Privacy at Scale: Local Differential Privacy in Practice

Graham Cormode
University Of Warwick
G.Cormode@warwick.ac.uk

Somesh Jha
University of Wisconsin-Madison
jha@cs.wisc.edu

Tejas Kulkarni
University Of Warwick
T.Kulkarni.2@warwick.ac.uk

Ninghui Li
Purdue University
ninghui@cs.purdue.edu

Divesh Srivastava
AT&T Labs-Research
divesh@research.att.com

Tianhao Wang
Purdue University
tianhaowang@purdue.edu

ABSTRACT

Local differential privacy (LDP), where users randomly perturb their inputs to provide plausible deniability of their data without the need for a trusted party, has been adopted recently by several major technology organizations, including Google, Apple and Microsoft. This tutorial aims to introduce the key technical underpinnings of these deployed systems, to survey current research that addresses related problems within the LDP model, and to identify relevant open problems and research directions for the community.

KEYWORDS

Data collection, privacy, differential privacy, local differential privacy

ACM Reference Format:
Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at Scale: Local Differential Privacy in Practice. In *SIGMOD'18: 2018 International Conference on Management of Data, June 10–15, 2018, Houston, TX, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3183713.3197390

1 MOTIVATION

The most impactful data science often relies on analysing data from individuals that is considered highly sensitive — medical history, personal interests and preferences, and opinions. In many cases it is not feasible to gather the necessary sensitive information without providing strong guarantees of privacy to the users in question. The model of differential privacy can provide such guarantees, and most recently the topic of local differential privacy (LDP) — where users randomly perturb their inputs to provide plausible deniability of their data without the need for a trusted party — has come to the fore.

Local differential privacy has been adopted by several major technology organizations, so the technology is used by hundreds of millions of users daily. These companies include Google through their RAPPOR system, to collect web browsing behaviour [12]; Apple's implementation, that allows Apple and app developers to collect usage and typing history [22]; and Microsoft's collection of a variety of telemetry data over time [10].

The aims of this tutorial are to introduce the key technical underpinnings of these deployed systems, and provide intuition on how they work; to survey current research that addresses related problems within the LDP model; and to identify open problems and research directions for privacy. For this tutorial, we use the deployed systems to exemplify and motivate the ideas that derive from algorithms and theory. Participants will learn how an idea from fifty years ago has found application in the 21st Century, and how major companies are scaling this up to Internet scale.

2 OUTLINE OF THE TUTORIAL

This tutorial is intended to introduce SIGMOD participants to the new but rapidly developing topic of Local Differential Privacy. Our approach is practice-led, inspired by the large-scale deployments of Locally Differential Private data collection by major technology companies, including Google, Apple and Microsoft [9, 10, 12]. We structure the core of the tutorial around these three deployed systems, using them to motivate the underlying algorithms, and connecting out to the research literature that underpins them. In more detail, our outline is given below.

2.1 Introduction and Preliminaries

We briefly motivate the need for tools for private data collection and analysis, and introduce the definitions of Differential Privacy and Local Differential Privacy (LDP). The first definition equivalent to LDP came from the database community as "amplification" [13], then came to prominence in the work of Duchi et al. [11]. We introduce the most basic LDP mechanism, randomized response [6, 27], which came from the survey design community, and masks a single bit by tossing a biased coin. We also introduce the mathematical tools to understand LDP, including unbiasedness, variance and confidence tail bounds.

2.2 State of the Art Deployments

We describe three practical realizations of LDP algorithms for collecting popularity statistics, and cover the development of these ideas through the computer science research literature, and subsequent enhancements that have been proposed.

(1) *RAPPOR* from Google, which combines Randomized Response with Bloom Filters to compactly encode massive sets [12]. The application is to identify popular web destinations (URLs),

SIGMOD'18, June 10–15, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-4703-7/18/06…$15.00
https://doi.org/10.1145/3183713.3197390

without revealing any individual user's browsing habits. Subsequent work from the same team has described how to efficiently extract the identities of popular destinations without prior knowledge of their URLs [14].

