Inferential Privacy Guarantees for Differentially Private Mechanisms

Arpita Ghosh[1] **and Robert Kleinberg**[2]

1 **Cornell University, Ithaca, NY, USA**
arpitaghosh@cornell.edu
2 **Cornell University, Ithaca, NY, USA**
robert.kleinberg@cornell.edu

Abstract

The following is a summary of the paper "Inferential Privacy Guarantees for Differentially Private Mechanisms", presented at the 8^{th} *Innovations in Theoretical Computer Science* conference in January 2017. The full version of the paper can be found on arXiv at the following URL: https://arxiv.org/abs/1603.01508.

1998 ACM Subject Classification K.4.1 Public Policy Issues – Privacy

Keywords and phrases differential privacy, statistical inference, statistical mechanics

Digital Object Identifier 10.4230/LIPIcs.ITCS.2017.9

1 Summary of the paper

Differential privacy [3, 4] has become the dominant theoretical framework for quantifying privacy loss, and has begun to make its way into policy and legal frameworks as a potential means for such a quantification. Running a differentially private mechanism on a dataset produces an outcome whose distribution is insensitive to removing one individual's data from the dataset or modifying her data. Consequently, differential privacy guards an individual against the possibility that observers of the mechanism's outcome will make strong inferences about her participation or non-participation (or, contingent on participation, inferences about her data).

Even without an individual's participation in a dataset, probabilistic inferences about her private data may be possible due to its correlation with the data of other individuals present in the dataset. In some cases, but not all, it may be desirable to protect individuals from such inferences. This paper aims to quantify the extent to which differentially private mechanisms guarantee such "inferential privacy protection", as a function of the prior belief (or set of potential priors) held by observers prior to the release of the mechanism's outcome.

We can illustrate the issues at play here using the oft-cited example of a study showing that smoking causes cancer [5]. If the data underlying the study were analyzed in a differentially private manner, readers of the study should not significantly update their beliefs about a given individual's participation in the dataset. However, if the individual were a known smoker, they should significantly revise their beliefs about her probability of developing cancer. Such belief revisions, in this circumstance, are generally not construed as a violation of privacy, because they merely reflect improved knowledge of an aggregate property of the population, not any knowledge specific to the individual. On the other hand, suppose that the hypothetical study focused on the more fine-grained question of how the likelihood of developing smoking-related cancers varies as a function of factors such as age, diet, lifestyle, and family history, and that the supplementary material accompanying the article included a

machine learning model (e.g., a random forest regressor) trained to predict these likelihoods. Even if the input-output behavior of the regressor were not construed as violating privacy, the regressor itself could be a highly complex object (e.g., a collection of more than a thousand decision trees) and its contents might reflect private information in the training data. If the training algorithm were differentially private, would it ensure that participants in the study were protected from "inferential privacy violations"? This paper abstracts away the specifics of the example and asks the basic question: when does a differential privacy guarantee for an algorithm imply an inferential privacy guarantee?

In the paper, we formally define the inferential privacy parameter of a mechanism with respect to a set of prior distributions. Informally, it measures the maximum amount by which an adversary might multiplicatively update his belief in the likelihood that a particular individual's entry in the database assumes a particular value, given that the adversary starts with one of the specified prior distributions and performs a Bayesian update on the mechanism's outcome. An easy application of Bayes' Law confirms the fact (known to many prior authors, e.g. [7]) that when individuals' private data are mutually independent, the differential privacy parameter of a mechanism equals its inferential privacy parameter. On the other hand, we have seen that when the adversary's prior incorporates uncertainty about certain aggregate statistics of the population (e.g., the frequency of cancer among smokers) the inferential privacy parameter of a mechanism may be much greater than its differential privacy parameter. Interpolating between these two extremes, one might guess that differentially private mechanisms are guaranteed to have a relatively small inferential privacy parameter when correlations between individuals are either weak or localized, i.e. when each individual has only a few others with whom her data correlates strongly.

