Differential Privacy in the Wild: A tutorial on current practices & open challenges

Ashwin Machanavajjhala
Duke University
Durham, NC, USA
ashwin@cs.duke.edu

Xi He
Duke University
Durham, NC, USA
hexi88@cs.duke.edu

Michael Hay
Colgate University
Hamilton, NY, USA
mhay@colgate.edu

ABSTRACT

Differential privacy has emerged as an important standard for privacy preserving computation over databases containing sensitive information about individuals. Research on differential privacy spanning a number of research areas, including theory, security, database, networks, machine learning, and statistics, over the last decade has resulted in a variety of privacy preserving algorithms for a number of analysis tasks. Despite maturing research efforts, the adoption of differential privacy by practitioners in industry, academia, or government agencies has so far been rare. Hence, in this tutorial, we will first describe the foundations of differentially private algorithm design that cover the state of the art in private computation on tabular data. In the second half of the tutorial we will highlight real world applications on complex data types, and identify research challenges in applying differential privacy to real world applications.

1. TUTORIAL OVERVIEW

Privacy concerns are a major obstacle to deriving the scientific insights now possible from increasing data collection and powerful new analysis techniques. The goal of privacy-preserving algorithms is to permit data mining and analysis to be carried out over a collection of sensitive records donated by individuals. Ideally, individuals receive a guarantee that the analysis does not lead to harmful disclosures about them. At the same time, data miners and scientists hope to study the data with little disruption to their methods and results. Differential privacy [7] has emerged as an important standard for protection of individuals' sensitive information. Its general acceptance by researchers has led to a flood of research across several communities: databases, data mining, theory, machine learning, security, programming languages, statistics and economics.

An algorithm satisfies ϵ-differential privacy if its output on a database of individuals is statistically indistinguishable (measured by parameter ϵ) from the output of the algorithm if any one individual had opted out of the database. These algorithms work by infusing noise into query answers, and more privacy (smaller ϵ values) require the infusion of larger amounts of noise. Over the past decade, there has been extensive work on designing sophisticated differential privacy algorithms to support answering batch and interactive workloads of counting queries, publishing synthetic data, and for supporting a number of data mining tasks like regression/classification, clustering and itemset mining. Additionally, recent work has also considered applying differential privacy to more complex data types like graphs and sequential data.

Despite its success in the research community, the adoption of differential privacy by practitioners in academia, industry, or government agencies has been startlingly rare. We believe that this failure in adoption stems from an undue focus on algorithm design in a simplified problem setup, a lack of understanding of the semantics of differential privacy for complex data types in terms of research, and a lack of awareness of the state of the art in differentially private algorithm design from the practitioners. This tutorial tries to address this gap.

In this tutorial, we cover the foundations of differentially private algorithm design as well as the challenges faced in interpreting and enforcing differential privacy in real applications that deal with complex data types. Thus, the tutorial will attract non-experts who would like to learn about differential privacy, as well as experts who may understand differential privacy, but are looking for new research problems in making differentially private algorithms work in practice. The tutorial will cover both landmark theoretical results in this area, as well as describe practical state of the art algorithms for a number of analysis tasks. Finally, the tutorial will be divided into modules, and each module will include a 'hands-on' segment. Attendees will have the opportunity to work on a exercise that reinforces the material covered in the module.

2. TUTORIAL OUTLINE

Our tutorial will consist of 6 modules each lasting 30 minutes. The modular organization will allow attendees to choose which parts of the tutorial they might be most interested in. The first three modules on 'Defining privacy,' 'Building blocks for differential privacy' and 'Answering counting queries on tabular data' will focus on the foundations of differentially private algorithm design. These modules (especially the first two modules) are intended for non-experts (e.g., graduate students interested in privacy research) and will provide intuition and essential concepts

Proceedings of the VLDB Endowment, Vol. 9, No. 13

that will be used in later modules. The last three modules 'Applications I,' 'Beyond tabular data' and 'Applications II' will focus on the theoretical and practical challenges faced in both defining privacy and designing algorithms in real world settings that involve complex data types. These latter modules will provide an overview of cutting edge research that may be of interest even to experts.

