On the Anonymity of Home/Work Location Pairs

Philippe Golle and Kurt Partridge

Palo Alto Research Center
{pgolle, kurt}@parc.com

Abstract. Many applications benefit from user location data, but location data raises privacy concerns. Anonymization can protect privacy, but identities can sometimes be inferred from supposedly anonymous data. This paper studies a new attack on the anonymity of location data. We show that if the approximate locations of an individual's home and workplace can both be deduced from a location trace, then the median size of the individual's anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census track and county respectively. The location data of people who live and work in different regions can be re-identified even more easily. Our results show that the threat of re-identification for location data is much greater when the individual's home and work locations can *both* be deduced from the data. To preserve anonymity, we offer guidance for obfuscating location traces before they are disclosed.

1 Introduction

Location-based services offer valuable applications to mobile users. To receive these services, users must disclose their location to service providers. This raises privacy concerns [6]. Location records, when analyzed, can reveal sensitive facts about an individual, such as business connections, political affiliations or medical conditions. Misuse of location data can lead to damaged reputation, harassment, mugging, as well as attacks on an individual's home, friends or relatives.

Privacy policies and legislation address some of these concerns. But protection mechanisms rooted in policy or law are only effective when data collectors are honest and trusted. They offer no protection against a dishonest collector, or one whose data is compromised by malware, laptop theft or a weak password.

To minimize privacy concerns, the best practice is to collect the minimum amount of information needed. For location-based services, this *principle of minimal collection* typically means collecting anonymous or pseudonymous location data [2]. A restaurant recommendation service, for example, can give adequate recommendations based on locations reported anonymously, or under a pseudonym linked to a profile of dining preferences.

Anonymity is a useful, but imperfect tool for preserving location privacy. The problem is that ostensibly anonymous location data may be traced back to personally identifying information with the help of additional data sources.

Krumm [5] showed how to recover the home address (with median error below 60 meters) and identity (with success above 5%) of subjects who disclosed two weeks' worth of GPS data collected in their car. In Krumm's experiment, re-identification was made possible by joining GPS traces with a reverse geocoder and a Web-based white pages directory. Krumm also showed how to prevent re-identification attacks. Location obfuscation (via the addition of noise, rounding or cloaking) defeats re-identification, at the cost of some loss in location precision.

To be effective, anonymity requires understanding and countering the threat of re-identification. This paper contributes a fuller understanding of this threat for anonymous location traces. Krumm's work, while groundbreaking, considers only the threat of re-identification that comes from identifying a subject's home. We extend his analysis by considering also a subject's *place of work*. After the home, the workplace is arguably the second most easily identifiable location in a trace. Obfuscation techniques which prevent re-identification based on (approximate) home location alone may not be adequate if the subject's (approximate) work location is also known. In fact, we show that home and work locations, even at a coarse resolution, are often sufficient to uniquely identify a person.

We rely on data from the U.S. Census Bureau to estimate the threat of re-identification based on home and work locations. This data, collected for the Longitudinal Employer-Household Dynamics (LEHD) program [9], contains the home and work locations of 96% of American workers in the private sector ($103,289,243$ individuals). This very large dataset gives a precise indication across the U.S. working population of the threat of re-identification, and of the effectiveness of defenses such as location obfuscation (for comparison, Krumm's analysis was based on a study of 172 participants).

We adopt a strong definition of privacy based on the concept of an *anonymity set*. The anonymity set associated with a location trace is the set of people from whom this trace may have been collected, given all information known to the data collector (see section 2). A large anonymity set offers strong privacy. Conversely, a small anonymity set is cause for concern. A unique (or nearly unique) trace may not always be linkable to an identity, but it is prudent to make the conservative assumption that if a unique link exists between a trace and an identity, that link may be discovered. This definition of privacy is stronger than that described by Krumm in [5], which considers re-identification successful only when a link is positively established. Our privacy definition is identical to that used by Sweeney [7, 8] in her well-known analysis of the uniqueness of simple demographic attributes in the U.S. population.

