Fifty Shades of Personal Data – Partial Re-Identification and GDPR

Jan Willemson[0000−0002−6290−2099]

Cybernetica AS, Narva mnt 20, 51009 Tartu, Estonia

Abstract. This paper takes a look at data re-identification as an economic game where the attacker is assumed to be rational, i.e. performs attacks for a gain. In order to evaluate expectancy for this gain, we need to assess the attack success probability, which in turn depends on the level of re-identification. In the context of GDPR, possibility of various levels of re-identification is a grey area – it is neither explicitly prohibited, nor endorsed. We argue that the risk-based approach of GDPR would benefit from greater clarity in this regard. We present an explicit, yet general, attacker model that does not fit well into the current treatment of GDPR, and give it a high-level game-theoretic analysis.

Keywords: Data re-identification, privacy attacks, cost-benefit analysis, GDPR

1 Introduction

Even though the European Union's General Data Protection Regulation (GDPR) became binding already in 2018, there are still active discussions ongoing about its interpretation and enforcement mechanisms. Among other notions, the core terms of *personal data* together with its counterparts of *pseudonymous* and *anonymous data* have definitions standing far from mathematical rigour.

As a result, the rules for deciding whether data should be considered personal (so that GDPR applies) or anonymous (so it does not) are heuristic and open to subjective assessment.

On one hand, GDPR Art. 4(1) gives a definition of personal data depending only on whether the person can be identified completely (even if this complete identification is indirect):

> *[. . .] 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*

On the other hand, Recital 26 talks about identification as a process that depends on some likelihoods, costs, etc.:

> *To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.*

Additionally, Art. 11 together with Recital 57 describe a situation where the controller's ability to identify a person may change over time. However, it is left unclear whether this change is considered to be gradual, or going from 0% to 100% or back in an instant.

In this paper we are going to study the question of the level of identifiability more closely. To put this study into context, we are going to consider identification not merely as a process on its own, but as part of an attack by a rational (i.e. economically incentivised) attacker.

On one hand, such an approach should fit relatively well into the current setup of GDPR as it is argued to take a risk-based approach to data protection [16, 11]. Indeed, Art. 32 is concerned with security of processing, Art. 35 states the need and approach to data protection impact assessment, and several recitals contain further guidelines for risk assessment. However, none of them refers to any attacker models (e.g. attacker motivation, capabilities, etc.).

This observation is surprising from the rational risk analysis point of view. By omitting a well-defined goal, it will be impossible to conclude whether the protection measures taken are efficient or not. A clear understanding of the targeted attacker profiles is a necessary precondition for a rational risk analysis process [3, 17].

As already mentioned above, GDPR does not explicitly refer to the possibility of partial identifiability (or related concepts like inference). Purtova [15] has reviewed the main recent interpretations of identification under the GDPR, but all of them assume 'singling out' a person one way or another. Out of the five considered typologic categories (look-up, recognition, session-related, classification and personalisation), classification and personalisation have potential to allow identification in an ambiguous manner as well. However, both of them are further explicitly limited to only give rise to identification if they point to a single person [15].

We argue that such an interpretation is too narrow to address the whole spectrum of potential privacy issues stemming from data re-identification. Of course, it is possible to take a stance that even partial identification is identification and GDPR should apply. However, as we will see in Section 3, under a reasonable definition, almost every dataset is partially identifiable for its subjects.

This is not just an artefact of a well-chosen theoretical construction. Probabilistic nature of data subject re-identification has been observed over and over again in practice [9, 5, 6]. On the other hand, identification does not have to be perfect in order to allow for attacks against the data subject. We will elaborate further on this idea in Section 2.

GDPR acknowledges identification as a process that involves effort and requires resources like extra data or time, (see Art. 11 and Recital 26). However, this effort is considered only in the context of identifying one person, and not as cost amortised over many subjects. As a result, it is unclear how to treat e.g. opportunistic attackers who get a hold of a dataset and try to see how many subjects they can identify without really targeting anyone in particular.

In the rest of the paper we will attempt to address the shortcomings observed above by defining an explicit attacker model and deriving a cost-benefit risk assessment framework to better understand how a data privacy target can be defined.

