Mobile Applications Tracking Wireless User Location

Sara Motahari, Hui Zang, Soshant Bali, Phyllis Reuther
Advanced Analytics Labs, Sprint Research
Sprint-Nextel
Burlingame, CA
{sara.gatmir-motahari,hui.zang,soshant.bali,phyllis.reuther}@sprint.com

***Abstract*-Location-based services enabled by broadband wireless access play an increasingly important role in people's daily navigation and coordination. Location-based applications frequently report user location to Internet servers, and location accuracy is essential to the utility of these services. However, *regular* and *accurate* location updates impact efficient usage of network resources and also users' privacy, which is not directly observed by the users. In this paper, we conducted a large scale measurement study to understand location accuracy and communication frequency of such applications. We found that while most location reports are accurate enough, some applications run in the background, reporting user locations with a higher accuracy and frequency than needed for the user's purpose. For example, while from the user perspective, hourly zip code updates are enough location information to obtain local weather conditions, some weather forecast applications report user location every five minutes or less and at GPS-level accuracy. We found that location reports from many phone applications are accurate enough and frequent enough to enable the inference of users home and work addresses, and potentially their identity, exacerbating user privacy concerns.**

I. INTRODUCTION

In 1996, the Federal Communications Commission (FCC) of the United States mandated all wireless service providers to provide E-911 service by which the mobile user location is reported when they dial an emergency call (911). The localization system developed for E-911 has enabled many Location Based Services (LBS). With the introduction and increasing popularity of smart phones, the development of location based applications has boomed in recent years. With LBS, users' current and past locations are used to enhance a service. For example, a weather service provides the weather forecast for the user location without requesting the user to input a city or zip-code; or a search service returns a list of pizza parlors sorted by the distance to the user's current location.

To enable LBS, wireless carriers provide location APIs that deliver the location of a mobile device to applications. The location provided could be obtained through GPS or network measurement. Such dissemination of users' location raises many concerns including inefficient use of network and bandwidth resources, battery drain, and privacy risks. Phone resources, such as battery life, have received increasing attention and discussion as many users begin to complain about the battery life of smart phones. For example, third party applications and their poor design are blamed for high battery consumption in Android phones [12]. While users can directly observe the battery drain on their phones, they are not aware of the impact of too accurate or too frequent location traces on the usage of bandwidth and their data plans and on their privacy. Specifically, location data collected through applications, smart phone applications in particular, have not received much attention. This is despite the fact that location history can pose serious privacy risks. For example, methods are developed in [1] to identify mobile users in cellular traces when sufficient history data is available for characterizing the mobility patterns before-hand. Therefore, the ability to locate users and track user movement will help an adversary to re-identify the users even if the users are anonymized. This is referred to as re-identification attacks. Attention has been given to techniques to improve the privacy while providing LBS [2,3,4,5,6]. The IETF Geopriv working group [11] has applied many techniques when designing protocols and APIs to enable privacy-preserving location data access. The question we are addressing by conducting this research is whether the applications running in people's phones today are properly designed to preserve their privacy. We are investigating to what extent the design of today's applications has focused on preserving user privacy and phone resources.
Towards this end, we examine location and user information contained in IP packets collected from a 3G data network. We discover that highly accurate location data is periodically transmitted to application servers, with each individual device uniquely identified. We explain that this poses serious privacy risks as users' home and work locations, and consequently their identity can be easily inferred by someone who can access the data collected by one such application.

To the best of our knowledge, this is the first measurement study trying to assess the amount of, and at what frequency, location information is accessed and reported by location based applications. The rest of the paper is organized as follows. Section 2 describes existing methods for localizing mobile devices. In Section 3 we explain how we carried out this investigation by inspecting the Internet traffic, and in Section 4 we report some of the results.

II. BACKGROUND: LOCATION ESTIMATION IN WIRELESS NETWORKS

In this section, we explain established techniques of localization and location estimation methods that are used in wireless mobile communications.

