CS229r: Mathematical Approaches to Data Privacy
Additional Topics and Project Ideas

March 13, 2013

1 Additional Topics

In this section, we list a variety of additional topics on differential privacy beyond what we have covered in the course so far, along with pointers to a few relevant papers in the literature. (The bibliography is far from comprehensive; you can find more by searching for references to and from the listed papers.) We will cover some of these topics in the second half of the course, depending in part on your interests (we will be polling you for your preferences).

This topic list may also help you in formulating a final project. On some of the topics, we have listed some concrete questions that might be interesting for you to investigate. However, you need not limit yourself to the topics or questions listed here—be creative and find something that really interests you.

Multiparty differential privacy and computational differential privacy. In differential privacy, we typically assume that there is a curator who holds the entire database. An interesting question is when we can eliminate the curator and have the users run a distributed protocol that guarantees differential privacy even when the other parties are untrusted.

In this setting, it is often useful to restrict to computationally bounded adversaries (unlike the standard, information-theoretic notion of differential privacy), and in this case cryptography can enable the parties to emulate the role of a trusted curator through a secure distributed protocol:

- Our Data, Ourselves. Dwork, Kenthapadi, McSherry, Mironov, Naor, 2006.
- Computational Differential Privacy. Mironov, Pandey, Reingold, Vadhan, 2009.

If we require information-theoretic differential privacy, then there are stronger limitations on the utility-privacy tradeoffs that can be achieved than in the standard single-curator model. Understanding these tradeoffs turns out to relate to communication complexity and randomness extractors.

- Distributed private data analysis: Simultaneously solving how and what. Beimel, Nissim, Omri 2008.
- The Limits of Two-Party Differential Privacy. McGregor, Mironov, Pitassi, Reingold, Talwar, Vadhan, 2010.

- Optimal Lower Bound for Differentially Private Multi-Party Aggregation. Chan, Shi, Song, 2012.

One open problem from the paper by McGregor et al. is to exhibit a function with global sensitivity 1 for which any two-party differentially private protocol (for (ε, δ)-dp) must incur error $\Omega(n)$. (The best known lower bound is about $\sqrt{n}$.)

It is also interesting to ask whether assuming computationally bounded adversaries can help for achieving more efficient sanitizers. For many kinds of functions, it is known that switching to computational privacy cannot help much.

- Limits of Computational Differential Privacy in the Client/Server Setting. Groce, Katz, Yerukhimovich, 2011.

Potential project topic: either find more general impossibility results (e.g. that apply to discrete and exponentially large output spaces), or to find a natural task for which computational differential privacy allows for better curators (either in terms of utility or efficiency).

Incorporating Differential Privacy in Learning Theory. Computational learning theory is a great source of interesting problems that would be nice to solve in a differentially private way, and there have been several works exploring whether there are efficiency gaps (statistical or computational) between private and non-private learning.

- Practical Privacy: The SuLQ Framework. Blum, Dwork, McSherry, Nissim, 2005.
- What Can We Learn Privately? Kasiviswanathan, Lee, Nissim, Raskhodnikova, Smith, 2007.
- Sample Complexity Bounds for Differentially Private Learning. Chaudhuri, Hsu, 2011.
- Characterizing the sample complexity of private learners. Beimel, Nissim, Stemmer, 2013.

Potential project topic: explore some of the remaining gaps between private learning and non-private learning (e.g. in sample complexity or computational complexity).

Learning vs. Private Data Release Many of the algorithms for answering many queries with differential privacy, such as SmallDB and Private Multiplicative Weights, are inspired by techniques in learning theory. This has led researchers to look for a general connection between learning and private data release, and indeed it is known that certain kinds of efficient learning algorithms imply efficient algorithms for private data release.

- Privately Releasing Conjunctions and the Statistical Query Barrier. Gupta, Hardt, Roth, Ullman, 2011.
- Private Data Release via Learning Thresholds. Hardt, Rothblum, Servedio, 2012.

However, these works leave open the possibility that private data release is easier than learning. It would be very interesting to show that private data release (for a simple family of queries, such as multivariate marginals/conjunctions) is at least as hard as some previously studied open problem in learning theory.

