Statistical disclosure control: moving forward

Josep Domingo-Ferrer[1], David Sánchez[1], and Krishnamurty Muralidhar[2]

[1] Department of Computer Science and Mathematics, CYBERCAT–Center for Cybersecurity Research of Catalonia, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Catalonia. E-mail {josep.domingo, david.sanchez}@urv.cat

[2] Department of Marketing and Supply Chain Management, Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman, OK 73019, USA. E-mail krishm@ou.edu

Acknowledgments. This research was funded by the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez and grant 2021SGR-00115), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grants PID2021-123637NB-I00 "CURLING" and PRE2019-089210, and the EU's NextGenerationEU/PRTR via INCIBE (project "HERMES" and INCIBE-URV Cybersecurity Chair).

1. Introduction

The very first privacy notions in official statistics can be found in the *randomized response* mechanism (RR) (Warner 1965), which protects query answers by reporting true answers with probability *p* and false answers with probability *1-p*, and the *Dalenius desideratum* (Dalenius 1977), which states that nothing about any specific individual should be learnable from a publicly released statistical database that cannot be learned without access to the database. These foundations ushered in the *statistical disclosure control* (SDC) discipline (Adam and Worthmann 1989; Willenborg and De Waal 1996; Hundepool et al. 2012), whose goal is to provide methods for data anonymization.

As the needs for privacy-preserving data sharing increased, it was soon realized that *i)* Dalenius-like privacy notions were too stringent to produce protected data with reasonable analytical utility, and *ii)* non-interactive data releases were much more useful and needed than interactive query answering. Hence, the first practical approaches to anonymization focused on the release of statistical databases and were *utility-first*, where anonymization parameters are iteratively tried until sufficient analytical utility is empirically achieved while reducing the observed disclosure risk below a certain threshold. Examples of these heuristic mechanisms are data suppression, sampling, noise addition, microaggregation, data swapping, or value generalization.

To overcome the limitations of the empirical nature of classical SDC methods, the computer science community approached the problem in the late 90s from the *privacy-first* angle by introducing the notion of *privacy model*, by which an *ex ante* privacy condition is enforced using one or several SDC methods. Privacy models have the advantage of setting the privacy level by design, without having to empirically evaluate it. The first privacy model was *k*-anonymity (Samarati and Sweeney 1998), which states that the probability of re-identifying a record in a protected database by leveraging available background knowledge on the individual to whom the record corresponds should be at most 1/*k*. *k*-Anonymity was followed by a number of extensions, such as *l*-diversity (Machanavajjhala et al. 2007), *t*-closeness (Li et al. 2010) and others, which

mitigate the disclosure risk of confidential values. All privacy models in the *k*-anonymity family were designed to anonymize microdata sets, that is, data sets formed by records that correspond to individual respondents.

In 2006, a new privacy model named ϵ-differential privacy (DP) was proposed in the cryptographic community, which aimed to bring to data sharing a privacy guarantee based on indistinguishability by reformulating the old Dalenius desideratum in an even more stringent manner: "the risk to one's privacy […] should not substantially increase as a result of participating in a statistical database" (Dwork 2006).

DP was originally proposed for interactive statistical queries to a database, where a randomized query function (that returns the query answer plus some noise) satisfies ϵ-DP if, for any two databases that differ in one record, the respective query answers do not differ by more than an exponential factor of ϵ (called the budget). This ensures the presence or absence of any single record does not significantly affect query answers. The smaller ϵ, the higher the protection, with $\epsilon \leq 1$ being considered 'safe' (Dwork 2011). The noise to be added to the answer to enforce a certain ϵ depends on the global sensitivity of the query to the presence or absence of any single record. One can see that the enforcing mechanism employed by DP is very similar to the RR forerunner, as both produce randomized answers to interactive queries with a predefined probability.

Thanks to its neat privacy guarantee, DP has been rapidly adopted by the research community, to the point that previous approaches tend to be regarded as obsolete. In fact, DP has become the *de facto* privacy standard in most data intensive areas, such as data analytics, statistics or machine learning (Wood et al. 2018). However, while mild noise may suffice to enforce DP on aggregated statistical queries (e.g., averages), its stringent privacy guarantee requires a very large noise to be added for identity queries returning the contents of a specific record. In those cases, the noise needed to enforce any safe-enough ϵ is so large that the DP-protected outcomes become nearly random and, therefore, analytically useless. The latter is precisely the case of microdata releases, which are more in demand than interactive queries (Domingo-Ferrer et al. 2021). The research community has devoted enormous efforts during the last 18 years trying to fit (or bend)

DP to those settings, with the arduous aim of reconciling the DP privacy guarantee with sufficient data utility.

