Research Statement
Jonathan Ullman
January, 2018

1 Summary

Every day, massive amounts of data are collected, analyzed, shared, and used to make high-stakes decisions. Our reliance on this data raises many questions about how to use it reliably. The majority of my research to date has centered around two central questions about reliable data analysis:

How can researchers use valuable but sensitive data to learn about a population without compromising the privacy of the individuals in that data?

How can researchers prevent false discovery and use data to learn meaningful facts about a population without overfitting to that data?

Although these questions have been studied for decades, they remain unsolved both in theory and in practice—leading to high-profile privacy breaches, and even what Gelman and Loken call a “statistical crisis in science.” My research develops the theoretical foundations of reliable data analysis by applying the tools and perspectives of cryptography, and complexity theory to problems in statistics and machine learning—charting the boundaries of what is and is not possible.

My research has contributed a number of foundational ideas to these questions. My work in privacy revolves around a rigorous privacy framework called differential privacy. In my PhD thesis [Ull13, BUV14], I discovered a key connection between fingerprinting codes and differential privacy, and used this connection to prove fundamental results about the limits of privacy for high-dimensional data. Building on this technique, I have answered a number of open questions about the limits of privacy [KMUZ16, SU17a, BU17, SU17b, BSU17, KMUW18], and designed novel, realistic attacks on allegedly private data releases [DSS+15] that give strong evidence that differential privacy truly captures the limits of privacy in theory and in practice. This technique has also impacted the work of other researchers [DTT14, BST14, TTZ15, BNS16b]. My most significant contribution to false discovery has been to show novel, tight connections between privacy and false discovery in adaptive data analysis—the realistic setting where the same dataset is reused repeatedly—leading to dramatically better methods for preventing false discovery [BNS+16a] and surprising new barriers [HU14, SU15].

2 Privacy

How can researchers use valuable but sensitive data to learn about a population without compromising the privacy of the individuals in that data? Unfortunately, protecting privacy has proven to be a very delicate task. Indeed, there have been several high-profile “re-identifications” (e.g. [NS08, HSR08]) of data sets “de-identified” using ad hoc methods that made incorrect
assumptions about the capabilities of the attacker, causing a lack of confidence [Pre14] in the most commonly used approaches.

Ten years ago, differential privacy [DMNS06] emerged as a strong notion of stability that puts privacy on the same kinds of rigorous, definitional foundations as cryptography. Differential privacy provides a robust individual privacy guarantee, ensuring that no adversary, regardless of their prior information, can learn much more about an individual user than they could have learned had that user’s data never been collected. Yet, despite its strength, there is a vast body research showing how to implement nearly any statistical algorithm privately, and differential privacy is now being deployed by Google [EPK14, BEM+17], Apple [TVV+17], and the U.S. Census Bureau [MKA+08, HMA+17].

My work in privacy breaks down into three main categories: (1) Understanding the inherent computational and statistical limitations of differentially private algorithms, and tradeoffs between computational and statistical costs. (2) Designing statistically accurate and computationally efficient private algorithms for foundational problems in statistics and machine learning. (3) Devising realistic privacy attacks against allegedly private data releases, which often shows that the strong notion of differential privacy is almost without loss.

2.1 Key Results: Fingerprinting Codes and the Price of Privacy

My research demonstrated for the first time that there is an inherent price of privacy in high-dimensional data. For context, the early results in differential privacy [DN03, BDMN05, DMNS06] showed how to privately and accurately estimate a modest number of statistics on a sensitive dataset by perturbing each answer independently with carefully calibrated noise. Later, a breakthrough result [BLR08] showed how to estimate an exponential number of statistics, and thereby showed that privacy comes at little cost in low-dimensional datasets. However the accuracy and running time of their algorithm scales very poorly in high-dimensional datasets.

I showed that these costs cannot be avoided in general. Specifically, under cryptographic assumptions, any algorithm whose running time is polynomial in the dimension can only estimate a modest number of statistics [Ull13], showing that adding independent noise to each statistic is optimal for efficient algorithms. Building on these techniques, I also proved that the accuracy of any differentially private algorithm (even a computationally inefficient one) must degrade significantly with the dimension even for the almost-trivial task of estimating the mean of a high-dimensional probability distribution, which was the first problem for which private algorithms provably required asymptotically more data than non-private algorithms [BUV14, SU17a]. The key technique underlying these results is a surprising connection between fingerprinting codes [BS98]—a seemingly unrelated cryptographic tool used for watermarking—and differential privacy, which has proven very useful:

The Complexity of Privacy. Recently, I revisited the question of the computational complexity of privacy. The central open question in the theory of differential privacy is whether or not there are efficient algorithms for privately estimating the sorts of simple, natural statistics that arise in practice, in contrast to my hardness results that apply to more complex families of statistics. One major obstacle to answering this question is that currently ruling out efficient algorithms for estimating any family of statistics requires essentially the strongest plausible cryptographic
assumptions, which forecloses using stronger assumptions to prove new results. I recently made substantial progress on this question, replacing the strong assumptions used in prior results with the minimal assumption of one-way functions [KMUZ16, KMUW18].