We will have an exercise in each of the modules that ties together all the concepts described in the module. The exercise will last about 5 minutes, and we describe a concrete example exercise in the section on 'Building blocks for differential privacy' (Section 2.2). The topics covered in each of the modules are described next and outlined in Table 1.

2.1 Defining Privacy

In this module, we motivate privacy in databases using examples of known privacy attacks on sensitive individual data. We formalize the database privacy problem and distinguish it from related technologies like query answering on encrypted databases or secure multiparty computation. Using examples, we will show how simple anonymization techniques do not work, and motivate the need for formal guarantees of privacy that ensure (1) security without obscurity, (2) privacy under post-processing, and (3) composition. We will define ϵ-differential privacy [7, 10] and show that it satisfies these privacy desiderata.

2.2 Building Blocks for Differential Privacy

In this module, we will cover classic differentially private algorithms including the Laplace mechanism [9], randomized response [9], and the exponential mechanism [31]. Attendees will learn how to compute sensitivity of queries and how to derive bounds on privacy loss and error. Composition theorems including sequential composition, parallel composition and post-processing will be covered to answer multiple queries. We will also cover the smooth sensitivity framework [33] for answering high sensitivity queries.

The exercise for this module will be to develop a differentially private algorithm for k-nearest neighbor clustering. This will invite attendees to build the algorithm using the aforementioned building blocks and prove privacy using composition theorems.

2.3 Answering Counting Queries on Tabular Data

This module will give an overview of a range of techniques that have been developed for answering counting queries over tabular data under differential privacy. Tabular data is a common type of data format, which uses a model of vertical columns (identified by name) and horizontal rows. Each row corresponds to an individual. Counting queries compute the number of rows in the table which column values satisfy certain properties such as Age > 10. Examples of counting queries are histograms, range queries, cumulative distribution functions, etc. Rather than being comprehensive, we will categorize prior work (see Table 1), and discuss representative algorithms for each category. Categories we will cover include: (i) answering queries vs publishing synthetic data, (ii) online vs offline query answering, (iii) techniques that work for low vs high dimensional data, (iv) algorithms that add noise that is data independent vs those that add noise that can depend on the input database, and (v) the gaps between theory and practice. Representative algorithms that we will mention here include [15, 17, 24, 25, 27, 34, 40].

2.4 Applications I

This module starts with a description of two success stories of differential privacy, where these techniques are currently in use in live products. We will discuss how differentially private algorithms power private data publication in a US Census Bureau product called OnTheMap [29] and the use of RAPPOR [11] algorithm (a variant of randomized response) in collecting browser characters from Google Chrome users. We will also briefly discuss some of the issues that arise when deploying differential privacy, such as choosing a value for ϵ, dealing with limits on the number of queries, and misperceptions about the limitations of differential privacy.

After this discussion of practical deployment, we will start an overview of research in important application areas. This module focuses on data mining tasks (e.g., regression / classification [12, 35, 37] and itemset mining [26, 39]).

2.5 Privacy beyond tabular data

Data in the real world does not always the fit the assumptions of differential privacy – namely, the data is tabular, each row corresponds to all the information about an individual that one would want to protect, and that the rows are independent of one another. In this module we will present methods to customize the differential privacy notion to fit the privacy requirements of real world applications. We will discuss the No Free Lunch Theorem in data privacy [22], and present alternate privacy definitions that can be customized to match the structure of the data [14, 19, 23].

2.6 Applications II

We will highlight the state-of-the-art, challenges, and open questions in deriving algorithms with formal privacy guarantees for complex data types like networks [20], data with multiple entities[13], and trajectories [1, 19]. Using trajectories and location privacy, we will also highlight how users may require the different levels of protection (the entire trajectory, or only trajectory on a single day, etc), as well present the challenges posed by constraints occurring in the data.

