Research Statement

Thomas Steinke

December 14, 2017

1 Introduction

As more and more data is being collected and used, *privacy* and *statistical validity* are becoming increasingly vexing issues for scientists, policymakers, and industry. Sound tools are needed, as "ad hoc" approaches have been prone to failure. My research focuses on providing rigorous techniques for *privacy-preserving data analysis* and *statistically valid adaptive data analysis*. These problems turn out to be intimately related and both are addressed by *differential privacy*.

I find working in these areas particularly compelling because they offer an exciting opportunity to put theory into practice. Google and Apple are already protecting user data with differential privacy and I believe that my research will help this technology mature and spread. Furthermore, differential privacy draws in a broad spectrum of researchers and ideas – from cryptography, learning theory, and information theory to statistics and law – and I enjoy the interdisciplinary nature of the field.

My work studies two sides of these problems: Firstly, I develop improved algorithms and analytical tools for private or adaptive data analysis. Secondly, I study the limitations of what is possible – this includes proving impossibility results that help guide the development of algorithms, and also includes developing "privacy attacks." Privacy attacks are an important part of motivating privacy research, as they highlight the very real dangers inherent when working with private data.

I also work in *pseudorandomness*, specifically constructing unconditional pseudorandom generators. Pseudorandom generators are powerful and ubiquitous tools; in computational complexity theory, they are used for *derandomization* – that is, converting efficient randomized algorithms into similarly-efficient deterministic algorithms. I am particularly interested in derandomizing computations that use very little space. My goal is to develop new techniques in this area. Specifically, my research focuses on the use of discrete Fourier analysis, which is a very powerful technique, but which had seen limited application to the small-space setting prior to my work.

2 Differential Privacy

More sensitive data is being collected about individuals than ever before. This data can be a goldmine for science – from medical research to machine learning. However, an increasingly important challenge is to balance the value of sharing or releasing information about this data with the risk of compromising the privacy of the individuals concerned.

Traditional "de-identification" techniques – in which the data is partially redacted in an effort to mask the identities of individuals – have suffered numerous high-profile failures. For example, researchers have "re-identified" information about individuals by cross-referencing supposedly de-identified data with other public datasets, such as voter registration records [Swe97] or online movie reviews [NS08], or by simply piecing together clues in the released data [BZ06]. Even aggregate statistics [HSR+08] and learned models [CKN+11] can expose an individual's private information. These cases have resulted in lawsuits, job losses, and policy changes. This checkered record highlights the need for a more rigorous approach to privacy-preserving data analysis, which led to the development of differential privacy.

My goal is to help develop principled approaches to data privacy and enable their real-world use. My research also delineates the limitations of what can be done privately by studying "privacy attacks." These demonstrate how information that may seem innocuous can compromise privacy.

Privacy attacks. Privacy is inherently a hard problem. Releasing too much information about a dataset will inevitably imperil privacy. The question is to pinpoint how much is too much. This helps set expectations for what can be achieved privately and demonstrates the need for rigorous privacy technologies like differential privacy. My work provides simple and general privacy attacks and also shows a sharp characterization of what is possible.

Re-identifications of de-identified data have demonstrated the infeasibility of privately releasing individual-level microdata. However, privacy attacks are possible even if only aggregate information is released. Early work in this area [DN03, et sequalae] showed that approximately releasing a rich set of statistics about a dataset can allow the dataset to be partially reconstructed. Later work has shown that even the simplest possible statistics can "leak" private information.

A key example is genome-wide association study (GWAS) data, in which the DNA of a "case group" of people is analyzed. The data is aggregated by simply publishing, for each genetic attribute (i.e. SNP), what fraction of people in the case group have that attribute; there may be millions of attributes, but only a few hundred individuals in the case group. Homer et al. [HSR+08] gave the first practical demonstration that even this limited information can identify members of the case group, which led to changes in how research data is shared. My work shows that this is unavoidable.

