Why the search for a privacy-preserving data sharing mechanism is failing

Theresa Stadler and Carmela Troncoso*
theresa.stadler@epfl.ch carmela.troncoso@epfl.ch

SPRING Lab, EPFL, Switzerland

April 29, 2022

Standfirst: The rapidly growing demand to share data more openly creates a need for secure and privacy-preserving sharing technologies. However, there are multiple challenges associated with the development of a universal privacy-preserving data sharing mechanism, and existing solutions still fall short of their promises.

Data-driven innovation is promised to bring enormous benefits for all, such as improved health care via personalised medicine, or better governmental services and more efficient and greener industrial production via data-driven resource allocation. Wide-spread access to data and the ability to use it are hence considered essential for future innovation and growth [1, 2].

Current aggressive data collection and analysis practices, however, raise alarms about the threat these techniques pose to societal values and fundamental rights [3]. How to widen access to data while safeguarding the confidentiality of sensitive, personal information has thus become one of the most prevalent challenges in unleashing the potential of data-driven technologies.

The promised way out: Privacy-enhancing technologies. Privacy-enhancing technologies (PETs) are seen by many as the "holy grail" that will open up access to valuable data while protecting individuals' right to privacy. PETs cover a wide range of data sharing scenarios and privacy requirements.

The most widely applied PETs are tools that help users to control private information sharing in online contexts, e.g., privacy preferences, or aim to enhance transparency, e.g., privacy mirrors [4].

A second class of PETs, which have rapidly advanced over the past few years, enable analysts to derive insights without access to sensitive data in the clear. These PETs comprise techniques, such as homomorphic encryption [5], secure multi-party computation [6], and differentially-private aggregation [7],

*Corresponding author

that operate on confidential inputs and *only reveal the final result of the computation* to the analyst. These techniques offer a valuable choice for data holders that are primarily interested in deriving population-level insights. For instance, those who want to publish the results of a one-off statistical analysis [7] or those who want to learn a statistical model from multiple datasets stored in several locations without the need to pool this data on a central server [6].

These two classes of PETs, however, do not solve one of the problems that is most relevant to practitioners: how to share *high-quality individual-level data* in a manner that preserves privacy but allows analysts to extract a dataset's full value. Sharing such fine-grained data, typically known as microdata, is considered crucial to foster innovation in many fields, such as medicine or finance, primarily due to the following reasons. First, microdata contains fine-grained patterns that provide insights which other types of data releases, such as differentially private aggregations, might not. For example, discovering medical anomalies or detecting financial fraud requires access to rare statistical signals that are hard to preserve in data derived through, for instance, differentially private aggregation. Second, as opposed to tools for private computation, the utility of microdata is *not constrained to a single analysis task*. Microdata has the advantage that it is well-suited for tasks such as data exploration and can be re-purposed to answer multiple research questions.

A history of failures. Early attempts to protect microdata from privacy breaches were based on the idea that to preserve privacy one simply had to remove certain data fields that might act as identifiers. Initial research showed that redacting direct identifiers such as names, social security or passport numbers, was not enough to prevent privacy breaches. Instead, these works suggested that to destroy the link between an individual's identity and their record in the published data it suffices to remove or blur combinations of attributes that might form a unique identifier [8]. Privacy notions such as k-anonymity, l-diversity, or t-closeness formalise this idea and all rely on the same paradigm: to predict *before publication* which data attributes could be used by privacy adversaries to single out or re-identify individuals, and then to *redact* this information through suppression, generalisation or perturbation [4].

The issue with this strategy is that due to the high dimensionality of most microdatasets, it is impossible to anticipate which combination of data attributes privacy adversaries might use to re-identify individuals and extract sensitive information [9]. High-dimensional datasets offer adversaries a myriad of attribute combinations that could act as potential identifiers. To mitigate the risk of re-identification, data holders need to either accurately predict which attributes are available to potential adversaries, and remove only these attributes from the shared data; or preemptively redact any attribute combinations that might lead to privacy violations. Neither strategy gives the desired high utility without residual privacy risks. The former in most cases fails to provide any meaningful privacy, as has been demonstrated by many real-world examples [9, 10], while the latter often destroys most of a dataset's statistical utility and undermines the major benefits of microdata sharing.

