Formal definition and verification of practical anonymization criteria
CAPPRIS postdoc project proposal
(in the context of the CNIL-Inria collaboration)

February 5, 2016

1 Motivation

In order to allow big data analytics while preserving privacy, it is often necessary to anonymize the datasets [1, 2]. Anonymization can be required for legal reasons (e.g. in the context of open data) and/or to minimize the risks for the data controller. However, anonymization often comes into conflict with the objective of preserving the utility of the data. In addition, the fact that a piece of data is anonymous is by essence a relative notion because it depends on the available auxiliary knowledge. This auxiliary knowledge may itself depend on many factors, in particular the exposition of a given individual in the media or the existence of public information (such as a voting register).

De-anonymization really happens in practice and the press has widely reported many examples such as the re-identification of the governor of Massachussets medical information in 1997 or the re-identification of several celebrities from the release of the New York Taxi and Limousine Commission in 2014. Another study [3] shows that anyone knowing at least 4 highly-visited locations (e.g., home, working place, etc.) of a data subject has a chance of 95% to learn all of his/her other visited locations (which might include, e.g., a religious place) from a call-data-record dataset[1].

These examples show how easy it is to get it wrong in terms of anonymization (and to confuse it with pseudonymisation) and they still fuel

[1]The study is based on a dataset where locations are specified hourly with a spatial resolution provided by mobile phone carrier's antennas.

debate between experts.

In order to clarify some of the issues surrounding data anonymization, the Working Party 29[2] has issued in April 2014 an Opinion [6] defining several criteria to be applied to evaluate the robustness of an anonymization technique (singling out, linkability, inference). However, these criteria themselves are subject to interpretation and have triggered new discussions among experts [4]. The clarification of these notions is necessary since Data Protection authorities will receive an increasing number of requests to validate anonymization algorithms. In France, the new law "Pour une République Numérique" provides that the CNIL[3] approves anonymization algorithms, in particular when they are used in the context of open data.

2 Project

The objective of the project is to address the above challenges through a collaboration between Inria and the CNIL. The project is broken down into 5 stages:

1. The goal of the first stage is to reach, through a set of examples and counte-rexamples, a common agreement about the intended meaning of the anonymization criteria of the WP 29 and to provide a refined, pedagogical, explanation of these criteria.

2. The goal of the second stage is to define the criteria in a formal way so as to remove any possible source of ambiguity and to pave the way for the design of testing tools.

3. The goal of the third stage is to design and implement a set of tools to test and validate the criteria (based on the definition provided in the second stage) of a given anomymization scheme.

4. The goal of the forth stage is to experiment the testing tools on a variety of datasets and to provide recommendations on their use. This experimentation phase may also lead to adjusments of the definitions of steps 1 and 2 and updates of the testing tools.

5. The goal of the fifth stage is to embed the anonymization process within a risk analysis framework. This stage is necessary to help the

[2]One of the missions of the Working Party 29, which includes representatives of the Data Protection Authorities of all EU Member States, is to make recommendations on matters relating to privacy and data protection in the EU.

[3]The French Data Protection Authority.

decision maker choose the acceptable thresholds (e.g. with respect to de-anonymization risks) depending on the context (e.g. access to the data, sensitivity, intended use, etc.) [5].

3 Required Skills

Minimal knowledge and motivation for privacy and software development.

4 Research Environment

- **Location and organization:** The project will take place in Grenoble or Lyon, within the research group PRIVATICS of the Inria Rhône-Alpes research unit in the context of the Inria project Lab CAPPRIS. The project will be conducted in close collaboration with the CNIL, as part of the CNIL-Inria agreement. The Inria supervisors will be mainly in charge of the scientific challenges raised by the project (formal characterization of anonymization criteria and testing through de-anonymization and analysis tools) whereas the CNIL experts will provide their interpretation of the WP 29 document and their knowledge of the needs fo Data Protection Authorities. Technically speaking, the project will build on the expertise and current work of the PRIVATICS team on data anonymization and risk analysis [1, 2].
  - https://team.inria.fr/privatics/
  - http://cappris.inria.fr
  - http://www.inria.fr/en/centre/grenoble
  - http://www.cnil.fr/english/
- **Supervisors:**
  - Claude Castelluccia[4]
  - Daniel Le Métayer[5]

5 Duration

1 or 2 years.

[4] http://planete.inrialpes.fr/c̃astel/
[5] http://planete.inrialpes.fr/people/lemetayer/

References

[1] G. Ács and C. Castelluccia. A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris. In *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, United States, Aug. 2014. ACM.

[2] R. Chen, G. Ács, and C. Castelluccia. Differentially private sequential data publication via variable-length n-grams. In *ACM Conference on Computer and Communications Security*, pages 638–649, 2012.

[3] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports, Nature*, March 2013.

[4] K. El Emam and C. Alvarez. A critical appraisal of the article 29 working party opinion 05-2014 on data anonymization techniques. *International Data Privacy Law*, 2014.

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

[6] Working Party 29. Opinion 05/2014 on anonymization techniques. Text adopted by the Article 29 Data Protection Working Party on 10 April 2014.