To what Extent does De-Anonymization of Mobile Datasets Compromise Privacy?

Lim Tze Ching Josephine

Abstract – **With the advent of mobile computing, there has been much concern regarding the privacy risks posed by the collection and dissemination of 'anonymous' datasets. Several studies have shown that many anonymized datasets collected from mobile phone data can be de-anonymized, raising concerns about re-identification of users and the fragility of anonymity in general. On the other hand, others have claimed that these de-anonymization techniques succeeded primarily due to the fact that the data was not appropriately 'anonymized' to begin with. This paper aims to examine both of these viewpoints by applying an evaluative re-identification framework to two studies regarding the de-anonymization of mobile phone datasets.**

I. INTRODUCTION

The importance of privacy in computer security has long been acknowledged by security professionals; loss of privacy is considered to be a breach of secrecy, which is in turn considered one of the four main tenets of computer security [1]. The advent of the Internet and mobile computing, however, has cast new light on the importance of privacy, as these relatively modern inventions pose new challenges to privacy that were previously technologically infeasible [2]. Large quantities of data are gathered on a daily basis from mobile devices, sometimes explicitly shared by the user, but often also without the user's awareness.

Perhaps of greater concern is the fact that these datasets are often publicly available, or at least available to a wide range of organizations [2]. These datasets are usually 'anonymized' to protect the privacy of the users, but many researchers have claimed that de-anonymization and subsequent re-identification of users is often not only plausible, but is in fact fairly easy to perform [2,3,4]. Potentially sensitive data is often contained in these datasets – mobility data that reveals location traces, for instance, or sensory data. Thus, if de-anonymization of data were truly as simple as claimed, the sharing of such datasets would expose users to a substantial breach of privacy.

Several papers have been written on the topic of de-anonymization of datasets, many of which contain alarming findings. For instance, Narayanan *et al.* [4] report having successfully de-anonymized the Netflix Prize dataset, using only the Internet Movie Database as their background knowledge source. With the identities of Netflix records of known users in hand, the authors claimed that they have managed to uncover the users' political preferences and other sensitive information.

In another paper, de Montjoye *et al.* [2] find that four spatio-temporal points in a simply-anonymized mobility dataset are sufficient to uniquely identify 95% of individuals in the dataset. Based on the formula that they derived for expressing the uniqueness of human mobility, they also conclude that coarsening the data spatially or temporally has minimal impact on reducing uniqueness or re-identification risk.

Lane *et al.*[3] performed a study on sensor data collected on mobile phones (*eg.* accelerometers, gyroscopes, magnetometers, and barometers), and concluded that although most of the emphasis in privacy-related research has been on location data with sensor data being considered 'harmless', sensor data can in fact lead to a new range of privacy threats. They claimed that sensor data can potentially be used to de-anonymize users and to obtain sensitive information about specific users from anonymized datasets.

However, El-Amam, in his article regarding the de-identification of health data [5], raises an interesting counterpoint to the findings of these studies. He suggests that it is overly simplistic to categorize datasets as 'identifiable' or 'not identifiable'. Rather, El-Amam states, we should consider identifiability as a spectrum, with various datasets falling on different points of the spectrum depending on ease and cost of re-identification as well as the risks posed by re-identification.

Although El-Amam's article was written from a standpoint of risk posed by de-identification of health data, the identifiability framework proposed by El-Amam may be a novel and interesting way of evaluating the genuine privacy risk posed by the 'de-anonymization' of datasets reported in the aforementioned studies.

In this paper, we attempt to apply El-Amam's evaluative framework to de Montjoye and Lane's studies regarding the de-anonymization of mobile phone datasets. In the next section, we explain in greater detail El-Amam's framework and the methods used by the two de-anonymization techniques. We then examine the de-anonymization techniques used in those

reports, and evaluate the privacy risks implied by de-anonymization of the datasets mentioned in the reports.

