Minimising data re-identification risks in the era of big data

Karlotine Weshe

June 17, 2025

Abstract

In the era of big data, unauthorised data re-identification has raised ethical considerations regarding data privacy. This paper uses literature review to show how the increase in publicly available data has contributed to the reversibility of re-identification processes and discusses the limitations of anonymization processes.

1 Introduction

Enabled by advanced data mining techniques, successful data re-identification attacks appears to demonstrate the limitations of current de-identification methods in ensuring data privacy. The sought-after privacy is also subject to the large amount of data made available through information generated by users of social networks, published by public agencies, and generated via connected devices. Considering that this context seems to be prone to violation of data protection principles, this raises concerns for both data donors and data holders.

This essay critically examines the issues surrounding the re-identification of individuals from anonymous data and how data science is evolving to address these challenges, based on the literature produced in recent years. Searches were performed in IEEE Xplore, ACM Digital Library, Google Scholar, JStor and PubMed. This work focuses only on cases where the units of observation are individuals and the databases contain Personal Identifiable Information. The remaining part of the paper proceeds as follows: Section 2 constitutes the main body of the text, give an overview of ethical questions surrounding data re-identification and the limitations of current anonymization models; Section 3 is the conclusion.

2 Issues and implemented solutions

2.1 Sources of data and classification

The connected society, fostered by technological advancements, has resulted in the production of large amount of complex data, increasing dataset options for adversaries[1]. In the private sphere, data is either published by individuals through social networks or shared by a private institution that has collected it in the context of the sale of a good, a service or in exchange for the use of an application. In parallel, in developed countries, governments have proceeded to integrate the information systems of different public institutions and agencies, published national databases that contain individual records on the use of public services and released open access data on different aspects of national life[1].

In the work related to anonymization, researchers generally consider three categories of attributes in these databases, where each raw correspond to a natural person. There are: a) the direct or explicit identifiers which are pieces of information unique to the person from whom they were collected (ex; full name, telephone number, national insurance number, IP address, email address); b)the quasi-identifiers which are information that are not specific to an individual but, when combined with other can lead to uniquely identify an individual and c) the sensible attributes such as disease and salary.[2, 1].In addition, the General Data Protection Regulation (GDPR) defines anonymous data as information which does not relate to an identified or identifiable natural person or personal data rendered anonymous in such manner that the data subject is no longer identifiable. Thus, the first step in anonymization process is to remove the direct identifiers[2].

2.2 Issues regarding de-identification/re-identification

In the literature, data anonymization is often referred to as a trade-off between data privacy and data utility; the aim being to minimise the risks of disclosure while maximising the utility of the data [3, 2]. Data that have the potential to optimize decision-making and processes have become highly valued in the public and private spheres and in scientific research. In companies, they can use data to learn more about customer preferences and behaviours and make personalised offers based on what this category of consumers is more likely to buy[1, 4]. In the medical field, big data is used, among other things, to prevent diseases and detect epidemics[4]. Thus, the usefulness of collecting and storing personal data cannot be dissociated from the responsibility to implement mechanisms to protect the privacy of data donors, such as sufficient anonymization. In developed countries, the processing of personal data is regulated by legal instruments such as the GDPR in the EU and the UK or the Health Insurance Portability and Account in the US.

Most researchers views individuals' re-identification as threat against privacy[3]. These studies have mentioned three categories of potential danger associated with this practice: social impacts, economic impacts, and mental impacts. Social repercussions may arise when the information revealed negatively affects the opinion of others on the data subject [5]. Economic aftermath are to be considered if the revealed information can be used to establish discriminatory practices in health insurance market and job market[5, 3] or undermine public's trust in data custodians [6, 7, 5]. It could also lead to mental distress as it affects self-image[5]). Furthermore, motivated by the high profitability (of the data market, data brokers also use this practice to build user profiles for commercial and business purposes[3]. This view is supported by researchers who are working to develop more robust anonymization techniques to enable society to take advantage of the potential economic growth generated by big data while minimising the risk of privacy violations.

Some analysts have attempted to draw fine distinctions between the usefulness of personal data in the medical field and the use of data for commercial purposes. On the other side of the spectrum, they draw our attention to the need to balance the privacy of patients with the benefit of being able to identify people with rare diseases[7]. They question whether identifiability should be considered a harm, or whether it should be analysed in the light of the potential dangers and benefits that may arise from it. They go on to explain that assigning a unique number to each individual through an encryption process can facilitate re-identification of research subjects [7], subject to the open consent of the patient[5].

2.3 Anonymization models

Much of the academic literature on data de-identification have been produced from the data custodian perspective since these papers have mainly focused on data publication or transmission. Indeed, during the two last decades, several models, combining generalization, suppression of records and perturbation, have been developed to anonymize data, among which the three methods that will be presented in this paper: k-anonymity, t-closeness and centralized differential privacy. K-anonymity has emerged as one most widely used methods[3]. After removing the explicit identifiers, k-anonymity is achieved if each unit of observation is indistinguishable from at least k - *1* other individuals in the release, with respect to the selected quasi-identifiers. However, this process could be reversed as the data is still vulnerable to attack[8], especially when the adversary has background knowledge on the records or when the sensible attributes lack diversity[2]. Another important source of uncertainty is the choice of the quasi-identifiers[8] to process since any attribute can be used as such by someone with prior knowledge[3]. t- closeness further develops the idea of k-anonymity, ensuring that the distribution of sensitive attributes in any data group with identical quasi-identifiers (equivalent class) is similar to that found in the overall database[2]. Centralized Differential Privacy (DP) differs from the two models mentioned above since it is a system for sharing patterns observed in the dataset instead of sharing anonymised data. The Dataset is perturbed with additive noise mechanisms such as Laplace mechanism [8].

