Cynthia Dwork – Evidence-Based Policymaking

Differential Privacy and its Properties

Differential Privacy is a definition of privacy, and a collection of supporting algorithmic techniques, tailored for privacy-preserving statistical analysis of large datasets.

Differential privacy is a mathematical guarantee that an individual data contributor will not be affected, adversely or otherwise, by allowing her data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are -- or will become – available. At their best, differentially private algorithms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views. Nonetheless, data utility will eventually be consumed: the Fundamental Law of Information Recovery states that overly accurate estimates of too many statistics can completely destroy privacy (Dinur and Nissim, 2003; Dwork et al, 2007; Homer et al, 2008, Dwork et al., 2015b). The Fundamental Law can no more be circumvented than can the laws of physics.

Every useful computation results in some loss of privacy. Differential privacy measures and controls privacy loss accumulating over multiple analyses. This signal capability makes it possible to “program” in a differentially private fashion. In ordinary, non-private computation, anything computable can be computed using only addition and multiplication, but this is not how programmers work. Algorithm design is the creative combining of appropriate computational primitives to carry out a sophisticated computational task, while minimizing the consumption of key resources, such as time and space. Similarly, differentially private algorithm design is the creative combining of simple differentially private primitives to perform a sophisticated analytical task, while also minimizing privacy loss and inaccuracy. As a rule, when the dataset is large the signal dominates the noise injected for privacy; when the dataset is small this is not the case. This is correct; think of the case of a dataset of size one: to ensure privacy the noise *must* dominate the signal. Designed to preserve the privacy of everybody – even the needles in the haystack – the goal is to elicit participation, without fear of repercussion, for a public good, such as learning that smoking causes cancer, and other facts of life. Indeed, it is often the outliers who most need protection.

Differential privacy also provably controls privacy loss accruing over computations on multiple, possibly overlapping, datasets, making it especially relevant to the kinds of analyses that will be needed for evidence-based policy making.

The Fundamental Law tells us that *meaningful* privacy guarantees come at a price. Other disciplines, such as ethics and economics, cannot be brought to bear without a measure of privacy loss. Differential privacy provides such a measure (Abowd and Schmutte, 2015).

Finally, differential privacy strengthens the scientific method in an unexpected way, even when privacy is not a concern. The rise of "Big Data" has been accompanied by increased risk of spurious scientific discovery. A great deal of effort has been devoted to reducing this risk, from the use of sophisticated validation techniques, to deep statistical methods for controlling the false discovery rate in multiple hypothesis testing. However, there is a fundamental disconnect between the theoretical results and the practice of data analysis: the theory of statistical inference assumes a fixed collection of hypotheses to be tested, selected *before* the data are gathered, whereas in practice data are shared and reused, with hypotheses and new analyses being generated on the basis of data exploration and the results of previous studies on the same dataset. This leads to overfitting, that is, learning about the dataset rather than about the population from which it is drawn. Differential privacy automatically protects against this source of false discovery (Dwork et al., 2015a).

Key Considerations

Differential privacy holds great promise but requires great effort. The Fundamental Law forces economic considerations in how data should be used, increasing the imperative for high quality differentially private algorithms, but the field is young and many of these will be the content of doctoral dissertations not yet written. The literature is silent on crucial preprocessing steps, such as imputation of missing fields and other aspects of data cleaning. Working with formal privacy guarantees requires a new skill set, foreign to most statistical agencies, social science researchers, and data scientists. Recent adoption of the approach by Google and Apple will draw talent away from the public and research sectors.

But what is the alternative? The distinction between Personally Identifiable Information (PII) and non-PII is not mathematically meaningful. In the words of the President's Council of Advisors on Science and Technology, "Anonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially" (PCAST, 2014). Traditional statistical disclosure limitation methods ruling out subtraction attacks do not defend against other attacks (Fellegi, 1972, and subsequent generalizations).