(2) *Apple's DP implementation* was announced in 2016, and is documented in a patent application [22] and subsequent white paper [9]. The technique combines several technical advances: using the Fourier transform to spread out signal information, and sketching techniques to reduce the dimensionality of the massive domain. In parallel, a rich stream of literature has abstracted the problem of *identifying Differentially Private Heavy Hitters*, progressively refining and optimizing these techniques [2–4, 20, 25].

(3) *Telemetry collection* from Microsoft, which makes use of histograms and fixed random numbers to collect data over time [10].

2.3 Theoretical Foundations

Informed by these practical deployments, we then more formally describe the LDP setting and give a brief overview of the history of this concept. In the LDP setting, there is a set of users. The i-th user possesses a private value v_i in some domain D. These users interact with an untrusted aggregator such that the aggregator learns statistical information about the the distribution of the private value in the user population, while the information leakage for each individual is bounded. More specifically, a user perturbs the private value v_i using an algorithm π and sends $\pi(v)$ to the aggregator. The aggregator then processes the collected reports to recover statistical information. The algorithm $\pi(\cdot)$ satisfies the following property:

Definition 2.1 (Local Differential Privacy). An algorithm π satisfies ϵ-local differential privacy (ϵ-LDP), where $\epsilon \geq 0$, if and only if for any input v and v', we have

$$\forall y \in Range(\pi): \; \Pr[\pi(v) = y] \leq e^{\epsilon} \; \Pr[\pi(v') = y],$$

where $Range(\pi)$ denotes the set of all possible outputs of the algorithm π.

We will discuss the history of this model and the relationship of LDP to the classic DP definition. Specifically, in the standard (or centralized) setting, each user sends raw data v to the aggregator, who obtains the true distribution, adds noise, and then publishes the result. In this setting, the aggregator is trusted to not reveal the raw data and is trusted to handle the raw data correctly. But in the local setting, each user perturbs the data locally and thus does not have to trust the aggregator. In brief, LDP has a stronger privacy model than DP, but entails greater noise. We will also cover the refinements of the definition, i.e., the relationship between bounded and unbounded definitions, node-based and edge-based definitions (in graph problems), pure and non-pure definitions. We will also describe composition theorems, including parallel, sequential, and advanced composition in the local setting.

2.4 Frequency Oracles

We describe an abstraction of a core problem in LDP which has attracted a lot of theoretical and practical attention. Protocols for enabling the estimation of the frequency of any value in the domain D are called *Frequency Oracles* (FO). A basic FO protocol generalizes the Random Response protocol in [27] (which is defined for binary values). It is specified as follows:

$$\forall_{y \in D} \; \Pr[\pi(v) = y] = \begin{cases} \frac{e^{\epsilon}}{e^{\epsilon}+|D|-1}, & \text{if } y = v \\ \frac{1}{e^{\epsilon}+|D|-1}, & \text{if } y \neq v \end{cases}$$

That is, we sample the true element with higher probability, and all other elements with lower uniform probability. This protocol's accuracy degrades as the size of the domain D increases, because the probability that a value is correctly reported, is approximately inversely proportion to $|D|$. Several other FO protocols have been proposed, including RAPPOR by Erlingsson et al. [12], Random Matrix Projection by Bassily and Smith [3], and a mechanism based on the Count sketch with Hadamard transform by Bassily et al. [2].

In recent work [24], authors of this tutorial introduced an abstract framework of FO protocols, and showed that most previously proposed protocols can be placed within it. Protocols in this framework share a common aggregation algorithm that produces unbiased estimates of values. They also enable a convenient derivation of the estimation variance, which describes the accuracy of the oracle. This enables us to compare different protocols; identify new, more optimized FO protocols; and provide precise conditions regarding which protocol one should use, depending on the domain size and the privacy budget ϵ. In this part, we will use the approach in [24] to organize and analyze FO protocols in the tutorial.

2.5 Current Research Directions

We will briefly describe some of the related quite recent results that have been published on applying LDP to other domains.