New proposals, same result. In an attempt to bypass these fundamental limits of redaction-based techniques, researchers and practitioners continue to propose new private data release mechanisms, such as synthetic data sharing or novel anonymisation techniques [11, 12]. Due to the high value ascribed to microdata releases these new proposals very quickly make it to the market [13–15], often with only little evidence to back up claims about their benefits [16]. The high dimensionality of most datasets, however, implies that the novel sharing mechanisms, presented as better alternatives that address the shortcomings of traditional techniques, are actually subject *to the same trade-offs* between preserving the privacy and utility of the shared microdata as their redaction-based predecessors. This fact is rarely identified early enough in the development and application process because of two main reasons.

1) Focus on absolute rather than relative trade-offs. The proponents of new data sharing techniques typically focus on quantifying the absolute privacy guarantees their new mechanism provides, i.e., how much sensitive information adversaries can extract from the mechanism's output. If, however, even publishing the raw data or the output of a simpler mechanism does not allow adversaries to make any such inferences, this approach overestimates the benefits the novel mechanism might bring. When evaluating a new proposal, the question to be answered is not only if the output of the mechanism protects a dataset's privacy but if it offers *a better trade-off* between privacy and utility than simpler techniques; or than releasing the raw data directly.

2) Lack of empirical adversarial evaluations. To argue a mechanism's privacy properties, proposals most often either rely on naive privacy notions, such as similarity metrics between a mechanism's in- and output, or on hard-to-achieve and difficult-to-interpret formal privacy definitions, such as differential privacy. Rarely, they include experiments that empirically quantify how well the mechanism withstands *strong* privacy adversaries and how well it protects a dataset's *most vulnerable records*, i.e., those most exposed to privacy violations when releasing the raw or anonymised data. Weaknesses in a mechanism's design or implementation often remain undiscovered either because evaluations are run under weak privacy notions or lack altogether [17].

The latest unfulfilled promise: Synthetic data. The latest example of these pitfalls is synthetic data. Synthetic data has been hailed as "the next best step in sanitised data release" [11] and was quickly put into application for a wide set of use cases. Synthetic data is often presented as a novel "data anonymisation solution" [14] that addresses the shortcomings of traditional, redaction-based techniques. Data holders are promised that publishing a synthetic in place of the raw dataset retains the data's full value but prevents the leakage of private information about individuals in the raw data previously observed under redaction-based techniques [9].

Upon scrutiny, synthetic data was shown to offer the same trade-offs as traditional anonymisation [18]. The records most vulnerable to privacy attacks under simple anonymisation techniques, statistical outliers that often belong to minority subpopulations, could only be protected from privacy breaches if the synthetic data published did not retain the full promised value of the original dataset – as was the case for earlier microdata anonymisation techniques.

The way forward: Restricting information release. Years of research have shown that sharing high-dimensional datasets in a manner that *preserves both privacy and high utility is close to impossible*. We thus argue that the continued search for a fully flexible, high-utility, strong-privacy data release mechanism comes close to chasing rainbows. As hard as it may be, both researchers and practitioners should finally accept the *inherent trade-off* between high flexibility in data utility and strong guarantees about privacy, even if may mean to reduce the scope of data-driven applications. Depending on the data used, the goals of the data sharing, and their privacy requirements, data holders will need to make explicit choices about the data sharing approach most suitable to their use case.

For use cases that require high utility and flexibility in analysis functions evaluated over the data, analysts must accept that technical privacy safeguards, such as microdata anonymisation or synthetic data sharing, will only offer weak protection. In these scenarios, privacy will hence depend upon legal protections that bind the published data to a particular purpose to guarantee compliance with relevant data protection regulations [3]. As an example, sharing high-quality, individual-level clinical trial data for secondary analysis offers enormous benefits because it enables researchers to re-purpose hard-to-obtain datasets and answer multiple research questions [19]. To draw new insights, the shared data must retain as many of its original statistical patterns as possible, including those previously undiscovered and not known to the data holder at the time of sharing. However, preserving enough utility for such exploratory analyses while at the same time providing strong guarantees about privacy is, as has been shown [9,18], an unattainable goal. Both traditional anonymisation techniques, as well as more recent alternatives, such as synthetic data sharing, have been shown to provide poor privacy-utility trade-offs. Data holders who seek to publish high-utility research data, such as clinical trial data, for secondary purposes should hence implement additional procedural controls that restrict the scope of the data sharing and minimise the risks of privacy breaches.

To use cases that come with well-defined, tightly scoped analysis tasks, PETs that derive specific insights from sensitive datasets under strong privacy guarantees, such as homomorphic encryption or differentially private computations, offer a promising solution. These technologies, by design, implement many of the relevant data protection principles, such as purpose limitation: they strictly limit the use of the data to a concrete analysis task and minimize information leakage. For example, frameworks for privacy-preserving analytics of genomic patient data enable analysts to answer a restricted number of common research questions [20]. These tools guarantee that analysts can obtain the desired study results but can not extract any information about individual patients or answer queries beyond the agreed analysis scope.

