NRC Data Collection and the Privacy by Design Principles

Imad Aad and Valtteri Niemi
Nokia Research Center, Lausanne
firstname.lastname@nokia.com

ABSTRACT

Nokia Research Center (NRC) in Lausanne, Switzerland has launched a rich data collection campaign during fall 2009, the purpose of which is to study user socio-geographical behavior, mobility patterns etc. of approximately 200 people. All sensors on the mobile devices (GPS, microphone, wireless interfaces etc.) were frequently activated in order to grab the most of the user contexts, to be generic enough and answer the needs of various researchers using the datasets.

The data is as rich as it could be without being too intrusive into the volunteers' lives. We therefore took particular care in preserving the privacy of the participants, while still keeping the data useful for the various future analyses. In this paper we describe the anonymization techniques that we applied to the data, how we met the principles of privacy by design, and the legal aspects with the participants and the researchers.

1. INTRODUCTION

In order to study user behavior, mobility, social interactions etc. NRC launched a data collection campaign in fall 2009 [1]. A number of volunteers were given high-end Nokia phones, equipped with a special software client capable of gathering rich data collected from the phone sensors 24/24 hours, 7/7 days, in order to get a deep insight of user activities. The logged data is automatically uploaded every day to a database server, anonymized, before being accessible to researchers.

The collected data includes GPS coordinates (in beginning of October 2010 we have around 9 million GPS entries in total), acceleration data(1 million samples), surrounding WLANs (458K unique WiFi access points; 43M access points seen in total), BlueTooth devices (414K unique BT addresses; 26M BT encounters), GSM cells (83K unique cell towers, 25K in Switzerland; 39M towers seen in total), incoming and outgoing call (327K calls) and SMS numbers (146K text messages), contacts list (99K phone book entries), calendar/memo, as well as the media played or being recorded (including 47K pictures/videos taken and 120K songs played). About half of the volunteers opted for sampling the audio features (497K audio samples). Altogether, we have almost 178M entries in the data base, 194 participants who have visited 62 countries and collected around one thousand person months of 24/7 data (as of beginning of October 2010).

On the other hand, no content of communication (call, sms), documents, nor URLs is recorded.

The research interests of the users of the (anonymized) data base cover a very wide range: from sociologists, to mobility modelers, socio-geographical analyzers, networking, security and privacy researchers. The number of new data users and new areas of interest have been increasing steadily since the launch of the campaign.

This richness in context data raises many privacy issues, and defining the anonymization techniques becomes challenging when it comes to keeping the data as useful as possible, even for future (not yet specified) research purposes. The details of our anonymization techniques, their compliance to the Privacy by Design principles [2], the encountered compromises to be made, and the envisioned improvements are the main contents of this paper.

The seven principles of "Privacy by Design" (PbD) [2] are:

1. *Proactive* not Reactive; *Preventative* not Remedial
2. Privacy as the *Default*
3. Privacy *Embedded* into Design
4. Full Functionality - Positive-Sum, not Zero-Sum
5. End-to-End Lifecycle Protection
6. Visibility and Transparency
7. Respect for User Privacy

These principles provide an excellent framework for system design, especially when privacy-sensitive data is involved. For most of the principles, the name of the principle gives already a pretty good idea about what is meant. The principle of "Positive-Sum" captures the important goal that privacy protection should not have negative impacts on other

properties of the system, e.g. usability or performance. The last principle "Respect for User Privacy" can be seen almost like a meta-level principle: applying all other principles from user point of view gives a good basis for complying with the seventh principle as well. More detailed descriptions of the seven principles can be found in [2].

In this paper, we use the framework of PbD principles to analyze how user privacy have been taken into account in the NRC Lausanne data collection campaign [1]. Section 2 contains a brief discussion of two facets: first, the positioning of our data set in the vast amount of various data sets used for research purposes, and secondly, challenges that are faced when anonymization techniques are used with privacy-sensitive data sets. In Section 3 we explain details of our anonymization methods that are in use for the collected data, and we also discuss some legal aspects relevant for our campaign. Section 4 is organized according to the seven PbD principles, and we discuss how each principle has been followed in the data collection. In Section 5, we present some enhancements to our anonymization mechanisms. These have not yet been put into use but could in principle be applicable to the present set of collected data. Finally we give some concluding remarks in Section 6.