2. Challenges

A shortcoming of the utility-first approach traditionally embraced by the SDC community, especially in official statistics, is that it requires (possibly several rounds of) empirical disclosure risk assessment to make sure the protected data are safe enough. Such an assessment is expensive, and it requires selecting plausible attack scenarios and parameters. In recent years, there has been a tendency to seek *ex ante* privacy guarantees. Whereas official statisticians have mostly opted for synthetic data generation (Burnett-Isaacs et al. 2022), computer scientists have favored privacy models, especially DP. We next sketch the challenges associated with those *ex ante* approaches.

Synthetic data are artificial data preserving some of the statistical features of the original data. Their main attractive is that, being artificial, it would seem they can be released without privacy concerns. However, they are not without issues. If they are generated with the only requirement that certain statistics be exactly preserved, their utility is limited to those statistics, whereas other statistics or subdomain analyses are not preserved. Then one might wonder why not just publish the statistics to be preserved rather than a synthetic data set. On the other hand, if synthetic data are generated by fitting a statistical or machine learning (ML) model to the original data and then sampling that model, the issue of potential overfitting appears. Indeed, sampling an overfitted model is likely to yield synthetic data extremely close to the original: while this is good for utility, it is very bad for privacy (reidentification and attribute disclosure are likely). Overfitting is especially worrisome when using complex ML models, such as deep neural networks, whose millions of parameters may encode a snapshot of the confidential original data they are trained on.

The deployment of privacy models has mostly focused on DP, whereas the *k*-anonymity family of models have so far enjoyed less popularity. Yet, *k*-Anonymity models modify the input

microdata, from which consistent statistical outputs can be computed; in contrast, DP adds noise independently to each statistical output, which is likely to result in inconsistencies among outputs (Citro et al. 2022). A relevant case study in official statistics is the application of DP to the 2020 U.S. Census. This application has faced operational and scientific challenges (Garfinkel et al. 2018), which have resulted in significant delays in data release, the use of relaxed DP with budgets as large as ϵ=39.907 (far above the $\epsilon \leq 1$ recommended by the DP inventors), and very substantial inaccuracies in the DP-protected data (Kenny et al. 2021). Other DP-deployments by the big companies (Apple, Google, Microsoft, Meta, LinkedIn) have also resorted to large ϵ budgets (whose privacy guarantee is ineffectual) or to questionable composition assumptions, in an attempt to preserve utility.

Regarding the combination of synthetic data and DP, it amounts to training the synthetic data generation models on the original data in a differentially private manner. Since training is an iterative procedure, the sequential composition property of DP applies, which means that the budget used increases with the number of iterations (Abadi et al. 2016) and hence the privacy level decreases. Despite this budget increase, the impact of DP-training on the utility of the trained model remains significant (Domingo-Ferrer et al. 2021), which means that the utility of the synthetic data obtained by sampling the model can be seriously affected. In fact, in Blanco et al. (2023) it is shown that anti-overfitting techniques applied during ML model training may yield a better privacy-utility trade-off than DP.

3. Directions

From the challenges above, it becomes clear that empirical disclosure risk assessment remains as unavoidable for synthetic or DP data as it was under the utility-first traditional approach to SDC. DP can only provide enough *ex ante* privacy to dispense with *ex post* disclosure risk assessment when the budget ϵ is really small (say below 1), but in this case utility is likely to be too low for most practical applications. When the budget is large, DP is little more than mild classical noise addition, and hence disclosure risk assessment is still needed. Future directions include:

- *Ex ante* privacy that does not require disclosure risk assessment remains a very desirable goal. Privacy models are the way to go but, given the difficulties of deploying DP in practice, exploring the other large family of privacy models (*k*-anonymity and its extensions) seems a good option. The *k*-anonymity family can provide reasonable utility in data releases and data collection with *ex ante* privacy guarantees that are easy to understand in terms of limiting identity disclosure (upper bound on the re-identification probability) and attribute disclosure (lower bounds on the diversity of confidential attributes). Further, *k*-anonymity and variants of it remain the best privacy models to anonymize some data types for which perturbative approaches would cause severe utility damage, such as stream data (Xiao and Tao 2007) or georeferenced data (Cremer et al. 2024).
- Coming up with *realistic* disclosure risk assessment in practical scenarios is necessary. Potential re-identification/disclosure attacks are sometimes invoked as an argument to justify the use of stronger SDC methods and privacy models. However, *no successful attacks against properly anonymized data* (e.g., those released by national statistical institutes) *have been reported*, which contradicts the pessimism in Ohm (2009). Much publicized re-identification attacks were mounted against data that had undergone mere pseudonymization or suppression of identifiers --- the reidentification of the Governor of Massachusetts (Sweeney 1997), the AOL attack (Barbaro and Zeller 2006), and the Netflix Prize Dataset attack (Narayanan and Shmatikov 2006)--- or against poorly anonymized data (De Montjoye et al. 2015; Sánchez et al. 2016). Also, the much discussed reconstruction of data compatible with the released statistical outputs does not in general imply re-identification (Muralidhar and Domingo-Ferrer 2023).
- When using ML models in data protection, especially to generate synthetic data, overfitting must be avoided. Overfitted models may memorize the entire original data and yield synthetic data extremely close to the original. One option is to use anti-overfitting methods. An alternative is to use only simple models for synthetic data generation, such as statistical models or basic ML methods (decision trees, random forests, etc.). The

bottom line is that models used for synthetic data generation should not be too accurate if we do not want the resulting synthetic data to be too close to the original data.