978-1-4673-0921-9/12/$31.00 ©2012 IEEE

A. Proximity

Proximity sensing is based on the closeness of a device to a reference point to estimate the location of the device. In this technique, the device location is usually estimated to be the location of the reference point [7]. In the case of wireless networks, since the mobile user is connected to a nearby cell tower, this technique reports the location of the cell tower as the location of the mobile device. Since a cell sector size can be hundreds of meters, this technique offers low accuracy.

B. Round trip time and signal attenuation

Round trip time (RTT) is the time it takes for the electromagnetic signal to travel from the transmitter to the receiver and back from the receiver to the transmitter. This time, coupled with the amount of attenuation in the signal power, can be used to estimate the distance of the mobile device from the cell tower [7]. For the common sectorized cellular network, the user location will be reported somewhere on the middle axis of the sector based on the calculated distance. This technique is more accurate than proximity estimation.

C. Trilateration and Triangulation

Trilateration computes the position of a device by measuring the distance between the device and a number of reference points at known locations, where the number of reference points required for computing the location is one greater than the number of the physical space dimensions [7]. This distance can be calculated from the round trip time or the angle of arrival. Triangulation uses the angle of arrival of signals traveling from a device to reference points. This technique is more accurate than proximity and round trip time estimation. GPS uses four-dimensional trilateration.

III. METHODOLOGY

The user location data was extracted from the IP usage traffic of a wireless service provider. An optical splitter was placed in a CDMA network between the Packet Data Serving Node and the Home Agent to forward a copy of the packets traversing the network. Data from users in the Southwest region of the USA was collected for the two weeks from Mar 7 2011, 15:28 PST to Mar 21 2011, 13:02 PST.

A. Location Extraction

We took a deep packet inspection approach and examined the URL in every HTTP GET and HTTP POST packet leaving the network. Whenever we found a pair of real numbers that could represent latitude and longitude of a location within mainland USA, we saved it as a potential location. We also examined the Geo-Position header in HTTP requests to extract user locations. The Geo-Position header is commonly used by Android phones to attach location information to Google search. We only examined URLs and Geo-Position headers, and not the rest of the packet payload. Our results are only based on the location information in these headers. We also collected call detail records during the same time period. These records contain information about the sector and cell tower that the user was connected to. Thus, they help us verify a user's actual location and ensure that the packet did not represent an incorrect or irrelevant location.

B. Collected Data

The data was collected for two weeks and belonged to a total of 10,500 users nationwide. About 0.01% of the records included location data, which we extracted for further analysis. We found many domains that report location data. The services offered by these domains fall into seven main categories: 1) Place search; 2) Proximity search; 3) Weather forecast; 4) Local news; 5) Media and games; 6) Shopping; 7) Advertisement and analytics.

Tables 1 and 2 display access statistics for the top 15 location consumers. Table 1 shows the percentage of users who accessed a domain. Table 2 shows the percentage of location-included requests made by those users.

As seen in the tables, services offered by weather.com and where.com are accessed by the most users and send the most frequent requests. However, ad.where.com comes in third for users accessing this site, but ws.geonames.org is third in percentage of requests.

C. Location Accuracy Determination

Consecutive estimations of a user location usually do not return the exact same longitude and latitude values. This can

TABLE 1: TOP 15 LOCATION CONSUMERS BY PERCENT USERS

Host	Type of service	Percent Users
wxdata.weather.com	Weather forecast	47.81%
api.where.com	Proximity search	27.82%
ad.where.com	Analytics and advertisement	16.02%
ws.geonames.org	Place search	7.19%
www.google.com	Place search- proximity search	5.35%
pws.ipoynt.com	Media and games	4.21%
api.scvngr.com	Media and games	3.47%
maps.google.com	Place search	1.93%
sense.dailymotion.com	Media and games	1.28%
web.mapquest.com	Place search	1.09%
direct.weatherbug.com	Weather forecast	1.09%
metrics.doapps.com	Local news - weather forecast	1.04%
app.shopkick.com	Shopping	0.99%
opml.radiotime.com	Media and games	0.89%
ads.mobclix.com	Analytics and advertisement	0.89%

