Predicting location-sharing privacy preferences in social network applications. In *Proceedings of the First Workshop on recent advances in behavior prediction and pro-active pervasive computing*, 2012.

[22] Matteo Camilli. Preserving co-location privacy in geo-social networks. *arXiv preprint arXiv:1203.3946*, 2012.

[23] Dario Freni, Carmen Ruiz Vicente, Sergio Mascetti, Claudio Bettini, and Christian S. Jensen. Preserving location and absence privacy in geo-social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 309–318, 2010.

[24] Ai Ho, A. Maiga, and E. Aimeur. Privacy protection issues in social networking sites. In *IEEE/ACS International Conference on Computer Systems and Applications*, pages 271–278, 2009.

[25] Krishna P. N. Puttaswamy and Ben Y. Zhao. Preserving privacy in location-based mobile social applications. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications*, pages 1–6, 2010.

[26] Xinxin Zhao, Lingjun Li, and Guoliang Xue. Checking in without worries: Location privacy in location based social networks. In *The 32st Annual IEEE International Conference on Computer Communications*, pages 3003–3011, 2013.