Our contributions. In summary, our contribution is threefold: 1) We study the threat of re-identification of anonymous location traces based on home and work locations; 2) We base our study on a very large dataset representative of the whole U.S. working population; 3) We adopt a strong and principled definition of privacy. Our analysis will help data collectors gain a better understanding of the sensitivity of location data, and the commensurate needs to obtain consent from users before collecting location data, to protect location data via obfuscation or access control, and to restrict the disclosure and publication of location data.

Organization. We review related work in the rest of this section. We discuss our model and assumptions in section 2. We present the statistical dataset that we used to estimate location privacy in section 3. We study the threat of re-identification for location traces, assuming approximate knowledge of home and work locations, in section 4. Finally, we conclude in section 5.

1.1 Related Work

Intentional degradation of location information quality, or obfuscation, is a well-known technique for preserving the anonymity, or pseudonymity, of location traces [2, 3], but the question of *how much* obfuscation is required to preserve anonymity is often sidestepped. Our paper answers this question for location traces from which home and workplace locations can be deduced.

Krumm's analysis of inference attacks on location traces [5] is closest to our work. The main distinction between our work and Krumm's is that we also take into account workplace locations. Using a stronger definition of privacy, and a much larger data set obtained from the U.S. Census Bureau, we show that the threat of re-identification for anonymous location traces based on home and work locations is more severe than the threat reported in [5] for home location alone.

Hoh et al. [4] propose a time-to-confusion metric to characterize the degree of privacy of location traces. Their approach to anonymity consists of periodically withholding location information long enough for a location trace to be confused with sufficiently many others. This approach is well-suited to vehicular network applications. But it has two drawbacks: 1) it requires coordination from a centralized server that tells mobile devices how long to withhold location data, and 2) it is not applicable to pseudonymous location traces, since fragments of a pseudonymous trace can easily be linked. Pseudonyms are important to many location-based services (e.g. to personalize service or prove membership). In contrast to [4], our analysis of privacy applies to pseudonymous traces.

Our work is indirectly related to Sweeney's well-known study [7] of the uniqueness of simple demographic attributes in the U.S. population. Sweeney's analysis of census data showed that 87% of the U.S. population is uniquely identifiable given their full date of birth (year, month and day), sex, and the ZIP code where they live. We adopt Sweeney's strong definition of privacy [8], called *k-anonymity*, which is based on the concept of an anonymity set (see section 2).

2 Assumptions and Privacy Model

Anonymous location traces. The value of location data dissociated from identity was illustrated in the introduction: a location-based restaurant recommender can offer adequate recommendations based on location and a pseudonymous profile of dining preferences. Many other location-based services (e.g. friend-finding services) can similarly operate using a registered pseudonym only, without learning users' real identities. From a technical point of view, location traces can be anonymized or pseudonymized with help from a trusted network proxy. Mobile subscribers, for example, may trust their network provider to forward their location data anonymously to third party location-based service providers.

Threat of re-identification. We study the threat of *re-identification* for anonymous location traces. We focus specifically on the threat of re-identification under the assumption that the approximate home *and* work locations of the subject can be deduced from the trace (for example with a reverse geocoder). Approximate home and work locations may then be joined with employment directories, tax records or any other public or private dataset available to the adversary to map pairs of home and workplace locations to identities.

Model of privacy. For the sake of example, assume that a subject is the only person in the U.S. who lives in a certain region A and works in a certain region B. The subject's location trace is the only one with the home/workplace pair (A, B). It does not necessarily follow that the trace can be linked to the subject, as there may be no directory that links the pair (A, B) with the subject's identity. But since the datasets that an adversary may use to re-identify location traces are not known a-priori, it is best to make the most conservative assumptions about them. Accordingly, we assume that if a unique link exists, it will be discovered. Our measure of privacy is the set of all people associated with the pair (A, B), called the *anonymity set* [8] of the pair. The larger the anonymity set, the larger the crowd one is indistinguishable from, and consequently the better the privacy protection one enjoys. Enlarging the regions A and/or B (e.g. via location obfuscation) increases the size of the anonymity set, and thus the quality of privacy protection. The rest of this paper analyzes the size of the anonymity set of home/workplace location pairs for different region sizes, based on the census data described in the next section.