2 An attacker model

We will consider scenarios where the attacker can get full access to a dataset that has been sanitised using some anonymisation technique in preparation for the release [12]. Note that we do not cover here the techniques that limit the attacker capabilities (allowing only pre-filtered queries, forcing computations over encrypted or secret-shared data, giving access to the queries only via controlled hardware environments, etc.).

From the risk analysis point of view, the exact process of obtaining the sanitised dataset is not so important. For example the data controller may have made it available in downloadable form for research purposes [14, 22, 8].

Consider the scenario where the attacker does not have a prior target, but is opportunistically interested in finding out private information about some subjects.

Based on the nature of the dataset and/or the obtained information, the attacker may decide upon his further actions. In order to facilitate rational risk analysis, assume that the information carries some (financial) incentive for the attacker – e.g. it could be sold to the yellow media, or the data subject could be blackmailed.

In order to discover something interesting, the attacker invests some resources like his own time and effort to analyse the dataset.

As an illustrating example, consider the attacker's task of identifying the subject based on location data. It is known that only a few data points from mobile phone location trace are sufficient to uniquely determine a data subject [8, 22, 21].

However, matching a unique trace to a natural person still requires additional effort. Assume the attacker has identified a person who passes through the same locations more or less regularly. A relatively simple way for identifying this person is to physically go to the traced locations and observe people passing by at the traced times. Note that observing the same locations on consecutive days on one hand means a bigger effort investment. On the other hand it also allows to narrow down the list of possible candidates more efficiently, translating into higher identification probability for every member in that list.

Of course, the attacker does not necessarily get the list down to a single element. However, this does not mean that the attack was unsuccessful.

For our example of a financially motivated attacker, partial information is useful, too. Say, he is able to narrow the list down to g people, one of them being a well-known politician. If the initial data analysis suggests that one member of the candidate list frequently visits the red light district, the attacker can make a guess even if he has not observed the politician directly in that area. Probability of the guess being correct is $\frac{1}{g}$. If the attacker decides to blackmail the politician requesting for price p, and assuming the victim pays when the attacker guessed correctly, the attacker's expected outcome of this economic game is still $\frac{p}{g}$ monetary units.

There are two lessons to learn here. First,

> *Re-identification does not have to be complete in order to facilitate successful attacks.*

And second,

> *Attacker's success in re-identification of the data subject(s) depends on the effort he is willing to invest.*

Thus, in order to adequately analyse attacker behaviour, it is not sufficient to consider just a single attack scenario with a specific amount of investment. Rather, a full spectrum of possible investments and returns needs to be evaluated.

3 Cost-benefit considerations

From the attacker's point of view, attacking involves various kinds of costs varying from direct cash and time expenses to potential penalties [7]. For the sake of simplicity, we will consider all the costs to have monetary units in this paper. Among other things, this approach allows us to compare the costs to the potential gains of the attack.

In order to assess these gains, we need to quantify the outcome of the re-identification attack. Different measures have been proposed in the literature [18, 19, 5, 10, 13]. In this paper, we will build on the approach proposed by Benitez and Malin [5].

They start from anonymised health datasets and match them to (semi)public records of voter lists based on sociodemographic parameters. Benitez and Malin proceed to estimate the expected number R of re-identifications and the cost C required for the analysis (which in their case means purchasing access to voter registration lists). The ratio $\frac{C}{R}$ then shows the expected cost per re-identification.

Let's augment this reasoning with benefit analysis. We note that not every re-identification is equally valuable for the attacker. Some high-profile persons are likely to yield a considerably higher outcome than most of the population.

In general, for every data subject S_i we assume a 'fair price' p_i that the attacker can get for their identification (e.g. by selling the re-identified dataset on a black market). Remember that the identification is not necessarily perfect, but may refer to a group of size g_i (achieving g_i*-distinction* in terms of Benitez and Malin). Thus the expected outcome for the attacker from the identified data subject S_i is $\frac{p_i}{g_i}$, and the total expected outcome over the whole population is

$$T = \sum_i \frac{p_i}{g_i} \; . \tag{1}$$

A rational attacker would only attack if this sum exceeds the cost of identification C. Note, however, that before the attack the attacker is only able to determine C, but not T, as he does not know *a priori* who the identified subjects will be, nor how small group sizes he will be able to identify. This may seem like an advantage for the party responsible for data anonymisation, but this is not necessarily the case.

