Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study

Hui Zang
Sprint
1 Adrian Ct
Burlingame, CA 94010, USA
hui.zang@sprint.com

Jean Bolot*
Technicolor
735 Emerson St
Palo Alto, CA 94301, USA
jean.bolot@technicolor.com

ABSTRACT

We examine a very large-scale data set of more than 30 billion call records made by 25 million cell phone users across all 50 states of the US and attempt to determine to what extent anonymized location data can reveal private user information. Our approach is to infer, from the call records, the "top N" locations for each user and correlate this information with publicly-available side information such as census data. For example, the measured "top 2" locations likely correspond to home and work locations, the "top 3" to home, work, and shopping/school/commute path locations. We consider the cases where those "top N" locations are measured with different levels of granularity, ranging from a cell sector to whole cell, zip code, city, county and state. We then compute the anonymity set, namely the number of users uniquely identified by a given set of "top N" locations at different granularity levels.

We find that the "top 1" location does not typically yield small anonymity sets. However, the top 2 and top 3 locations do, certainly at the sector or cell-level granularity. We consider a variety of different factors that might impact the size of the anonymity set, for example the distance between the "top N" locations or the geographic environment (rural vs urban). We also examine to what extent specific side information, in particular the size of the user's social network, decrease the anonymity set and therefore increase risks to privacy. Our study shows that sharing anonymized location data will likely lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain (meaning the data is collected over short periods of time, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level). In both cases, the utility of the anonymized location data will be decreased, potentially by a significant amount.

*Part of this work was done while the author was working at Sprint.

MobiCom'11, September 19–23, 2011, Las Vegas, Nevada, USA.

Categories and Subject Descriptors

C.2.m [**Computer-Communication Networks**]: [Miscellaneous]; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement

Keywords

Cellular Data, Location, Privacy, k-anonymity

1. INTRODUCTION

Decades ago, people used to worry about leaving their keys or wallets behind when they left their house. Nowadays, we worry about leaving our cell phones behind (and we still check for keys and wallets...) Indeed, cell phones have becomes an increasingly important part of life and it is no wonder that nationwide penetration rates trend towards 100% in many developed countries, and are already significantly above 100% in several others.

Because they are ubiquitous and because they have become a natural part of daily life (meaning that users are not expected to behave abnormally when they carry their phone), cell phones have become a powerful tool to analyze human behavior, in particular as it relates to physical places through the study of mobility patterns, and the research interest in this area has increased dramatically over the past few years [3, 27, 9, 28, 24]. The availability of mobility and location data also drives a vibrant ecosystem of location-based services ranging from navigation to proximity advertising (mobile coupons) with a plethora of new services introduced daily. All those services need access to some kind of location and mobility data. Cellular operators collect such location information, in particular Call Details Records (CDRs) in particular for billing and troubleshooting purposes. CDRs contain information about every call carried by the cellular network, including time of the call, location, and identities of both parties involved in the call. Thus, they enable the study of human mobility at a very large scales. With the increasing need or desire to publish or share CDRs or CDR-like call logs to third parties, the issue of privacy has become a major concern. The default approach has often been to anonymize CDRs or other call logs and replace user identities with random identifiers. A key question then is: Can these anonymized records be

shared or published? That is precisely the question we are trying to answer in this paper.

Privacy breach occurs when users are re-identified from anonymized data. Past studies have shown that the majority of US population can be uniquely identified by the combination of gender, zip-code, and birth-date [25]. The fraction of identifiable individuals ranges from 87% based on the 1990 census data [25] to 63% based on the 2000 census data [7]. In a related study, the medical records of a past governor of Massachusetts were identified from anonymized health insurance data because his birthdate, gender and home address are available through voter's data [26]. More recently, Golle and Partridge [8] showed that a fraction of the US working population can be uniquely identified by their home and work locations even when those locations are not known at a fine scale or granularity. Given that the locations most frequently visited by a mobile user often correspond to the home and work locations, the risk in releasing locations traces of mobile phone users appears very high.

Our work is inspired by Reference [8] but we look at the problem from a different perspective. Instead of focusing on home and work locations, we consider the *"top N"* locations visited by each user. It is natural to think of the top 2 locations as home and work for a large fraction of the population, but more generally the number N of top preferential locations determines the power of an adversary and the safety of a user's privacy. The more top locations an adversary knows about a target, the easier it is to identify the target. The fewer top locations a user has, the safer they are in terms of privacy.

In this paper, we study a data set of 30 billion CDRs from a nationwide cellular service provider in the United States which contains location information about 25 million mobile phone users collected over a three-month period. We consider the risk of releasing location data from such CDRs with anonymized user identifiers. We identify important factors that impact the anonymity of released location data, such as the value of N, the granularity level at which locations are reported, whether the top locations are ordered or unordered, the availability of additional social information about users, and geographical regions. We consider six granularity levels in this study, namely sector, cell, zip-code, city, county, and state. To the best of our knowledge, a study at this magnitude of scale has never happened before.

Our results show that releasing anonymized location data in its original format, i.e. at the sector level or cell level, poses serious privacy threats as a significant fraction of users can be re-identified from the anonymized data. They also show that different geographical areas have different levels of privacy risks, and at a different granularity level this risk may be higher or lower than other areas. In general, we find that preserving privacy requires that data be published either at very coarse granularity level, namely city level or above, or at very short time durations, e.g. a day. Both solutions will compromise the utility of the location data, were the data to be used for modeling mobility. To cite an argument from [2]: "even modest privacy gains require almost complete destruction of data-mining utility". The published data may only be of use to very high-level studies and/or product developments such as mobile advertisements.

The impact of our work is multi-fold. We provide guidelines to cellular operators interesting in publishing or sharing location data. We suggest spatial and time domain treatments to make the sharing of location data privacy-preserving. We show that traces collected from different geographical regions may be treated differently because they exhibit different levels of anonymity by nature. While our work may be somewhat discouraging regarding the release of location data in its original form (namely at a fine granularity such as sector or cell level), we do provide important guidelines on how location data *can* be published, even at the cost of reduced utility. The lessons learned from this work can also be of reference when other types of location data is to be published, in particular location data collected through mobile applications that involve location updates for example as done with Foursquare [5]. In general, though, we strongly recommend that the community be extremely cautious when publishing anonymized location data.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 provides background knowledge on data and techniques used in this study. We conduct an in-depth study on various factors impacting anonymity in Section 4 and explore two types of methods to anonymize location data and discuss their privacy-utility trade-off in Section 5. Section 6 concludes the paper.

2. RELATED WORK

Our work is inspired by Reference [8]. While [8] relied on US Census data, we rely on call data from a US cellular provider. Instead of using reported home and work locations, we infer the top N preferential locations from the call data and perform similar tests (as well as many others) as those in [8]. We show that releasing location data also faces the threat of re-identification attacks.

Our work is also inspired by the concept of k-anonymity proposed in [26]. k-anonymity quantifies the degree of privacy of an anonymized data set, specifically with k-anonymity, each individual is indistinguishable from at least $k-1$ others, i.e., is "hiding in crowd of k". When applied to location traces, k-anonymity means that the mobility behavior of a user has to be similar to that of at least $k-1$ others. More recent work has shown that k-anonymity fails to provide strong privacy protection [23] which led to the introduction of l-diversity [20] and t-closeness [15]. We choose k-anonymity to evaluate the privacy risks of published location traces for two reasons. First, despite being unsatisfactory, k-anonymity is still widely used in many studies including [8]. Second, although k-anonymity is not a good enough metric to ensure privacy, we found it to be a good enough to evaluate privacy vulnerability.