It is a consequence of the Fundamental Law that all privacy-preserving computations must introduce some error. Current statistical disclosure limitation (SDL) techniques also introduce errors. For example, one paper states, "These algorithms [with formal privacy guarantees] are currently being implemented on data gathered at a national statistical agency on which we empirically evaluate the utility of our algorithms. We show that for reasonable values of the privacy loss parameter … the error introduced by our provably private algorithms is comparable or better than the error introduced by existing SDL techniques" (Haney et al., 2015). Similarly, a study on privacy in massive open online courses (MOOCs) data from MITx and HarvardX on the edX platform reported that standard anonymization methods force changes to datasets that "threaten replication and extension of baseline analyses" (Daries et al., 2014). When the errors are introduced in a principled way, as in differential privacy, the analyst can better interpret the results.

Even synthetic data (Rubin, 1993) can be problematic. Once constructed, synthetic data may be queried ad libitum, with no risk of further privacy loss, using the analyst's choice of techniques. However, privacy is *not* an automatic consequence of the data being synthetic, but depends crucially on the process by which the synthetic data are constructed. (The Census Bureau uses synthetic data generated with a variant of differential privacy in OnTheMap, a website providing information on where people work and where workers live (Machanavajjhala et al., 2008).)

Recommendations

I close with three policy recommendations. First, *Publish Your Epsilons*. Differentially private algorithms are equipped with a privacy parameter, usually called epsilon, capping their privacy loss. In a non-private algorithm epsilon is infinite. But what is the “meaning” of a given value of epsilon? By maintaining a registry of privacy loss, akin to a toxic release registry, we can observe the accuracy/privacy tradeoffs actually made and stimulate competition to obtain better analyses at lower privacy costs, engaging those who traffic in the data of individuals in the effort to protect their privacy.

Second, *Establish a list of approved private data analysis techniques and appropriate applications, and keep it current*.

Third, *Consider Restraint*. In a data-rich world, the challenges revolve around the trade-off between what can be done and acceptance of the fundamental truth that overly accurate estimates of too many statistics can destroy privacy. If we are interested in privacy, sometimes restraint might be the right approach.

References

J. M. Abowd and I. Schmutte, Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods, http://digitalcommons.ilr.cornell.edu/ldi/22/, 2015

J.P. Daries, J. Reich, J. Waldo, E.M. Young, J. Whittinghill, A.D. Ho, A. D., D.T,. Seaton, and I. Chuang. Privacy, anonymity, and big data in the social sciences. Communications of the ACM, 57(9), 56-63, 2014.

I. Dinur and K. Nissim, Revealing information while preserving privacy. In Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pages 202-210, 2003

C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. Science 349(6248), 2015a.

C. Dwork, F. McSherry, and K. Talwar, The price of privacy and the limits of lp decoding. In Proceedings of the 39th ACM Symposium on Theory of Computing, pages pp. 85-94, 2007

C. Dwork, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. Robust traceability from trace amounts. In Proc. 56th IEEE Annual Symposium Foundations of Computer Science (FOCS), 2015b

U. Erlingsson, V. Pihur, and A. Korolova, “RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Proc. ACM SIGSAC Conference on Computer and Communications Security (CCS), 2014.

I. P. Fellegi, On the Question of Statistical Confidentiality. *Journal of the American Statistical Association* 67, no. 337, 1972

S. Haney, A. Machanavajjhala, M. Kutzbach, M. Graham, J. Abowd, L. Vilhuber. Formal privacy protection for data products combining individual and employer frames. Presented at UNECE/Eurostat Statistical Data Confidentiality Work Session, 2015

N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. Pearson, D. Stephan, S. Nelson, and D. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. PLoS Genet, 4, 2008

A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In Proceedings of the IEEE 24th International Conference on Data Engineering, 2008

President's Council of Advisors on Science and Technology (PCAST), Big Data and Privacy: A Technological Perspective, May 2014.

D. B. Rubin, Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), pp.461-468, 1993

A.D. Sarwate, S.M. Plis, J. Turner, and V.D. Calhoun. Sharing privacy-sensitive access to neuroimaging and genetics data: a review and preliminary validation. Frontiers in Neuroinformatics, 2014

S. Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309), 1965