Our work makes this intuition precise and confirms it rigorously. To do so, we identify a surprisingly tight relationship between our questions about privacy and inference and a corresponding set of questions in mathematical physics regarding the magnitude of the change in a Gibbs measure when an external field is applied to the system. While it may initially seem surprising that the two fields are connected in this way, one can see the first intimations of the connection in the discussion about weak, localized correlations at the end of the preceding paragraph—this is none other than the *correlation decay* property that is the hallmark of statistical mechanical systems at high temperature.

Our first main theorem pertains to cases in which the data are binary-valued and the adversary's prior distribution satisfies *positive affiliation*, meaning that conditioning on all but two entries in the database, the remaining two entries can never be negatively correlated. This assumption is satisfied by the priors commonly ascribed to biological data—where heredity and contagion lead to positive correlations among individuals' attributes, but never or rarely lead to negative correlations—and to social data, where positive correlations result from homophily and social contagion. (If one interprets the adversary's prior as a Gibbs measure, the positive affiliation property says that the system is *ferromagnetic*.) Our theorem gives a precise formula for the worst-case inferential privacy parameter of ϵ-differentially private mechanisms in terms of the magnetization of the corresponding ferromagnetic system in an external field of strength $\frac{\epsilon}{2}$. The proof of the theorem shows that the mechanism attaining this worst-case bound is not a contrived mechanism; in fact it is one of the most commonly occurring differentially private mechanisms: adding Laplace noise to the sum of the values in the database! Thus, when data are positively affiliated, any inferential privacy guarantee that one can prove for the Laplace mechanism *automatically* carries over to arbitrary differentially private mechanisms.

Because our theorem provides an exact formula for worst-case inferential privacy in terms of magnetization (and not merely an upper bound) it allows us to translate results about the physics of magnets directly into results about inferential privacy. In particular, existing results about phase transitions in Ising models (e.g., for the infinite d-regular tree [1]) imply that inferential privacy parameters can be surprisingly sensitive to variations in other parameters of the model. For example, if we vary the differential privacy parameter of the Laplace mechanism, starting at $\epsilon = 0$ and increasing from there, the mechanism's inferential privacy parameter increases gradually until ϵ crosses a critical value, at which point it can increase very precipitously, approaching a step discontinuity as the number of individuals tends to infinity. In other words, tiny variations in the differential privacy parameter of a mechanism can potentially lead to enormous variations in inferential privacy, a privacy-theoretic manifestation of the physical phenomenon of phase transitions in spin systems.

Our second main result applies when the data are not binary-valued, or when the prior violates positive affiliation. It shows that the inferential privacy parameter of a mechanism is bounded above by a function of the mechanism's differential privacy parameter and the spectral norm of an *influence matrix* encoding the strength of pairwise correlations between individuals. The theorem again manifests the relationship between inferential privacy and statistical mechanics; its statement and proof constitute an adaptation of the Dobrushin Comparison Theorem [2, 6, 8] from the traditional setting of additive approximation to multiplicative approximation.

Acknowledgements. The authors gratefully acknowledge helpful discussions with Christian Borgs, danah boyd, Jennifer Chayes, Cynthia Dwork, Kobbi Nissim, Adam Smith, Omer Tamuz, and Salil Vadhan.

Much of this research was completed while Robert Kleinberg was a researcher at Microsoft Research New England. Both authors were partially supported by NSF award AF-1512964. Arpita Ghosh was partially supported by NSF award III-1513692.

References

1 Rodney J Baxter. *Exactly solved models in statistical mechanics*. Courier Corporation, 2007.

2 Roland L Dobrushin. Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486, 1970.

3 Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming (ICALP)*, 2006.

4 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, 2006.

5 Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2013.

6 H. Föllmer. A covariance estimate for Gibbs measures. *Journal of Functional Analysis*, 46:387–395, 1982.

7 Shiva P Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1, 2014.

8 H Künsch. Decay of correlations under Dobrushin's uniqueness condition and its applications. *Communications in Mathematical Physics*, 84(2):207–222, 1982.