TABLE 2: TOP 15 LOCATION CONSUMERS BY PERCENT REQUESTS

Host	Type of service	Percent requests
wxdata.weather.com	Weather forecast	50.72%
api.where.com	Proximity search	25.80%
ws.geonames.org	Place search	17.49%
ad.where.com	Proximity search	1.80%
pws.ipoynt.com	Media and games	0.56%
clip.doapps.com	Local news and weather forecast	0.54%
www.google.com	Place search- proximity search	0.47%
opml.radiotime.com	Media and games	0.42%
api.scvngr.com	Media and games	0.39%
www.radarnow.net	Weather forecast	0.28%
mt.hiplogic.com	Media and games	0.18%
maps.google.com	Place search	0.15%
direct.weaterbug.com	Weather forecast	0.10%
api.the_nd.com	Analytics and advertisement	0.09%
metrics.doapps.com	Local news and weather forecast	0.08%

be due to two reasons: 1) the accuracy of location measurement is limited. For example, location estimations based on signal travel time have a varying error caused by multipath RF propagation; and 2) the user may be mobile and changes their location between consecutive measurements. For determining location accuracy we are interested in variations in location that are due to estimation error and not to mobility. Thus, we will only focus on stationary users.

A user who is stationary now might start moving later. Therefore, we filtered the collected location records to those with 30 samples to represent a short period during which the user was either mobile or stationary with a high probability. We call each interval of 30 samples a 'trace'.

The ideal way to measure estimation accuracy of a user's location would be to use the accuracy value provided by the location-based application on the device. However, we found that only a few applications, such as Google search which includes estimated positional uncertainty (epu), attach an accuracy value to location reports. Most other applications do not provide location accuracy values. However, we know that our collected location samples were estimated based on one of the techniques explained in Section 2: GPS measurements (4-dimensional trilateration), signal travel time from the cell tower (RTT), or cell tower replacement (proximity). Therefore, we use metrics previously shown to be correlated with location accuracy to classify the traces into different classes. The classification was done by clustering as explained below, and is meant to separate GPS (high accuracy) location data from round trip time and proximity (low accuracy) techniques. We will then show that different classes actually show different patterns which can relate to the margin and nature of their estimation error. For example, location estimations obtained from GPS are scattered around a center and vary only a few meters, while locations estimations that are just based on the serving cell tower, may jump between cell towers and their variations can be hundreds of meters. Moreover, mobile location traces show a different pattern from stationary ones.

The algorithm used for clustering was k-nearest neighborhood and leave-one-out cross validation was used for evaluation of the classification. We used the following metrics to separate mobile traces from stationary ones and classify them based on their accuracy:

a. Location entropy: Location entropy shows the amount of variation and randomness in a user's location. If a user visited K locations and the number of appearances at location i is x_i, 0<i<K, and the user appeared M times in total, then the frequency at which the user visited location i is x_i/M. Let X be the random variable representing the user's location. The entropy of X can be calculated as follows:

$H(X) = \sum^{K} [x_i/M.\log(x_i/M)]$

Location entropy is affected by the movement of the user and also by the accuracy of location estimation, i.e., for users that are not moving, location entropy can differentiate how accurate the location estimation is. Reporting the cell tower location instead of the actual user location is expected to result in low entropy. GPS localization is expected to have the highest entropy because the results from repeated measurements are usually different, although close to each other. To differentiate GPS- localized stationary users from moving ones, we introduce block-level entropy.

b. Block level entropy: We divide the map into non-overlapping 20m by 20m blocks, and compute entropy over these blocks. If a user visited L blocks and the number of appearances at block i is y_i, 0<i<K, and the user appeared N times in total, then the frequency at which the user visited block i is y_i/N. Let Y be the random variable representing the user's blocks. The entropy of Y can be calculated as follows:

$H(Y) = \sum^{L} [y_i/N.\log(y_i/N)]$

Block level entropy can detect motion. Users that are moving are expected to have higher block level entropy than static users. Therefore, changes seen in their location are due to actual movement and not estimation error.