- *Heavy Hitter Identification.* Here the goal is to identify the values that are frequent. When the size of the domain is small, this can be solved with an FO protocol. One simply queries the frequency of every value in the range, and then identifies the frequent ones. However, if the domain is very large (e.g., 2^{128} or larger), finding the most frequent values this way is computationally infeasible. The method proposed by Thakurta et al. [22] identifies frequent byte at each location, and uses semantic analysis to filter out meaningless combinations. There are several related works in the pure LDP setting: Hsu et al. [15] and Mishra et al. [18] provide efficient protocols for heavy hitters, but the error bound proved is higher than for the method proposed by Bassily and Smith [3]. The follow-up work by Bassily et al. [2] proposed TreeHist, which is shown to be more efficient and accurate than that of Bassily and Smith [3]. Bun et al. [4] proposed PrivateExpanderSketch with state-of-the-art theoretical performance. Concurrent work [25] gave the first implemented protocol PEM.
- *Itemset Mining.* This problem considers the setting where each user's value is a set of items from the item domain. Such a set-valued setting occurs frequently in the situation where LDP is applied. For example, when Apple wants to estimate the frequencies of the emojis typed everyday by the users, each user has a set of emojis that they typed [23]. The problem is quite challenging even when one just tries to find frequent items. Encoding each transaction as a single

value in the domain (power set of the original domain), and using existing FO protocols does not work. An FO protocol can identify only values that are very frequent in the population, because the scale of the added noise is proportional to square root of the population size [5]. It is quite possible that each particular transaction appears relative infrequently, even though some items and itemsets appear very frequently. When no value in new domain is frequent enough to be identified, using a direct encoding an aggregator can obtain only noise.

For example, assume that the transactions are $\{a, c, e\}$, $\{b, d, e\}$, $\{a, b, e\}$, $\{a, d, e\}$, and $\{a, f\}$, respectively. While no transaction appears more than once, items a and e each appears 4 times, and the itemset $\{a, e\}$ appears 3 times. Thus the three most frequent itemsets are $\{a\}$, $\{e\}$, $\{a, e\}$.

To solve the problem, the LDPMiner protocol in [20] uses a technique called "padding and sampling". That is, each user first pads her set of values with dummy items to a fixed size ℓ, then randomly samples one item from the padded set, and finally uses an FO protocol to report the item. When estimating the frequency of an item, one multiples the estimation from the FO protocol by ℓ. Without padding, the probability that an item is sampled is difficult to assess, making accurate frequency estimation difficult.

Very recent work [26] proposes a protocol that provides much better accuracy than LDPMiner within the same privacy constraints. The advantage comes from several key observations including privacy amplification under sampling, which is known to hold in the centralized DP setting [16].

- *Private location collection* Data can often be represented as points in multidimensional space—as a simple example, consider user locations in two-dimensional space. Sketching frequencies within multidimensional spaces, allowing rectilinear counting queries to be answered approximately, and identifying "hot spots" are primitives that could then be used to build more sophisticated user activity models. Initial work on this problem has extended LDP private frequency collection [7]. It is open to extend this to build more sophisticated user movement models.
- *Marginal distributions of multidimensional data.* Given users represented as points in multidimensional space, a natural question is to extract distributions over subsets of dimensions. Naively, we could materialize all possible subsets and apply existing approaches, but this rapidly degrades the accuracy. Instead, taking projections of the data via a Fourier basis allows better reconstructions [8].
- *Graph algorithms and synthetic graph modeling.* Much sensitive individual data is best represented as a graph—either a simple graph between users, or a bipartite graph between users and other entities. Recent work has aimed to build accurate graph models under LDP [21].
- *Language modeling.* An application of private data collection is to build better prediction models e.g., for typing on mobile devices. Recent work has shown how to accurately and privately train sophisticated deep neural network models [17].

Finally, we will briefly connect to other models of privacy and security, that adopt different assumptions and trust models. These include contrasting with the centralized differential privacy model, and achieving privacy by adding centralized noise via encrypted data collection. We will also discuss other approaches from secure multi-party computation, homomorphic encryption and private information retrieval, amongst others, that achieve different trade-offs.

2.6 Open Problems and New Directions

We will point to a number of directions for future work, based on emerging trends in the literature.

