# Anonymous versus personal data: from a binary view to a rigorous risk-based approach*

Gergely Ács[†], Claude Castelluccia[‡], Daniel Le Métayer[§]
Inria, France

October 30, 2015

## 1 Gap between binary legal views and the reality of anonymization

There is a broad consensus on the fact that data minimization is a key principle of privacy protection: not disclosing personal data (or disclosing the minimum amount of data) is the first way to protect privacy rights in the digital world. In some cases, data minimization can be implemented by PETs[1] which make it possible, through the use of cryptography, to implement the functionalities of a system without any disclosure of personal data (or the disclosure of the minimum amount of data). For example, cryptographic techniques such as homomorphic encryption and cryptographic commitments can be used to ensure that the correct fee is computed by the operator of a smart metering system without knowing the individual consumption values of the customers.

However there are also many situations in which the objective is not to avoid data disclosure but to make it possible to collect data and to analyze it in order to provide new services, either commercial or for the benefits of society. Indeed, with the advent of big data and the development of the open data movement in many countries, ever larger amounts of data are

*Contribution to the High Level Conference of the European Parliament on Protecting on-line privacy by enhancing IT security and EU IT autonomy.

[†]Email:gergely.acs@inria.fr
[‡]Email:claude.castelluccia@inria.fr
[§]Email:daniel.le-metayer@inria.fr

[1]Privacy Enhancing Technologies.

available and will be published or made available in the future. Even if not all these data are personal, huge amounts of personal data are already collected by various stakeholders and the exploitation of these data for a variety of purposes (including knowledge-based decision making, forecasting, medical research, fight against terrorism, intelligence, etc.) becomes a priority for many companies and governments. Even more important, it could also greatly contribute to the social well-being and public good. For example, the exploitation of health data can lead to a better understanding of diseases, help to identify appropriate treatments and support public health policies.

In order to allow such big data analytics while preserving privacy, it is necessary to anonymize the datasets. 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[2].

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 debate between those who claim that anonymization cannot work and those who believe that anonymization can actually work for the vast majority of people (ordinary people with limited media exposure) and, even if possible in theory, de-identification is practically impossible for most people. In any case, this shows that it is more and more difficult to draw a clear border between personal data and non-personal (truly anomymous) data.

The same observation can be made about sensitive data [5]. From a legal point of view, sensitive data (for which express consent is required in the European Directive 95-46 EC) includes race, political opinions, religion, sexual behaviour and health, among others. But geo-location information,

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

which is not defined as sensitive legally speaking, can potentially be a perfect indicator of many sensitive attributes.

The situation is all the more worrying in that regulations (even though they involve more gradual concepts such as proportionality or "appropriate level of security") tend to take a dual approach to the characterisation of personal and sensitive data: in the European Directive 95-46 EC a data is either personal or not[3] and the list of sensitive data is defined explicitly[4].

The General Data Protection Regulation adopted by the European Parliament [4] constitutes a significant step in the right direction. In particular, it places emphasis on risk analysis and data protection impact assessment[5]. However, it still defines a closed list of "special categories of data" and puts forward processing operations that are "likely to present specific risks". In addition, it does not require risk analyses or data protection impact assessments to be systematically conducted or validated by independent third parties. Therefore, it does not pass the rigor criteria discussed in the next section. Last but not least, even if legal obligations were strong enough, this would not necessarily be sufficient because much progress has to be made also on the technical side to make robust anonymization and rigorous risk analysis possible.

Clearly regulations and technologies are not in line with today's practices and this discrepancy will become more and more critical with the demands of the society and the economy for data analytics. We believe that the only way forward is to go beyond this dual vision and to rely on a rigorous risk-based approach.

## 2 Need for a rigorous risk-based approach

Considering that data cannot be easily classified as personal or non-personal, or as sensitive or non-sensitive, it is counterproductive to adopt a binary approach in regulations because it can lead both to (i) inadequate protection of data subjects in certain situations (as shown above by the restricted definition of sensitive data) and (ii) inacceptable burden for industry for certain types of treatments (for example if similar measures had to be taken for personal data and for anonymized data considering that they still present a residual risk of de-anonymization).

Therefore, we argue that the only way forward is to follow a more pro-

[3] In which case it is not covered by the Directive.

[4] Article 8.

[5] In particular in Article 33.

gressive, nuanced approach based on a rigorous analysis of the potential risks and benefits associated with data processing [6]. However, as stated by the Working Party 29 [8], the risk-based approach should not weaken the rights of individuals in respect of their personal data: "Those rights must be just as strong even if the processing in question is relatively 'low risk'. Rather, the scalability of legal obligations based on risks addresses compliance mechanisms". Most importantly, the risk-based approach should not be used as an argument to reduce the perimeter of personal data protected by the law.

For the risk-based approach to really improve the protection of data subjects, a number of conditions have to be met. First and foremost, the analysis has to be rigorous, both from the technical point of view and from the procedural point of view:

1. **Rigor from the technical point of view:** the methodology used for the analysis should be clearly defined, as well as the assumptions about the context (auxiliary information, types of attackers and motivation, etc.), the potential consequences of an attack for all "stakeholders" including, obviously harms to the data subjects, but also any detrimental impact on groups of individuals and society as a whole.

2. **Rigor from the procedural point of view:** the analysis process itself should be transparent and audited by a neutral third party; it should be renewed periodically and in the occurrence of any event that could change any initial assumption.

The risk analysis can lead to the application of a set of appropriate measures, justified by the results of the analysis and the potential benefits of the processing. These measures can include any useful technical, organizational and legal measures (e.g. anonymization, access control, encryption, non disclosure agreements, etc.). In extreme situations, it could also call into question the deployment of the system itself.

The risk analysis approach is closely linked to the accountability principle: because it makes it possible to trace the motivations for all decisions, it takes part in the overall accountability process. In addition, the accountability of the processing should make it possible to trace any occurrence of de-anonymization and to check that appropriate measures have been taken (deletion, application of a new risk analysis and anonymization process, etc.).

## 3 Recommended policy measures

To make the above approach viable, it should be supported by the following policy measures:

- Legal measures: **strong accountability requirements for data controllers**; risk analyses should be conducted or audited by **neutral third parties**; accountability of practices should also be audited on a regular basis by neutral third parties.

- Legal measures and economic incentives: **encourage the development of a viable ecosystem of risk analysis, anonymization and certification**; examples of measures include requiring certifications or audits (by third parties) in certain situations or putting in place official accreditation schemes (e.g. inspired by the Common Criteria in security).

- Research funding and standardization: **support for the development and the standardization of a rigorous privacy risk analysis framework**. Indeed, existing proposals are either too vague or too security centric (e.g. they do not handle anonymisation appropriately) and more work is needed to define a general purpose and widely recognized privacy risk analysis method.

- Research funding and standardization: **support for the development, the evaluation and the standardization of stronger anonymization algorithms** (with guidelines about their application in specific sectors). Many anonymization algorithms exist [1, 2, 7] but they are defined on a case-by-case basis and much more work is needed to be able to design and evaluate them more systematically.

## 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] European Parliament. Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation). Text adopted by the European Parliament on 12 March 2014.

[5] P. Ohm. Sensitive information, 2014. Southern California Law Review, Vol. 88, 2015, Forthcoming.

[6] J. Polonetsky, O. Tene, and J. Jerome. Benefit-risk analysis for big data projects, 2014. Future of Privacy Forum.

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

[8] Working Party 29. Statement on the role of a risk-based approach in data protection legal frameworks. Text adopted by the Article 29 Data Protection Working Party on 30 May 2014.