Privacy in Sparse Datasets. Real high-dimensional datasets are sparse, so it typically suffices to find only the relatively small set of relevant statistics. For example, in genomic datasets most mutations are rare, and typically we are only interested in the relatively small number of frequent mutations. In a pair of joint works, one with then-undergraduate Mitali Bafna [BU17] and one with Steinke [SU17b] we extended the fingerprinting codes technique to pin down the price of privacy even for analyzing sparse high-dimensional datasets.

Further Impact. The fingerprinting technique has also been used by others to prove optimal lower bounds on the accuracy of private algorithms for solving several fundamental problems such as PCA [DTTZ14], convex optimization [BST14, TTZ15], and learning [BNS16b].

2.2 Privacy and the Complexity of Simple Statistics

In light of my negative results, progress in differential privacy will only come from focusing on specific use cases, and designing tailored, optimized differentially private algorithms for these cases. Indeed, I have designed a variety of tailored differentially private algorithms that bypass the limitations on the accuracy and efficiency of general-purpose algorithms.

Learning High-Dimensional Distributions. In a work in preparation with Kamath, Li, and my PhD student Vikrant Singhal [KLSU18], we have obtained nearly optimal algorithms for two fundamental problems in statistical estimation: learning a high-dimensional Gaussian distribution and learning a high-dimensional product distribution. In contrast to many of my negative results, we show that privacy comes almost for free when solving these problems. Our new technique—private recursive preconditioning—is likely to find more applications.

Private Marginal Tables. Finding efficient summaries for marginals with optimal sample complexity has been one of the central open problems in differential privacy for over a decade [BCD+07]. Marginals are statistics “What fraction of the individuals in then dataset have each of the attributes \(A_1, \ldots, A_w\)?” (e.g. “What fraction have published papers on both privacy and false discovery?) Despite their simplicity, marginals capture essential properties of a dataset like correlations among attributes, are sufficient statistics for least-squares regression and log-linear models, and are favored by statistical agencies like the U.S. Census Bureau for public data releases. Although my computational hardness results do not apply to marginals, finding effective differentially private algorithms for releasing the marginals of the dataset has proven challenging, and I have shown some significant barriers ruling out most known approaches [UV11, GHRU11]. However, my work also introduced a connection between privately releasing marginals and agnostic learning, yielding the first algorithms with subexponential running time and sample complexity [GHRU11, TUV12, CTUW14].

New Private Algorithms. I have designed private algorithms for a number of interesting problems, including privately estimating the cut-structure of a graph or network [GRU12], privately finding equilibrium solutions of zero-sum games [HRU13, KPRU14], and finding optimal tolls to regular congestion in networks [RRUW15]. In addition to enlarging the set of
differentially private algorithms, each of these results has introduced new approaches to the
design of differentially private algorithms more generally. Most notably, my work on private
equilibrium computation [KPRU14] introduced a novel relaxation called \textit{joint differential privacy}
that has made it possible to give meaningful private guarantees for problems where differential
privacy is not appropriate. I have also designed novel algorithms for simultaneously solving
many convex optimization problems on a single dataset [UL15]

Privacy Tools. I am a senior researcher on Harvard University’s Privacy Tools Project. As such,
I have contributed to the PSI (Ψ) tool [GHK+16], which enables users without privacy expertise
to select and apply differentially private algorithms and receive guidance to the user about the
implications of their choices. In addition to my algorithmic work, Rogers, Roth, and Vadhan, and
I [RRUV16] developed a new composition theorems that allow differentially private algorithms
to be stitched in even more general ways than prior work. Other groups [WCHR17, LNR+17]
have incorporated this result into their work.

2.3 Realistic Privacy Attacks

My work has established strong limits on how accurately we can hope to analyze data
while promising differential privacy. However, because differential privacy provides guarantees
against arbitrary attackers, lower bounds for differential privacy do not always correspond to
attacks that could be performed by realistic attackers.

