Understanding De-identification Guidance and Practices for Research Data

Wentao Guo
University of Maryland

Aditya Kishore
University of Maryland

Paige Pepitone
NORC at the University of Chicago

Adam J. Aviv
The George Washington University

Michelle L. Mazurek
University of Maryland

Abstract

Publishing de-identified research data is beneficial for transparency and the advancement of knowledge, but it creates the risk that research subjects could be re-identified, exposing private information. De-identifying data is difficult, with evolving techniques and mixed incentives. We conducted a thematic analysis of 38 recent online de-identification guides, characterizing the content of these guides and identifying concerning patterns, including inconsistent definitions of key terms, gaps in coverage of threats, and areas for improvement in usability. We also interviewed 26 researchers with experience de-identifying and reviewing data for publication, analyzing how and why most of these researchers may fall short of protecting against state-of-the-art re-identification attacks.

1 Introduction and methods

Publishing research data has numerous benefits for the research community and the public [12]; as a result, researchers are increasingly expected to do so not only by their peers, but also by public [8, 9, 13] and private [2] research funders, and by journals and other publication venues [15]. When research data is about human subjects, though, these benefits and expectations must be balanced with the need to protect potentially sensitive information. Linking data back to individuals can result in various harms: data about sexual behavior may be socially stigmatizing, data about reproductive care may result in legal consequences, and data about political opinion in conflict zones may lead to physical violence. Researchers aim to reduce this risk by *de-identifying* data—modifying data, or the interface for viewing it, to make it more difficult to re-identify or learn information about individuals.

Techniques for de-identifying data are many, and striking a balance between risk and utility can be difficult given various attacks on de-identification [4, 11, 14]. Many traditional approaches are time-consuming [1] and offer questionable protection [4, 11]. However, newer techniques such as differential privacy [5] are no panacea either: differential privacy may be unfamiliar or unacceptable to practitioners [3, 10]—especially for datasets where it may not be possible to achieve a satisfactory balance between privacy and utility—and there is an unmet need for accessible, fully featured tools for implementation [6, 7]. Compounding these challenges, in many cases de-identification is carried out as an afterthought by researchers with limited time and resources. We aim to help researchers de-identify data efficiently and effectively. To date, we have approached this goal from two angles.

Analyzing de-identification guides. Well-designed guidance could teach researchers de-identification approaches that are proven, context-appropriate, and accessible. Indeed, the Internet abounds with de-identification advice, ranging from short corporate blog posts to government-written guides numbering hundreds of pages. To assess the quality and consistency of these existing resources, we systematically collected online de-identification guides from the last five years and conducted thematic analysis on a sample of 38. We investigate the content they contain, particularly with regard to techniques and attacks, as well as how they are designed to help readers decide on a de-identification strategy and carry it out.

Interviewing practitioners and curators. To best help researchers de-identify data, we should first understand how they currently approach de-identification, what kinds of re-identification risk remain (and why), and what they perceive as challenges. With IRB approval, we conducted paid, hour-long remote interviews with 18 practitioners who have de-identified research data for publication, as well as 8 curation

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

staff at data repositories and funding agencies who review data submitted by others. We recruited most participants systematically from online data repositories (focusing on the areas of health and healthcare, crime and criminal justice, and international development), and a few through purposive sampling to cover a greater diversity of research organizations and de-identification methods. We investigate participants' perceptions of threats, their process for de-identification, and their needs and wants.

2 De-identification guide analysis findings

We find that de-identification guides cover some basic techniques such as deletion, pseudonymization, and generalization near-universally, while they mention technically complex approaches such as differential privacy less frequently—especially in guides for researchers, as opposed to guides for government agencies or businesses. Sowing potential for confusion, terms such as *anonymization*, *aggregation*, and *differential privacy* are defined inconsistently across guides. And we observe notable gaps in threat coverage, including claims that variables such as salary and medical diagnosis are non-identifying, when in fact these types of information are often for sale, available online, or known to specific people such as healthcare workers. Gaps also include patchy coverage of reverse engineering attacks, which encompass using knowledge or assumptions about how specific de-identification techniques were applied to undo them,[1] and reconstructing missing data by reasoning about available data.[2]

As for helping readers decide on a de-identification strategy and carry it out, we find that guides do discuss trade-offs to help readers choose between de-identification techniques; however, much of this discussion is vague, merely stating that choosing one approach over another will change the balance between privacy and utility. Several guides contain encyclopedic tables of techniques or instructions, which have the potential to be useful resources but also overwhelming or overly prescriptive. While just half of the guides we analyzed contain examples of data to help illustrate de-identification approaches, we find that even fewer guides provide examples that help readers think through de-identification in the context of multiple variables in a dataset. Similarly, few guides provide case studies of disclosure in the real world, which could serve as crucial motivation.