3. INTENDED AUDIENCE

The tutorial assumes basic knowledge of probability (including distributions, means and variances, concentration theorems). The tutorial will not assume prior knowledge of cryptography or differential privacy. The tutorial will assume some background in databases and data mining, equivalent to that obtained in an introductory undergraduate or graduate class.

4. INTENDED LENGTH

The tutorial spans 2 sessions (3 hours). The first session will focus primarily on the foundations of differentially private algorithm design on tabular data, while the second session will focus on extending differential privacy to applications on different data types (networks, trajectories, etc.).

Module	Topic
Defining privacy	Motivating database privacy
	Problem formulation and desiderata
	Definition of ϵ-differential privacy
	Discussion
Building blocks for DP	Laplace mechanism
	Randomized response
	Exponential mechanism
	Bounds on privacy loss and error
	Composition theorems
	Smooth sensitivity
Answering counting queries for tabular data	Example: histograms & range queries
	Query answering vs publishing synthetic data
	Online vs offline query answering
	Low dimensional vs high dimensional data
	Data independent vs data dependent noise infusion
	Theory vs practice
Applications I	Real deployments (OnTheMap & RAPPOR)
	Regression/classification
	Frequent itemsets
Beyond tabular data	Neighboring databases
	Constraints or prior knowledge
	No-free lunch theorem
	Pufferfish privacy
	Blowfish privacy
Applications II	Network data
	Relations with multiple entities
	Trajectories and location privacy

Table 1: Tutorial Outline. Each section will conclude with a short exercise.

5. PRESENTERS

Ashwin Machanavajjhala is an Assistant Professor in the Department of Computer Science, Duke University and an Associate Director at the Information Initiative@Duke (iiD). Previously, he was a Senior Research Scientist in the Knowledge Management group at Yahoo! Research. His primary research interests lie in algorithms for ensuring privacy in statistical databases and augmented reality applications. He is a recipient of the National Science Foundation Faculty Early CAREER award in 2013, and the 2008 ACM SIGMOD Jim Gray Dissertation Award Honorable Mention. Ashwin graduated with a Ph.D. from the Department of Computer Science, Cornell University and a B.Tech in Computer Science and Engineering from the Indian Institute of Technology, Madras. His early work on ℓ-diversity [30] has been very influential in the field of data privacy and has been cited over 2500 times (according to Google Scholar). He also helped design one of the first real data publication powered by formal privacy guarantees in collaboration with the US Census Bureau in 2008 [29]. He has published in PODS, SIGMOD, VLDB, ICDE, WWW and WSDM, and has given tutorials on privacy at IEEE SSP 2009, ICDE 2010, and on entity resolution at AAAI 2012, VLDB 2012 and KDD 2013.

Xi He is a PhD student at Computer Science Department, Duke University. Her research interests lie in privacy-preserving data analysis and security. She has also received an M.S from Duke University and double degree in Applied Mathematics and Computer Science from University of Singapore. Xi has been working with Prof. Machanavajjhala on privacy since 2012, and has published in SIGMOD and VLDB.

Michael Hay is an Assistant Professor in the Department of Computer Science, at Colgate University. Before that he was a Computing Innovation Fellow at Cornell University and completed his PhD at UMass Amherst in 2010.

His research interests include privacy-preserving data analysis, data management, data mining, social networks, and privacy. His PhD thesis titled "Enabling Accurate Analysis of Private Network Data" is the recepient of the 2011 ACM SIGKDD Dissertation Award. His ICDM 2009 paper titled "Accurate estimation of the degree distribution of private networks" received the Best Student Paper award. He has given a tutorial on privacy and graphs at SIGMOD 2011.