3 The LEHD Origin-Destination Dataset

The Longitudinal Employer-Household Dynamics (LEHD) program, run by the U.S. Census Bureau, compiles information about where people work and where they live, together with reports on their age, earnings and distribution across industries. LEHD includes all jobs covered by the reporting requirements of the states' unemployment insurance system. According to [1], "The prime exclusions are agriculture and some parts of the public sector, particularly federal, military, and postal works. Coverage varies across states and time, although on average, 96% of all private-sector jobs are covered."

LEHD lets us study the privacy implications of revealing (intentionally or indirectly) coarse-grained location information, such as the county or ZIP code where one lives and works. A person revealing this information allows her identity to be narrowed down to the set of people who live and work in the same geographic areas. The size of this *anonymity set*, which can be estimated with the LEHD dataset, is a good measure of the privacy loss that revealing coarse home and work locations entails.

The raw LEHD data is not publicly available for download, due to privacy concerns. But the Census Bureau releases privacy-preserving synthetic data derived from the raw data. Bayesian techniques are used "to synthesize workers' place of residence conditional on disclosable counts of workers by place of work, industry, age, and earnings categories." [1]. According to [1] (p. 6), "The key statistical property to preserve in the synthetic data is the joint distribution of workers across home and work areas." The synthetic data thus appears suitable for our privacy analysis of home/work location pairs. Three implicates (independent draws from the synthesizing algorithms) are available for download from [10]. To confirm the validity of our results, we repeat our analysis with all three implicates. We obtain nearly identical results, as described in section 4.

We focus on the "Origin/Destination" dataset, which reports where workers live and work at the granularity of census blocks. The most recent dataset available is from 2004. It includes information on workers from 42 states (the eight missing states are Arizona, Connecticut, Massachusetts, Nebraska, New Hampshire, New-York, Ohio and South Dakota). In total, the 2004 Origin/Destination dataset includes information on $103,289,243$ individual workers.

The Origin/Destination dataset reports the home and workplace locations of workers along the following hierarchical geographic scale:

- **State.** The state is one of the 42 states included in the LEHD dataset.
- **County.** There are $2,784$ counties and county equivalents (boroughs, parishes) in the 42 states included in the LEHD dataset.
- **Census Tract.** Census tracts are county subdivisions defined by the U.S. Census Bureau. In terms of size and number of residents, they are roughly comparable to ZIP codes. The LEHD dataset contains $64,881$ tracts where workers live and $52,852$ tracts where they work. The mean number of workers living in a tract is $1,592$ and the median is $1,479$. The mean number of people *working* in a tract is $1,954$ and the median is 951 (it ranges from zero in completely residential tracts to $166,680$ in a tract of Los Angeles County).
- **Census block.** Blocks are the smallest area of census geography. They are subdivisions of census tracts, typically alongside streets. In urban areas, census blocks typically coincide with individual city blocks. In rural areas, blocks may cover many square kilometers.

4 Anonymity Set of Home/Work Location Pairs

In this section, we study the size of the anonymity set for workers who reveal where they live and where they work at various degrees of granularity (census block, census tract, or county). The knowledge of home and work locations at the census block granularity is information that could be learned from a lightly obfuscated location trace (with noise or rounding on the order of a city block or less). With more obfuscation (on the order of a kilometer or so), a location trace would reveal only the census tract where the person lives and works. Heavy obfuscation (on the order of tens of kilometers) may only allow for inference of the county or counties where a person lives and works.

Location precision	Size of anonymity set			
	Median	10th percentile	5th percentile	bottom percentile
Census block	1	1	1	1
Census tract	21	3	1	1
County	34,980	446	92	6

Table 1. Size of the anonymity set for workers who reveal where they live and work.