2. BACKGROUND

Driven by the increasing success of social networking, the various businesses behind it, and the increasing capabilities of smart-phone capabilities in sensing user context [7], many research groups, operators, and ISPs are now exploring the potentials of mining rich context data [9, 6]. Even the general public begins to be aware of the astronomical amounts of data that exist in various data bases.

Collecting and opening such data bases for research purposes definitely comes with considerable work on anonymizing it in order to preserve the user privacy [5] since it often comes without the users consent, especially when it is a large scale data such the ones done by phone operators or ISPs on their clients [8].

Many of the existing data collection campaigns have specific focuses like collecting phone usage statistics, identifying groups of people, human behavior etc. separately. In contrast to those, the data collected in our campaign is as generic as it can be, limited only by the (already high) sensing capacities of the phones deployed. From the users side, no specific research group or research interest was predefined, leaving it open potentially to many groups of researchers with various topics of interest. More details on our data collection can be found in [1].

Such rich data, even when anonymized, typically leave some identifiers easily usable to trace back the identities of users allowing security or privacy researchers to get credit in deanonymizing them [10], nevertheless resulting in privacy scandals such as [3].

Building upon the experiences learned from others, and to enforce privacy-preservation of the users, technical and legal techniques were put in place for our campaign. Throughout the data anonymization work we used legacy anonymization techniques and primitives, without having to design new ones. The challenging part, however, turned out to be the degree to which we should apply anonymization, and still keep the data useful.

All individual anonymization techniques used here such as keyed hashing, coordinate truncation etc. can be commonly found in the literature [5], and are used to anonymize individual databases. In our case the rich set of data types imposed specific combinations of anonymization primitives, applied to a certain degree, in order to keep the data useful. To fill the gap between usable and perfect anonymization, legal agreements are done with the researchers, mainly because of protecting the users' privacy.

3. ANONYMIZING PIs AND PIIs

In this section we describe our anonymization approach for Personal Information (PI) and Personally Identifiable Information (PII) in the data base.

3.1 What is anonymized and how

With relevance to anonymization, the collected data can be treated as three different types: GPS coordinates that get truncated, textual data that get hashed, and acoustic data that get sampled and shuffled.

By "hashing" an info we mean concatenating the message as (Key1||info||Key2) then hashing it using SHA256 function. The use of a hash function provides one-way property: it is infeasible to compute "info" from the hashed version. This is a keyed hash construction; without access to the keys it is neither possible to check whether a specific "info" (obtained, e.g., by an educated guess) leads to a given hashed version. We did not have the need to use a (slightly more complex) HMAC construction [12] because we have a fairly restricted range of use for the function and the data format and length of all entries in the data base is well understood and controlled.

"Info" is converted to lowercase beforehand. Note that this keeps the data consistent, in the sense that data entries that have differences only for the case will anyway result in equal hashes. For implementation of these hash functions we used an SQL library called pgcrypto.sql [11].

GPS coordinates are stored with three different precision levels; we give each research group access to the one that is sufficient for their purposes. The different precisions levels are: complete GPS coordinates, removing the last 2 digits and rounding (which, in Switzerland, results in an accuracy of around 110 m in latitude and 80 m in longitude), removing the last 3 digits and rounding (accuracy of roughly 1 km for Switzerland). The truncated coordinates result in step-like paths which increase the ambiguity level. The resulting ambiguity level depends on the initial geographical area: in rural areas, the step-like paths can be easily mapped back to the (only?) road, and the path ends to the (only?) house. Whereas in dense city centers such truncation results in high ambiguity levels, proportional to the number of streets/flats within the output path "step". An adaptive approach is discussed in Section 5.

Phone numbers (in phonebooks and caller/callee lists) have the last 7 digits hashed, while the first ones are kept in

clear. Such prefixes are useful to identify the regions and to distinguish mobile phone numbers from landline ones. All names (of users, contacts in contacts list, caller, callee etc.) are hashed

MAC addresses (of WLAN, BlueTooth devices) have their last 6 digits hashed. The first 6 are left in clear text since they point to the chip manufacturer etc. This provides still a high ambiguity about the user ID while indicating, for instance, what kind of devices are in the neighborhood.