Location privacy, the risks associated with the release of location data and the techniques to quantify and mitigate those risks, have gained much attention in the literature [11, 10, 12, 14, 1, 13]. A general theme is that reducing the granularity of location information can help improve privacy protection. In this paper, we rely on large-scale data to demonstrate the effect of reducing granularity on the level of privacy protection, or specifically on the obtainable values of k for k-anonymity.

Privacy threats with published mobility traces have been considered in the literature. For example, methods are developed in [3] to identify mobile users in cellular traces when sufficient history data is available to characterize the mobile users a priori. Their study is similar to ours in spirit but with a fundamental difference. Our work does not assume the availability of detailed mobility traces, which was required in

[3], to generate the transition probability matrix or the stationary Markov distribution of all visited cells per user. Our study assumes that an adversary would know the preferred locations of each user and correlate them for example with publicly available information about users' home and work locations. Reference [18] considers traces obtained from public transportation systems (cabs and buses). In Reference [19], the authors propose a technique to anonymize the commuting data of U.S. workers while satisfying privacy requirements. Differential privacy is described in [4] to provide privacy guarantees in statistical databases. The technique works by adding the appropriate amount of noise to the result of queries, the benefit being that a user's presence or absence from the database cannot be determined by queries. A method is proposed in [16] to connect differential privacy and k-anonymity by adding a random sampling step before "safe" k-anonymization. Random sampling or random noise insertion impacts the utility of data and the tradeoff between utility and privacy need to be carefully considered [21, 17, 2]. A framework is proposed in [17] for evaluating the change of utility and privacy for various privacy models and provide guidelines for choosing the right tradeoff. A good survey of privacy issues in data publishing can be found in [6].

Human mobility has been shownn to be highly predictable in [28] and [24]. Both studies rely on cellular network traces to study the mobility patterns of mobile users. Specifically, reference [28] shows that the entropy of locations does not increase much beyond 14 days (meaning that it is enough to observe users for 14 days to analyze their behavior) whereas reference [24] shows that with 93% probability one can correctly predict an individual's location independent of how far that person travels among the preferred locations. Related work based on mobility traces of 100,000 mobile phone users [9] shows that human mobility is regular in both temporal and spatial domains and with a high probability each user returns to a few preferred locations. Finally, in another study of the locations visited by 3G users, the authors of [27] find that users spend a significant fraction of time in their top three locations only.

3. BACKGROUND

In this section, we describe the data used in this paper and relevant background knowledge on privacy and anonymity.

3.1 Dataset

We use the Call Data Records (CDR) from a nationwide US cellular provider collected over three entire months, February, March, and April 2010. Our dataset consists of about 25 million distinct users and over 30 billion call records. The users span all fifty states and tens of thousands of base stations.

A call record is created when a call originates or terminates on the cellular network and it contains various fields of information regarding that call. Table 1 lists several fields from an example CDR.

We separate the trace into three month-long segments and study them separately. We process CDRs from each day and identify the locations visited by each user. We create daily location lists for each user with all locations and the number of appearances at that location. Then for each month, we aggregate the daily location lists into a monthly location list and order the locations by the frequency of appearances.

Table 1: Selected fields from a sample CDR

Field		Value
Mobile ID		987-654-3210
Time of call		2010 02 02 12 33 02
Call duration		300 (seconds)
Start location	Cell ID	153
	Sector ID	2
End location	Cell ID	157
	Sector ID	1
Call direction		incoming
Caller ID		987-012-3456

The top N locations for each user can be easily identified from these location lists.

Some mobile users made or received very few calls during the observation period which made it difficult to infer their top N locations accurately. Therefore we filtered out those users who made or received fewer than 30 calls per month, which still left us with about 20 million users.

We study the location data at various levels of granularity. The original location data in a CDR represents location in terms of sectors of cells. There are typically two or three sectors in a cell (meaning that a sector covers a 120-degree sector in a cell). With knowledge of the location of each base station, we can convert the data from one level of granularity to a coarser level: for example, we aggregate data across all 3 sectors of a cell to get cell-level data from sector data. We consider six levels of granularity: sector, cell, zip code, city, county, and state.

Our work differs from past studies in i) the different levels of granularity we consider (past work has mostly focused on cell-level mobility) and ii) the scale of the data we analyze. Regarding scale, the work in [3] is based on mobility data collected from one hundred (100) instrumented smart phones, in [27] on a trace consisting of 281,394 3G users in one metropolitan area of 1,900 square miles, in [9] on the records of 100,000 mobile phone users over a six month period, in [28] on call records of two million (2,000,000) mobile phone users from three large cities collected over a month, and in [24] on a 3-month-long trace of the most active 50,000 users among a set of 10 million mobile users.

3.2 Anonymization techniques

Anonymization techniques have been developed for both data query and data publishing. Our work falls in the category of data publishing in which the published data contains information about individuals. For example, Census data contains birth date, gender, address, and other information for each individual. These attributes form "quasi-identifiers" and individuals whose "quasi-identifiers" are the same form an *anonymity set*.

Common techniques to achieve privacy-preserving data publishing are generalization and suppression. Generalization replaces precise attribute values with less precise ones or value ranges. For example, Sweeney [26] uses the first four digits for zip code instead of all five digits because the latter, combined with gender and birth date information will lead to $k = 1$ for k-anonymity. Suppression is to remove some (often extreme) cases from the dataset. For example, if there is only one person using an identifier and the attribute value of that person is very different from others'

(such as a 3-meter tall human), it would be more difficult to "generalize" the value than to remove that person from the published data. In this paper, we consider generalization in terms of location granularity levels.

4. FACTORS AFFECTING ANONYMITY

In this section, we analyze the dataset described above and consider the different factors that impact the size of the anonymity set, i.e. the number of users uniquely defined by a set of top k preferential locations. We first consider the impact of location granularity in Section 4.1, then the impact of the distance between top locations in Section 4.2. The benefit (to an adversary) of extra social knowledge is examined in Section 4.3 and the anonymity of users in different geographical regions in Section 4.5. We study anonymity sets with variable-length quasi-identifiers in Section 4.6 and examine the stability of location traces in Section 4.7. Finally we study the preferential locations of a group of known users with known preferential locations to consider the issue of ground truth in Section 4.8.

4.1 Impact of location granularity

The location information in CDRs consists of cell ID and sector ID, and (implicitly) the ID of the switch which created the CDR. Therefore, we can represent a location using the format x-y-z, where x is the switch ID, y is the cell ID and z is the sector ID. This is the finest granularity available for location data in our study.

Alternatively, we can ignore the sector ID and represent a location as x-y. This the "cell" granularity level.

We know the zip-code, city, county and state information of each base station in the network. Therefore, instead of using the x-y-z or x-y approach, we can use these higher-level information to represent the location. We refer to those as the zip code, city, county, and state granularity respectively.