There have, however, been real demonstrations of the privacy risks of releasing statistical
information about a dataset (in contrast to the firmly established privacy risks of releasing
“de-identified” datasets). The most infamous is the attack of Homer et al. [HSR+08] that breached
privacy in certain genomic datasets using essentially just the mean of the dataset, causing the
NIH to change its policy on access to data from their studies. This attack is clearly realistic, but
can be defeated by ad hoc methods that do not meaningfully ensure privacy.

This disconnect between theory and practice falsely suggests that there are two extremes—
releasing many, extremely accurate statistics resulting in practical attacks on one end, and
rigorous privacy protections that severely limit data utility on the other end—and a chasm
in between that is beyond rigorous analysis. My joint work with Dwork, Smith, Steinke, and
Vadhan [DSS+15] challenged this belief and showed that the fingerprinting technique can be
generalized to yield a simple, robust, and realistic privacy attack called the \textit{fingerprinting attack}.

The fingerprinting attack is best understood as a robust version of Homer et al.’s attack. Their
attack uses the mean of this data to infer whether a specific individual is present in the dataset.
Mere presence in the dataset can be highly sensitive information—for example, in studies of
drug users. However, their attack requires the \textit{exact} mean and \textit{nearly exact} knowledge of the data
distribution, making them brittle. In contrast, the fingerprinting attack requires only a \textit{very noisy}
approximate mean and \textit{very limited knowledge} of the data distribution.

An interesting property of the fingerprinting attack is that, as we introduce more noise into
the statistics, the attack succeeds right up to the point where the noise is sufficient to ensure
differential privacy. Thus, for releasing the mean of a high-dimensional dataset, differential
privacy appears to be almost \textit{without loss} if we want to provide meaningful privacy in practice. I
believe this phenomenon is much more general, and much of my work on privacy is guided by
the following informal conjecture: *The gap between the accuracy at which privacy can be compromised in practice and the accuracy that can be achieved using rigorous privacy methods is very small.*

### 2.4 Additional Contributions

My research on differential privacy has been quite broad, and I have made several additional contributions that don’t fit into the “core” threads above. Some of these contributions have even extended beyond differential privacy into other applications:

- **Building on the game-theoretic work above, my joint works with Roth, Slivkins, and Wu [RUW16, RSUW17] gave novel algorithms for optimization from revealed preferences.** Our algorithms are capable of efficiently learning revenue-maximizing or welfare-maximizing prices in settings where the producer does not have direct access to the buyers’ utility functions, but can only see the buyers’ decisions when confronted with certain prices. We have also resolved other purely game-theoretic questions using our techniques [PRU17].

- **With Liberty, Mitzenmacher, and Thaler [LMTU16], we repurposed techniques for private lower bounds in differential privacy to prove optimal lower bounds on the space required to store frequent itemset information about a dataset.**

- **With my Ravi Sundaram and my PhD student Albert Cheu [CSU18], we introduced a novel variant of the best-arm-identification problem for multi-armed bandits that we call skyline identification, and gave optimal algorithms for this problem.**

### 2.5 Future Direction: Distributed Private Data Analysis

An increasingly important challenge for differential privacy is to work with distributed data. The recent wave of commercial systems for distributed private data analysis [EPK14, TVV+17] work in the so-called *local model* of differential privacy. In a local model protocol, each user separately perturbs their data using a differentially private algorithm, and then sends this perturbed data to a potentially *untrusted* aggregator who can then extract some useful information about the set of users as a whole. Local model protocols have many advantages, but provably require very high error (see e.g. [BEM+17] for realistic examples where local protocols would require billions of samples). In contrast, most differentially private algorithms are designed for a *central model*, in which all users send their data in the clear to an aggregator who is *trusted* to reveal only the output of some differentially private algorithm run on the dataset. However, finding such an aggregator is infeasible in many settings.

**Distributed Privacy Beyond Local Protocols.** In principle, there is no dilemma between the central and local models—any central model algorithm can be implemented without a trusted aggregator using *secure multiparty computation (MPC)*. However, despite dramatic progress on secure computation, existing techniques require large overheads in terms of computation and communication costs, and rounds of interaction between the users and aggregator. I am beginning to investigate bespoke protocols that retain the positive features of local model algorithms while providing the power of central model algorithms.
One particular limitation of distributed private protocol like those in use at Apple and Google is that they typically incur a significant privacy loss each day, and do not provide users with privacy guarantees over time. One reason why this is the case is because, as I recently showed [Ull18] the many powerful central-model techniques for preventing the degradation of privacy over time—the exponential mechanism and the above threshold mechanism—provably have no known analogues in the local model. In another work in preparation with Joseph, Roth, and Waggoner [JRUW18], we have shown how to obviate the need for these techniques by slightly relaxing the notion of accuracy, however it appears that the most promising avenue for addressing this problem is to cryptographically implement more central-model algorithms.

