PhD Position Offer
Safe Anonymization Techniques for Data Call Records

Proposal

All network access made by mobile phones leave some traces. Mobile network operators are used to store those traces for billing purpose in the form of call detail records (CDR) that includes mainly communication time, duration, macroscopic location and destination of the call. Besides, many handset applications (e.g. True Caller, Number Book,...) capture also, enriched CDRs that includes personal contacts, precise location, communication content, and recipients contacts. CDR are classically anonymized before being used [BEC+12, XT06]. The analysis of anonymous results can lead to interesting behaviours for researchers [ICWG14, BCH+13] but also for companies wishing to develop new markets. However, a privacy breach can occur when the anonymized data is disclosed and combined with additional knowledge. For example, analysis of call patterns of individuals (e.g., call frequency, call location) can be used to infer the caller's identity. Simply deleting identifiers (phone numbers, for example) is not enough at all. Many privacy breaches have been observed as a result of failed anonymization techniques [BaBNG16], even in the limited context of call data graphs [SD13, SD14].

Because of the risk of breaking anonymity, many telephone operators place limitations on the the amount and type of data they are ready to sell, share, or make publicly available: for example, Orange, with its FluxVision[1] solution, "only" produces general statistics on the presence of a population class during specific events. This data is derived from connections to the operator's telephone antennas.

This doctoral thesis aims at providing a model of anonymization that ensures that these data will not allow re-identification and cannot be used to link individuals to their sensitive information, when disclosed to third parties. However, these data should also remain useful for aggregate analysis and exploitation. This model must first of all make it possible to mathematically prove that it is impossible to re-identify individuals and link them to their sensitive information. It will also allow immediate, if not extremely fast, execution on data usually derived from large telephone call graphs.

Realization

Supervisions: Jean-François COUCHOT, Associate Professor HDR FEMTO-ST, Béchara AL BOUNA Associate Professor, Antonine University, Xiao XIAOKUI Associate Professor National University of Singapore, Singapore. This work will be carried out in collaboration with the Orange partner, developer of the Flux-Vision solution located in Belfort.

Most of the PhD work will be done at the FEMTO-ST laboratory, DISC department, Belfort, France. Travel to Antonine University and Singapore is planned.

Work plan

The first practical work will consist in validating the data set built in partnership with Orange to allow the dissemination of methods and the evaluation of their implementation.

The first 4 months of the doctorate will be devoted to continuing the analysis and appropriation of work relating to knowledge extraction, a task initiated by the supervisors. This synthesis work will be submitted to an international journal.

In parallel to the two previous tasks, the student will study privacy attacks on data modelized in the form of graphs (social networks) and will model the adversary/opponent.

With this adversary model established, the next step will consist in developing a privacy protection model that will quantify the risk of publication of the data from the call graph with the knowledge of the

[1] https://www.orange-business.com/fr/produits/flux-vision

adversary we are considering. To this end, we plan to adopt a Bayesian statistical approach to infer the knowledge that may later be acquired by the attacker after observing the anonymized graph. We will then quantify the risk of privacy breach based on this knowledge induced by the adversary. Finally, we will express properties regarding the privacy risk that need to be verified by the anonymization method: this will provide a formal privacy protection framework in this context.

The next step will be to develop an anonymization method within this context and implement it. Existing algorithms on graph anonymization focus mainly on simple graphs. They use only simple anonymization methods, such as edge insertion/removal or node looping. These methods are unsatisfactory when telephone graphs are associated with heterogeneous data from social networks. To solve this problem, we aim to develop a new anonymization method method for this type of graph, using the properties of call graphs and incorporating the best of the different existing approaches. The main challenge we will address is preserving the usefulness of the data subject to confidentiality constraints. The effectiveness of the implementation will be at the heart of the subject.

An intermediate objective already identified will be the definition of a set of relevant metrics capable of measuring information loss due to data suppression/anonymization. This assessment of information loss will also be based on the different types of attackers defined in advance. The main idea is to maximize the utility of the anonymized data as much as possible, while minimizing privacy threats and leaks. After formalizing objectives as a constrained optimization problem, using previously formalized appropriate metrics, we will investigate traditional optimization methods, in order to find the best way to dissociate data by reducing information loss and satisfying minimum diversity constraints for each data set, k-anonymity and heterogeneity.

Candidates Profile and Application

The candidates should have a master degree in applied mathematics or in computer science, with proven skills in anonymization, applied mathematics, statistics. Proficiency in English is important, and the candidates shall master writing and presenting scientific work.

The application consists of one PDF file comprising:

- a CV,
- a letter of motivation justifying the interest for this particular PhD subject,
- a recommendation letter from the supervisor of the master's thesis, with contact details,
- a short summary of the master's thesis,
- the transcript of records of the license and master degree (or equivalent), with rank and size of the promotion.

The application should be sent by e-mail to `couchot@femto-st.fr`, `bechara.albouna@UA.EDU.LB`, and `xkxiao@nus.edu.sg`. The closing date for applying is June 12th 2018.

References

[BaBNG16] Sara Barakat, Bechara al Bouna, Mohamed Nassar, and Christophe Guyeux. On the evaluation of the privacy breach in disassociated set-valued datasets. In Christian Callegari, Marten van Sinderen, Panagiotis G. Sarigiannidis, Pierangela Samarati, Enrique Cabello, Pascal Lorenz, and Mohammad S. Obaidat, editors, *Proceedings of the 13th International Joint Conference on e-Business and Telecommunications (ICETE 2016) - Volume 4: SECRYPT, Lisbon, Portugal, July 26-28, 2016.*, pages 318–326. SciTePress, 2016.

[BCH+13] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. Human mobility characterization from cellular network data. *Communications of the ACM*, 56(1):74–82, 2013.

[BEC+12] Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *arXiv preprint arXiv:1210.0137*, 2012.

[ICWG14] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[SD13] Kumar Sharad and George Danezis. De-anonymizing d4d datasets. In *Workshop on Hot Topics in Privacy Enhancing Technologies*, 2013.

[SD14] Kumar Sharad and George Danezis. An automated social graph de-anonymization technique. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, pages 47–58. ACM, 2014.

[XT06] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 139–150. VLDB Endowment, 2006.