Workshop Proposal:
Statistical and Learning-Theoretic Challenges in Data Privacy

Organizing Committee:

Cynthia Dwork, Microsoft Research
Stephen Fienberg, CMU Statistics
Aleksandra Slavkovic, Penn State Statistics
Adam Smith, Penn State Computer Science & Engineering

June 19, 2009

This is a proposal for a week-long workshop to be held in early 2010 at IPAM.

1 Introduction

Privacy is a fundamental problem in modern data analysis. Collections of personal and sensitive data, previously the purview of governments and statistical agencies, have become ubiquitous. Increasing volumes of personal and sensitive data are collected and archived by health networks, government agencies, search engines, social networking websites, and other organizations. The potential social benefits of analyzing these databases are significant: better informed policy decisions, more efficient markets, and more accurate public health data, just to name a few. At the same time, releasing information from repositories of sensitive data can cause devastating damage to the privacy of individuals or organizations whose information is stored there. The challenge is to discover and release global characteristics of these databases, without compromising the privacy of the individuals whose data they contain.

A number of high-profile breaches of privacy due to the release of supposedly anonymized information [31, 4, 27] have raised awareness among data holders of the importance of careful consideration of privacy and, in particular, of the fact that anonymization by removing all obviously identifying information is not sufficient for ensuring confidentiality. One study [31] identified the governor of Massachusetts in a medical database anonymized in this simple-minded manner. A New York Times article [4] exposed a searcher in supposedly anonymized AOL search data where each username was replaced by a random ID. Finally, users of Netflix were identified from similarly anonymized movie-preference data [27], by combining the anonymized data with reviews posted by users on a popular film review website. In all three cases, there was *no* break into the database by hackers: the organization holding the database violated privacy by incorrectly assuming that it was publishing only safe information.

The problem of identifying which information in the database is safe to release has generated a vast body of work, both in statistics and computer science. Until recently, there were two nearly disjoint fields studying the data privacy problem: "statistical disclosure limitation" (also known as "data confidentiality"), initiated by the statistics community in 1960s, and "privacy-preserving data mining", active in the database community during the 1980's and rekindled at the turn of the 21st century by researchers in data mining.

The literature in both fields is too vast to survey here. For some pointers to the broader literature in statistics, see [34, 10, 9, 35, 15, 28, 30, 19, 13, 14, 21, 20, 33, 29]. For early work in computer science, see the survey in [1]. Recent work in data mining was started by [2] and led to an explosion of literature. For (partial) references, see [8, 23, 32].

2 Scientific Challenges

The proposed workshop will bring together researchers from a variety of areas in statistics, machine learning, cryptography and data mining. The focus of the workshop will be establishing a coherent theoretical foundation for research on data privacy. This implies work on (1) how the conflicting goals of privacy and utility can or should be formulated mathematically; and (2) how the constraints of privacy—in their various incarnations—affect the accuracy of statistical inference and machine learning.

(1) *Definitions and Modeling.* Questions about data privacy are inevitably tied to distinguishing global statistical properties of a data set from individual information. Numerous attempts have been made to quantify this divide, with varying levels of rigor and precision. One of the significant difficulties is handling *side information* adequately. Roughly, side information is anything available to an entity interested in breaking privacy beyond what is published by the agency running the sanitization algorithm. Examples includes partial information about specific individuals (such as movie reviews posted on the web [27]), and anonymized versions of related data sets (as in "composition attacks" [22]).

One goal of the workshop is to discuss rigorous notions of privacy that provide meaningful security in the presence of unknown side information. *Differential privacy*, which emerged from a line of work in theoretical computer science [12, 18, 7, 16, 17], provides an example of such a notion; it makes assumptions neither about what kind of attack might be perpetrated based on the released statistics, nor about what additional information the attacker might possess. Part of the workshop will consider alternate formulations of privacy and resistance to active attacks (in which an adversary actively manipulates the contents of the database, as in [3]).

A complementary, implied goal is to discuss clear and mathematically meaningful formulations of *utility*, that is, inference and learning goals most appropriate for sensitive data.

(2) *Methodology and impossibility results.* Different models of privacy imply different constraints on statistical analysis. This raises the question: *How do confidentiality constraints affect the accuracy of statistical inference and machine learning?*

Little is known about the implications of rigorous approaches to privacy for statistical validity and learning.[1] Although much applied research has been done in the statistics community, the techniques for disclosure protection that were studied lack rigorous analysis. In contrast, computer scientists have considered more precise definitions, but typically approach *utility* in terms of functional approximation: given a database x and a function f, how well does the released information allow one to approximate $f(x)$?

[1] At a high level, statistical inference and machine learning refer to the same task, namely, producing a concise model of a given data set, but the terms are associated with different techniques and communities; we use both here since both communities deal with sensitive data, albeit from different domains.

The thesis underlying this proposal is that *the relation between privacy and statistical validity is connected to fundamental statistical, learning-theoretic and algorithmic problems, and that bringing researchers from these diverse communities together will lead to new interactions between, and progress in, their respective scientific fields.*

For example, differential privacy is based explicitly on the principle that global properties are exactly those that do not depend heavily on any particular individual in the population. In retrospect, previous approaches appear to be driven by a similar idea. This suggests a strong connection to ideas such as "robustness" of statistical procedures, noise-tolerance, and algorithmic stability. However, that connection has only begun to be explored because of the small number of researchers that understand all the relevant fields.

Specific types of questions Many technical questions arise from the tension between confidentiality and statistical inference. Here are a few that inspired the current proposal.