Note that the attacker does not need the exact prior knowledge of T, but only the inequality $T \geq C$. An experienced attacker can make educated guesses based on his previous experience with the given type of data, the anonymisation mechanism used, and current black market prices. Assuming he runs re-identification attacks as a part of a larger systematic venture, he can also accept an occasional loss as long as he is profitable across the whole business.

We already saw in Section 2 that the outcome T, in general, depends on the investment C. We can now concretise this observation as follows.

The expected outcome T depends on the investment C in a monotonously increasing manner.

Justification of this claim is straightforward. If the attacker has a strategy that gives him the outcome T as a result of investment C, he still has the same strategy available with resources $C_1 > C$. This means that with the optimal strategy, he can only get a better result $T_1 \geq T$.

Even though monotonous, this increase does not have to be strict. In fact, we can expect $T(C)$ to be a stepwise function, see example depicted in Figure 1. Some initial level of positive outcome may already be achievable with essentially zero investment (e.g. by downloading freely available public datasets and running some easy analysis). However, in order to substantially improve the result, the attacker needs a qualitatively improved approach that costs him a non-trivial amount of resources (e.g. purchasing an additional dataset from a black market). There may be several such strategy improvements until the maximal achievable level of outcome T is reached.

Figure 1 also shows the line $C = T$ corresponding to the break-even strategies. Everything above this line is beneficial to the attacker, whereas the strategies below incur a loss.

Even though Figure 1 is just a sketch and does not correspond to any real analysis, there are two interesting observations to make here.

First,

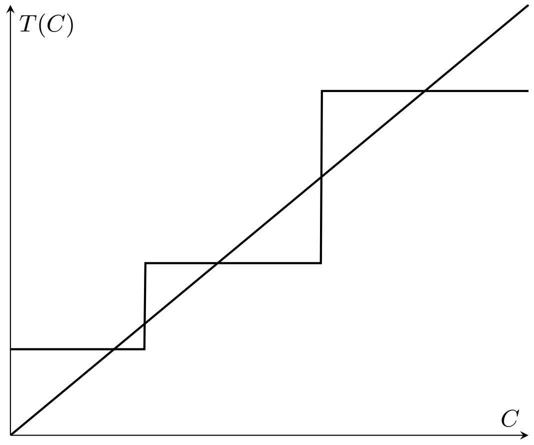

Fig. 1. Sample behaviour of outcome of the re-identification attack game

> *the attacker does not necessarily have just one profitable strategy, but he may have several.*

Thus, in the risk analysis, it is not sufficient to cover just one or two strategies, but the whole spectrum must be considered. This is of course far from being trivial, and this is what really makes risk analysis a hard task.

Second,

> *if the attacker is able to achieve even a marginal partial re-identification with zero cost, and there exists a subject S_i with positive attacker profit p_i, the attacker already has a profitable strategy.*

Indeed, note that in the equation (1) we have $g_i > 0$ for every i. Thus, if there exists a subject S_i with positive attacker profit p_i, we also have $T > 0$. We can conclude that existence of profitable attacker strategies is practically inevitable.

4 Discussion

Of course, the economic game model presented above is greatly simplified. Besides just inferring personal information, the attacker would also need to use some resources to turn this information into real profit (e.g. actually blackmail someone). Such a need incurs extra costs for the attacker both in therms of his own efforts and potential penalties. This in turn may influence the cost-benefit considerations presented in Section 3.

On the other hand, publishing sanitised datasets is not done just for fun, there is at least a social benefit in there, and we should account for that as well. Wan *et al.* have studied this setting and found that it is possible to find an equilibrium in this game that actually allows publishing more data than just following the default regulations (U.S. Health Insurance Portability and Accountability Act (HIPAA) in their case) [20].

