Use Privacy in Data-Driven Systems

Theory and Experiments with Machine Learnt Programs

Anupam Datta
CMU
danupam@cmu.edu

Matt Fredrikson
CMU
mfredrik@cs.cmu.edu

Gihyuk Ko
CMU
gihyukko@cmu.edu

Piotr Mardziel
CMU
piotrm@gmail.com

Shayak Sen
CMU
shayaks@cmu.edu

This paper presents an approach to formalizing and enforcing *use privacy* in data-driven systems. It addresses a suite of privacy harms that arise from use (rather than knowledge) of personal information in data-driven systems that employ machine learning and other statistical methods. The increasing adoption of these systems in a wide swath of sectors, including advertising, education, healthcare, employment, and credit, underscores the critical need to address these privacy concerns [9, 1].

We start with a set of examples to motivate these threats to privacy and identify the key research challenges that this paper will tackle to address them. In 2012, the department store Target drew flak from privacy advocates and data subjects for using the shopping history of their customers to predict their pregnancy status and market baby items based on that information [5]. While Target intentionally inferred the pregnancy status and used it for marketing, the privacy concern persists even if the inference were not explicitly drawn. Indeed, the use of health condition-related search terms and browsing history—*proxies* (i.e., strong predictors) for health conditions—for targeted advertising have been the basis for legal action and public concern from a privacy standpoint [10, 2, 8].

Use privacy: To address these threats, this paper articulates the problem of protecting *use privacy* in data-driven systems.

Use privacy constraints restrict the use of protected information types and some of their proxies in data-driven systems.

A use privacy constraint may require that health information or its proxies not be used for advertising. Indeed there are calls for this form of privacy constraint [10, 4]. In this paper, we consider the setting where a data-driven system is audited to ensure that it complies with such use privacy constraints. The auditing could be done by a co-operative data processor who is operating the system or by a regulatory oversight organization who has access to the data processors' machine learning models and knowledge of the distribution of the dataset. In other words, we assume that the data processor does not act to evade the detection algorithm, and provides accurate information.

In this setting, it is impossible to guarantee that data processors with strong background knowledge are not able to infer certain facts about individuals (e.g., their pregnancy status) [6]. Even in practice, data processors often have access to detailed profiles of individuals and can infer sensitive information about them [5, 13]. Use privacy instead places a more pragmatic requirement on data-driven systems: that they simulate ignorance of protected information types (e.g., pregnancy status) by not using them or their proxies in their decision-making. This requirement is met if the systems (e.g., machine learning models) do not infer protected information types or their proxies (even if they could) or if such inferences do not affect decisions.

Recognizing that not all instances of proxy use of a protected information type are inappropriate, our theory of use privacy makes use of a normative judgment oracle that makes this inappropriateness determination for a given instance. For example, while using health information or its proxies for credit decisions may be deemed inappropriate, an exception could be made for proxies that are directly relevant to the credit-worthiness of the individual (e.g., her income and expenses).

Proxy use: A key technical contribution of this paper is a formalization of *proxy use* of protected information types in programs. Our formalization relates proxy use to intermediate computations that occur in a program. We begin with a qualitative definition that identifies two essential properties of the intermediate computation (the proxy): *1)* its result perfectly predicts the protected information type in question, and *2)* it has a causal affect on the final output of the program.

In practice, this qualitative definition of proxy use is too rigid for machine learning applications along two dimensions. First, instead of demanding that proxies are perfect predictors, we use a standard measure of association strength from the quantitative information flow security literature to define an ϵ-*proxy* of a protected information type; here $\epsilon \in [0, 1]$ with higher values indicating a stronger proxy. Second, qualitative causal effects are not sufficiently informative for our purpose. Instead we use a recently introduced causal influence measure [3] to quantitatively characterize influence. We call it the δ-*influence* of a proxy where $\delta \in [0, 1]$ with higher values indicating stronger influence. Combining these two notions, we define a notion of (ϵ, δ)-proxy use.

We arrive at this program-based definition after a careful examination of the space of possible definitions. In particular, we prove that it is impossible for a purely semantic notion of decomposition to support a meaningful notion of proxy use as characterized by a set of natural properties or axioms. The

program-based definition arises naturally from this exploration by replacing semantic decomposition with decompositions of the program.

Detection: We present a program analysis technique that detects instances of proxy use in a model, and provides a witness that identifies which parts of the corresponding program exhibit the behavior. Our algorithm assumes access to the text of a program that computes the model, as well as a dataset that has been partitioned into analysis and validation subsets. The algorithm is program-directed and is directly inspired by the definition of proxy use. We prove that the algorithm is complete in a strong sense — it identifies every instance of proxy use in the program. We provide three optimizations that leverage *sampling*, *pre-computation*, and *reachability* to speed up the detection algorithm.