c. Location revisit ratio (LRR): Location revisit ratio is the ratio of the number of revisits to the total number of locations in the trace. This metric indicates whether the general motion of the user is fixed and is good for determining stationary vs. mobile status. It is also useful for determining location accuracy as cell tower replacement will have a high LRR.

d. Radius of Gyration: Radius of gyration measures the size of the footprint of a user [9]. This metric shows how big user movements are and how far in different directions they travel. Radius of Gyration is the linear size that a user occupies and is calculated as the root mean square distance of the location points from the centroid of the trace:

$$r_g(K) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\vec{r_k} - \vec{r_c})^2}$$

where the vector $\vec{r_k}$ represents the k=1…K locations recorded for the trace and $\vec{r_c}$ is the centroid of the trajectory.

D. Localization Frequency Determination

To investigate how often different applications report a user location, we looked into the time interval between two consecutive location reports from the same host to the same destination. For this purpose, we analyzed the data from the entire experiment duration. The results of this analysis and the above clustering analysis are explained in the next section.

IV. RESULTS

A. Location Classes

Location traces that fall under different classes (clusters) show noticeably different patterns. They also show different levels of location entropy, block entropy, and location revisit ratio. Figure 5 shows the probability distribution of block entropy for moving classes versus stationary classes. As the figure shows, the distribution is concentrated around zero for stationary classes, which means that the location samples were confined within one block during the picked time interval.

Based on their visual patterns, we named the *stationary user* clusters as follows: Spray Paint, Dancer, and Hopper.

a) Spray Paint. Spray paint is a type of stationary location trace which has the highest accuracy because it shows

high location entropy, lower block-entropy and a low LRR (which means location estimations remain in the same proximity and do not jump between cells). A sample of such location trace is shown in Figure 1. We think that among round trip time, GPS and proximity, this location class is related to GPS estimation.

b) Dancer Dancer is a type of location trace where the user hops around between a number of locations. This class has low block-entropy, medium-high LRR, and medium-high location entropy. Therefore, it must relate to stationary users whose locations are estimated with a lower accuracy than Spray Paint, but with higher accuracy than Hopper. We think that this cluster comes from round trip time measurement as the method of location estimation. An example of Dancer is shown in Figure 2.

c) Hopper Hopper hops between multiple fixed locations which are the actual locations of the cell towers. This is the case when the device simply reports the cell tower location as its own location. The reason for this apparent hopping is that the mobile device connects to the cell tower with the best signal strength at any given time. When the user is close to the borders of the cells, the device can switch back and forth between cell towers. This class has low block entropy, low location entropy, and very high LLR (as it jumps between fixed locations). An example of Hopper class is shown in Figure 3.

Figure 4 shows the histogram of the location entropy for these three stationary classes. We see that as the accuracy increases, the average level of location entropy increases.

d) Mobile classes. Mobile classes have high block entropy, high radius of gyration, and high location entropy. As explained before, we will not focus on them in this paper. An example of a location trace from a mobile user is shown in Figure 6.