Some papers attempt to expand beyond this aspect of data processing to introduce additional safeguards during the data collection stage. These researchers have proposed theories and protocols to implement local privacy by using randomized responses in surveys[9], by randomizing position to delocalize individual records [3]. Local Differential Privacy (LDP) has become the most widely adopted of those models which aim to ensure that each records contains little personal identifiable information

by adding noise while collecting data from the individual[10]. LDP is deployed in services offered by Google, Apple, Uber[10] and appears to be a promising model, especially with the increasing use of connected objects and the amount of data they generate. These methods, with their respective levels of robustness, are designed to make it more difficult to associate published data with data donors and thus prevent data re-identification.

2.4 Likeliness of re-identification of personal data

Most published successful re-identification attacks have been conducted by experienced computer scientits and on small databases for demonstration purposes[6, 7, 1]. Under the Data Protection Act 2018, re-identification is referred to as the act of re-identifying information that is de-identified personal data. This can be done, among other methods, by integrating several dataset or using background knowledge. In 2011, El Emam et al found that 26 percent of the attacks carried out in the US, particularly on health data, were successful. He points out that this high proportion of successes to be mainly due to non-compliance with existing standards and does not indicate the failure of de-identification. Furthermore, his results suggest that the proportion of records re-identified is much lower when it comes to large databases compare to small database[6]. In contrast, other researchers argue that augmentation of available public datasets could fuelled re-identification[1]. Throughout the world, the number of successful re-identification attempts on publicly available datasets has increased significantly since 2009; At least two to three attributes are required for successful re-identification[1]. These successes suggest that any publicly available dataset or information from a data breach[8] can be used for this purpose.

This appears to indicate that the multiplication of anonymisation methods has not yet succeeded in eliminating the risk of re-identification. For algorithms based on k-anonymity, several researches underline the difficulty of setting the optimal value of k, pointing out that this is sometimes done in an arbitrary way[11]. Yet the level of information loss and security gains depend heavily on the value of this parameter. Although the choice of the value of $k$has been explored in the literature, to date there is no consensus on the best approach. While some researchers have focused on minimising the information loss for a given value of k, others have developed methodologies for determining the actual value of k, depending on the context[11]. As regards Differential privacy, studies have highlighted that models based on that methodology are able to protect published datasets against most attacks as it guarantees robust privacy-preserving results [8]. Nevertheless, developers continue to build new models, often incorporating k-anonymity or DP, to facilitate the publication of data while reducing the risk of re-identification.

3 Conclusion

Current legislation in the European Union and United Kingdom requires institutions to protect themselves against risk of data re-identification, or even at the cost of anticipating new technologies that could be used by adversaries with harmful motives. Although the latest results show promising avenues, further research should be done on the de-identification of data collected through connected objects.

References

[1] J. Henriksen-Bulmer and S. Jeary, "Re-identification attacks—a systematic literature review," *International Journal of Information Management*, vol. 36, no. 6, pp. 1184–1192, 2016.

[2] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*, pp. 106–115, IEEE, 2006.

[3] M. Rodriguez-Garcia, M.-A. Cifredo-Chacon, and A. Quiros-Olozabal, "Cooperative privacy-preserving data collection protocol based on delocalized-record chains," *IEEE Access*, vol. 8, pp. 180738–180749, 2020.

[4] A. Khanan, S. Abdullah, A. H. H. Mohamed, A. Mehmood, and K. A. Z. Ariffin, "Big data security and privacy concerns: a review," in *Smart Technologies and Innovation for a Sustainable Future: Proceedings of the 1st American University in the Emirates International Research Conference—Dubai, UAE 2017*, pp. 55–61, Springer, 2019.

[5] J. E. Lunshof, R. Chadwick, D. B. Vorhaus, and G. M. Church, "From genetic privacy to open consent," *Nature Reviews Genetics*, vol. 9, no. 5, pp. 406–411, 2008.

[6] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PloS one*, vol. 6, no. 12, p. e28071, 2011.

[7] M. G. Hansson, H. Lochmüller, O. Riess, F. Schaefer, M. Orth, Y. Rubinstein, C. Molster, H. Dawkins, D. Taruscio, M. Posada, *et al.*, "The risk of re-identification versus the need to identify individuals in rare disease research," *European Journal of Human Genetics*, vol. 24, no. 11, pp. 1553–1558, 2016.

[8] A. Zaman, C. Obimbo, and R. A. Dara, "An improved differential privacy algorithm to protect re-identification of data," in *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, pp. 133–138, IEEE, 2017.

[9] J. Ding and B. Ding, "Interval privacy: A framework for privacy-preserving data collection," *IEEE Transactions on Signal Processing*, vol. 70, pp. 2443–2459, 2022.

[10] S. Wang, Y. Qian, J. Du, W. Yang, L. Huang, and H. Xu, "Set-valued data publication with local privacy: Tight error bounds and efficient mechanisms," *Proc. VLDB Endow.*, vol. 13, p. 1234–1247, apr 2020.

[11] R. Dewri, I. Ray, I. Ray, and D. Whitley, "On the optimal selection of k in the k-anonymity problem," in *2008 IEEE 24th International Conference on Data Engineering*, pp. 1364–1366, IEEE, 2008.