We start by considering the size of the anonymity sets obtained at different levels of granularity. For each mobile phone user, we calculate the PDF (probability density function) of locations visited by this user, and sort the locations based on the frequency, most visited first. We then map the user to an anonymity set identified by the top N locations, at various granularity levels. We choose N to be $1, 2$, and 3 in this study.

For example, a user's location list may look like this: 1-10-2, 1-11-1, 2-70-3, ... If we choose $N = 1$, the quasi-identifier "1-10-2" identifies the anonymity set with all users whose most visited location is "1-10-2" (group 1). If we choose $N = 2$, the quasi-identifier is "1-10-2, 1-11-1", and identifies all users whose top two locations are "1-10-2" and "1-11-1", respectively (group 2). Apparently group 2 is a subset of group 1. Therefore, the more locations are used to identify an anonymity set, the smaller the anonymity set is. This is demonstrated in Fig. 1. At each granularity, users are partitioned into anonymity sets. The size of the anonymity set, or the k-anonymity value, characterizes how safe that user is from being re-identified. We sort the users based on decreasing k-anonymity and plot the maximal k-anonymity for various fraction of users. The percentiles are calculated over users rather than anonymity sets.

We note that the position and shape of the top two curves in Fig. 1(e) (county level) are close to the equivalent curves in the county sub-figure in Fig. 1 of [8]. The difference between the two figures is that our data does not have contextual meanings as in home/office and only indicates the relative order of preference. Similarly, the top two curves in Fig. 1(b) are close to their counterparts in the census tract sub-figure. This indicates that the size and significance of a cell is comparable to a census tract.

Table 2: Anonymity set with top 1 location

Location granularity	Size of anonymity set			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	28	71	111	372
Cell	92	220	331	967
Zip code	184	557	909	3125
City	162	487	874	7638
County	802	2972	6272	55649
State	60139	1.5e+05	2.6e+05	7.2e+05

Table 3: Anonymity set with top 2 locations

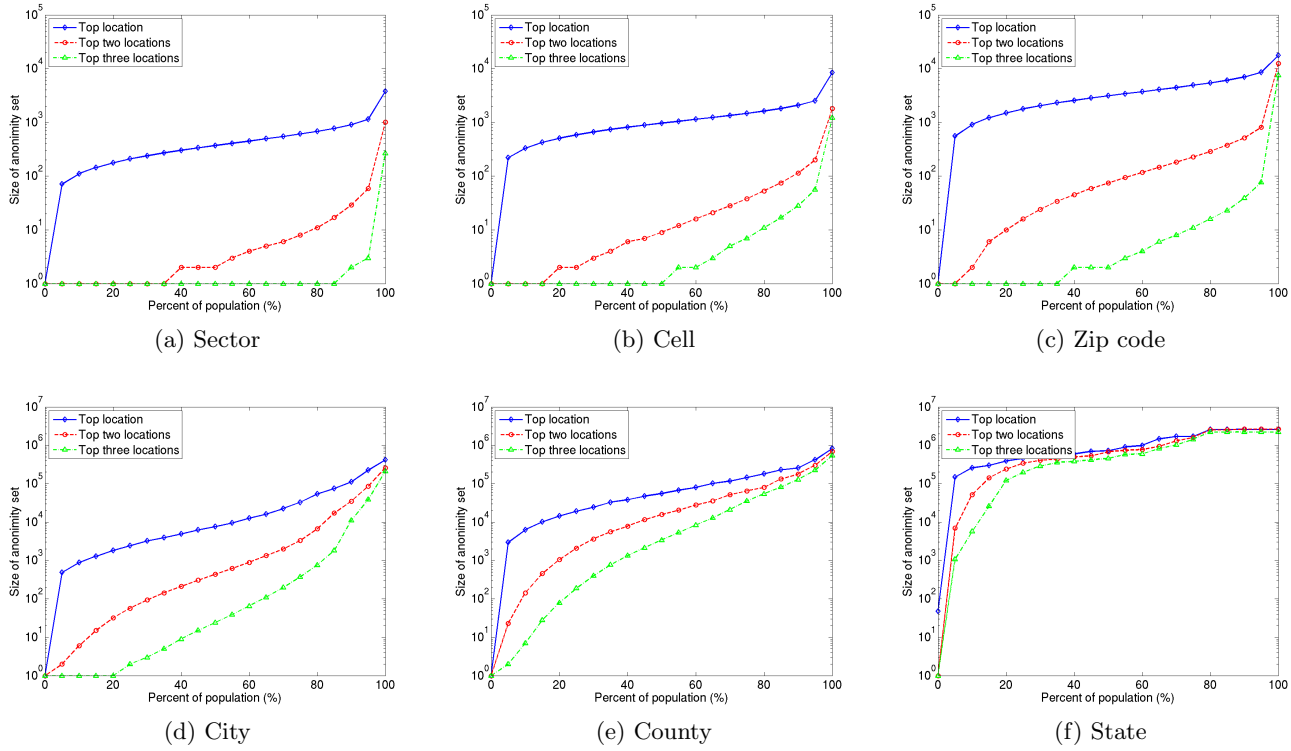
Location granularity	Size of anonymity set			
	1^{st} %ile	5^{th} %ile	10^{th} %ile	Median
Sector	1	1	1	2
Cell	1	1	1	9
Zip code	1	1	2	75
City	1	2	6	437
County	2	23	143	15628
State	530	6912	51291	6.8e+05

Table 4: Anonymity set with top 3 locations

Location granularity	Size of anonymity set			
	1^{st} %tile	5^{th} %tile	10^{th} %tile	Median
Sector	1	1	1	1
Cell	1	1	1	1
Zip code	1	1	1	2
City	1	1	1	24
County	1	2	7	3407
State	40	1074	5671	4.6e+05

We summarize the 1^{st} percentile, the 5^{th} percentile, the 10^{th} percentile and the median of users' k-anonymity values in Tables 2 through 4, for $N = 1, 2, 3$, respectively. Entries in the tables with a "1" indicate the fraction of users which are uniquely identified i.e., they have 1-anonymity. We find that more than 50% of the users at the granularity of sector and cell are uniquely identified by their top 3 locations; between 10% and 50% of the users at the zip code and city levels, and between 1% and 5% at the county level. When the top two locations are used, between 10% and 50% users are uniquely identifiable at the sector and cell levels; more than 5% at the zip code level and 1% at the city level. In other words, if we know the cities corresponding to the top two locations for each user, and given that the total population under study is 20 million, we can uniquely identify 200,000 users.

We plot the median k-anonymity value (i.e. the median size of the anonymity set) from the three tables in Fig. 2. Each curve represents, for a different value of $N = 1, 2, 3$, how location granularity changing from very fine level "sector" to very coarse level "state" affects the median size of anonymity sets. These curves depicts the *radius of uncertainty* for anonymity sets formed based on the top N locations. For $N = 1$, the radius of uncertainty varies from 372 on the sector level to $7.2e + 5$ on the state level. For $N = 2$, it varies from 2 for the sector level to $6.8e + 5$ for the state



(a) Sector (b) Cell (c) Zip code

(d) City (e) County (f) State

Figure 1: Size of anonymity set when top 1, 2, or 3 locations are revealed for different granularity levels

level. For $N = 3$, it stays at 1 for sector and cell levels, goes to 2 for zip code level, and quickly rises to $4.6e+5$ for the state level. We note that for all three values of N, the anonymity set for state level is always above $1e+5$. This is because the majority of the population have only one state in their location lists (we will see this again in Section 5.1).