3 Interview findings

Many interview participants perceive a tension between the importance of protecting their research participants against severe consequences of re-identification and the belief that, realistically, no one would ever try to re-identify them. They are reasonably well informed about the kinds of factors that affect the re-identifiability of data, such as the presence of values that uniquely identify data subjects in combination, and the availability of external linking data.

Despite their high-level understanding of risk factors, most participants follow a de-identification process that is dominated by informal thought experiments and discussions, mediated by an instinct for what they think attackers are likely to be capable of, rather than clear standards measuring risk across the dataset as a whole. As a result, they fall short of providing guarantees against re-identification. Some are aware of this limitation and justify it based on their belief that threats are unlikely; others seem uncertain or unaware.

Participants face both technical and structural challenges to effective and efficient de-identification. Technically, some struggle with issues such as not knowing how to optimally balance privacy and utility, or struggling to interpret vague expectations and standards. Structurally, participants feel that de-identification is treated as an afterthought, by research organizations and also by funding agencies, which require data publication but do not always allocate funding for de-identification. While curators can help practitioners improve de-identification through a review process, communication can be a pain point going both ways: some practitioners described situations in which curators pushed for weaker de-identification, while curators expressed frustration with practitioners who have poor or no communication after the initial data submission.

4 Discussion

We have several suggestions to improve de-identification guides. To address inconsistent terminology, we recommend a mix of transparency (e.g., being specific about what the goals of de-identification are) and standardization (e.g., avoiding the use of *aggregation* to describe grouping values into broader buckets). To address gaps in threat coverage, we recommend that reverse engineering be treated as a family of attacks, rather than mentioned in ad hoc examples. To address potential limitations in usability, we recommend that guides include more examples and case studies, especially examples that consider risk across multiple variables at a time.

Based on our interviews with practitioners and curators, we believe that deploying tools to assess re-identification risk and automate de-identification could help researchers de-identify data more strongly, while potentially also saving time. However, a crucial first step is to understand the acceptability of such tools and resources in the eyes of various stakeholders, especially given that increased privacy would likely come with a commensurate sacrifice in data utility, compared with current outcomes.

[1] E.g., unmasking pseudonyms that were assigned non-randomly, downcoding k-anonymous data [4], and brute-forcing improperly hashed values.

[2] E.g., deducing an individual's redacted employer using retained data about their work, or guessing the identity of a pseudonymized city from the demographics of individuals in the dataset who live in that city.

References

[1] Olivia Angiuli, Joe Blitzstein, and Jim Waldo. How to de-identify your data: Balancing statistical accuracy and subject privacy in large social-science data sets. *Queue*, 13(8), 2015-09/2015-10. https://doi.org/10.1145/2838344.2838930.

[2] Bill & Melinda Gates Foundation. Open Access policy. https://openaccess.gatesfoundation.org/, 2024.

[3] Danah Boyd and Jayshree Sarathy. Differential perspectives: Epistemic disconnects surrounding the U.S. Census Bureau's use of differential privacy. *Harvard Data Science Review*, (Special Issue 2), June 2022. https://hdsr.mitpress.mit.edu/pub/3vj5j6i0/release/3.

[4] Aloni Cohen. Attacks on deidentification's defenses. In *Proceedings of the 31st USENIX Security Symposium*, August 2022. https://www.usenix.org/conference/usenixsecurity22/presentation/cohen.

[5] Cynthia Dwork. Differential privacy. In *Proceedings of the 2006 International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science. Springer, 2006. https://doi.org/10.1007/11787006_1.

[6] Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, WPES'18. Association for Computing Machinery, October 2018. https://doi.org/10.1145/3267323.3268949.

[7] Simson L. Garfinkel and Philip Leclerc. Randomness concerns when deploying differential privacy. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society*. Association for Computing Machinery, September 2020. https://doi.org/10.1145/3411497.3420211.

[8] John P. Holdren. Increasing access to the results of federally funded scientific research. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf, February 2013.

[9] Eric S. Lander. Public access congressional report. https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf, November 2021.

[10] Priyanka Nanayakkara and Jessica Hullman. What's driving conflicts around differential privacy for the U.S. Census. *IEEE Security & Privacy*, 21(05), 2023. https://doi.org/10.1109/MSEC.2022.3202793.

[11] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, May 2008. https://doi.org/10.1109/SP.2008.33.

[12] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, and Committee on Toward an Open Science Enterprise. *Open Science by Design: Realizing a Vision for 21st Century Research*. National Academies Press (US), July 2018. https://www.ncbi.nlm.nih.gov/books/NBK525421/.

[13] Alondra Nelson. Ensuring free, immediate, and equitable access to federally funded research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf, August 2022.

[14] Latanya Sweeney. Simple demographics often identify people uniquely. Technical report, Carnegie Mellon University, 2000. https://dataprivacylab.org/projects/identifiability/.

[15] TOP Factor. All journals. https://topfactor.org/journals?factor=Data+Transparency.