Privacy for Streaming Algorithms. The algorithms we've studied so far run once and terminate. On the other hand, there are some algorithms, e.g. streaming algorithms, that are designed to run over an extended window of time and produce output continually. These algorithms raise questions such as the privacy of the internal state of the algorithm and the accumulation of privacy loss over a long window.

- Pan-Private Streaming Algorithms. Dwork, Naor, Pitassi, Rothblum, Yekhanin, 2010.
- Privacy Under Continual Observation. Dwork, Naor, Pitassi, Rothblum, 2010.
- Pan-Private Algorithms via Statistics on Sketches. Mir, Muthukrishnan, Nikolov, Wright, 2011.

There are still open problems about the optimality of the parameters achieved by some of these algorithms.

Privacy for Graph Analysis. Many datasets do not naturally fit the model of individual records that we have used in class. One such example is when the data is represented by a graph, e.g. a social network. So far there have been some interesting attempts to look at natural data analysis tasks on graphs but there are a lot of open problems.

- Iterative Constructions and Private Data Analysis. Gupta, Roth, Ullman, 2011.
- Private Analysis of Graph Structure. Karwa, Yaskhodnikova, Smith, Yaroslavtsev, 2011.
- Graph Analysis with Node-Level Privacy. Kasiviswanathan, Nissim, Raskhodnikova, Smith, 2013.
- Differentially Private Analysis of Social Networks via Restricted Sensitivity. Blocki, Blum, Datta, Sheffet, 2013.
- Beyond Worst-Case Analysis in Private Singular Vector Computation. Hardt, Roth, 2013.

Potential Project Topics: 1) Differentially private algorithms for other kinds of graph or social network analyses, beyond those that have been studied so far. 2) Are there efficient algorithms for releasing a private synthetic graph approximately preserving the size of every cut? What about on natural families of graphs? What about the related problem of detecting communities in a social network? 2) Iterative mechanisms for answering cut queries (e.g. Private Multiplicative Weights) run in time $O(n^2)$ per query, even when the number of edges is $m \ll n^2$. Can the running time be improved to $O(m)$ with similar utility guarantees? 3) Can we release the answers to all cut queries or do spectral analysis with *node-level* (rather than edge-level) privacy?

Differential Privacy and Mechanism Design. Differential privacy has many interesting connections to mechanism design. On one hand, differentially private algorithms lead to a certain kind of truthfulness that can be difficult to achieve otherwise. On the other hand, interesting new questions in mechanism design arise when we incorporate privacy concerns into agents' utility functions. This problem has been looked at from many angles and there are many interesting remaining technical and modeling questions.

- Selling Privacy at Auction. Ghosh, Roth, 2011.
- Take it or leave it: running a survey when privacy comes at a cost. Ligett, Roth, 2012.
- Approximately Optimal Mechanism Design via Differential Privacy. Nissim, Smorodinsky, Tennenholtz, 2012.
- Is privacy compatible with truthfulness? Xiao, 2013.
- Privacy-Aware Mechanism Design. Nissim, Orlandi, Smorodinsky, 2012.
- Truthful Mechanisms for Agents that Value Privacy. Chen, Chong, Kash, Moran, Vadhan, 2011.
- Mechanism Design in Large Games—Incentives and Privacy. Kearns, Pai, Roth, Ullman, 2012.
- A Theory of Pricing Private Data. Li, Li, Miklau, Suciu, 2013.

Alternative Privacy Definitions. Differential privacy can be criticized for being both too strong—limiting utility over unrealistic privacy concerns—and too weak—not guaranteeing intuitive notions of privacy. There have been many attempts to incorporate these critiques into new privacy definitions.

- Crowd-Blending Privacy. Gehrke, Hay, Lui, Pass, 2012.
- Zero-Knowledge Privacy. Gehrke, Lui, Pass, 2011.
- Noiseless Database Privacy. Bhaskar, Bhowmick, Goyal, Laxman, Thakurta, 2011.
- The Privacy of the Analyst and the Power of the State. Dwork, Naor, Vadhan, 2012.
- Differential Privacy for the Analyst via Private Equilibrium Computation. Hsu, Roth, Ullman, 2013.