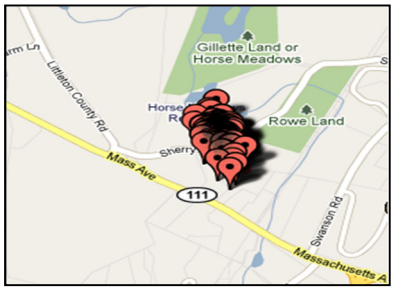

Fig. 1. Spray paint class

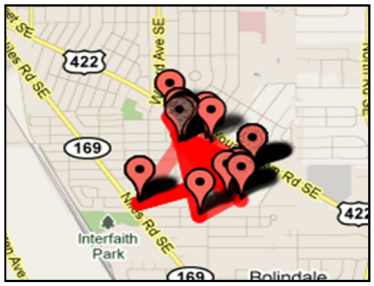

Fig. 2. Dancer

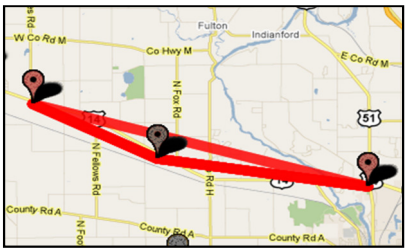

Fig. 3. Hopper

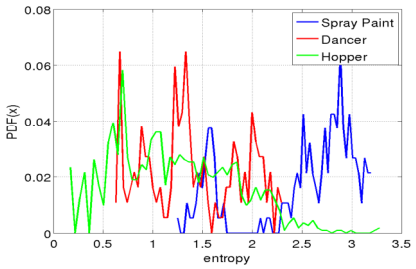

Fig. 4. Histogram of entropy for stationary clusters.

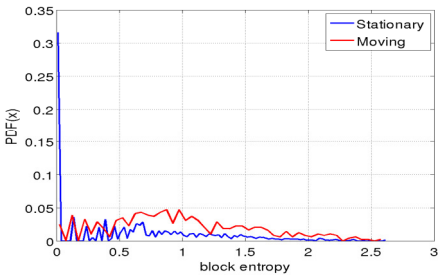

Fig.5. Block entropy for stationary vs. moving classes

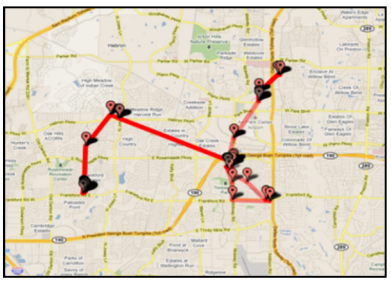
Fig. 6. Location trace of a mobile user.

B. *Classification Statistics*

Table 3 shows the statistics of the resulting classes after classifying the collected data. The first three rows show stationary classes and the last row shows the percentage of users in mobile classes. As Table 3 shows, 36.4% of all records belong to the high accuracy stationary class (Spray Paint). This class contains more than half of all stationary records. We also extracted the records that belonged to weather.com (Tables 1 and 2show that most records belong to this domain). Classification results for weather.com are shown in Table 4. Focusing on stationary classes, about 30% of locations collected by this domain are high accuracy localizations.

TABLE 3. CLASSIFICATION RESULTS FOR ALL DOMAINS.

Class	Percentage
Spray Paint	36.4%
Dancer	19.6%
Hopper	11.2%
Mobile Classes	32.7%

TABLE 4. CLASSIFICATION RESULTS FOR WEATHER.COM.

Class	Percentage
Spray Paint	19%
Dancer	37.3%
Hopper	9.2%
Mobile Classes	34.6%

C. *Localization Frequency*

The per user inter-location time probability density function (PDF) is shown in logarithmic scale in Figure 7. About half the inter-location reports are less than 1 second apart. The distribution, centered around a mean of 0, clearly shows that while location reports are bursty in nature, there are reports that are sent periodically, creating spikes in the PDF. Applications that periodically report the location regularly are probably running as background processes collecting location data, while other application only report when a user explicitly accesses them.

We should note that this method does not capture all applications running in the background since some applications do run in the background but send the updates at low, irregular intervals and some are designed to run in the background but may not be captured in the IP traffic because of transmission errors or application bugs.

Figures 8 and 9 show the histogram of inter-localization time for the domains weather.com and where.com respectively (the top two applications in Table 1). These figures are also plotted in logarithmic scale. As the figures show, the histogram is weighted around zero, which means location updates are sent very frequently. The median of inter-location time is 3.1 seconds for weather.com and 2.1 seconds for where.com.