We compute the anonymity set for U.S. workers assuming knowledge of where they live and work at these three levels of granularity. Table 1 summarizes our findings. It shows the median size of the anonymity set, as well as the 10th, 5th and bottom percentiles of anonymity sets sorted by size. The table shows that revealing where one lives and works at the granularity of census blocks is uniquely identifying for a majority of the U.S. working population. Revealing where one lives and works at the relatively coarse level of census tracts is uniquely identifying for 5% of U.S. workers, and offers little privacy to the majority. Revealing the counties where one lives and works poses little threat to privacy.

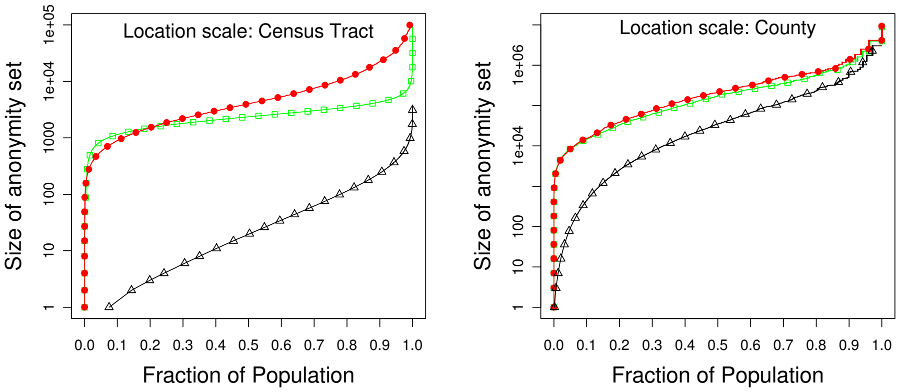

Fig. 1. Size of anonymity set under disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is either census tract (left graph) or county (right graph). Note the different scales on the Y-axes.

The graphs in Fig. 1 give more detail on the size of the anonymity set of workers who reveal the location where they live, the location where they work, or both. The graphs show what fraction of the U.S. working population (on the X-axis) falls in an anonymity set of less than a given size (on the Y-axis) after revealing location information (where they live, where they work, or both) at different precisions (census tract granularity for the graph on the left; county granularity for the graph on the right).

The red circle curve plots the anonymity of workers who reveal only where they work. The green square curve plots the anonymity of workers who reveal only where they live. The black triangle curve plots the anonymity of workers who reveal both where they live and where they work. Both at the granularity of census tracts and at the granularity of counties, we observe that revealing both the locations where one lives and works is strikingly more identifying than revealing only one of them (note that the Y-axis follows a logarithmic scale). For example, disclosing both the census tracts where one lives and where one works places 24.2% of the working population in an anonymity set that contains 5 or fewer individuals, and 7.4% in a set of 2 or fewer individuals.

While these statistics are based on synthetic data, we obtained the same results in Fig. 1 with all three synthetic implicates of the 2004 LEHD dataset (the curves are so similar as to appear indistinguishable).

4.1 Factors Influencing Anonymity

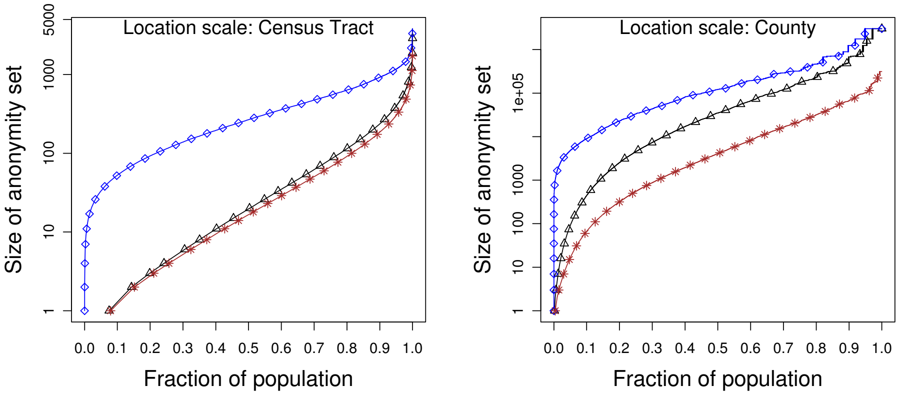

Fig. 2. Size of anonymity set under disclosure of home and work locations, for workers who live and work in the same location (blue diamonds) or in different locations (brown stars). For reference, black triangles show the anonymity set of all workers as in Fig. 1. Location granularity is census tract (left graph) or county (right graph). Note the different scales on the Y-axes.