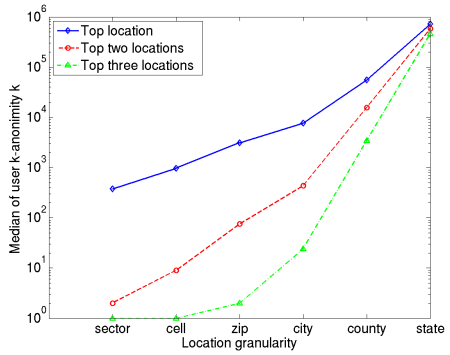


Figure 2: Median size of anonymity set at various granularity levels

4.2 Impact of distance

The authors in [8] show that the anonymity level differs dramatically between individuals whose home and work locations are in the same region and those whose are not. Inspired by this observation, we examine the impact on

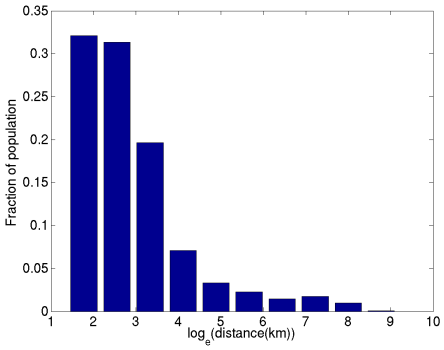


Figure 3: Distribution of distance between the top two preferential locations

anonymity of the distance between the top locations. We focus on the the case of $N = 2$ (top two locations) since it is most relevant and straightforward. We group users according to the distance between their top two locations. Figure 3 shows the distribution of that distance. We then plot the median and mean k values for different distances in Fig. 4. As expected from [8], the longer distance between the top two locations, the smaller the anonymity set. The figure also shows that the size of the anonymity set decreases approximately inversely to the distance between the two locations (the exact fit is $y = 224x^{-1.1}$ on the figure). This is not surprising: given a distance d between the locations and a

granularity level which we represent by a circle of radius R (thus R would represent the average radius of a cell or county), the number of non-overlapping circles that fit at distance d from a specific location is $\pi d/R$ and the number of users in a circle is proportional to $R/\pi d$. The argument above makes hypotheses such as uniform user density but still gives a reasonable explanation for a decrease of the order of $1/d$.

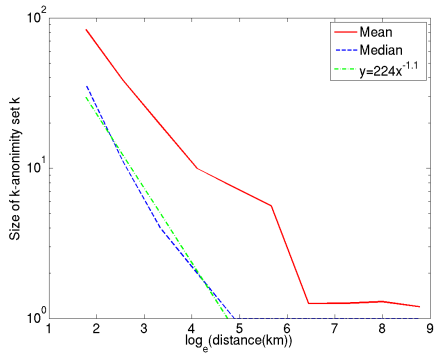


Figure 4: Median and mean size of anonymity set vs. distance between top locations

4.3 Impact of additional social information

We consider the case where more information than just location is released from CDRs, specifically information about the social networks of users. In addition to locations and associated frequency of visits, the number of calls, the number of different "friends" each user called, the number of "friends" that called the user, the total number of "friends" the user talked to, or the number of calls associated with each "friend", might be published as well. We now examine the risks associated with releasing both location as well as that additional social information.

In the following, we assume that the adversary has a small amount of extra information, namely the size of a user's social network measured by the number of calls made by the user in a month. We will assume that the adversary does not know the exact size of the network but it knows whether it is *large* or *small*. In social context, one might be able guess this information based on a user's personality. This extra information can be represented by a single bit representing whether the size of a user's social network is larger than a predefined threshold.

To determine the threshold, we first study the distribution of the size of users' social networks in terms of the number of "friends" they talk to over the given time period. The nodal degree of the call graph has a complex distribution [22] with a long tail. To ease visual representation, we select users with fewer than 150 friends and plot the degree distribution in Fig. 5. We observe that about 50% of the users have social networks with 20 friends or fewer.

Therefore we pick 20 as the threshold and we partition users into two sets: those who made calls to more than 20 friends and those who did not. We then add an extra bit to the quasi-identifiers representing this information. For example, with "top 2" locations, the quasi-identifier $\{loc_0, loc_1\}$ changes to $\{loc_0, loc_1, S\}$ where S is a bit indicating whether

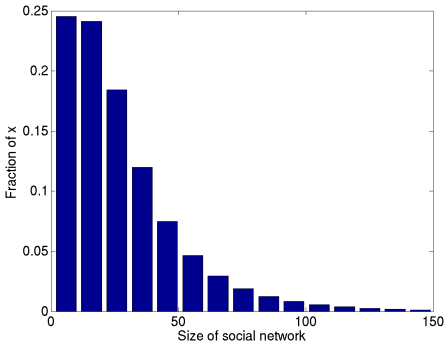


Figure 5: Degree distribution of call graph

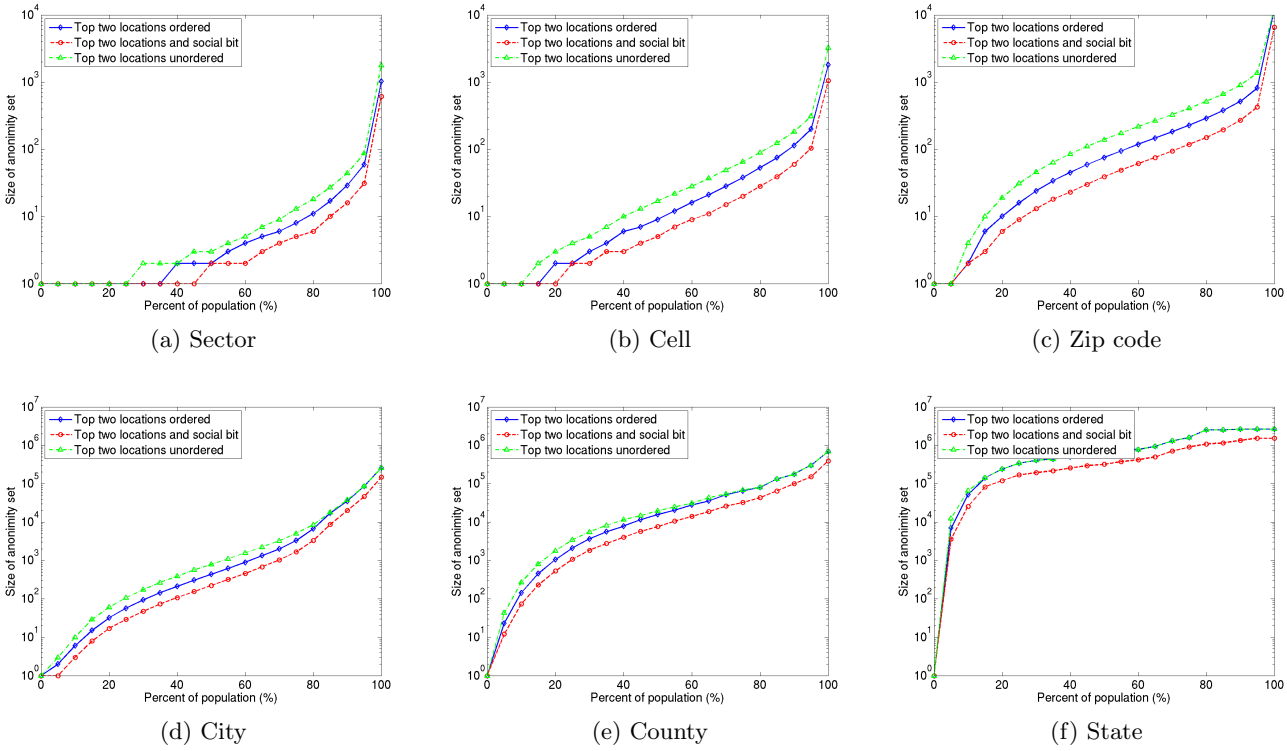
the user's social network is larger than 20 or not. The impact on the size of the anonymity set (in the case of 2 top locations) is shown in Fig. 6. The blue curves denote the anonymity sets corresponding to $\{loc_0, loc_1\}$ and the red curves correspond to the anonymity sets corresponding to $\{loc_0, loc_1, S\}$. We observe that with this extra bit of information, the anonymity sets are smaller at every granularity level. The drop is about 50% in most cases, which demonstrates that social behavior is usually orthogonal to mobility behavior and that additional knowledge about the users' social patterns is helpful to re-identify those users.