II. MATERIALS AND METHODS

De Montjoye et al. (2013): Unique in the Crowd: The privacy bounds of human mobility [2]

De Montjoye *et al.*[2] performed their study on a 'simply anonymized' mobility dataset containing 15 months of mobility data for 1.5 million people in a small European country. Their definition of a 'simply anonymized' dataset referred to a data set in which all 'obvious identifiers' such as name, home address and phone numbers, were removed. These datasets were collected by the mobile phone operator. Each time the user initiates or receives a call or a text message., the location of the connecting antenna was recorded. This data collection took place from April 2006 to June 2007. 114 interactions per user were recorded each month on average, within the space of 6500 antennas across the country.

From this longitudinally sparse and discrete dataset, the authors evaluated the uniqueness of traces by extracting from the dataset, the subset of trajectories $S(I_p)$ that match a set of spatio-temporal points I_p. A trace is found to be unique if $S(I_p) = 1$, indicating that only one trace matches that particular set of spatio-temporal points.

The authors then plot the fraction of unique traces against the number of available points p, leading to their conclusion that four arbitrary points are sufficient to uniquely characterize 95% of the users, while two arbitrary points uniquely characterize more than 50% of the users. They claim that the uniqueness of a user's mobility trace places a lower bound on the risk of deductive disclosure and dictates the likelihood of a brute force re-identification to succeed.

The authors do not attempt actual re-identification of the users in the simply anonymized dataset, however. They have only provided a theoretical model for the assessment of uniqueness of a mobility trace, based on the premise that a unique trace would be easily re-identified using outside information such as publicly available workplace and home addresses or geo-localized tweets or pictures.

Lane et al. (2012): On the feasibility of user de-anonymization from shared mobile sensor data [3]

Lane *et al.* [3] attempted to quantify the feasibility of de-anonymization a sensor dataset collected from commodity smartphones. They performed their study on a large representative activity recognition dataset from the ALKAN project. Their experiments were built on the basis that structurally sparse datasets have been historically proven to be easily de-anonymized. Rather than designing their own algorithm for de-anonymization, they build on popular algorithms used for de-anonymization in non-mobile datasets by demonstrating that mobile sensor data contains the same structural sparsity properties as the aforementioned non-mobile datasets, thus rationalizing the application of those algorithms to mobile sensor data.

In their experiments, Lane *et al.* consider two de-anonymization scenarios: de-anonymization of a particular user with auxiliary information, and de-anonymization by linking a user's identities from two separate datasets. They used two application-specific, fundamental representations of shared sensor data – event-oriented publishing of data such as Nike and other exercise applications that allow users to publish their exercise information (routes, time, duration, etc.), and periodic publishing of behavior summary such as Snapshot discount programs which require the user to allow sharing of their usage with the company. However, they require that the mechanism for sharing of sensor data within the application must 'maintain a level of user anonymity appropriate for (1) the sensitivity of data and (2) the type of relationship between the user and the people with whom the data is shared.

Lane *et al.* used two datasets for their experiments; first, a public activity recognition dataset from the ALKAN project, which contains data from >200 users and 35,000 activities, from a combination of iOS and Android devices. This data includes a diverse range of physical, social, and transportation activities, with the events either being user-labeled or extracted automatically. The second dataset was the AOL query logs, used as a baseline to compare against the results of the ALKAN dataset. These logs contained the anonymized search query logs of 650,000 users over 3 months, which has been successfully de-anonymized in other literature and shown to be very sparse.

The authors first permuted the representation of activities within their ALKAN dataset, to ensure that their results were not dependent on a specific application scenario. They then proved

that their ALKAN data was equally sparse as AOL data, and thus hypothesized that it was vulnerable to the same de-anonymization techniques that have been successfully used on AOL data.

El-Amam (2010): Risk-based de-identification of health data. [5]

El-Amam defined an identifiability continuum on which datasets can be placed, allowing for the objective measure of a dataset's identifiability, and proposed a threshold decision rule to assist a data custodian in deciding whether or not disclosure of a particular dataset should be allowed.

El-Amam's identifiability continuum is best described by the figure below (Fig. 1). It comprises of five discrete levels, each referring to the initial level of anonymity present in the dataset, and its corresponding risk of re-identification. This continuum is intended as a descriptive scale, with each level inheriting any de-identification traits of the lower levels.