With collaborators, I showed [DSS+15] that, if there are sufficiently many attributes, then even approximately releasing attribute frequencies makes it possible to identify members of the case group. We showed that, under reasonable assumptions, we can determine membership of the case group by simply looking at the correlation (i.e. inner product) between a person's attributes and the approximate case group statistics; to analyze this we unified the statistical approaches [HSR+08, SOJH09] with the cryptography literature on "fingerprinting" [BS98, Tar08, BUV14]. This privacy attack is both simpler than previous work and more universal – no accurate algorithm can circumvent it. Our privacy attack becomes feasible at the same point where differentially private algorithms begin to break down, which shows that differential privacy provides the best possible accuracy. Our survey [DSSU17] discusses this and other privacy attacks. Recently, I have been extending these attacks to settings where the data or the released statistics are "sparse" [SU17b].

Refining differential privacy. Differential privacy is a formal standard for privacy-preserving data analysis, which provides a basis for developing such algorithms. A key theme in my research is understanding foundational questions in differential privacy, with the aim of improving the analytical tools underlying all differentially private algorithms and deepening connections to areas like information theory and statistical inference.

Intuitively, differential privacy requires that *anything that can be "learned" about an individual from the output of the algorithm could also be learned if their data is not part of the algorithm's input.* Formally, a frequentist (rather than Bayesian) definition is used. Namely, the algorithm's output distribution should not depend too much on any individual's data:

Definition 1 ([DMNS06, DKM+06])**.** *A randomized algorithm M satisfies (ε, δ)-differential privacy if, for any pair of inputs x and x' differing only by the addition, removal, or replacement of a single individual's data,*

$$\mathbb{P}\left[M(x) \in E\right] \leq e^{\varepsilon}\mathbb{P}\left[M(x') \in E\right] + \delta \tag{1}$$

for all possible events E over the output of M.

The literature on differential privacy has flourished in the decade since its inception and it has been broadly accepted as an effective privacy standard. Google has recently deployed differentially

private data collection in the Chrome browser [EPK14], Apple (since iOS 10) is now collecting user information in a differentially private manner [App], and the United States Census Bureau is developing differentially private tools for deployment with the 2020 census.

This definition is surprisingly versatile. However, the hybrid multiplicative-additive bound (1) can be mathematically inelegant and often does not allow a tight analysis of the privacy properties of algorithms. Given the central importance of Definition 1, I believe that it is vital practically and theoretically to identify the best way to quantify privacy.

Bun and I [BS16] (building on [DR16]) gave a cleaner information-theoretic definition, which replaces the multiplicative-additive bound (1) with a bound in terms of Rényi divergence. Our approach – dubbed *concentrated differential privacy* – yields tighter quantitative results that achieve a sharper and more precise tradeoff between privacy and utility. In other words, we provide a better tool to analyze the privacy properties algorithms. In recently submitted joint work [BDRS17], I have further refined these ideas to provide analytical tools that sharply capture all the important properties of differential privacy, which I believe will make a real difference practically.

Optimal algorithms. Differential privacy establishes a standard for privacy, but we must also develop algorithms meeting that standard. My work has produced optimal or near-optimal algorithms for several well-studied problems.

The simplest algorithmic task is to answer several queries of the form "what fraction of the dataset has property P?" The standard algorithm for this task is to independently distort each answer using Gaussian or Laplacian noise. However, Jonathan Ullman and I showed [SU17a] that this well-known algorithm is suboptimal by presenting two new algorithms that use non-independent noise to achieve better accuracy. Furthermore, we showed that our algorithms are nearly-optimal by proving matching lower bounds.

Most of the differential privacy literature provides "worst-case" accuracy guarantees that do not depend on the data. In recent work, Feldman and I showed that it is possible to have the accuracy scale with the standard deviation of the values in the dataset. This is the natural scale for accuracy and, in many circumstances, this is significantly better than the worst-case bound.

If the queries are structured (such as having a low-dimensional domain), there are many more sophisticated algorithms (e.g. [HR10]) that can exploit the dependencies between the queries to attain better guarantees. My collaborators and I have developed an algorithm (based on [NTZ13, et sequalae]) which, in certain parameter regimes, provides the best possible accuracy for any set of queries; we are working to extend this to more parameter regimes and develop a geometric or combinatorial characterization of the distortion needed for answering any given set of queries.

There is plenty of room for further research, particularly when it comes to developing computationally efficient algorithms. I am hoping to import algorithmic ideas from sum-of-squares literature [BM16] or use heuristic algorithms like MCMC or SAT solvers [GAH+14]. I am also keen to collaborate to find better application-specific algorithms.