There are a number of earlier privacy definitions that are further from the style of differential privacy.

- "k-anonymity". Generalizing Data to Provide Anonymity when Disclosing Information. Samarati, Sweeney, 1998.
- ℓ-diversity: privacy beyond k-anonymity. Machanavajjhala, Gehrke, Kifer, 2007.
- t-closeness: privacy beyond k-anonymity and ℓ-diversity. Li, Li, Venkatasubramanian, 2007.
- Security-control methods for statistical databases: A comparative study. Adam, Wortmann, 1989.

Some more recent work has explored taking an "axiomatic" approach to defining privacy—stating a set of self-evident properties of privacy, and attempting to derive specific privacy definitions from these properties— and has used this to start to systematize and compare different definitions of privacy.

- Towards and Axiomatization of Statistical Privacy and Utility. Kifer, Lin, 2010.
- Pufferfish: A Framework for Mathematical Privacy Definitions
- Privacy in Search Logs. Michaela Goetz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, Johannes Gehrke

Potential project topics related to definitions:

- Provide some new comparisons of existing privacy definitions—which ones are stronger, weaker, or incomparable? Are there separations that may be significant in practice? Which ones better capture our intuitive notions of privacy?
- Are there equivalent (e.g. Bayesian) definitions that are equivalent to some of the other definitions?
- Formulate new privacy definitions appropriate for a context not covered by existing definitions, and explore the privacy–utility tradeoffs offered by your definitions.
- If you were to take an axiomatic approach to privacy, what axioms would you choose? Do they lead to existing definitions, or to new ones?

Differential Privacy and Statistics/Machine Learning There has been a lot of interesting work designing differentially private algorithms for data analysis tasks commonly used in machine learning and statistics, e.g. logistic regression, SVMs, maximum likelihood estimation. Many of these algorithms have strong theoretical guarantees and good practical performance.

- Differentially Private Empirical Risk Minimization. Chaudhuri, Monteleone, Sarwate, 2011.
- Privacy-Preserving Statistical Estimation with Optimal Convergence Rates. Smith, 2011.
- Private Convex Empirical Risk Minimization and High-Dimensional Regression. Kifer, Smith, Thakurta, 2012.
- Differentially Private Feature Selection via Stability Arguments and the Robustness of the Lasso. Smith, Thakurta, 2012.

Programming Frameworks and System Design for Differential Privacy. Since differentially private algorithms are often difficult to design and analyze, an attractive possibility would be to design a "differentially private programming language"—on in which we could ensure that every program written satisfies differential privacy, even if the programmer has never heard of differential privacy. There have been several interesting attempts at designing such a language, as well as designing larger systems that incoporate differential privacy.

- Privacy integrated queries: an extensible platform for privacy-preserving data analysis. Mcsherry, 2009.
- Distance Makes the Types Grow Stronger: A Calculus for Differential Privacy. Reed and Pierce, 2010.

- Airavat: Security and Privacy for MapReduce. Roy, Setty, Kilzer, Shmatikov, and Witchel, 2010.
- GUPT: Privacy Preserving Data Analysis Made Easy. Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler, 2012.
- Linear Dependent Types for Differential Privacy. Marco Gaboardi, Andreas Haeberlen, Justin Hsu, Arjun Narayan, Benjamin C. Pierce, 2013.

Potential project topic: present a general design for a larger system that incorporates differential privacy for some setting not addressed by the existing work, analyze the system-level privacy guarantees, and the tradeoffs between privacy, utility, usability, and efficiency your system offers.

Workload Optimal Algorithms. Many of the results on differentially private query release focus on giving worst-case error bounds that hold for arbitrary sets of queries (as a function of the size or VC-dimension of the set of queries), and these worst-case bounds are nearly optimal for some sets of linear queries. However, for some sets of queries, better error bounds can be achieved, and there are sophisticated algorithms that can approach the optimal bound for every set of queries.

- On the Geometry of Differential Privacy. Hardt and Talwar, 2010.
- The Geometry of Differential Privacy: The Approximate and Sparse Cases. Nikolov, Talwar, Zhang, 2012.
- An Adaptive Mechanism for Accurate Query Answering under Differential Privacy. Li, Miklau, 2012.