D. *Identification of Frequent Locations*

Our analysis of the data collected from *one week* shows that location records obtained from many of the observed domains, including weather.com and where.com can enable anyone with access to these records to identify the two locations that a user visits most often. These locations are likely to be their home and work locations [15]. Previous studies have suggested that

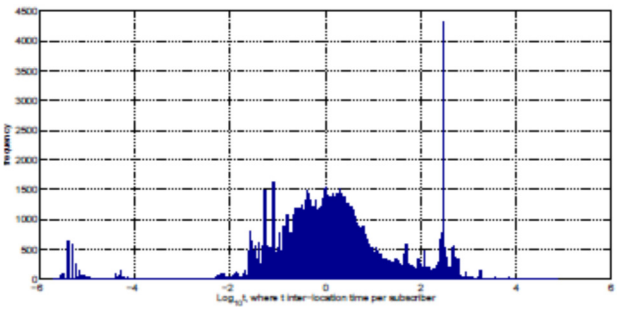

Fig. 7. Histogram of inter-localization time for all domains.

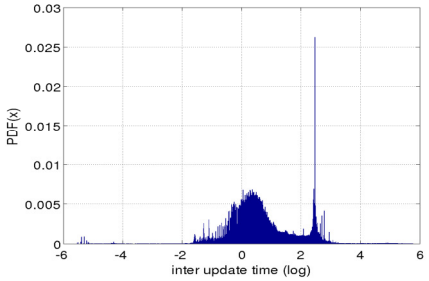

Fig. 8. Histogram of inter-localization time for where.com

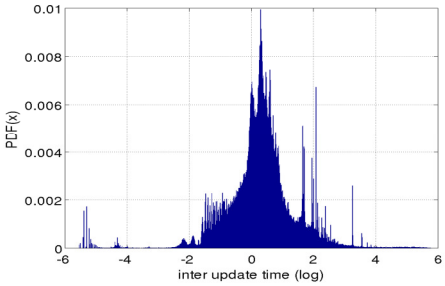

Fig. 9. Histogram of inter-localization time for weather.com.

a minimum of two-weeks of location records are usually needed to identify user mobility patterns [15]. However, our analysis shows that when it comes to the two most frequent locations, that aggressive domains such as weather.com and where.com could identify the top two frequent locations of a user within a week. Since the top two locations that people frequent are usually their home and work place, the immediate privacy concerns relate to the possibility of associating home address and work places to IP address or other identifiers in the location reports. However, as we'll explain in the next section, even if these records were anonymized, the user identity could still be inferred from them.

V. DISCUSSION

Table 3 shows that most of the location-collecting applications are very accurate in estimating user location. In addition, Figure 7 shows that some domains collect location information very frequently. For some applications, such as live navigation, this accuracy and frequency is needed. However, Tables 1 and 2 show that most of them are weather forecasts, local news, and entertainment applications. Indeed, most of location requests from weather.com happen about every three seconds and almost 30% are estimated with a few meters, even though the weather condition reports only change by zip-code and within many minutes, not every few seconds. Therefore, from the user perspective, location accuracy is collected at a rate which is far more detailed than their needs.

Location reports that are more accurate and/or more frequent than needed for their original purpose can pose a threat to user privacy. Currently, anonymization efforts rely on not disclosing personally identifiable information, such as name, SSN, etc. However, having any anonymized unique identifier along with other pieces of information revealed by the application could enable the *inference* of a user's identity. This can happen when the anonymity degree [10] (the number of candidates who match the revealed information) is so low

that the actual user can be correctly identified with a very high probability. The inference problem as a threat to privacy has been investigated in a few studies [12]. For example, it has been shown that 87% of the U.S. population is uniquely identifiable knowing their gender, zip code, and date of birth [14]. Accurate and frequent *location reports* can also lead to identity inference. This is because only a very small portion of people share both their work and home address, which leads to a very low anonymity degree. We conducted a study with the purpose of measuring the anonymity degrees of the users having their frequent locations [16]. The results show that even when the precision of location reports have low accuracy (cell tower sector), the median of the anonymity degree is 2 and almost 40% of the users are uniquely identifiable. Therefore, location traces collected by domains such as weather.com can enable the data collectors to infer users' home and work addresses, and potentially, their identities.

VI. Conclusion And Future Work

Location-based applications can report or store location data with differing precision or frequency. Privacy implications of the granularity and frequency of location reports are hidden issues with wide implications.