For all use cases, empirical adversarial evaluations remain necessary to uncover flaws in a mechanism's design or implementation, as well as to ensure that the sharing mechanism in use does not create any disparate impact on population minorities [21]. Even PETs that largely reduce the exposure of private information and bind data to a fixed use case might produce outputs that lead to unexpected inferences [17]. Data-oriented, empirical risk assessments enable data holders to detect such unexpected leakage and assert that the data published is in accordance with relevant regulations and protect individual privacy.

Conclusions. We conclude that going forward privacy researchers and policy makers should rethink their current approach to support data holders in their goal to share data in a privacy-preserving manner. As a first step, both groups should abandon the futile search for a silver bullet solution to all-purpose-utility high-privacy sharing of fine-grained data. Instead, we argue, data holders need to accept that the set of use cases solvable under strict privacy guarantees may be restricted, and thus so the data-driven business models linked to them. Privacy researchers should hence refocus their efforts on developing tools that help data holders to identify those use cases that can be tackled under good privacy and good utility simultaneously. Finally, we recommend that policy makers, together with technical experts, develop guidelines that assist data holders in navigating the complex landscape of PETs. These guidelines should focus not only on matching uses cases to their suitable sharing technologies but also comprise recommendations for empirical evaluation methods that can assure the public that any loss in privacy is weighed off by the promised societal benefits.

Acknowledgments. This work was partially funded by the Swiss National Science Foundation with grant 200021-188824 (TS).

Author contributions. C.T. wrote the paper; T.S. wrote the paper and contributed to the analysis of synthetic data privacy properties.

Ethics declaration. This work does not have any ethical implications.

Competing interests. The authors declare no competing interests.

References

[1] European Commission. European data strategy. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy, 2020. Accessed 2020-06-01.

[2] US Federal Departement of Commerce US Federal Office of Management & Budget and Science & Technology. Federal data stategy. https://strategy.data.gov/, 2020. Accessed 2020-06-01.

[3] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. https://ec.europa.eu/justice/article-29/

documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

[4] C. Troncoso. Privacy & Online Rights. In *The Cyber Security Body of Knowledge*. University of Bristol, 2019.

[5] N. Smart. Cryptography. In *The Cyber Security Body of Knowledge*. University of Bristol, 2019.

[6] D. W. Archer, D. Bogdanov, Y. Lindell, L. Kamm, K. Nielsen, J. I. Pagter, N. P. Smart, and R. N. Wright. From keys to databases—real-world applications of secure multi-party computation. *The Computer Journal*, 2018.

[7] C. Dwork and A. et al. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

[8] L. Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS*, 2002.

[9] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of "personally identifiable information". *Communications of the ACM*, 2010.

[10] C. Culnane, B. I. P. Rubinstein, and V. Teague. Health data in an open world. *CoRR*, 2017.

[11] S. M. Bellovin, P. K. Dutta, and N. Reitinger. Privacy and synthetic datasets. *Stan. Tech. L. Rev.*, 22, 2019.

[12] Cristian Augusto, Jesús Morán, Claudio De La Riva, and Javier Tuya. Test-driven anonymization for artificial intelligence. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, 2019.

[13] Mostly-ai. https://mostly.ai/.

[14] Statice. https://statice.ai/.

[15] Hazy. https://hazy.com/.

[16] V. Bernardo. Tech Sonar 2021 – 2022 Report – Synthetic Data. Technical report, European Data Protection Supervisor, 2021.

[17] F. Houssiau, L. Rocher, and YA. de Montjoye. On the difficulty of achieving differential privacy in practice: user-level guarantees in aggregate location data. *Nature Communications*, 2022.

[18] T. Stadler, B. Oprisanu, and C. Troncoso. Synthetic data–anonymisation groundhog day. In *USENIX*, 2022.

[19] Khaled El Emam, Lucy Mosquera, and Chaoyi Zheng. Optimizing the synthesis of clinical trial data using sequential trees. *Journal of the American Medical Informatics Association*, 2020.

[20] David Froelicher, Juan R Troncoso-Pastoriza, Jean Louis Raisaro, Michel Cuendet, Joao Sa Sousa, Jacques Fellay, and Jean-Pierre Hubaux. Truly privacy-preserving federated analytics for precision medicine with multi-party homomorphic encryption. *Nature Communications*, 2021.

[21] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov. Differential privacy has disparate impact on model accuracy. In *NeurIPS*, 2019.