3 Statistically Valid Adaptive Data Analysis

A key problem in machine learning and empirical sciences is to infer properties of a large population given only a small, random sample from that population. The challenge is ensuring statistical validity – that is, ensuring that patterns observed in the sample generalize to the whole population, rather than occurring in the sample by chance. There is a vast literature addressing this problem. However, often the same sample dataset is reused for multiple analyses, where the outcomes of earlier analyses can inform later analyses. This "adaptivity" introduces the danger that later analyses may be tailored to the sample, thereby producing results that "overfit" the dataset in ways that are difficult to account for.

Adaptivity has been identified as a major problem that leads to flawed research in the empirical sciences (sometimes referred to as "p-hacking" or "data dredging") [Ioa05, SNS11, GL14].

Tight connection to differential privacy. Dwork et al. [DFH+15] provided the first nontrivial results for handling adaptive data analysis by showing that differential privacy provides protection against overfitting. With collaborators [BNS+16], I improved and extended the results of Dwork et al. to obtain a quantitatively tight connection.

Furthermore, the above mentioned algorithms offer fixed accuracy guarantees that do not depend on the data distribution. Feldman and I [FS17] showed that it is possible to have the accuracy scale with the standard deviation of the values in the sample, which often yields significantly better accuracy. Our results required altering the way differential privacy is applied and we are continuing to simplify these results by looking at relaxations of differential privacy and developing an information-theoretic approach to this area.

Overall, my goal is to push these results to obtain the strongest possible theoretical guarantees, while also making the algorithms clean and simple so that they can easily be applied. I believe that this connection between "stability" (in the form of differential privacy) and generalization may also be useful for understanding generalization in other settings like deep learning [BE02, HRS16].

Sharp lower bounds. How much more difficult is adaptive data analysis than pre-specified non-adaptive data analysis? Ullman and I [SU15] (improving [HU14]) gave a sharp impossibility result, showing that the aforementioned algorithmic results are asymptotically optimal. In particular, this gives an exponential separation between what is possible in the adaptive and non-adaptive settings. These results require the use of cryptographic hardness assumptions and a combinatorial object which we call interactive fingerprinting codes. In ongoing work, we are refining these results and extending them to other settings.

4 Pseudorandomness

A fundamental tool – both practically and theoretically – is a *pseudorandom generator*. That is, an efficient algorithm that takes as input a small number of random bits (the "seed") and produces many output bits that "look random." Namely, no efficient algorithm should be able to distinguish the generator's output from truly uniform bits. In general, we can only construct pseudorandom generators under unproven computational hardness assumptions [NW94]. Thus we consider weaker classes of distinguishers. My work constructs generators whose output appears random to *small-space distinguishers*.

Nisan [Nis92, INW94] showed that a $O(\log^2 n)$-bit seed is sufficient to produce n bits that appear random to distinguishers using only $O(\log n)$ bits of memory space. This is a powerful result; however, despite two decades of work, we have not succeeded in reducing the seed length to $O(\log n)$. Although, there has been some progress for the further restricted classes of *regular or permutation branching programs* [BRRY10, BV10, KNP11, De11], including my work [Ste12].

Given that these techniques seem to hit a barrier at $O(\log^2 n)$, I believe that new techniques are needed and my research aims to provide them.

Fourier-analytic pseudorandom generators. With Reingold, Vadhan, and Wan [RSV13, SVW14], I developed pseudorandom generators for small space based on Fourier-analytic techniques. Discrete Fourier analysis has been successful in constructing pseudorandom generators for other classes of distinguishers [NN93, Bra08, Vio08, DGJ+10, GMR+12], but has had limited application to small-space [BDVY13]. One advantage of our Fourier-analytic approach is that the pseudorandomness property is invariant under permutations of the output bits of the pseudorandom generator, which is a property that Nisan's generator provably lacks [Tzu09].

My work has inspired recent progress [CHRT17]. Namely, recently my conjecture [RSV13] about the Fourier spectrum of functions computable in small space was proved. This implies that the pseudorandom generator from my work can be applied to all constant-space computations. I hope that this is a catalyst for further progress and I intend to push these techniques as far possible.