These algorithms approximate the optimal error bound up to factors that depend (poly)logarithmically on the data universe size, the number of queries, and/or δ (in the case of (ε, δ) differential privacy). It would be interesting to tighten the bounds. Also, most or all of the results don't say anything about error bounds above $\sqrt{n}$. For that regime, it seems necessary to also incorporate ideas related to fingerprinting codes, like in:

- Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. Ullman, 2013.

Experimental Evaluation of Differential Privacy There have been some efforts to experimentally study and heuristically improve both the utility and computational efficiency of differentially private algorithms.

- A Simple, Practical Algorithm for Differentially Private Data Release. Hardt, Ligett, McSherry, 2012.
- Differentially Private Empirical Risk Minimization. Chaudhuri, Monteleone, Sarwate, 2011.
- Accurate and efficient private release of datacubes and contingency tables. Cormode, Procopiuc, Srivastava, Yaroslavtsev, 2013.
- An Adaptive Mechanism for Accurate Query Answering under Differential Privacy. Li, Miklau, 2012. (And related papers by Miklau et al.)

Project Ideas: Either try to improve an existing implementation of an algorithm, or implement and experiment with a different algorithm from the literature.

Privacy problems outside the scope of differential privacy There are many privacy problems outside the scope of differential privacy, such as behavior tracking, targeted advertising, and discrimination. There has been some work using ideas from differential privacy to help address or model these problems.

- Fairness through awareness. Dwork, Hardt, Pitassi, Reingold, Zemel, 2012.
- Privad: Practical Privacy in Online Advertising. Guha, Cheng, Francis, 2011.

Potential project idea: explore how the ideas in the Fairness through awareness paper might be used in practice to address the discrimination problems raised by Sweeney (http://dataprivacylab.org/projects/onlineads/).

Private Data Release vs. Traitor-Tracing Schemes. In class, we have seen hardness results for generating differentially private synthetic data that allows for answering many queries, and we have seen that these hardness results can sometimes be bypassed by using other forms of data summaries (e.g. a multivariate polynomial approximation). Can we give any hardness results that are independent of the form of the summary (analogous to representation-independent hardness results in learning theory)? It turns out that this is closely related to the problem of constructing "traitor-tracing schemes" in cryptography:

- The complexity of differentially private data release: efficient algorithms and hardness results, Dwork et al., 2009.
- Answering $n^{2+o(1)}$ counting queries with differential privacy is hard. Ullman, 2013.

It is still wide open to give representation-independent hardness results for privately answering "natural" queries, such as k-way marginals.

2 Additional Project Ideas

Here we list some additional topics that may be suitable for projects, but not for coverage in class.

Stability and Clustering. One widely-used form of data analysis is clustering. As we've seen (in Ch.7 of Dwork–Roth), various forms of "stability" of solutions are useful for accurate differentially private analysis.

- Smooth Sensitivity and Sampling in Private Data Analysis. Nissim, Raskhodnikova, Smith, 2007.
- Differential Privacy and Robust Statistics. Dwork, Lei, 2008.

There has also been a recent line of work using various forms of stability to design more efficient clustering algorithms.

- Approximate Clustering without the Approximation. Balcan, Blum, Gupta, 2009.)
- Stability Yields a PTAS for k-Means and k-Median. Awasthi, Blum, Sheffet, 2010.
- Clustering under Perturbation Resilience. Balcan, Liang, 2012.

Can these techniques be combined to yield an efficient, accurate, and private clustering algorithm?

Privacy-preserving Data Analysis Contests. Data analysis contests have become a popular way to improve and test machine learning and data mining techniques. Many of these contests (see e.g. Netflix, Hubway, the Kaggle Social Network Challenge, and the Heritage Health Prize) have suffered privacy breaches, in which individuals in the published test data were reidentified.

- Robust De-anonymization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset). Narayanan, Shmatikov, 2008.
- Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge. Narayanan, Shi, Rubinstein, 2011.

How would you propose to implement data analysis contest in a privacy-preserving way? Is it even possible to do so? Some papers that have explored such questions in the past are:

- Differentially-private recommender systems: Building privacy into the Netflix Prize contenders. Mironov, McSherry, 2009.