4.4 Unordered top locations

We have so far treated the top N locations as separate attributes: users with first location loc_0 and second location loc_1 are in a different anonymity set than users with first location loc_1 and second location loc_0. While for some users, the top location is usually fixed, for others, the top two locations can change from time to time. As we will see in Section 4.8, a user may make most phone calls from home in a month, and from work in another month. Therefore, the top N locations can form *one* attribute regardless of their order. We use the *unordered pair* of the top 2 locations as the quasi-identifier and the size of the corresponding anonymity set is shown in Fig. 6 (green curves). We see that with the unordered pair as the identifier, the anonymity sets are about twice the size of those in the ordered case for finer granularity levels such as sector, cell, and zip code, while the difference reduces at coarser levels. At state level, the difference disappears mainly because the top two locations of the majority of users are in the same state and ordering does not matter.

4.5 A study of different geographical regions

We next investigate anonymity sets for users in different geographical regions. We choose users whose top location is in four selected states: California, New York, Texas, and Illinois. Consider first cell- level granularity. Note that we call a user *region-x-based* or say the user is in region x if the user's top location is in region x. As shown in Fig. 7, the value of k is much larger in Texas than in the other three states. California and New York have very similar behaviors regarding the k values, while Illinois has slightly higher k values in the middle of the curve, i.e., for people whose k rank is between the 40- and 95-percentile. Figure 7 suggests



(a) Sector (b) Cell (c) Zip code

(d) City (e) County (f) State

Figure 6: Anonymity set with additional social information (red) and with unordered top locations (green)

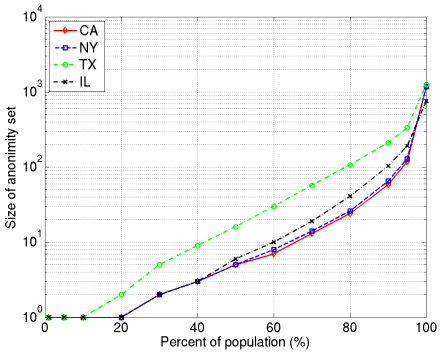


Figure 7: Size of anonymity set for users with top location in California (CA), New York (NY), Texas (TX), and Illinois (IL)

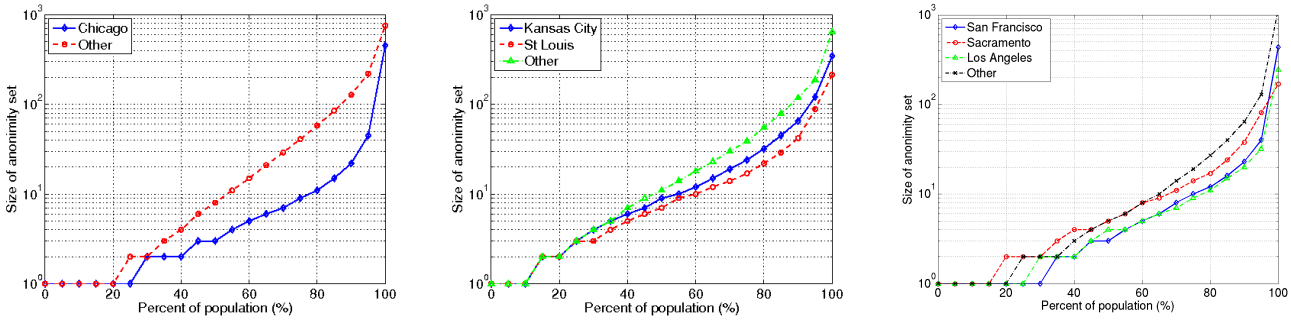
that people in California and New York have very diversified mobility patterns, while people in Texas have much "synchronized" mobility patterns, i.e., users who live close to each other tend to work close to each other too.

To better understand the mobility behavior of users in Illinois, we divide them into two groups: those in the city of Chicago and the rest. The size of the anonymity set is depicted in Fig. 8(a). We see that users in Chicago have much smaller anonymity sets than users in the rest of Illinois. This observation suggests a difference in k-anonymity caused by urban/rural lifestyles: in urban environments, we expect users' lifestyles to be more diversified and hence the anonymity sets to be smaller. We observe a similar behavior for Colorado between users in Denver and users elsewhere in the state.

We then turn to Missouri and California. There are two large cities in Missouri, Kansas City and St. Louis. We therefore partition Missouri-based users into three sets, those based in Kansas City, those based in St. Louis, and the rest. The size of their anonymity sets are plotted in Fig. 8(b), showing that users in St. Louis are quite different from the those in the other areas, whereas users in Kansas City fall somewhere in the middle. If the distinction comes from the rural versus urban life styles, this suggests that St. Louis has more of a urban lifestyle while most areas of the state including Kansas City (although being the largest city in Missouri) are of a rural type.

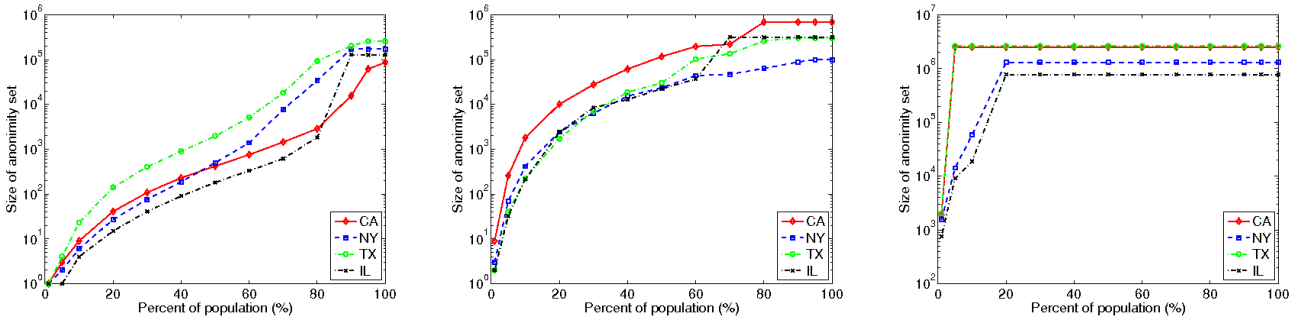
For California, we picked three cities: San Francisco, Sacramento, and Los Angeles. As shown in Fig. 8(c), the anonymity sets of users in San Francisco and Los Angeles are similar. The rest of California is less at re-identification risk than these two cities.