5 Future Work

My research has several threads and I intend to continue developing all of these. I am also continually seeking to develop new directions for research, which means always looking out for interesting questions and, most importantly, enthusiastic collaborators. Below I outline where I see the main threads of my work going in the future.

Differential privacy. Differential privacy is already having a real-world impact, but this is just the tip of the iceberg. The aim of my research agenda is to help it reach its full potential.

As mentioned above, there are plenty of fruitful directions for theoretical work in differential privacy – more accurate algorithms, more computationally efficient algorithms, sharper analytical tools, tailoring algorithms to the structure of queries or distribution of the data, proving impossibility results, understanding different application models (such as local differential privacy) – and I intend to further pursue these directions.

However, applied work is equally important. Differential privacy needs to be incorporated into larger systems and we need to develop more readily-useable code to help non-experts easily get started in the area. I am particularly keen to organize inter-disciplinary collaborations (like the PrivacyTools project I was involved in at Harvard). Such collaborations bring in new perspectives and skills that help understand and overcome the challenges of deploying cutting-edge technology.

Another key task for differential privacy researchers is outreach. One of the biggest barriers to adoption is convincing decision makers that differential privacy is the appropriate tool for their problem (and convincing them that they have a problem, e.g., by demonstrating privacy attacks). This requires reaching out to people with very different backgrounds, such as lawyers, medical researchers, statisticians, and business executives. I have been engaged in this at Harvard and at IBM and I hope to keep doing so. This can also be very enlightening as it requires understanding different viewpoints and being realistic about the challenges we face in practice.

Generalization & adaptivity. Understanding generalization is a central question in machine learning and statistics and the connection to differential privacy has added a new dimension to it. I intend to delve further into this, particularly examining how it relates to previous methods for understanding generalization (e.g. VC dimension) and how we can look a weaker notions of stability, e.g., by using information theoretic approaches instead of differential privacy.

The application to adaptive data analysis is only one of the potential uses of differential privacy. We can view differential privacy as providing generalization guarantees "for free," which adds further impetus to the use of differentially private algorithms. I am keen to find other uses. Connections to algorithmic game theory and fairness in machine learning have already been explored and I would like to explore these further

Pseudorandomness. There are many deep and longstanding open questions in pseudorandomness. My favorite question is developing pseudorandom generators for small space (ideally strong enough to imply $\mathsf{RL} = \mathsf{L}$). My research in this area has focused on applying new techniques to this problem, namely discrete Fourier analysis (which has been very successful elsewhere in the pseudorandomness literature). I believe that this approach will yield further results. Indeed, there has been some very recent progress building on my work. I hope to be involved in continuing this line of research.

References

[App] Apple. Apple previews ios 10, the biggest ios release ever. `http://www.apple.com/newsroom/2016/06/apple-previews-ios-10-biggest-ios-release-ever.html`.

[BDRS17] Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. 2017. In submission.

[BDVY13] Andrej Bogdanov, Zeev Dvir, Elad Verbin, and Amir Yehudayoff. Pseudorandomness for width-2 branching programs. *Theory of Computing*, 9(283-293):1, 2013.

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[BM16] Boaz Barak and Ankur Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In *29th Annual Conference on Learning Theory*, pages 417–445, 2016.

[BNS+16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1046–1059. ACM, 2016.

[Bra08] Mark Braverman. Polylogarithmic independence fools ac0 circuits. *J. ACM*, 57(5):28:1–28:10, June 2008.

[BRRY10] Mark Braverman, Anup Rao, Ran Raz, and Amir Yehudayoff. Pseudorandom generators for regular branching programs. *FOCS*, pages 40–47, 2010.

[BS98] Dan Boneh and James Shaw. Collusion-secure fingerprinting for digital data. *IEEE Transactions on Information Theory*, 44(5):1897–1905, 1998.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer Berlin Heidelberg, 2016.

[BUV14] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *STOC*, pages 1–10. ACM, May 31 – June 3 2014.

[BV10] Joshua Brody and Elad Verbin. The coin problem and pseudorandomness for branching programs. In *FOCS*, pages 30–39, 2010.

[BZ06] Michael Barbaro and Tom Zeller. A face is exposed for aol searcher no. 4417749. *New York Times*, August 2006.

