Privacy-preserving data publishing: attacks, countermeasures, and risk analysis

PhD thesis proposal

Supervisors: Tristan Allard (Univ. Rennes, IRISA), Barbara Fila (INSA Rennes, IRISA)

Contact: tristan.allard@irisa.fr and barbara.fila@irisa.fr

Keywords: privacy-preserving data publishing, differential privacy, membership inference attacks, risk analysis, formal methods.

1 Context and goal

Health data, social networks, electricity consumption... Vast quantities of personal data are collected today by private companies or public organizations. Various legal[1], monetary[2], or visibility incentives push data holders to envision sharing *anonymized versions* of the collected datasets or *machine learning models* trained over them. Indeed, sharing data or models at large, *e.g.,* as *open data*, is expected to bring strong benefits (strengthening, *e.g.,* scientific studies, innovation, public policies).

Privacy-preserving data publishing techniques are dedicated to the *sanitization* of personal data before sharing in order to enjoy (hopefully strong enough) privacy guarantees. Substantial progresses have been made during the last two decades about privacy-preserving data publishing techniques, leading to a wealth of privacy models and algorithms, the differential privacy family being most prominent today [2]. While this gives to data holders a large range of choices for sanitizing their datasets, the rich body of work makes it hard for non expert users to understand clearly the privacy implications of choosing a specific privacy model or algorithm, or of selecting suitable values for their privacy parameters (*e.g.,* the values of ϵ and δ for an (ϵ, δ)-differentially private algorithm). The privacy levels usually considered safe often lead to large utility loss prompting data holders to give up on high privacy levels in order to increase utility. While, at the same time, this increases the likeliness of successful attacks on the sanitized dataset/model, **analyzing the risks remains hazardous**, even when the chosen privacy-preserving data publishing techniques are sound. This often contributes to a strong reluctance to sharing data.

In parallel to privacy-preserving data publishing techniques, the study of **attacks on sanitized data or models** has grown tremendously [7]. Attacks on sanitized data and models have been used for auditing real-life implementations of differentially private algorithms [5]. However, although privacy auditing can help obtain empirical estimates of the privacy guarantees offered by an algorithm and its privacy parameters, **using privacy auditing techniques for risk analyzis is non trivial because of the large number of possible attackers and of the large costs of each attack**.

The main goal of this PhD thesis is to use today's attacks on sanitized data or models for allowing a data holder to analyze the risks related to one or more publications. This will require **to formalize the attackers** (*e.g.,* adversarial goals, background knowledge, reasoning methods, impacts and costs of the attacks, vulnerable algorithms), **to structure the space of attackers** (*e.g.,* generalization/specialization of attackers, implications), and **to design algorithms for exploring efficiently the resulting space**.

[1] See for example the EU Directive 2019/1024 on open data and the re-use of public sector information.

[2] For example, through data marketplaces such as Innodata (https://innodata.com/ai-data-marketplace/) or Defined.ai (https://www.defined.ai/).

First, formalizing the attackers and structuring the research space will be needed (*e.g.,* by designing privacy games). Moreover an adversary builds its background knowledge by acquiring data possibly from various sources (*e.g.,* open data, personal data leaked, outputs from other attacks). As a result, it will be necessary to identify, model, and reason about adversarial background knowledge and sequences of steps leading to successful attacks (*e.g.,* by relying on formal models dedicated to representing and analyzing security issues [3]). Finally, given a sanitized dataset, the space of attackers must be explored efficiently, *e.g.,* based on multi-criteria optimization algorithms, in order to find a small set of attackers that illustrates the possible privacy risks.

The sanitization algorithms considered will vary from ad-hoc techniques (*e.g.,* aggregation, pseudonymization, roundings) to elaborate algorithms (*e.g.,* **differentially private synthetic data generation algorithms** [6]). The attacks considered will include in particular **membership inference attacks** [4, 8]. The data models considered will include **tabular data and time series**.

In addition to the core tasks of the project, the successful candidate will also contribute to the **organisation of competitions** where the privacy guarantees of sanitization algorithms are challenged[3] [1].

2 Profile of the candidate

- The candidate must have obtained, or be about to obtain, a master degree in computer science or in a related field.
- The candidate must be curious, autonomous, and rigorous.
- The candidate must be able to communicate in English (oral and written). The knowledge of the French language is not required.
- The candidate must have a strong interest in cybersecurity.
- Skills in machine learning will be appreciated.

3 General information

Laboratory, Team IRISA laboratory (UMR 6074), Spicy team[4].

Supervisors Tristan Allard, Barbara Fila.

Start As soon as the position is filled.

Duration 36 months.

Location Rennes, France.

Funding PEPR Cybersecurity IPoP project[5].

4 Contact and application

For all inquiries please contact Tristan Allard and Barbara Fila (`firstname.lastname@irisa.fr`). Informal inquiries are welcome.

To apply, please send us by email:

- Your detailed CV.

[3] See for example the SNAKE$_1$ challenge: `https://snake-challenge.github.io/`.
[4] `https://www-spicy.irisa.fr/`
[5] `https://files.inria.fr/ipop/`

- A short letter explaining your motivation for working on this project.
- The grade transcript of all university-level courses taken.
- Your master thesis.

References

[1] Tristan Allard, Louis Béziaud, and Sébastien Gambs. Snake challenge: Sanitization algorithms under attack. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, 2023.

[2] Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020(2):288–313, 2020.

[3] Barbara Kordy (Fila), Ludovic Piètre-Cambacédès, and Patrick Schweitzer. Dag-based attack and defense modeling: Don't miss the forest for the attack trees. *Comput. Sci. Rev.*, 13-14:1–38, 2014.

[4] Hongsheng Hu, Zoran A. Salcic, Lichao Sun, Gillian Dobbie, P. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54:1 – 37, 2021.

[5] Fred Lu, Joseph Munoz, Maya Fuchs, Tyler LeBlond, Elliott V. Zaresky-Williams, Edward Raff, Francis Ferraro, and Brian Testa. A general framework for auditing differentially private machine learning. In *Advances in Neural Information Processing Systems (NeurIPS '22)*, 2022.

[6] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data, 2021.

[7] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy (S&P '23)*, pages 327–345, 2023.

[8] Antonin Voyez, Tristan Allard, Gildas Avoine, Pierre Cauchois, Élisa Fromont, and Matthieu Simonin. Membership inference attacks on aggregated time series with linear programming. In *Proceedings of the 19th International Conference on Security and Cryptography (SECRYPT '22)*, 2022.