SSIDs of WLANs are hashed, since it is common practice that families or companies set their wireless network SSIDs similar to their own names.

Other data such as calendar titles and location (text), filenames of media generated (e.g. pictures), names of folders (Boxes) for text messages are entirely hashed since they are typically personalized, therefore likely to reveal PIIs. Phone IMEIs (i.e. serial numbers) are also entirely hashed.

Acoustic data are recorded in order to help identifying various environments (e.g. noisy, quiet...) of individual users, or to distinguish between different locations/rooms of different users in geographical proximity. This data is read every 10 minutes for a duration of 30 seconds, utilizing Mel-Frequency Cepstral Coefficients (MFCC) [13]. These same coefficients are typically used for speech recognizers and they do not provide alone high enough privacy. In order to increase the privacy level, we randomly shuffle the time order of the individual parts so that the content or identity of the speaker can not be detected anymore. In contrast with the other data types that get anonymized after upload onto the data base, acoustic data is scrambled (therefore anonymized) on the user device itself prior to the upload. After randomization, certain statistical properties of the acoustic sample are still preserved, and they are sufficient to provide information about the environment.

3.2 What is not anonymized

Names of media played (music, album, track number etc.) are kept in clear text since it is valuable for user profiling only when kept so, and these reveal no privacy-sensitive information in practice. Hashing such information would imply big losses in contextual data for negligible improvements in privacy.

Cell tower IDs that the mobile sees are kept in clear, and so is the level of received signal power. Other local system data such as battery status and level, running application(s), screensavers etc. is kept in clear.

No transformation is applied to acceleration data.

3.3 Legal commitments from the researchers, to the participants

As one may infer from the previous description of the techniques, single data types provide little personal information about the participants, but reverse-engineering gets easier when more data types are combined in order to reveal the participant identities or private information.

In order to complement the technical anonymization functions, researchers are tied with legal commitments not to reverse-engineer the data and keep the participants' privacy preserved. Note that unlike many other user data bases that grant access to any user, accessing NRC data collection is tied to "Data Sharing Agreements" between the research institution, the individual users, and Nokia. Prior to granting access, the purpose of the research is discussed, and the commitments on preserving participants privacy is made clear, then the legal agreements are finalized. Of course, these restrictions are somewhat unfortunate because they restrict the open access to the data set that would of course be beneficial to the scientific community.

On the other hand, prior to filling and signing the consent forms, participants were carefully informed about the research targets, data collection, storage, transfer, and anonymization methodologies used, as well as the awaited benefits. Furthermore they were trained on how to visualize their (raw and statistical) data, share it with friends, or how to delete it, using a dedicated and easy to use web page. Regular events took place, during which the participants were updated with the above information, the campaign news and statistics.

The participants have the right to leave the campaign at any time (which a few of them did, mainly due to the short battery lifetime or because of leaving the country.) Nokia reserved the right to exclude participants from the campaign in the event of non-compliance with the protocol, which never happened so far.

3.4 The compromises during anonymization

In this section we discuss various compromises done in order to find the right balance between anonymity and utility. The two extremes can be briefly described by the two examples:

- High anonymity levels could be achieved by removing all kinds of identifiers from the GPS coordinates, call logs etc. This would make it hard to construct e.g. paths, and therefore reverse engineering of the identities would also be hard to perform. However, this would drastically reduce the linkability between calls, events, coordinates etc. hence degrading the usability for mobility models, socio-geographical analysis etc.
- High utility and usability levels could be achieved by leaving contextual data in clear text. Indeed, this would put the data into a perfect shape for context analysis (e.g. social interactions). However, reverse engineering would become an easy task for finding people's identities and their whereabouts.

To avoid drawbacks illustrated by the above examples, some subtle compromises were to be done for anonymizing GPS coordinates, and many textual data that can be used as PIIs.

GPS coordinates

One easy option to strongly anonymize the GPS coordinates would be to (key-) hash them, similar to what was done for textual data. This still provides identical outputs for identical inputs, while securing against reverse-engineering of the

private locations / IDs. Apart from preserving the "sameness" property, hashing results in "randomized" coordinates: inputting a user's path would output random geographical jumps, therefore losing the information about speed, proximity, and visited Points of Interests (PoI).