[CHRT17] Eshan Chattopadhyay, Pooya Hatami, Omer Reingold, and Avishay Tal. Improved pseudorandomness for unordered branching programs through local monotonicity. *Electronic Colloquium on Computational Complexity (ECCC)*, pages TR17–171, 2017.

[CKN+11] J.A. Calandrino, A. Kilzer, A. Narayanan, E.W. Felten, and V. Shmatikov. "you might also like:" privacy risks of collaborative filtering. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 231–246, May 2011.

[De11] Anindya De. Pseudorandomness for permutation and regular branching programs. In *CCC*, pages 221–231, 2011.

[DFH+15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *STOC*. ACM, June 14–17 2015.

[DGJ+10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A. Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology - EUROCRYPT*, pages 486–503, St. Petersburg, Russia, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, March 4-7 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM, June 9-12 2003.

[DR16] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DSS+15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *56th Annual IEEE Symposium on Foundations of Computer Science (FOCS'15)*, October 2015.

[DSSU17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4, 2017.

[FS17] Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory*, pages 728–757, 2017.

[GAH+14] Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. In *ICML*, pages 1170–1178, 2014.

[GL14] Andrew Gelman and Eric Loken. The statistical crisis in science data-dependent analysis?a ?garden of forking paths??explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6):460, 2014.

[GMR+12] Parikshit Gopalan, Raghu Meka, Omer Reingold, Luca Trevisan, and Salil Vadhan. Better pseudorandom generators from milder pseudorandom restrictions. In *FOCS*, pages 120–129, 2012.

[HR10] Moritz Hardt and Guy Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proc. 51st Foundations of Computer Science (FOCS)*, pages 61–70. IEEE, 2010.

[HRS16] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*. IEEE, October 19-21 2014.

[INW94] Russell Impagliazzo, Noam Nisan, and Avi Wigderson. Pseudorandomness for network algorithms. In *STOC*, pages 356–364, 1994.

[Ioa05] John P. A. Ioannidis. Why most published research findings are false? *PLoS Medicine*, 2(8):124, August 2005.

[KNP11] Michal Koucký, Prajakta Nimbhorkar, and Pavel Pudlák. Pseudorandom generators for group products. In *STOC*, pages 263–272, 2011.

[Nis92] Noam Nisan. $\mathcal{RL} \subset \mathcal{SC}$. In *STOC*, pages 619–623, 1992.

[NN93] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Computing*, 22:838–856, 1993.

[NS08] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008.

[NTZ13] Aleksandar Nikolov, Kunal Talwar, and Li Zhang. The geometry of differential privacy: the sparse and approximate cases. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 351–360. ACM, 2013.

[NW94] Noam Nisan and Avi Wigderson. Hardness vs randomness. *J. Comput. Syst. Sci.*, 49(2):149–167, October 1994.

[RSV13] Omer Reingold, Thomas Steinke, and Salil Vadhan. Pseudorandomness for regular branching programs via fourier analysis. In *APPROX-RANDOM*, pages 655–670, 2013.

[SNS11] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[SOJH09] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[Ste12] Thomas Steinke. Pseudorandomness for permutation branching programs without the group theory. *ECCC*, 19:83, 2012.

[SU15] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory (COLT'15)*, July 2015. Full version available at http://arxiv.org/abs/1410.1228.

[SU17a] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 2017.

[SU17b] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 552–563, Oct 2017.

[SVW14] Thomas Steinke, Salil Vadhan, and Andrew Wan. Pseudorandomness and fourier growth bounds for width 3 branching programs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 28, pages 885–899, 2014.

[Swe97] Latanya Sweeney. Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics*, 25(2-3):98–110, 1997.

[Tar08] Gábor Tardos. Optimal probabilistic fingerprint codes. *J. ACM*, 55(2), 2008.

[Tzu09] Yoav Tzur. Notions of weak pseudorandomness and GF(2^n)-polynomials. Master's thesis, Weizmann Institute, 2009.

[Vio08] Emanuele Viola. The sum of d small-bias generators fools polynomials of degree d. *CCC*, 0:124–127, 2008.

[EPK14] lfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security*, Scottsdale, Arizona, 2014.