Next we study the anonymity sets of the four states at other granularity levels. Figure 9 shows the distribution of k-anonymity at city, county and state levels, respectively. The result, together with Fig. 7, is rather interesting in the sense that one state may have higher anonymity (larger k) at one level, but lower anonymity (smaller k) at another level. For example, users in California have by far the highest anonymity at the county level (Fig. 9(b)), but have the lowest anonymity at the cell level (Fig. 7). Users in Texas



(a) Illinois (b) Missouri (c) California

Figure 8: Anonymity set for users with top location in Illinois, Missouri and California (with specific cities)



(a) City (b) County (c) State

Figure 9: Size of anonymity set at different granularity levels for users with top location in California (CA), New York (NY), Texas (TX), and Illinois (IL)

Table 5: The ratio of same-region commuters in different states at three granularity levels

	CA	NY	TX	IL
City	0.20	0.19	0.41	0.21
County	0.75	0.47	0.65	0.62
State	0.96	0.90	0.96	0.89

have higher anonymity than the other states at the cell level, but lower anonymity than the others at the county level.

This phenomenon can be linked to the fraction of users who live and work at the same location in each state at different granularity levels. We call these users "same-region commuters". If the top two locations are in the same location region, the anonymity set is much larger than when the two locations are in different regions [8]. Therefore, the fraction of same-region commuters directly affects the distribution of the anonymity sets. Table 5 summarizes the fraction of same-region commuters on the three levels. At state level, both California and Texas have 96% of the users as same-region commuters, which explains why the two curves are so close to each other and also above the rest in Fig. 9(c). At county level, 75% of California users are same-region commuters, 65% for Texas, and only 47% for New York. Texas has by far the highest ratio of same-region commuters at city level and that is why the green curve is far above the others in Fig. 9(a). The high ratio of intra-city commuters might be explained by the affordability of real-estate properties. Houses in Texas are more affordable and users might afford to live in the same city where they work. We are investigating housing affordability indices and their relationships to commuting distance to understand this more clearly.

4.6 Variable-length quasi-identifiers

We have so far only considered a fixed number of top locations as quasi-identifiers. For example, an anonymity set can be represented by a sector, $\{1-5-1\}$, two cells, $\{1-1, 1-5\}$, or three cities, {San Francisco, New York City, Dallas}. Some users mostly visit two locations, for example home and work; others usually stay at one location, and yet others may visit three locations frequently. Therefore the number of "preferred" locations varies between users. In this subsection, we investigate quasi-identifiers with variable length, which depend on the number of preferred locations of each user.

For each user, we sort the preferential locations in decreasing frequency and keep adding locations to the quasi-identifier until the total frequency reaches or exceeds 0.5, i.e., the user spends at least 50% of their time at these loca-

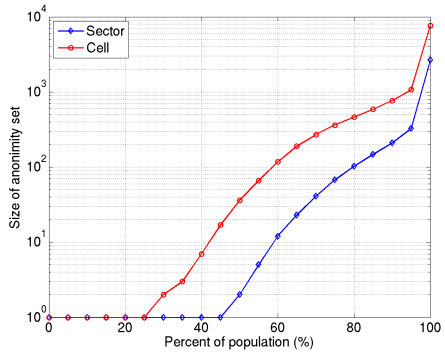


Figure 10: Size of anonymity set using variable-length quasi-identifiers

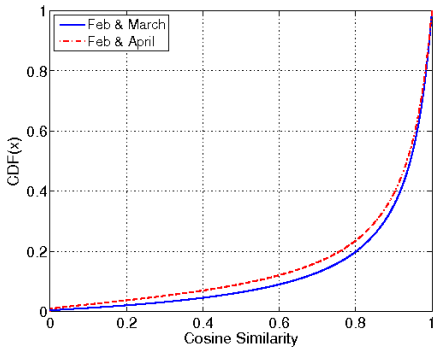


Figure 11: Cosine similarity between two month-long location traces

tions. We call these locations the "preferred" locations of this user. This way, a user with one preferred location is in an anonymity set identified by this single location; a user with two preferred locations will be in an anonymity set identified by the two locations. If an adversary knows the person's life pattern regarding the number of preferred locations and what the preferred locations are, she can re-identify the person if the associated anonymity set has cardinality one.

Figure 10 shows the size of the anonymity set at the sector and cell levels using the variable-length quasi-identifier. The distributions of the two curves are close to their counter parts with top 2 locations (Fig. 1), while being slightly lower. This indicates that most people have two preferred locations while some of them have more. We hence studied the number of "preferred" locations of users and the results did confirm that the majority of users have two or three such locations. The detailed are omitted due to space constraints.

4.7 Similarity of data from two time periods

In this section, we examine whether call logs identify preferential locations in a consistent manner, month after month. Therefore, we examine the probability that a user appears at different cells in two consecutive months and calculate the cosine similarity of the PDF vectors of the two time periods. The length of the vector is the number of cells in the network (tens of thousands) and each element in the vector corresponds to the fraction of time a particular user appeared in the cell. Cosine similarity between vectors A and B is calculated as the dot product of A and B divided by the product of the magnitude of A and the magnitude of B. A cosine similarity equal to 1 indicates two identical vectors over the cells.

Figure 11 shows the cumulative density function (CDF) of the cosine similarity between the months of February and March, and between the months of February and April, respectively. We find that the cosine similarity between February and April is slightly smaller than between February and March. However, the closeness of the two distribution curves suggests that the similarity does not degrade much over one-month periods. At the same time, we notice that between 75% and 80% of the users have cosine similarity greater than 0.8, which means that the location patterns of most users revealed by two month-long call logs are similar month over month. This is in agreement with earlier work in [28].

4.8 Ground truth of preferred locations

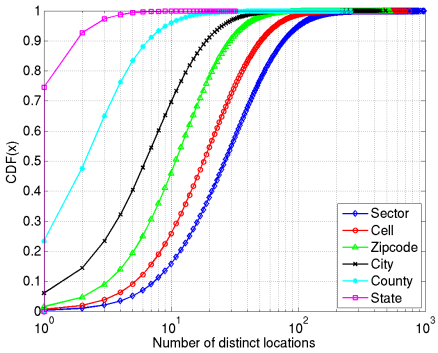
In order to better understand the role of preferred locations, we selected a group of 12 subscribers whom we personally know (and who revealed their exact home and work locations, all 12 have full-time jobs) and studied the correlation between their top 2 locations (which we assume are their home and work locations) and their actual home and work locations. For four of them, their top 2 locations correspond to home and work locations, respectively, for all three months. For four others, their top 2 locations correspond to work and home locations, respectively, for all three months. For two others, their top 2 locations correspond to home and work locations respectively in the first month, but correspond to work and home locations respectively in the subsequent two months. The last two (users A and B) both have a non-home non-work location as their 2^{nd} top location in one of the months. We refer to these non-home non-work locations as "other" locations. For user A, the top locations is always "home" and the second top location is "other" in the first month, and "work" in the second and third months; For user B, the top location is always "work". In the first month, the 2^{nd} top location is "home", and in the second and third months, the 2^{nd} top location is both "other".