Another option where speed and proximity can be preserved is the use of linear transformation of the GPS coordinates: a given path input to the anonymization function results in a translated/rotated/scaled path geographically distant or different from the initial one, therefore anonymizing "private" coordinates. This approach has the following drawbacks:

- Most of the movements of people are along roads and highways rather than arbitrary paths in forests/lakes etc. Those road shapes are easily identifiable, often even visually, in the output space, hence making them easy to reverse and map to the original coordinates.
- Transforming a set of coordinates into another, located somewhere else on the globe removes all information of PoIs, which is a useful information component for many research areas. For instance, a user going from work, to a bar, then to a cinema, then to home obviously has a different profile from one going from point A, to B, to C then D in the middle of the pacific ocean.

Truncating the GPS coordinates provides a good (and very simple) balance between anonymity and usability. Public PoIs can be identified as such, then tagged, and stored in the database.[1]

Textual data

Leaving the data in clear text easily could make the data base users, intentionally or not, break the privacy of the participants. On the other hand randomizing it makes it completely useless. The adopted hashing technique preserves sameness and the resulting data showed to be still highly useful to most researchers. However, the similarity between calendar entries *"meeting with John"* and *"meeting with Laura"* is lost after hashing, and "meetings" are not identifiable neither. Advanced anonymization techniques that tackle this problem are discussed in Section 5.

Researchers who are also participants

Another type of compromise is encountered because of the fact that quite many (although only a small minority of) campaign participants are EPFL staff/students, among whom several are also data base users at the same time. These persons have access to their own clear text data on their phones or over the web interface, and they also have access to the same data in anonymized form in their role as a researcher. In principle, this enables them to easily create a mapping between certain data items and their anonymized counterparts. As explained in the previous subsection, their role as a researcher prevents them from doing such reverse-engineering (by contractual means).

The issue is not quite as simple, though. For many research topics, such a mapping would be quite useful. An example is the name of a static WLAN access point. On the other hand, de-anonymization of such a static AP in a public place leads to a minimal privacy violation because, similarly, proximity to public PoIs is visible in the data base. To avoid situations where a researcher has a temptation to try de-anonymization for the purpose of progress in his research work, a reverse table that maps the anonymized data items back to the original form is provided to the researchers in cases where it is shown that no privacy violations are introduced because of this. In particular, for the case of WLAN APs, such APs located at the campus of EPFL can still be identified with their SSIDs, because a reverse table is provided to the researchers.

[1]Work in progress.

4. HOW WELL WE MEET THE PbD PRINCIPLES?

1. Proactive not reactive: the most important point is that all privacy-sensitive data is indeed anonymized. Because we have wanted to impose minimal restrictions to the nature of the research problems that could be addressed using the collected data, we have tried to avoid anonymizing too much. On the other hand, this kind of "future-proofing" has implied that breaking the anonymization is possible for a skillful person with sufficient amount of local knowledge about Lausanne and its inhabitants. Therefore, we have been forced to use legal type of protection against reverse-engineering: data access is only provided for researchers who commit themselves to NOT trying to break the anonymization. This is also the main reason why we cannot release the full data set to completely public usage.

2. Privacy as the Default: anonymization is indeed automatically enabled all the time. As explained in the previous section, there are various levels of anonymization, especially for location. The default level of anonymization is always the strongest and could be relaxed if the research problem necessitates it.

3. Privacy Embedded into Design: anonymization is a key feature of the system architecture and the whole campaign design.

4. Full Functionality: anonymized data is sufficient for research purposes but we cannot exclude the possibility that some valuable research opportunities are lost because of it. For instance, providing content of communication in the data base would certainly have opened many new vistas for studying users' contexts (and in general much better view on the social life of campaign participants).

5. End-to-End Lifecycle: anonymization and access control will be enforced throughout the lifetime of the data base.

6. Visibility and Transparency: users have full view to their own data and they are able to delete anything they want. On the other hand, individual accesses to the anonymized data by researchers is not visible to individual participants of the data collection campaign. In principle that kind of transparency could have been arranged also but it is hard to see what kind of purpose it would serve.

