FULL-STACK OUTPUT PRIVACY RISK ASSESSMENT[†]

Shlomi Hod[1], Jayshree Sarathy[2]

[1]Institute for Employment Research, Germany [2]Northeastern University, USA

Statistical analyses on sensitive datasets pose privacy risks to individuals in the dataset [3]. In this work, we consider *output privacy* risks, or the risks that released outputs—including synthetic data, published statistics, trained models, and query responses—reveal sensitive information about individuals in the underlying data [5]. Assessing output privacy risks is crucial for informing decision-making and implementing appropriate controls, but such assessments are understudied in the literature and underspecified in frameworks commonly used in practice (e.g. [10]).

A general principle in risk assessment is *risk decomposition*: separating risk into the likelihood of occurrence and the severity of an adverse impact. In cybersecurity, this principle is well-established in frameworks such as NIST SP 800-30 [7], where likelihood is determined, for example, by threat characterization (attack vectors and adversary capabilities). Output privacy deserves equivalent rigor, yet practitioners currently lack the tools to conduct systematic risk assessments. This gap is particularly acute for *differential privacy* (DP) [4] deployments, where numerous design choices—considering whether or not to release statistics about datasets, choosing mechanisms to use for releases, defining privacy units, setting privacy-loss parameters, post-processing the statistics—each affect the privacy-utility tradeoff [1]. Following the spirit of DP as a rigorous framework for privacy-preserving statistical analysis, we believe that these design choices around DP should also be informed by systematic risk assessments, rather than ad-hoc judgment.

To address this gap, we propose a **full-stack framework for output privacy risk assessment**. We conceptualize this as a *stack* of complementary frameworks and tools, progressing from high-level ethical and legal considerations down to empirical, quantified measurement. At the top, *threat modeling* and *impact assessment* provide processes for identifying technically feasible attacks and connecting these attacks to real-world harms. These processes can guide organizations releasing data on how best to engage diverse stakeholders and prioritize risks. In the middle, a shared language provides *taxonomies for attacks and harms*, serving both as a specification for defining these concepts concretely in a given context and as a vocabulary for deliberation and communication among technical and policy stakeholders. At the bottom, an *attack engine* operationalizes these attack specifications, quantifying identified threats and producing empirical evidence that directly informs deployment decisions. Below we describe existing and ongoing work across this stack.

Threat Modeling and Impact Assessment. We start by building a high-level understanding of the gaps between how technical and non-technical stakeholders understand privacy threats. This is essential for creating frameworks to bridge the different perceptions and needs of data curators, data subjects, and data users, and for supporting organizations who must make decisions about privacy controls to protect their data systems

We will first conduct literature review and interviews with privacy decision-makers who are non-experts in DP, to understand perceptions about data privacy and current practices of threat modeling for data privacy. Next, we will develop a framework for real-world threat modeling for data privacy. To validate and refine this framework, we will run threat modeling workshops with data privacy experts to document processes of reasoning by experts and develop insights around the gaps between expert and non-expert reasoning about data privacy threats. Finally, using the insights about threat modeling from non-expert and expert perspectives, we will develop and evaluate a playbook that contains an actionable framework for conducting data privacy threat modeling with robust stakeholder input in specific contexts. This playbook will provide a high-level overview and guidance for organizations to conduct threat modeling and impact assessment. By considering both the perspectives of stakeholders and the expertise of technical privacy professionals, this guidance will effectively address ethical, legal, and technical aspects of threat modeling.

Taxonomies for Attacks and Harms. The middle level of the stack is to develop shared language for describing and analyzing privacy attacks. This is a surprisingly complex task due to the high-dimensional feature space of attacks, including aspects such as the type of attack, attacker's auxiliary information, success metrics, and many more.

[†]This abstract is based on our work with multiple collaborators: Rachel Cummings, Jörg Drechsler, Marcel Neunhoeffer, Marika Swanberg, and Jonathan Ullman.

In recent work, we have developed a taxonomy of attacks [2] by identifying the relevant features and characterizing existing attack papers in terms of this taxonomy. Beyond prior surveys of attacks [8, 9], our taxonomy considers several dimensions of attacks that have been previously understudied, enabling robust and expressive threat modeling and grounding discussions around privacy risks. We show how to apply our taxonomy to describe and reason about privacy threats, using the example of the high-profile DP deployment to release a birth dataset by Israel's Ministry of Health [6]. The taxonomy can be used to do fine-grained threat modeling, reason about privacy risks, and enable system designers to make more informed choices about privacy parameters that take into account stakeholder needs and risk assessments from above.

Attack Engine. At the most granular level of the stack, we propose building a modular, open-source Python package for quantifying output privacy risks of synthetic data. The package will operationalize the attack taxonomy described above and includes a library of standard implementations of state-of-the-art attacks. Users will be able to specify: (1) how the private dataset and attack targets are constructed; (2) attacker goals (e.g. membership inference and attribute inference); and (3) attacker knowledge (e.g. auxiliary information about the data-generation process, full or partial knowledge of the the columns of the data, and access level). The engine will run configured attacks via adapters that translate user specifications into concrete attack instantiations, and then evaluate results against carefully constructed baselines and oracles. It is important to note, however, that the attack engine will only quantify risk with respect to the best known attacks today—it cannot account for future, more powerful attacks. We therefore still advocate for DP with reasonable parameters and for using the attack engine to complement, rather than replace, formal guarantees by grounding risk assessment in empirical, current-state evidence.

References

[1] Rachel Cummings and Jayshree Sarathy. "Centering policy and practice: Research gaps around usable differential privacy". In: *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE. 2023, pp. 122–135.

[2] Rachel Cummings et al. "Attaxonomy: Unpacking differential privacy guarantees against practical adversaries". In: *arXiv preprint arXiv:2405.01716* (2024).

[3] Irit Dinur and Kobbi Nissim. "Revealing information while preserving privacy". In: *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Ed. by Frank Neven, Catriel Beeri, and Tova Milo. PODS '03. 2003, pp. 202–210.

[4] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Proceedings of the Theory of Cryptography Conference*. TCC '06. 2006, pp. 265–284.

[5] Cynthia Dwork et al. "Exposed! A Survey of Attacks on Private Data". In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 61–84.

[6] Shlomi Hod and Ran Canetti. "Differentially Private Release of Israel's National Registry of Live Births". In: *IEEE Symposium on Security and Privacy (S&P)*. 2025.

[7] Joint Task Force Transformation Initiative. *Guide for Conducting Risk Assessments*. Tech. rep. 800-30 Rev. 1. National Institute of Standards and Technology (NIST), Sept. 2012. URL: `https://doi.org/10.6028/NIST.SP.800-30r1`.

[8] Maria Rigaki and Sebastian García. "A Survey of Privacy Attacks in Machine Learning". In: *ACM Computing Surveys* 56.4 (2024), 101:1–101:34.

[9] Ahmed Salem et al. "SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning". In: *Proceedings of the 44th IEEE Symposium on Security and Privacy*. S&P '23. 2023, pp. 327–345.

[10] Kim Wuyts, Laurens Sion, and Wouter Joosen. "Linddun go: A lightweight approach to privacy threat modeling". In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE. 2020, pp. 302–309.