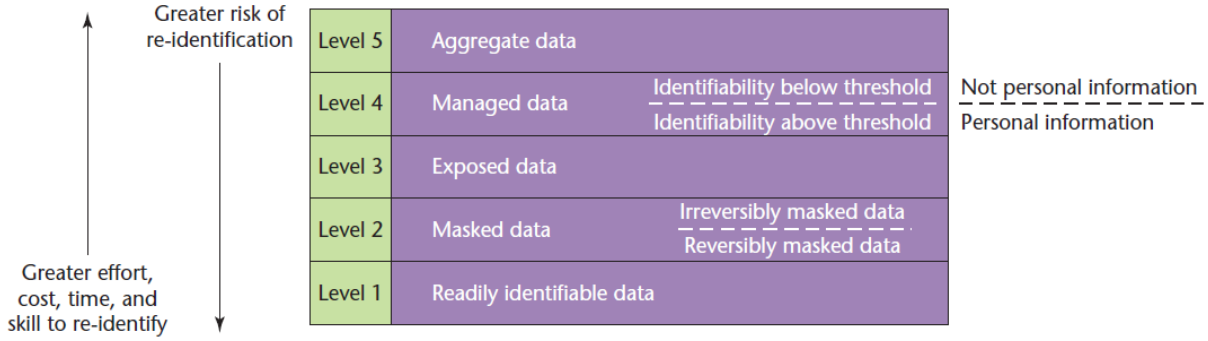

Figure 1: El-Amam's continuum of identifiability

A summary of the different levels and the corresponding datasets is as follows:

- **Level 1:** Dataset contains clearly identifiable data, such as names and social security numbers. No effort is needed to re-identify an individual.
- **Level 2:** Masked data, involving manipulation of direct identifiers such as removing names, or creating reversible or irreversible pseudonyms. Does not remove or obfuscate quasi-identifiers such as dates, locations, and socio-economic information.

- **Level 3:** Exposed data, involving attempts at obfuscating quasi-identifiers as well as direct identifiers. However, data identifiability and risk of exposure is not objectively measured.
- **Level 4:** Managed data, whereby the data custodian has objectively measured the data's identifiability and can offer substantial evidence that it is above or below a certain threshold.
- **Level 5:** Clearly unidentifiable information – for example, unstratified counts, frequencies, or rates. Information contains no direct or quasi-identifiers.

El-Amam claims that most of the re-identified datasets in current literature are level 2 data, which should not have been considered de-identified or anonymous to begin with. Thus, he suggests, their re-identification 'isn't a surprise' and should not be used as proof of de-anonymization. He also suggests that only level 4 data and above can data truly be considered de-identified.

Additionally, El-Amam states that there are three types of re-identification risk, that can be objectively measured with probability metrics. They are:

- **Prosecutor risk:** The adversary is attempting to re-identify a specific individual, has background knowledge on him or her, and knows for certain that he or she is in the dataset.
- **Journalist risk:** The adversary is attempting to re-identify a specific individual but doesn't know for certain whether or not the individual is actually in the dataset.
- **Marketer risk:** The adversary is attempting to re-identify as many people as possible in the data set.

The author states that prosecutor risk is numerically higher than journalist risk, which will be higher than marketer risk. Thus, in deciding which risk metric represents a plausible attack scenario for a particular dataset, the data custodian should start from the top and only manage the highest plausible risk.

'

III. RESULTS

We are thus able to evaluate the de-anonymization studies presented in [2] and [3] based on El-Amam's identifiability and risk assessment framework.

De Montjoye et al. (2013): Unique in the Crowd: The privacy bounds of human mobility [2]

The authors have mentioned that their dataset consists of 'simply anonymized' mobility records with 'obvious identifiers' such as name, home address, and phone numbers removed. This corresponds to Level 2 of El-Amam's identifiability continuum – *ie.* masked data with direct identifiers removed but quasi-identifiers such as location left intact. This is consistent with El-Amam's claim that most of the re-identified datasets in current literature are level 2 data.