5. POSSIBLE SOLUTIONS

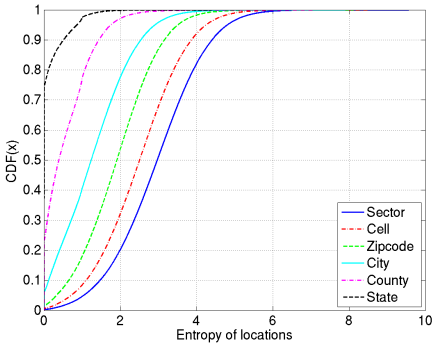
To publish location data without compromising privacy, either of the following conditions needs to be met: 1) The traces are at coarse granularity levels so that the anonymity sets are large enough to preserve privacy, for example the location granularity is at city level or above; or 2) The traces are so short that it is difficult to infer the "top N" locations correctly. The former is a spatial-domain approach while the latter is a time-domain approach. We investigate the privacy-and-utility trade-off of both solutions in this section.

5.1 Spatial domain solution

The spatial domain solution refers to the approach by which location granularity is coarsened to levels that guarantee large enough anonymity sets. Section 4 is dedicated to the privacy implications of this approach. The results revealed the privacy vulnerability of traces at the zip code and finer granularity levels. For example, with top 3 locations, 85% of the users are identifiable at the sector level, 50% at



(a) Number



(b) Entropy

Figure 12: Number and entropy of distinct locations visited by users at different granularity levels

the cell level, and 35% at the zip code level. The results suggest data release at the city level or above.

In this subsection, we focus on the tradeoff between privacy (expressed in terms of granularity levels) and utility. We explore the number and entropy of locations from traces as a measure of the utility for each granularity level (refer to [24] for more detailed study of mobility entropy).

Figure 12 shows the cumulative density function (CDF) of the number and the entropy, respectively, of the locations visited by each user at the six granularity levels. Notice that the x-axis is in logarithmic scale in Fig. 12(a). For the six granularity levels from sector to state, the median of the number of distinct locations visited decreases from 27, 19, 11, 7, 3, to 1.

The entropy of locations indicates the difficulty in predicting a user's locations. If a user u visited K locations and each of these locations appeared x_j times, $0 \leq j \leq K-1$, and user u appeared M times in total, i.e., $\sum_{j=0}^{K-1} = M$, then the user u appears at location j with frequency x_j/M. Let X be the random variable representing the locations visited by user u. The entropy of X can be calculated as follows:

$$H(X) = \sum_{j=0}^{K-1} (x_j/M) log \frac{1}{x_j/M} \quad (1)$$

As shown in Fig. 12(b), the median entropy is 2.97 at the sector level, 2.51 at the cell level, 1.91 at the zip-code level, 1.24 at the city level, 0.39 at the county level, and 0 at the state level. This means that more than half the users visited only one state in a month, which agrees with Fig. 12(a). The quick drop in entropy indicates that the information contained in the location data reduces significantly at coarser granularity levels, and a location trace at a coarser granularity level will be a much *simpler* trajectory than it would be at finer granularity levels. At the state level, most users will appear stationary. At the county level, they may look like ping-pong balls bouncing back and forth between two points. City-level data do have more information, however, the utility of such location data is still significantly limited.

5.2 Time domain solution

The time domain solution refers to the approach by which location traces are truncated into segments of limited duration which do not accurately reveal the "top N" locations. We use sector level location to preserve maximum information in the spatial domain. Our task is to assess the risks associated with location traces of various time durations. We examine, for each user, the extraction of the "top N" locations from traces of m days starting Feb. 2, 2010, $m = 1, \ldots, 27$, and whether the extracted "top N" locations differ from the "top N" locations extracted from the month-long trace in Feb. 2010.

We compute the fraction of users whose "top N" locations are correctly identified through the m-day trace over the range of m and plot it in Fig. 13(a). At the end of the 1^{st} day, the top location of about half of the users is correctly identified, the top 2 locations of 15% of the users are correctly identified, and the top 3 locations of only 3% of the users are correctly identified. After a week (on Feb. 8), the percentages are 75%, 45% and 20%. After two weeks (on Feb. 15), they are 86%, 64% and 40%. This suggests that when day-, week-, or two-week long sector-level location traces are published, those fractions of users are at *potential* risk of being re-identified.

Figure 13(a) is obtained over the entire set of users. Among these users, only a fraction of them belong to small anonymity sets with $k = 1$, i.e., are uniquely identifiable if their top locations are revealed. It is important to know how many of these "re-identifiable" users' top locations are correctly extracted from the day, week, or two-week long traces. To do this, we focus on the set of users who can be re-identified by the top two locations, which is about 35% of the entire population (20 million), and examine the top two locations revealed over the same m-day period (Fig. 13(b)). After the first day, the top two locations of 13% of users are revealed; after a week, 40%; and after two weeks, almost 60%. These fractions serve as an upper bound in the fraction of users being re-identified because incorrectly extracted top locations of other users may interfere with the re-identification of these users.

Let us now assume that we want to publish a day-long trace. We would like to know exactly how many users will be re-identified through the top two locations. Users are not re-identifiable either because the top N locations are incorrectly revealed or because of larger k values for k-anonymity. Therefore, we examine users who are re-identifiable from the month-long trace, and see how many of them, *from the day-long trace*, not only have 1-anonymity *but also* are the top two locations correctly extracted. We choose a day, February 2 2010, for this purpose and the fraction of correctly

re-identified users is 8%, which is significantly lower than the fraction of users for whom the top two locations are correctly revealed (13%). Note that this ratio may change for another day, and if the published day-trace is from a weekend, the ratio may be even lower because people tend to have unusual trajectories over weekends [28]. For example, we repeat the test for Feb. 25 and the ratio is then 7.5%; it drops to 4.4% on Valentine's Day(Feb. 14). The ratio is for 27% the week-long trace (Feb. 2 - Feb. 8) and 34% for the two-week-long trace (Feb. 2 - Feb. 15). Since the total "re-identifiable" users constitute 35% of the total users, fewer than 3% of the total users are re-identified through a day-long trace (Feb. 2), about 10% are re-identified through the week-long trace, and about 12% are re-identified through the two-week-long trace.

These results suggest that the day-long trace is a much safer choice from the privacy perspective. Overall, a trace longer than two weeks reveals more than 50% population's top two locations and should raise privacy concerns. Two weeks is also the time period identified by [28] to capture enough information to create location profiles for mobile users for paging and localization purposes.