There are of course more aspects to this game than just the search for equilibrium. The benefits of the game are social (e.g. higher-quality research and better policy decisions), whereas the consequences of the data inference attacks must be carried by the persons (e.g. in the form of being subjected to blackmailing attacks). This suggests that, as a part of a complete data release policy, the society may also need to establish recovery mechanisms for the breach victims. Possible measures include insurance and setting some of the extracted social outcome aside for compensation.

5 Conclusions and future work

GDPR fails to explicitly mention the issue of partial identification, nor does it refer to clear guidelines what to do in this case. Of course, one can take a viewpoint that even a partial identification is identification and GDPR should apply to the full extent. However, we argue that such a viewpoint is not rational as it would efficiently render any practical dataset as containing identifying data. This in turn would contradict the spirit of Recital 26 that speaks of taking objective factors (like potential attacker effort) into account when deciding about the likelihood of identification.

In practice, there are many possible levels of identifiability, and this situation should be addressed explicitly in the GDPR framework. One option would be to move away from the current binary approach where GDPR either does not apply at all or applies to the full extent. In case of partially indentifiable data, it should be possible to apply GDPR only partially as well.

It is worth noting that under the previous European data protection regulation (Directive 95/46/EC), Article 29 Data Protection Working Party released an opinion on anonymisation techniques [2]. This opinion explicitly considered the threat of inferring some partial information about the data subjects, even though no attempt was made to actually quantify the level of inference. For GDPR, similar guidelines are still in the planning phase at the time of this writing (early 2022) [1]. A position paper published in 2021 jointly by the European Data Protection Supervisor and Spanish Data Protection Agency is working in this direction stating that anonymisation is not a binary concept, and it is wrong to assume that it always reduces the risk of re-identification to zero [4].

The current paper did not propose a specific cost-benefit analysis methodology. Agreeing on one presumes a lot of discussions and is ultimately a political decision. The main message of the current paper is that this decision should take into account incentives of the actors involved. Most importantly, decisions about protection mechanisms should be based on the potential attack scenarios and not just some general rules-of-thumb.

The attack scenarios, in turn, depend on the value of the data to the attacker. This value is a continuous parameter that does not require subject identification to be 100% reliable. Even marginal success probability may be sufficient to mount a profitable attack. This consideration should be explicitly addressed by the relevant data protection regulations including GDPR.

Acknowledgements The author is grateful to Triin Siil, Tiina Ilus, Tanel Mällo and Kati Sein for fruitful discussions. The paper has been supported by the Estonian Research Council under the grant number PRG920.

References

1. EDPB Work Programme 2021/2022, The European Data Protection Board, https://edpb.europa.eu/system/files/2021-03/edpb_workprogramme_2021-2022_en.pdf
2. Opinion 05/2014 on Anonymisation Techniques (April 2014), Article 29 Data Protection Working Party, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
3. Common Methodology for Information Technology Security Evaluation. Evaluation methodology (April 2017), Version 3.1, Revision 5, CCMB-2017-04-004, https://www.commoncriteriaportal.org/files/ccfiles/CEMV3.1R5.pdf
4. AEPD-EDPS joint paper on 10 misunderstandings related to anonymisation (2021), https://edps.europa.eu/data-protection/our-work/publications/papers/aepd-edps-joint-paper-10-misunderstandings-related_en
5. Benitez, K., Malin, B.: Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association **17**(2), 169–177 (03 2010). https://doi.org/10.1136/jamia.2009.000026
6. Buchmann, E., Böhm, K., Burghardt, T., Kessler, S.: Re-identification of Smart Meter data. Pers. Ubiquitous Comput. **17**(4), 653–662 (2013). https://doi.org/10.1007/s00779-012-0513-6
7. Buldas, A., Laud, P., Priisalu, J., Saarepera, M., Willemson, J.: Rational Choice of Security Measures Via Multi-parameter Attack Trees. In: López, J. (ed.) Critical Information Infrastructures Security, First International Workshop, CRITIS 2006, Samos, Greece, August 31 - September 1, 2006, Revised Papers. Lecture Notes in Computer Science, vol. 4347, pp. 235–248. Springer (2006). https://doi.org/10.1007/11962977_19
8. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: The privacy bounds of human mobility. Scientific reports **3**(1), 1–5 (2013). https://doi.org/10.1038/srep01376
9. El Emam, K., Jonker, E., Arbuckle, L., Malin, B.: A systematic review of re-identification attacks on health data. PloS one **6**(12), e28071 (2011)
10. Elamir, E.A.H.: Analysis of Re-identification Risk Based on Log-Linear Models. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases: CASC Project International Workshop, PSD 2004, Barcelona, Spain, June 9-11, 2004. Proceedings. Lecture Notes in Computer Science, vol. 3050, pp. 273–281. Springer (2004). https://doi.org/10.1007/978-3-540-25955-8_21
11. Finck, M., Pallas, F.: They who must not be identified—distinguishing personal from non-personal data under the GDPR. International Data Privacy Law **10**(1), 11–36 (03 2020). https://doi.org/10.1093/idpl/ipz026