With regards to type of re-identification risk, we start assessing from the higher possible risk and work our way downwards, as El-Amam suggests. Prosecutor risk is not relevant to this case, as there is no plausible way for an adversary to know for certain who is or is not in the dataset. Journalist risk, on the other hand, is relevant to this study, as an adversary might be attempting to identify a specific individual in the data set, as opposed to only attempting to re-identify groups of people. Thus, the re-identification risk of this dataset is journalist risk.

Lane et al. (2012): On the feasibility of user de-anonymization from shared mobile sensor data [3]

The position of Lane *et al.*'s dataset on the identifiability continuum is not as clear-cut as de Montjoye *et al.*'s case, for a few reasons. Lane *et al.* have used two datasets, one from the ALKAN project, and another from the AOL query logs which have been 're-identified' by previous studies. They have not mentioned specifically what information is or is not contained in their datasets. The AOL query logs are not the focus of Lane *et al.*'s study, as they were covered by previous research and only used as a baseline comparison. Thus, we evaluate Lane *et al.*'s study based on the ALKAN dataset.

Based on literature regarding the ALKAN project [6], we find that ALKAN was designed as a large-scale activity data gathering system. It functions via smart phones that are equipped

with accelerometers, that record and uploads activity data to the ALKAN server. The ‘subjects’ in the ALKAN dataset are 216 university students and staff who agreed to participating in the project, and were given iPod-Touches as the device containing the ALKAN client. The researchers acquired activity data that was uploaded by these volunteers. It was not mentioned whether or not the names were removed, but the quasi-identifiers (activity data) was not obfuscated or aggregated, so this dataset lies at Level 2 of the identifiability continuum at the very most.

As for the type of re-identification risk, prosecutor risk is relevant to the ALKAN dataset, as records of the ALKAN project would presumably inform us for certain who is and is not in the dataset. This is the highest possible risk level in El-Amam’s framework; thus, the re-identification risk of this dataset is prosecutor risk.

IV. DISCUSSION

We thus see that it is fair for security professionals to be concerned with the widespread dissemination of datasets collected from mobile phone usage. Many of these datasets have only the direct identifiers removed, thus placing them at Level 2 of El-Amam’s identifiability continuum, and are likely easy to re-identify given some background information. The risk increases if it is possible for an adversary to find out with certainty who is and isn’t in the dataset, as in the case of the ALKAN data.

However, the fact that these datasets were only at level 2 of the identifiability continuum to begin with, raises a consideration that perhaps the problem is not that ‘anything can be de-anonymized’ or that ‘sharing of mobile phone data compromises user privacy’ per se, but rather with the fact that those particular datasets were not adequately anonymized to begin with.

Thus, rather than viewing the problem of privacy breach through the lens of a binary construct (anonymity vs. no anonymity), it may be beneficial instead to view anonymity as a continuum on which we should strive to improve by obfuscation of quasi-identifiers, as El-Amam also suggests. De Montjoye’s and Lane’s findings lend support to this perspective, as

their reports on re-identification risk are based on the premise of quasi-identifiers being available.

Further research in attempting and evaluating the de-anonymization of Level 3 and especially Level 4 datasets would be beneficial to assess the plausibility of re-indentification of strongly-anonymized data as opposed to weaker anonymization methods.

REFERENCES

[1] Lampson, B. W. (2004). Computer security in the real world. *Computer*, *37*(6), 37-46.

[2] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*,*3*.

[3] Lane, N. D., Xie, J., Moscibroda, T., & Zhao, F. (2012, November). On the feasibility of user de-anonymization from shared mobile sensor data. In*Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones* (p. 3). ACM.

[4] Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*(pp. 111-125). IEEE.

[5] El Emam, K. (2010). Risk-based de-identification of health data. *Security & Privacy, IEEE*, *8*(3), 64-67.

[6] Hattori, Y., Inoue, S., & Hirakawa, G. (2011, June). A large scale gathering system for activity data with mobile sensors. In *Wearable Computers (ISWC), 2011 15th Annual International Symposium on* (pp. 97-100). IEEE.