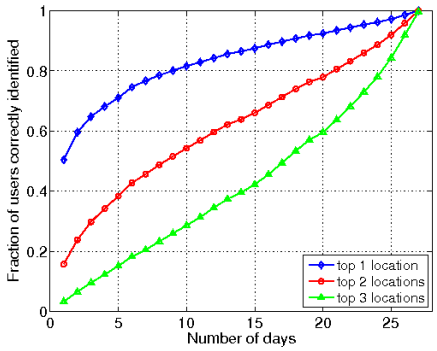
From a utility perspective, if we were to use a mobility trace, we would like it to reflect the user's trajectory as much as possible. We use the same metrics as those in Section 5.1, namely the number and entropy of locations, to evaluate the utility. The results are shown in Fig. 14. We find that the median number of locations visited per user drops from 27 for the month-long trace to 21 for the two-week-long trace and to 13 for the week-long trace. The median number of locations is only 7.5 in the day-long trace (Feb. 2), The median entropy also reduces dramatically from the month-long trace (2.97) to the day-long trace (1.46), and is 2.44 and 2.73 for the week-long and two-week-long traces, respectively. Both the number and entropy of locations show significantly reduced utility in the traces of shorter time durations, in particular in the day-long trace. Unfortunately, the reduced utility is the price to pay for the increased privacy as discussed above.

5.3 Discussion

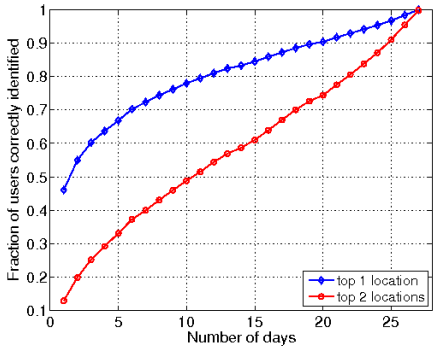
The time domain approach is essentially a sampling process to introduce noise into the data, which is similar to the concept of differential anonymity [4]. The spatial domain approach is essentially a generalization approach which replaces a data value with a less precise one or a range to maintain a larger k for k-anonymity. Both approaches are not mutual exclusive and they can be combined into a "hybrid" approach [16]. However, both approaches will significantly reduce the utility of the location traces. As argued by Brickell and Shmatikov in [2], "even modest privacy gains require almost complete destruction of data-mining utility". We are afraid the quote applies quite accurately to location data. Although possible to be released at coarse granularity or short time durations, the utility of location traces will be significantly reduced to satisfy privacy constraints.

6. CONCLUSIONS

In this paper we conducted a large scale study on the risk of re-identification attacks with published location data obtained through call records. Our study shows that publishing or sharing anonymized location data will likely lead to privacy risks and that, at a minimum, the data needs to



(a) Over all users



(b) Over users with 1-anonymity

Figure 13: Fraction of users whose top locations are correctly extracted

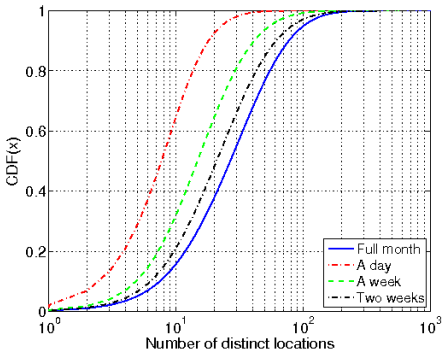
be coarse in either the time domain (meaning the data is collected over short periods of time of the order of a day, in which case inferring the top N locations reliably is difficult) or the space domain (meaning the data granularity is strictly higher than the cell level). In both cases, the utility of the anonymized location data will be decreased by a significant amount, which we quantified using information-theoretic measures.

While our work may be a bit discouraging regarding the release of location data in its original form (namely at a fine granularity such as sector or cell level), we believe our results provide important guidelines on how location data *can* be published, even at the cost of reduced utility. In general, though, we strongly recommend that the community be extremely cautious when publishing anonymized location data.

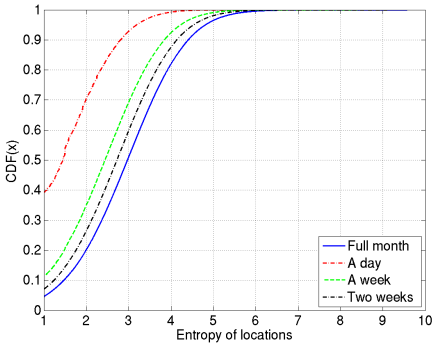
We hope this work also provides incentives for researchers to investigate and develop methods beyond simple anonymization that ensure privacy-preserving sharing or publishing of location data. This is a topic of growing importance since ever larger amount of location data is being collected and in great needs to be shared, for example with academic institutions (for research purposes) and with the providers of location-based services.

7. REFERENCES

[1] A. R. Beresford and F. Stajano. Location privacy in pervasive computing. *IEEE Pervasive Computing*, 2(1):46–55, 2003.



(a) Number



(b) Entropy

Figure 14: Number and entropy of distinct locations visited by each user in traces of different lengths

[2] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*, pages 70–78, New York, NY, USA, 2008. ACM.

[3] 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*, pages 23–32, New York, NY, USA, 2008. ACM.

[4] C. Dwork. Differential privacy. *Automata, Languages and Programming*, 4052:1–12, 2006.

[5] Foursquare. http://foursquare.com.

[6] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42:14:1–14:53, June 2010.

[7] P. Golle. Revisiting the uniqueness of simple demographics in the US population. In *WPES '06: Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80, New York, NY, USA, 2006. ACM.

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

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[10] 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*, pages 31–42, New York, NY, USA, 2003. ACM.

[11] M. Gruteser and X. Liu. Protecting privacy in continuous location-tracking applications. *IEEE Security and Privacy*, 2(2):28–34, 2004.

[12] J. Krumm. Inference attacks on location tracks. In *PERVASIVE'07: Proceedings of the 5th international conference on Pervasive computing*, pages 127–143, Berlin, Heidelberg, 2007. Springer-Verlag.

[13] J. Krumm. A survey of computational location privacy. *Personal Ubiquitous Comput.*, 13(6):391–399, 2009.

[14] L. Kulik. Privacy for real-time location-based services. *SIGSPATIAL Special*, 1(2):9–14, 2009.

[15] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)*, pages 106 –115, April 2007.

[16] N. Li, W. H. Qardaji, and D. Su. Provably private data anonymization: Or, k-anonymity meets differential privacy. *Technical report, Purdue university*, 2011.

[17] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '09)*, pages 517–526, New York, NY, USA, 2009. ACM.

[18] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the 16th annual international conference on Mobile computing and networking*, MobiCom '10, pages 185–196, New York, NY, USA, 2010. ACM.

[19] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering(ICDE)*, 2008.

[20] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1, March 2007.

[21] V. Rastogi, D. Suciu, and S. Hong. The boundary between privacy and utility in data publishing. In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 531–542. VLDB Endowment, 2007.

[22] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 596–604, New York, NY, USA, 2008. ACM.

[23] R. Shokri, C. Troncoso, C. Diaz, J. Freudiger, and J.-P. Hubaux. Unraveling an Old Cloak: k-anonymity for Location Privacy. In *ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2010.

[24] C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[25] L. Sweeney. Uniqueness of simple demographics in the U.S. population, 2000.

[26] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[27] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In *Proceedings of IMC '09*, pages 267–279, New York, NY, USA, 2009. ACM.

[28] H. Zang and J. Bolot. Mining call and mobility data to improve paging efficiency in cellular networks. In *Proceedings of the 13th annual international conference on Mobile computing and networking*, MobiCom '07, pages 123–134, New York, NY, USA, 2007. ACM.