- *Consistency and convergence.* Supposing that the data set is generated according to a particular probabilistic model, under what conditions do privacy-preserving inference procedures converge to the underlying model as the amount of data increases? At what rate? When is the convergence of private procedures similar to, or (more interestingly) provably slower than, that of their optimal non-private analogues? In the language of machine learning, can generalization bounds for private procedures match their non-private analogues? Are their tasks, such as outlier detection, that are inherently at odds with privacy?

- *Relationships with robustness, noise-tolerance and stability.* Information leakage seems closely related to how sensitive a given procedure is to small changes in the data. Intuitively similar conditions have been studied extensively in several areas of statistics (e.g. the field of robust statistics [26]) and machine learning (e.g. work on noise-tolerance [24] and algorithmic stability [11, 25, 6, 5]).

- *The "private" curse of dimensionality.* High-dimensional, many-faceted models are notoriously difficult to handle computationally. They also raise a particular challenge for private data analysis: the more complex the fitted model, it seems, the higher the chance that it reveals information specific to particular individuals. How, then, do complexity and privacy interact? Intuition suggests that the point where models start showing features of individual data points is exactly the point at which inference becomes difficult, but there currently is no formal substantiation of that intuition.

- *Differently structured data: graphs, text, genomic data, trace data.* Data privacy has become more important recently precisely because collected data has become vastly more varied and valuable. How should one reason about privacy in settings where the lines between individuals' data are not cleanly drawn, as with, say, data from a social network or genomic sequences ? What about settings in which individuals contribute many transactions over time, as with a search engine, credit card database, or intrusion detection system?

- *Generation of synthetic data.* In many cases, users of "anonymized" data would be more comfortable with real-seeming synthetic data that shares the important properties of the original data, as opposed to a simply listing of those properties or a server that provides them in response to specific queries. To what extent is synthetic data generation possible, and how should one reason about the information leaked by synthetic data sets?

3 Invitees and Audience

The topics above have begun to be investigated in a number of different communities, and there have been some limited opportunities for interaction through joint workshops. However, so far very little of the interaction has been focused on theoretical foundations, concentrating instead on specific applications such as health statistics (e.g. a recent workshop at the National Center for Health Statistics).

The workshop would aim to gather together experts from relevant theoretical fields with a mix of tutorial presentations, aimed at establishing a common language for the workshop, and research talks.

[...]

References

[1] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 25(4), 1989.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *SIGMOD Conference*, pages 439–450. ACM, 2000.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography. In *Proc. 16th Intl. World Wide Web Conference*, 2007.

[4] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749. *The New York Times*, August 2006.

[5] Shai Ben-David, Dávid Pál, and Hans-Ulrich Simon. Stability of k -means clustering. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 20–34. Springer, 2007.

[6] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In Gábor Lugosi and Hans-Ulrich Simon, editors, *COLT*, volume 4005 of *Lecture Notes in Computer Science*, pages 5–19. Springer, 2006.

[7] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *PODS*, 2005.

[8] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y. Zhu. Tools for privacy preserving data mining. *SIGKDD Explorations*, 4(2):28–34, 2002.

[9] T. Dalenius and S.P. Reiss. Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, (6):73–85, 1982.

[10] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, (5):35–64, 1977.

[11] L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Tran. Info. Theory*, 25(5):601–604, 1979.

[12] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210, 2003.

[13] J. Domingo-Ferrer and V. Torra, editors. *Privacy in Statistical Databases (PSD)*, volume 3050 of *Lecture Notes in Computer Science*. Springer, 2004.

[14] Josep Domingo-Ferrer and Luisa Franconi, editors. *Privacy in Statistical Databases, CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006, Proceedings*, volume 4302 of *Lecture Notes in Computer Science*. Springer, 2006.

[15] P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors. *Confidentiality, Disclosure and Data Access. Theory and Applications for Statistical Agencies.* Elsevier, New York, 2001.

[16] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *ICALP*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[17] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[18] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *CRYPTO*, pages 528–544, 2004.

[19] S. E. Fienberg and A. B. Slavkovic. Making the release of confidential data from multi-way tables count. *Chance*, 17(3), 2004.

[20] S.E. Fienberg and A.B. Slavkovic. A survey of statistical approaches to preserving confidentiality of contingency table entries. In Charu C. Aggarwal and Philip S. Yu, editors, *Privacy-preserving Data Mining: Models and Algorithms*. Kluwer Academic, 2008.

[21] Jonathan J. Forster and Emily L. Webb. Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(5):551–570, 2007.

[22] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. Arxiv Report 0803.0032 `http://arxiv.org/abs/0803.0032`, 2008.

[23] Johannes Gehrke. Models and methods for privacy-preserving data publishing and analysis (tutorial slides). In *Twelfth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD 2006)*, 2006.

[24] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

[25] Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In Adnan Darwiche and Nir Friedman, editors, *UAI*, pages 275–282. Morgan Kaufmann, 2002.

[26] Ricardo Maronna, Doug Martin, and Victor Yohai. *Robust Statistics: Theory and Methods*. Wiley, 2006.

[27] Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. ArXiV report cs.CR/0610105, `http://arxiv.org`, November 2006.

[28] T.E. Raghunathan, J.P Reiter, and D.B. Rubin. Multliple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19:1–16, 2003.

[29] Chris Skinner and Natalie Shlomo. Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association*, 103(483):989–1001, 2008.

[30] A.B. Slavkovic. *Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables.* Ph.D. Thesis, Department of Statistics, Carnegie Mellon University, 2004.

[31] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[32] Latanya Sweeney. Privacy-enhanced linking. *SIGKDD Explorations*, 7(2):72–75, 2005.

[33] Daniel Ting, Stephen E. Fienberg, and Mario Trottini. Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security*, (2):86–105, 2008.

[34] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[35] L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Lecture Notes in Statistics,155. Springer, New York, 2000.