We collected IP traffic on a large scale to study what percentage of the Internet traffic packets and what types of applications collect user location data and with what accuracy they estimate the locations. Since most of those applications do not provide estimated accuracy values, we used an unsupervised learning method to classify the location data into high accuracy, medium accuracy, and low accuracy records. While based on the features that we used, the resulting clusters show obvious patterns of stationary versus mobile users and low accuracy versus high accuracy location estimations, further steps can be taken to verify this clustering. For example, controlled studies or test data with known accuracy can offer a better test of our classification. We are also looking for other ways of detecting the source and technique of location estimation.

Our observations of frequency and accuracy of location estimations indicate a mismatch between application requirements from the user point of view, users' concerns, and the real implementation of applications. Many of the applications that run on data enabled devices are not designed with the aim of providing the best privacy protection for the users. Many applications send too many queries and track a user too precisely. To identify potential ways of improving their performance as the next step, it is important to collect more knowledge about the design and implementation of existing location-based applications. Analyzing today's devices and applications, best design practices, users' privacy concerns, and the obtained location traces from the Internet traffic can lead to providing recommendations and guidelines on the limits and the type of location data reported by the applications.

In addition to privacy concerns, it is likely that the frequency of location reports impacts the amount of resources the applications consume, such as battery power and bandwidth. While this impact has been visible to users, few applications are developed to provide the least resource consumption. Future work may consider investigating this resource consumption issues for location updates.

VII. References

1. Y. De Mulder, G. Danezis, L. Batina, and B. Preneel, ``Identification via location-profiling in gsm networks," in WPES '08: Proceedings of the 7th ACM workshop on Privacy in the electronic society. New York, NY, USA: ACM, 2008, pp. 23--32.
2. M. Gruteser and X. Liu, "Protecting privacy in continuous location-tracking applications," IEEE Security and Privacy, vol. 2, no. 2, pp. 28–34, 2004.
3. M. Gruteser and D. Grunwald, "Anonymous usage of location-based services through spatial and temporal cloaking," in MobiSys '03: Proceedings of the 1st international conference on Mobile systems, applications and services. New York, NY, USA: ACM, 2003, pp. 31–42.
4. J. Krumm, "Inference attacks on location tracks," in PERVASIVE'07: Proceedings of the 5th international conference on Pervasive computing. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 127–143.
5. L. Kulik, "Privacy for real-time location-based services," SIGSPATIAL Special, vol. 1, no. 2, pp. 9–14, 2009.
6. A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," IEEE Pervasive Computing, vol. 2, no. 1, pp. 46–55.
7. Location in Ubiquitous Computing, Alexander Varshavsky, Shwetak Patel, *Ubiquitous Computing Fundamentals, 2009, pp. 285-320*
8. N. Bulusu, J. Heidemann, and D. Estrin, GPS-less low cost outdoor localization for very small devices, IEEE Personal Communications Magazine, vol.7, no.5, pp. 28--34, October 2000.
9. M. C. González, C. A. Hidalgo, A.L. Barabási, Understanding individual human mobility patterns. *Nature* 453, 779-782, 2008.
10. L. Sweeney, "k-anonymity: a model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl.-Based Syst., vol.10, no.5, pp. 557--570, 2002.
11. IETF Geopriv working group. http://datatracker.ietf.org/wg/geopriv/charter/.
12. S. Motahari, S.G. Ziavras, Q. Jones, Online Anonymity Protection in Computer-Mediated Communication, IEEE transactions on Information Forensics and Security, 5(3), 570-580, 2010
13. C. Davis, http://androidcommunity.com/google-android-apps-to-blame-for-subpar-battery-life-20100519/, May 2010
14. L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population, LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.
15. H. Zang and J. C. Bolot, Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks, *Proceedings of ACM Mobicom '07*. Sept. 2007.
16. H. Zang and J. C. Bolot, Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study. *Proceedings of ACM Mobicom* '11, Sept. 2011.