Personal Big Data, GDPR and Anonymization

Josep Domingo-Ferrer

Universitat Rovira i Virgili
Department of Computer Science and Mathematics
CYBERCAT-Center for Cybersecurity Research of Catalonia
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. Big data are analyzed to reveal patterns, trends and associations, especially relating to human behavior and interactions. However, according to the European General Data Protection Regulation (GDPR), which is becoming a *de facto* global data protection standard, any intended uses of personally identifiable information (PII) must be clearly specified and explicitly accepted by the data subjects. Furthermore, PII cannot be accumulated for secondary use. Thus, can exploratory data uses on PII be GDPR-compliant? Hardly so.
Resorting to anonymized data sets instead of PII is a natural way around, for anonymized data fall outside the scope of GDPR. The problem is that anonymization techniques, based on statistical disclosure control and privacy models, use algorithms and assumptions from the time of small data that must be thoroughly revised, updated or even replaced to deal with big data.
Upgrading big data anonymization to address the previous challenge needs to empower users (by giving them useful anonymized data), subjects (by giving them control on anonymization) and controllers (by simplifying anonymization and making it more flexible).

Keywords: Big Data, GDPR, Anonymization

Last century, Kafka, Orwell, Huxley and Böll wrote novels on dystopian societies. They were premonitory of what can be achieved and is achieved with big data in our century (*e.g.* social credit system in China, [8]). Without going that far, even in liberal democracies, big data can be very privacy-invasive [4, 13]. To protect citizens, the European Union has promoted the General Data Protection Regulation (GDPR, [5]), that is quickly being adopted as a *de facto* global privacy standard by Internet companies [7, 6]. GDPR limits the collection, processing and sharing of personally identifiable information (PII) and requires a privacy-by-design approach on the controllers' side [1, 3].

Nonetheless, the surge of big data analytics has brought a lot of progress and opportunities and is here to stay: one can hardly expect the private (and even the public) sector to refrain from harnessing big data on people for a great deal of

secondary purposes (other than the purpose at collection time). These include data exploration, machine learning and other forms of knowledge extraction. Satisfying the GDPR legal obligations towards subjects is very difficult in such a scenario where a host of controllers exchange and merge big data for secondary use to extract knowledge.

According to GPDR, anonymization is the tool that allows turning PII-based big data into big data *tout court*, and hence legitimately circumventing the legal restrictions applicable to PII. As repeatedly aired in the media [10, 2, 12], just suppressing direct identifiers (names, passport numbers, etc.), let alone replacing them by pseudonyms, is not enough to anonymize a data set. Anonymizing for privacy requires further data modification beyond identifier suppression, which may decrease utility. On the other hand, attaining good levels of utility and privacy for PII-based big data is essential to conciliate law with reality.

In this talk, I will first survey the main current limitations of the state of the art in big data anonymization:

1. *Unjustified* de facto *trust in controllers.* Twenty years ago, National Statistical Institutes (NSIs) and a few others were the only data controllers explicitly gathering data on citizens, and their legal status often made them trusted. In contrast, in the current big data scenario, a host of controllers collect PII and it is no longer reasonable to assume the subject trusts all of them to keep her data confidential and/or anonymize them properly in case of release [9].
2. Ad hoc *anonymization methods.* Many privacy models have been proposed (k-anonymity, l-diversity, t-closeness, ϵ-differential privacy, etc.) and each privacy model is satisfied using a specific statistical disclosure control (SDC) method, or a few specific ones. For example, k-anonymity is reached via generalization or microaggregation, and DP via noise addition. Life would be easier if a unified masking approach existed that, under proper parameterization, could be used to attain a broad range of privacy models. This would empower controllers in centralized anonymization and subjects in local anonymization.
3. *Difficulty of merging and exploring anonymized big data.* Even if subjects decide to accept centralized anonymization by the controllers, none of the main families of privacy models in use manages to satisfy all the desiderata of big data anonymization that we identified in [11]: (i) *protection* against disclosure no matter the amount of background information available to the attacker; (ii) *utility* of the anonymized microdata for exploratory analyses; (iii) *linkability* of records corresponding to the same or similar individuals across several anonymized data sets; (iv) *composability*, that is, preservation of privacy guarantees after repeated application of the model or linkage of anonymized data sets; and (v) *low computational cost.* Utility and linkability are needed to empower the *users/data analysts*, protection and composability are desired by *subjects*, and low cost is desired by *controllers.* On the other hand, it is hard for controllers holding data sets to engage in joint exploratory analysis without disclosing any of their data to other controllers. Note that