Repair: If a found instance of proxy use is deemed inappropriate, our repair algorithm uses the witness to transform the model into one that provably does not exhibit that instance of proxy use , while avoiding changes that unduly affect classification accuracy. We leverage the witnesses that localize where in the program a violation occurs in order to focus repair there. To repair a violation, we search through expressions local to the violation, replacing the one which has the least impact on the accuracy of the model and at the same time reduces the association or influence of the violation to below the (ϵ, δ) threshold.

Evaluation: We empirically evaluate our proxy use definition, detection and repair algorithms on four real datasets used to train decision trees, linear models, and random forests. Our evaluation demonstrates the typical workflow for practitioners who use our tools for a simulated financial services application. It highlights how they help them uncover more proxy uses than a baseline procedure that simply eliminates features associated with the protected information type. For three other simulated settings on real data sets—contraception advertising, student assistance, and credit advertising—we find interesting proxy uses and discuss how the outputs of our detection tool could aid a normative judgment oracle determine the appropriateness of proxy uses. We evaluate the performance of the detection algorithm and show that, in particular cases, the runtime of our system scales linearly in the size of the model.

Closely related work: The emphasis on restricting use of information by a system rather than the knowledge possessed by agents distinguishes our work from a large body of work in privacy (see Smith [11] for a survey). The privacy literature on use restrictions has typically focused on explicit use of protected information types, and not on proxy use (see Tschantz et al. [12] for a survey).

Differential privacy [7] protects against a different type of privacy harm by restricting explicit use of individuals' contributions to a data set. One way to understand the difference in guarantees is the following: differential privacy ensures that a single row of the database does not affect the output too much; use privacy guarantees that a protected information type (e.g., a single column of the database), or one of its proxies, does not affect the output too much. Indeed, in many settings we may want both guarantees.

Contributions: In summary, we make the following contributions:

- An articulation of the problem of protecting *use privacy* in data-driven systems. Use privacy restricts the use of protected information types and some of their proxies (i.e., strong predictors) in automated decision-making systems.
- A formal definition of *proxy use*—a key building block for use privacy–and an axiomatic basis for this definition.
- An algorithm for detection and tracing of proxy use in a machine learnt program.
- A repair algorithm that provably removes violations of proxy use in a machine learning model
- An implementation and evaluation of our approach on popular machine learning algorithms applied to real datasets.

References

[1] The President's Council of Advisors on Science and Technology. *Big Data and Privacy: A Technological Perspective*. Tech. rep. Executive Office of the President, May 2014.

[2] Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination". In: *Proceedings on Privacy Enhancing Technologies (PoPETs)*. De Gruyter Open, 2015.

[3] Anupam Datta, Shayak Sen, and Yair Zick. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems". In: *Proceedings of IEEE Symposium on Security & Privacy 2016*. 2016.

[4] Wendy Davis. *FTC's Julie Brill Tells Ad Tech Companies To Improve Privacy Protections*. Accessed Nov 11, 2016. 2016. URL: http://www.mediapost.com/publications/article/259210/ftcs-julie-brill-tells-ad-tech-companies-to-impro.html.

[5] Charles Duhigg. *How Companies Learn Your Secrets*. Accessed Aug 13, 2016. 2012. URL: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[6] Cynthia Dwork. "Differential Privacy". In: *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*. Ed. by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. Vol. 4052. Lecture Notes in Computer Science. Springer, 2006, pp. 1–12. ISBN: 3-540-35907-9. URL: http://dx.doi.org/10.1007/11787006_1.

[7] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography Conference*. Springer, 2006, pp. 265–284.

[8] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. “Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence”. In: *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*. CCS ’15. Denver, Colorado, USA: ACM, 2015, pp. 554–566. ISBN: 978-1-4503-3832-5.

[9] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA, USA: Harvard University Press, Jan. 2015. URL: http://www.hup.harvard.edu/catalog.php?isbn=9780674368279.

[10] Findings under the Personal Information Protection and Electronic Documents Act (PIPEDA). *Use of sensitive health information for targeting of Google ads raises privacy concerns*. Accessed Aug 13, 2016. 2014. URL: https://www.priv.gc.ca/cf-dc/2014/2014_001_0114_e.asp.

[11] Geoffrey Smith. “Recent Developments in Quantitative Information Flow (Invited Tutorial)”. In: *Proceedings of the 2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*. LICS ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 23–31.

[12] Michael Carl Tschantz, Anupam Datta, and Jeannette M. Wing. “Formalizing and Enforcing Purpose Restrictions in Privacy Policies”. In: *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. (SP ’12). Washington, DC, USA, 2012, pp. 176–190.

[13] Joseph Turow. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. Yale University Press, 2011. ISBN: 9780300165012.