12. Kassem, A., Ács, G., Castelluccia, C., Palamidessi, C.: Differential Inference Testing: A Practical Approach to Evaluate Sanitizations of Datasets. In: 2019 IEEE Security and Privacy Workshops, SP Workshops 2019, San Francisco, CA, USA, May 19-23, 2019. pp. 72–79. IEEE (2019). https://doi.org/10.1109/SPW.2019.00024
13. Kikuchi, H., Yamaguchi, T., Hamada, K., Yamaoka, Y., Oguri, H., Sakuma, J.: Ice and Fire: Quantifying the Risk of Re-identification and Utility in Data Anonymization. In: Barolli, L., Takizawa, M., Enokido, T., Jara, A.J., Bocchi, Y. (eds.) 30th IEEE International Conference on Advanced Information Networking and Applications, AINA 2016, Crans-Montana, Switzerland, 23-25 March, 2016. pp. 1035–1042. IEEE Computer Society (2016). https://doi.org/10.1109/AINA.2016.151
14. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (S&P 2008). pp. 111–125. IEEE (2008)
15. Purtova, N.: From Knowing by Name to Personalisation: Meaning of Identification Under the GDPR. Available at SSRN 3849943 (2021)
16. Quelle, C.: Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-based Approach. European Journal of Risk Regulation **9**(3), 502–526 (2018). https://doi.org/10.1017/err.2018.47
17. Rocchetto, M., Tippenhauer, N.O.: On Attacker Models and Profiles for Cyber-Physical Systems. In: Askoxylakis, I., Ioannidis, S., Katsikas, S., Meadows, C. (eds.) Computer Security – ESORICS 2016. Lecture Notes in Computer Science, vol. 9879, pp. 427–449. Springer International Publishing, Cham (2016)
18. Skinner, C., Holmes, D.J.: Estimating the re-identification risk per record in microdata. Journal of Official Statistics **14**(4), 361 (1998)
19. Truta, T.M., Fotouhi, F., Barth-Jones, D.C.: Disclosure Risk Measures for Microdata. In: Proceedings of the 15th International Conference on Scientific and Statistical Database Management (SSDBM 2003), 9-11 July 2003, Cambridge, MA, USA. pp. 15–22. IEEE Computer Society (2003). https://doi.org/10.1109/SSDM.2003.1214948
20. Wan, Z., Vorobeychik, Y., Xia, W., Clayton, E.W., Kantarcioglu, M., Ganta, R., Heatherly, R., Malin, B.A.: A game theoretic framework for analyzing re-identification risk. PloS one **10**(3), e0120592 (2015). https://doi.org/10.1371/journal.pone.0120592
21. Yin, L., Wang, Q., Shaw, S.L., Fang, Z., Hu, J., Tao, Y., Wang, W.: Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. PloS one **10**(10), e0140589 (2015)
22. Zang, H., Bolot, J.: Anonymization of location data does not work: a large-scale measurement study. In: Ramanathan, P., Nandagopal, T., Levine, B.N. (eds.) Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MOBICOM 2011, Las Vegas, Nevada, USA, September 19-23, 2011. pp. 145–156. ACM (2011). https://doi.org/10.1145/2030613.2030630