3 False Discovery from Adaptive Data Analysis

Suppose we have a dataset sampled from some population. How can we prevent false discovery and ensure that the conclusions drawn based on the dataset generalize to the population? For decades, statisticians have been developing methods to prevent false discovery, such as the Bonferroni correction [Bon36, Dun61] and Benjamini-Hochberg procedure [BH95]. And yet, false discovery remains a vexing problem in the scientific community, leading to provocatively titled scientific articles like “Why Most Published Research Findings are False” [Ioa05].

While there are many causes of false discovery, one that is receiving increasing attention is adaptive data analysis—the common scenario where datasets are re-used across multiple analyses, and the choice of one analysis depends on the outcomes of previous analyses. Adaptivity invalidates statistical methods for preventing false discovery, which assume that the choice of analysis is independent of the data, and has been implicated in a “statistical crisis in science” [GL14].

Adaptivity can take many forms. It can arise in a “structured” way from using sequential algorithms, and these structured procedures often allow for direct analysis. For example, statisticians have devised numerous procedures for “postselection inference”, in which the same dataset is used both to select relevant features and fit a model on those features.

In contrast, the modern practice of large-scale data analysis is “unstructured”—researchers reuse the same datasets for multiple publications, data analysts tune learning algorithms for optimal performance, and competitors in data science contests repeatedly train classifiers on a fixed dataset to improve their score. This unstructured adaptivity leads to a “garden of forking paths” [GL14], and makes it nearly impossible to even specify the statistical procedure, and thus challenging to prevent false discovery without severely compromising statistical accuracy.

3.1 Preventing False Discovery via Differential Privacy

The general problem of unstructured adaptive data analysis was formalized and investigated in a pair of concurrent works, one by Hardt and myself [HU14] and one by Dwork et al. [DFH+15], revealing a close connection between interactive data analysis and differential privacy. Consider a dataset $X$ sampled i.i.d. from a population $P$. Unless $X$ contains a huge number of samples, there will exist some statistic such that $X$ and $P$ look very different for that statistic. However, no data analyst who interacts with $X$ only through differentially private algorithms can find such a statistic. Therefore, any answer that is accurate with respect to the dataset $X$ must
generalize to the underlying population $P$. In a joint work with Bassily, Nissim, Smith, Steinke, and Stemmer [BNS+16a], we proved a quantitatively optimal version of this fact and extended it to more general types of queries, which currently yields the best guarantees of its kind for adaptive data analysis.

The connection to differential privacy allows us to repurpose differentially private algorithms to reduce the sample complexity of adaptive data analysis. A straightforward approach of using a fresh sample to estimate each of $k$ statistics requires at least $\Omega(k)$ samples, differentially private algorithms are able to estimate $k$ statistics using $n = O(\sqrt{k})$ or fewer!

However, since there are many limitations of differentially private algorithms, one might hope that there are even better methods for adaptive data analysis that bypass these limitations by eschewing differential privacy for some other approach. While this remains a fascinating question, my work with Hardt and Steinke show that at least some of the limitations of differential privacy in high-dimensional datasets do apply to adaptive data analysis in general [HU14, SU15], which established surprising computational and statistical bottlenecks for adaptive data analysis. In work in progress with Nissim, Smith, Steinke, and Stemmer [NSS+18], we have shown even stronger bottlenecks for a broad class of algorithms, showing that significantly different techniques will be needed for progress.

### 3.2 Future Directions: Bridging Computer Science and Statistics

Since my initial work on the subject and the work of Dwork et al., there has been a flurry of work on this topic by researchers in statistics, machine learning, and computer science. Thus, while many fundamental technical questions about interactive data analysis remain open and deserving of attention, I am most excited about the possibility of integrating my approach to false discovery with different approaches from statistics and machine learning. I am particularly interested in incorporating Bayesian methods, which at least superficially circumvent many of the problems of adaptive data analysis that my work has identified. Developing a full understanding of the relationship between the two schools of thought seems essential for the long term development of this research agenda.

### References