- Non-perturbative SDC methods (such as data suppression, generalization or sampling, see Hundepool et al. (2012)) are probably a better choice than perturbative methods in settings where perturbed data (e.g., DP data obtained via noise addition) might be unacceptable or invalidate analytical conclusions (e.g., in the medical domain or official statistics). Non-perturbative methods keep protected data truthful, and they are sufficient to enforce formal *k*-anonymity privacy.
- Alternative data sharing mechanisms that are privacy- and utility-preserving-by-design also deserve attention. Let us mention federated learning, anonymous channels, or secure multiparty computation as an alternative to sharing anonymized data.

References

M. Abadi, A. Chu, L. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. 2016. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

N. R. Adam and J. C. Worthmann, J.C. 1989. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys, 21(4): 515-556.

M. Barbaro and T. Zeller. 2006. A face is exposed for AOL searcher no. 4417749. The New York Times, Aug. 9.

A. Blanco-Justicia, D. Sánchez, D., J. Domingo-Ferrer, and K. Muralidhar. 2022. A critical review on the use (and misuse) of differential privacy in machine learning. ACM Computing Surveys, 55(8), pp.1-16.

K. Burnett-Isaacs, C. Girard, K. Sallier, G. Raab, A. Ramsden, M. Slokom, C. Task, M. Salamh, A. Wardlaw, E. Ghashim, R. Ostic, P. Belanger et al. 2022. Synthetic Data for Official Statistics: A Starter Guide. United Nations Economic Commission for Europe.

C. Citro, R. Prevost, and I. Youngs. 2022. U.S. Political and Statistical Geography and Census Small-Area Statistics: A Primer. Working Draft 2.1. Georgetown University. https://mdi.georgetown.edu/census-geographies-project/

S. Cremer, L. Jehmlich, and R. Lenz. 2024. Masking georeferenced health data-an analysis taking the example of partially synthetic data on sleep disorder. In International Conference on Privacy in Statistical Databases (PSD 2024), 297-309. Springer.

T. Dalenius. 1977. Towards a methodology for statistical disclosure control. Statistik Tidskrift 15: 429-444

J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia. 2021. The limits of differential privacy (and its misuse in data release and machine learning). Communications of the ACM, 64(7):33–35.

C. Dwork. 2006. Differential privacy. In Proceedings of the 33rd Intl. Colloquium on Automata, Languages and Programming-ICALP 2006. 1–12.

C. Dwork. 2011. A firm foundation for private data analysis. Communications of the ACM 54 (1): 86–95.

S. L. Garfinkel, J. M. Abowd, and S. Powazek. 2018. Issues encountered deploying differential privacy. In Proceedings of the 2018 Workshop on Privacy in the Electronic Society. 133-137.

A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.-P. De Wolf. 2012. Statistical Disclosure Control. John Wiley & Sons.

C. T. Kenny, S. Kuriwaki, C. McCartan, E. T. R. Rosenman, T. Simko, and K. Imai. 2021. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census. Science Advances, 7(1):eabk3283.

N. Li, T. Li, and S. Venkatasubramanian. 2010. Closeness: a new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering 22(7): 943–956.

A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramaniam. 2007. l-Diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data 1(1): 3.

Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland. 2015. Unique in the shopping mall: on the reidentifiability of credit card metadata. Science, 347(6221):536-539.

K. Muralidhar and J. Domingo-Ferrer. 2023. Database reconstruction is not so easy and is different from reidentification. Journal of Official Statistics, 39(3):381-398.

A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (SP 2008), 111-125). IEEE.

P. Ohm. 2009. Broken promises of privacy: responding to the surprising failure of anonymization. UCLA L. Rev. 57:1701.

P. Samarati and L. Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report. SRI International.

D. Sánchez, S. Martínez, and J. Domingo-Ferrer. 2016. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". Science, 351(6279):1274-1274.

L. Sweeney. 1997. Weaving technology and policy policy together to maintain confidentiality. Journal of Law, Medicine and Ethics 25:98-110.

S. L. Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60(309): 63–69.

L. Willenborg and T. De Waal. 2012. Elements of statistical disclosure control. Springer Science & Business Media.

A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D.R. O'Brien, T. Steinke, and S. Vadhan. 2018. Differential privacy: A primer for a non-technical audience. Vand. J. Ent. & Tech. L. 21:209.

X. Xiao and Y. Tao. 2007. M-invariance: towards privacy preserving re-publication of dynamic datasets. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, 689-700.