Anonymity differs dramatically between individuals who live and work in the same region, and individuals who work in a different region from where they live. Fig. 2 plots these differences. It shows that workers who live and work in different locations (brown stars) have a much smaller anonymity set (i.e. are less anonymous) than those who live and work in the same location (blue diamonds). This holds true both for locations revealed at census tract (left graph) and county granularity (right graph).

The traces of workers who cross location boundaries to go to work are particularly vulnerable to re-identification attacks. Across the states included in the LEHD dataset, 94.1% of workers live and work in different *census tracts*. The percentages range from a high of 97.2% in California to a low of 80.3% in Wyoming. These numbers show that the extra anonymity afforded by living and working in the same census tract is uncommon. The fraction of workers who live and work in different *counties* is 43.5%. The percentages range from 63.8% in Virginia to 10.7% in Hawaii. Living and working in the same county is more frequent, and therefore had a bigger effect on the national average (see Fig. 2).

5 Conclusion

We studied the threat of re-identification of anonymous location traces in the U.S. based on obfuscated home and work locations. We showed that this threat is substantially greater with the disclosure of *both* home and workplace locations then either one alone. This result is important, because the workplace is arguably the second most easily identifiable location in a subject's trace, after the home.

Obfuscation techniques which prevent re-identification based on home location alone [5] may not be adequate if the subject's work location is also known. When both home and work locations can be deduced, our results show that a considerable amount of location obfuscation (at the granularity of counties) is required to protect the anonymity of location traces. An alternate approach to privacy would be to maintain different"home" and "work" personas, in applications where these personas can be kept strictly separate and unlinkable.

Our study distinguishes itself from previous work in that it adopts a principled definition of privacy based on the concept of anonymity sets, and relies on a large dataset that is representative of the whole U.S. working population. While our analysis is based on U.S. data, we speculate that our results apply to other developed countries with broadly similar densities and commute patterns.

Acknowledgements

We thank the anonymous reviewers and our shepherds James Scott and Alastair Beresford for their valuable comments and suggestions.

References

1. F. Andersson, M. Freedman, M. Roemer and L. Vilhuber. LEHD OnTheMap Technical documentation. February 21, 2008.
2. A.R. Beresford and F. Stajano. Location privacy in pervasive computing. In *IEEE Pervasive Computing*, 2(1):46-55, 2003.
3. M. Duckham and L. Kulik. A formal model of obfuscation and negotiation for location privacy. In *Proc. of the 3rd International Conference on Pervasive Computing (PERVASIVE 2005)*, pp. 152–170. LNCS vol. 3468.
4. B. Hoh, M. Gruteser, H. Xiong and A. Alrabady. Preserving Privacy in GPS Traces via Density-Aware Path Cloaking. In *Proc. of ACM Conference on Computer and Communications Security (CCS)*, 2007.
5. J. Krumm. Inference Attacks on Location Tracks. In *Proc. of Fifth International Conference on Pervasive Computing (Pervasive 2007)*, pp. 127–143.
6. B. Schilit, J. Hong and M. Gruteser. Wireless Location Privacy Protection. In *Computer*, vol. 36, no. 12, pp. 135–137, Dec. 2003.
7. L. Sweeney, Uniqueness of Simple Demographics in the U.S. Population. Laboratory for International Data Privacy, Carnegie Mellon University, 2000.
8. L. Sweeney. K-anonymity: a Model for Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
9. U.S. Census Bureau. Longitudinal Employer-Household Dynamics. On the web at http://lehd.did.census.gov/led/
10. VirtualRDC OnTheMap Data. http://www.vrdc.cornell.edu/onthemap