- *Multiple Rounds.* Most deployed LDP protocols require the user to follow a fixed protocol over their data, and send their (perturbed) response for aggregation. More generally, we could allow multiple rounds of interaction, where the aggregator poses new queries in the light of previous responses. This approach has been proposed for building machine learning models [19]. It is open to understand the power of multiple rounds, compared to what is possible in a single round.
- *Hybrid models.* LDP gives a very strong protection to users, at the expense of lower accuracy compared to a centralized model with a trusted aggregator. Recent work has proposed a hybrid model where some users follow LDP and some users submit to a trusted aggregator, and both sets are "blended" together [1].
- *Theoretical underpinnings.* Several works on LDP have started to appear in the theory literature, addressing questions about the power of LDP [3, 4, 11]: what are the lower bounds on the accuracy guarantees (as a function of privacy parameter and population size); is there any benefit from adding an additive "relaxation" δ to the privacy definition; and minimizing the amount of data collected from each user to a single bit.

3 INTENDED AUDIENCE AND BACKGROUND KNOWLEDGE

We intend to make this tutorial accessible to all participants in SIGMOD and PODS. Although there has been a vast amount of research on the topic of privacy, even when narrowing to work on Differential Privacy, the topic of this tutorial is quite accessible, and does not require any familiarity with prior work. The emphasis is on the design of scalable algorithms, with some consideration of how these can be built into robust systems. To appreciate the correctness and accuracy guarantees of the algorithms, some statistical tools are needed. These are at the level of an introductory statistics course: computing the variance of a discrete random variable, and using this to provide confidence bounds. We will give a brief refresher on the necessary tools, but will not provide detailed proofs of the algorithms; rather, we will try to provide our insights into what the guarantees mean, and where the different terms in the guarantees arise from.

4 BIOGRAPHY OF THE PRESENTERS

Graham Cormode is a Professor in Computer Science at the University of Warwick in the UK, where he work on research topics in data management, privacy and big data analysis. Previously, he was a principal member of technical staff at AT&T Labs-Research. He is a University Liaison Director at the Alan Turing Institute, and in 2017 he was the co-recipient of Adams Prize for Mathematics for his work on Statistical Analysis of Big Data.

Somesh Jha earned the B.Tech. degree in electrical engineering from Indian Institute of Technology, New Delhi in 1985, and Ph.D. in Computer Science from CMU in 1996. He is currently a Sheldon B. Lubar Professor of computer science in University of Wisconsin, Madison. Prior to joining University of Wisconsin in 2000, he was a postdoctoral fellow at the School of Computer Science in Carnegie Mellon University. His research interests are in security and software engineering. He is currently the chair of the steering committee for the ACM CCS. He is on the editorial boards of ACM TOPS and Journal of Computer Security. He is an ACM Fellow.

Tejas Kulkarni is a PhD student at the University of Warwick, with a focus on Local Differential Privacy. He completed his Masters at Indian Institute of Technology, Madras.

Ninghui Li received the B.Eng. degree in computer science from the University of Science and Technology of China in 1993, and Ph.D. degrees in Computer Science from New York University, in 1998 and 2000. He is currently a professor of computer science in Purdue University. Prior to joining Purdue University in 2003, he was a research associate at the Computer Science Department in Stanford University. His research interests are in security and privacy in information systems. He is currently the chair of the ACM Special Interest Group on Security, Audit and Control (SIGSAC), and was the Program Chair of ACM Conference on Computer and Communications Security (CCS) in 2014 and 2015. He is on the editorial boards of ACM Transactions on Privacy and Security (TOPS) (formerlly TISSEC), ACM Transactions on Internet Technology, and Journal of Computer Security. He is a senior member of the IEEE and an ACM distinguished scientist.

Divesh Srivastava is the head of Database Research at AT&T Labs-Research. He is a Fellow of the ACM, the Vice President of the VLDB Endowment, and the managing editor of the Proceedings of the VLDB Endowment (PVLDB). His research interests and publications span a variety of topics in data management. He received his Ph.D. from the University of Wisconsin, Madison, and his B.Tech. from the Indian Institute of Technology, Bombay.

Tianhao Wang is a PhD student at Purdue University, with a focus on Local Differential Privacy. He completed his Bachelor's at Fudan University.