7. Respect User Privacy: privacy has been the key element in the whole campaign design; the data is (also) used for creating better privacy protection mechanisms.

5. CAN WE DO BETTER ANONYMIZATION?

In this section we give a couple of examples about how our anonymization techniques could be enhanced while still keeping them applicable to our setting. Certainly, there are also other possible directions for improvements, but in this paper we focus only on these examples.

Regarding GPS coordinates, the future improvement step could be to adapt the truncation of GPS coordinates to the densities of roads, houses, and population of the various visited areas, so as to maintain a constant level of ambiguity, regardless of whether the area is rural or densely populated. This, however, requires a rough knowledge of demographics of the visited areas.

A more challenging improvement is the one for hashing textual data as hinted in Section 3. So far entire data entries (e.g. "meeting with John") were hashed, therefore if another "meeting with Bob" shows in the calendar, or "John's Birthday", nothing can be found in common after anonymization: "meetings" are hard to identify in the agendas, and so is "John" if his name appears among another text. In order to improve this and increase utility for researchers, while maintaining equal levels of privacy, hashing individual words seems to be appropriate and the hash("John") can be identified in the anonymized data base, whether for "meeting" or "Birthday". This comes, however, with several challenges:

- Sentences (i.e. data entries) should not have common links because of hashes of common words like "with", "and", "or" etc. Therefore such publicly common words should be kept in clear, or a dictionary of hashes of common words is written (somehow in analogy with tagging public PoIs in the truncated GPS coordinates). This task requires natural language processing techniques, applied to the various languages spoken by people in the campaign (at least six languages).
- Even the same word ("meeting ...") may have different meanings in different contexts such as in business or social life. Same word in different contexts results in different privacy-sensitivities, and therefore anonymization techniques should be applied accordingly.
- Especially in the extreme case where every word is hashed separately we would run into the problem well known with so-called Electronic Code Book (ECB) mode of encryption. Indeed, in a large data base, it would be easy to distinguish more commonly used words from more rarely occurring words. Furthermore, after every new word that is inverted/decrypted, the task of inverting/decrypting the rest becomes easier.

These improvements would clearly help in better understanding of user context (increasing utility of the data set) without degrading anonymity. However, it comes at the cost of designing proper natural language processing techniques and applying it to the existing data.

6. CONCLUSION

We showed how we anonymized the data of our rich data collection campaign. The data set is highly privacy-sensitive and therefore privacy protection is needed. Because there is a wide range of research areas that could potentially utilize our rich data, the anonymization and utility have been carefully balanced. Some legal counter-measures were also needed against reverse-engineering efforts. We also discussed potential enhancements to the current anonymization techniques.

We hope our findings and approach can be useful for other researchers anonymizing other collected data.

7. REFERENCES

[1] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez and J. Laurila, "'Towards rich mobile phone datasets: Lausanne data collection campaign", in Proceedings of ICPS 2010.

[2] http://www.privacybydesign.ca

[3] http://en.wikipedia.org/wiki/AOL_search_data_scandal

[4] Barbaro, M., Zeller Jr., T.: A face is exposed for AOL searcher no. 4417749. New York Times (August 9, 2006), http:www.nytimes.com20060809technology09aol.html

[5] Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. pp. 111-125 (2008)

[6] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin, Diversity in Smartphone Usage, in MobiSys'10, June 2010

[7] N. Eagle (2010), "Mobile Phones as Social Sensors", The Handbook of Emergent Technologies in Social Research, Oxford University Press (in press).

[8] J. Blumenstock, D. Gillick, and N. Eagle (2010), "Who's Calling? Demographics of Mobile Phone Use in Rwanda", AAAI Spring Symposium 2010 on Artifical Intelligence for Development (AI-D) (in press)

[9] http://reality.media.mit.edu/

[10] A. Narayanan, V. Shmatikov, De-anonymizing Social Networks, in Proceedings of S&P 2009

[11] http://anoncvs.postgresql.org/cvsweb.cgi/pgsql/contrib/pgcrypto/pgcrypto.sql.in

[12] M. Bellare, R. Canetti and H. Krawczyk, "Keying hash functions for message authentication", in Proceedings of CRYPTO 1996

[13] http://en.wikipedia.org/Wiki/Mel-frequency_-cepstral_coefficient