cryptographic secure multi-party computation (MPC) is of limited use here, because it is intended for specific calculations planned in advance, rather than exploratory analyses. Furthermore, while MPC ensures input confidentiality, it gives exact outputs that can lead to disclosure by inference (for example if the outputs are the mean and the variance of the inputs, and the variance is very small, it can be inferred that the inputs are very close to the mean).

Thus, the grand challenge is to obtain anonymized big data that can be validly used for exploratory analyses, knowledge extraction and machine learning while *empowering subjects, users and controllers*:

- Subjects must be given control and even agency on how their data are anonymized. Local anonymization gives maximum agency to the subject. However, it is ill-suited for privacy models relying on hiding the subject's record in a group of records, such as k-anonymity and its extensions, because these need to cluster the contributions of several subjects. If we obviate this shortcoming and go for local anonymization, randomized response and *local* DP are natural approaches. Unfortunately, the current literature on both approaches focuses on obtaining statistics on the data from subjects, rather than multi-dimensional full sets of anonymized microdata that are valid for exploratory analysis. The latter is precisely what is wanted. Centralized DP can indeed produce anonymized data sets preserving some dependences between the original attributes, but the challenge is to avoid centralizing anonymization at the *untrusted* controller.
- Users should receive anonymized data that are analytically useful.
- Controllers should be given more unified approaches to anonymization, allowing them to engage in multi-controller exploratory computation.

I will conclude the talk with some hints on how to tackle the above grand challenge.

Acknowledgment and Disclaimer

Partial support to this work has been received from the European Commission (project H2020-700540 "CANVAS"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705), and from the Spanish Government (project RTI2018-095094-B-C21). The author is with the UNESCO Chair in Data Privacy, but the views in this paper are his own and are not necessarily shared by UNESCO.

References

1. G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye and A. Bourka. *Privacy by Design in Big Data — An overview of privacy enhancing technologies in the era of big data analytics.* European Union Agency for Network and Information Security (ENISA), 2015.

2. M. Barbaro and T. Zeller. A face is exposed for AOL searcher no. 4417749. *New York Times*, Aug. 9, 2006.
3. G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Métayer, R. Tirtea and S. Schiffner. *Privacy and Data Protection by Design — From Policy to Engineering*. European Union Agency for Network and Information Security (ENISA), 2015.
4. C. Duhigg. How companies learn your secrets. *New York Times Magazine*, Feb. 16, 2012.
5. *General Data Protection Regulation*. Regulation (EU) 2016/679. https://gdpr-info.eu
6. *General Data Protection Regulation (GDPR)*. Google Cloud Whitepaper, May 2018.
7. N. Lomas. Facebook urged to maked GDPR its "baseline standard" globally. *Techcrunch*, Apr. 9, 2018.
8. A. Ma. China has started ranking citizens with a creepy 'social credit' system – here's what you can do wrong, and the embarrassing, demeaning ways they can punish you. *Business Insider*, Apr. 8, 2018.
9. P. Rogaway. The moral character of cryptographic work. Invited talk at *Asiacrypt 2015*. http://web.cs.ucdavis.edu/ rogaway/papers/moral.pdf
10. O. Solon. 'Data is a fingerprint': why you aren't as anonymous as you think online. *The Guardian*, Jul. 13, 2018.
11. J. Soria-Comas and J. Domingo-Ferrer. Big data privacy: challenges to privacy principles and models. *Data Science and Engineering* 1(1):21-28, 2015.
12. L. Sweeney. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Work Paper 3. Pittsburgh 2000.
13. S. Yu. Big privacy: challenges and opportunities of privacy study in the age of big data. *IEEE Access*, 4:2751-2763, 2016.