Acknowledgements

GC is supported in part by European Research Council grant ERC-2014-CoG 647557 and The Alan Turing Institute under the EPSRC grant EP/N510129/1. TK is supported in part by AT&T, Marie Curie CIG 618202, a Warwick collaborative postgraduate research scholarship, and The Alan Turing Institute under the EPSRC grant EP/N510129/1.

REFERENCES

[1] B. Avent, A. Korolova, D. Zeber, T. Hovden, and B. Livshits. BLENDER: enabling local search with a hybrid differential privacy model. In *26th USENIX Security Symposium, USENIX Security 2017, Vancouver, BC, Canada, August 16-18, 2017.*, pages 747–764, 2017.

[2] R. Bassily, K. Nissim, U. Stemmer, and A. G. Thakurta. Practical locally private heavy hitters. In *NIPS*, pages 2285–2293, 2017.

[3] R. Bassily and A. Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 127–135. ACM, 2015.

[4] M. Bun, J. Nelson, and U. Stemmer. Heavy hitters and the structure of local privacy. *arXiv preprint arXiv:1711.04740*, 2017.

[5] T.-H. H. Chan, M. Li, E. Shi, and W. Xu. Differentially private continual monitoring of heavy hitters from distributed streams. In *Privacy Enhancing Technologies*, volume 7384, pages 140–159. Springer, 2012.

[6] A. Chaudhuri and R. Mukerjee. *Randomized response: Theory and techniques.* Marcel Dekker, 1988.

[7] R. Chen, H. Li, A. K. Qin, S. P. Kasiviswanathan, and H. Jin. Private spatial data aggregation in the local setting. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 289–300, 2016.

[8] G. Cormode, T. Kulkarni, and D. Srivastava. Marginal release under local differential privacy. In *SIGMOD Conference*. ACM, 2018.

[9] Differential Privacy Team, Apple. Learning with privacy at scale. 2017.

[10] B. Ding, J. Kulkarni, and S. Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems 30*, December 2017.

[11] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*. IEEE, 2013.

[12] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067. ACM, 2014.

[13] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222. ACM, 2003.

[14] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 2016(3):41–61, 2016.

[15] J. Hsu, S. Khanna, and A. Roth. Distributed private heavy hitters. In *International Colloquium on Automata, Languages, and Programming*, pages 461–472. Springer, 2012.

[16] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *ASIACCS*, 2012. The initial official proceedings includes only a 2-page summary for every paper. Complete paper was later added to ACM DL, and is available at http://delivery.acm.org/10.1145/2420000/2414474/p32-li.pdf.

[17] H. B. McMahan, D. Ramage, K. Talwar, and L. Zhang. Learning differentially private language models without losing accuracy. In *ICLR*, 2018.

[18] N. Mishra and M. Sandler. Privacy via pseudorandom sketches. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 143–152. ACM, 2006.

[19] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, and J. Shin. Collecting and analyzing data from smart device users with local differential privacy. *CoRR*, abs/1606.05053, 2016.

[20] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 192–203. ACM, 2016.

[21] Z. Qin, T. Yu, Y. Yang, I. Khalil, X. Xiao, and K. Ren. Generating synthetic decentralized social graphs with local differential privacy. In *CCS*, pages 425–438. ACM, 2017.

[22] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudiger, V. R. Sridhar, and D. Davidson. Learning new words, Mar. 14 2017. US Patent 9,594,741.

[23] A. G. Thakurta, A. H. Vyrros, U. S. Vaishampayan, G. Kapoor, J. Freudinger, V. V. Prakash, A. Legendre, and S. Duplinsky. Emoji frequency detection and deep link frequency, July 11 2017. US Patent 9,705,908.

[24] T. Wang, J. Blocki, N. Li, and S. Jha. Locally differentially private protocols for frequency estimation. In *USENIX'17: Proceedings of 26th USENIX Security Symposium on USENIX Security Symposium.* USENIX Association, 2017.

[25] T. Wang, N. Li, and S. Jha. Locally differentially private heavy hitter identification. *arXiv preprint arXiv:1708.06674*, 2017.

[26] T. Wang, N. Li, and S. Jha. Locally differentially private frequent itemset mining. In *IEEE Symposium on Security and Privacy*, pages 578–594, 2018.

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