Differential Privacy in Your Area. Explore how differential privacy (or a variant) might be relevant to a problem in your areas of interest/research. You can either combine existing techniques or develop your own.

Attacks on Aggregate Data. Many reidentifications have been performed in practice on anonymized datasets.

- Re-identification of DNA through an automated linkage process. Malin, Sweeney, 2001.
- Identifiability of de-identified clinical trial data. Sweeney, 2009.
- Robust De-anonymization of Large Sparse Datasets (How to Break Anonymity of the Netflix Prize Dataset). Narayanan, Shmatikov, 2008.
- Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge. Narayanan, Shi, Rubinstein, 2011.
- Hubway challenge re-identification. `http://aboutmyride.org/more.html` Abrams, Davis, Lackner, Lodha, Sweeney, Zeng, 2012.

There are a number of theoretical results showing that it is possible in principle to attack aggregate data if insufficient noise is added (as we've seen in Dwork-Roth Ch. 8, and HW3).

- Revealing Information while Preserving Privacy. Dinur, Nissim, 2003.
- The Price of Privacy and LP Decoding. Dwork, McSherry, Talwar, 2007.
- New Efficient Attacks on Statistical Disclosure Control Mechanisms. Dwork, Yekhanin, 2008.

However, relatively few attacks on aggregate data have been carried out in practice. Some notable exceptions are:

- Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. Homer et al., 2008.

- Privacy in a Demographic Database. Nissim et al., 2013. http://tinyurl.com/abuwzr7

Are there other settings where attacks like these can be applied? Are there other novel and/or practical attacks against aggregated data that can be carried out in practice (e.g. on systems such as Google Trends)? For an exploration of the privacy protections in Facebook's advertiser platform, see:

- Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study. Chin and Klinefelter, 2012.

Algorithms with Heuristic Utility. We have mostly focused on algorithms that have provable guarantees of both privacy and utility. What if we don't require provable utility? Many algorithms in practice do not have provable utility guarantees, but are extremely effective on real data. Take or design such an algorithm, show how to make it provably differentially private, and assess (either experimentally or analytically) the loss of utility for ensuring privacy. One particular area of interest would be query release (answering many queries with differential privacy), where we have seen that there are computational complexity barriers for achieving provable utility guarantees in the worst case. In particular, practical heuristics for generating differentially private synthetic data would be exciting.

Characterizing the Computational Complexity of Sanitization. In class, we have seen some computational hardness results for differential private sanitizaton/query release based on various cryptographic assumptions (such as the existence of one-way functions, or traitor-tracing schemes). What are the minimal assumptions under which sanitization is hard? That is, where does the complexity of sanitization (or other privacy tasks, such as private learning) lie in relation to P vs. NP, the existence of one-way functions, the hardness of learning, etc.? For a similar systematic study of the complexity of learning, see:

- On Basing Lower-Bounds for Learning on Worst-Case Assumptions. Applebaum, Barak, Xiao, 2008.
- On Basing ZK $\neq$ BPP on the Hardness of PAC Learning. Xiao, 2009.

This topic is related to (but not equivalent to) the question raised in the topic Differential Privacy vs. Learning above (where the focus is on simple/natural families of queries, whereas here we are referring to arbitrary efficient queries).

Hardness Results for Other Privacy Tasks. We have seen computational hardness results for the query release/sanitization problem for counting queries (based on digital signatures and/or traitor-tracing schemes). There are several other privacy tasks for which the best known algorithms are inefficient, and it would be very interesting to give evidence that this inefficiency is inherent. For answering more than n^2 low-sensitivity queries (not just counting queries), the known algorithms run in time exponential in n (even when the data universe is small):

- Boosting and Differential Privacy. Dwork, Rothblum, Vadhan, 2010.

Adding noise depending on local sensitivity (as in Dwork–Roth Chapter 7) also can yield inefficient algorithms (as it requires computing the distance to the nearest input of high local sensitivity):

- Smooth sensitivity and sampling in private data analysis. Nissim, Raskhodnikova, Smith, 2007.
- Differential privacy and robust statistics. Dwork, Lei, 2009.