6. RELATED WORK

We have identified five tutorials [4, 5, 16, 28, 38] on differential privacy in the past five years, which are mainly from SIGMOD, KDD, and WIFS. Compared to tutorials before 2013 [5, 16, 28, 38], the tutorial proposed for this venue will highlight recent techniques, as well as focus on the application of differential privacy to real problems and complex data types. While the building blocks of differentially private algorithms was the focus of [5], our tutorial has a larger scope of understanding the promise and limitations of differential privacy in real applications. While [4, 16, 28] only focused on one specific application such as network data, or machine learning, we will also cover relational databases and trajectories. Moreover, will also show how to customize differential privacy to meet the privacy requirements of these applications with complex data.

7. REFERENCES

[1] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *CCS*, 2013.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: The sulq framework. In *PODS*, 2005.

[3] J. Brickell and V. Shmatikov. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *KDD*, 2008.

[4] K. Chaudhuri and A. D. Sarwate. Differential privacy for signal processing and machine learning. In *WIFS*, 2014.

[5] G. Cormode. Building blocks of privacy: Differentially private mechanisms, 2013. Invited tutorial talk at Privacy Preserving Data Publication and Analysis (PrivDB) workshop.

[6] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, 2003.

[7] C. Dwork. Differential privacy. In *ICALP*, volume 4052 of *Lecture Notes in Computer Science*, 2006.

[8] C. Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1), 2011.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[10] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2013.

[11] U. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

[12] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Sec*, 2014.

[13] S. Haney, M. Kutzbach, M. Graham, J. Abowd, and L. Vilhuber. Formal privacy protection for data products combining individual and employer frames. In *UNECE/Eurostat Statistical Data Confidentiality Work Seesion*, 2015.

[14] S. Haney, A. Machanavajjhala, and B. Ding. Design of policy-aware differentially private algorithms. *PVLDB*, 2015.

[15] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *CoRR*, abs/1012.4763, 2010.

[16] M. Hay, K. Liu, G. Miklau, J. Pei, and E. Terzi. Privacy-aware data management in information networks. In *SIGMOD*, 2011.

[17] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *VLDB*, 2010.

[18] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava. Dpt: Differentially private trajectory synthesis using hierarchical reference systems. *VLDB*, 2015.

[19] X. He, A. Machanavajjhala, and B. Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *SIGMOD*, 2014.

[20] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev. Private analysis of graph structure. *ACM Trans. Database Syst.*, 2014.

[21] S. P. Kasiviswanathan, K. Nissim, S. Raskhodnikova, and A. Smith. Analyzing graphs with node differential privacy. In *TCC*, 2013.

[22] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, 2011.

[23] D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 2014.

[24] C. Li, M. Hay, G. Miklau, and Y. Wang. A data- and workload-aware algorithm for range queries under differential privacy. *VLDB*, 2014.

[25] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, 2010.

[26] N. Li, W. Qardaji, D. Su, and J. Cao. Privbasis: Frequent itemset mining with differential privacy. *VLDB*, 2012.

[27] N. Li, W. Yang, and W. Qardaji. Differentially private grids for geospatial data. In *ICDE*, 2013.

[28] K. Liu, G. Miklau, J. Pei, and E. Terzi. Privacy-aware data mining in information networks. In *KDD*, 2010.

[29] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, 2008.

[30] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *KDD*, 2007.

[31] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

[32] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *SP*, 2008.

[33] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

[34] W. Qardaji, W. Yang, and N. Li. Understanding hierarchical methods for differentially private histograms. *VLDB*, 2013.

[35] A. D. Sarwate and K. Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Processing Magazine*, 2013.

[36] J. Ullman. Private multiplicative weights beyond linear queries. In *PODS*, 2015.

[37] X. Wu, M. Fredrikson, W. Wu, S. Jha, and J. F. Naughton. Revisiting differentially private regression: Lessons from learning theory and their consequences. *CoRR*, 2015.

[38] Y. Yang, Z. Zhang, G. Miklau, M. Winslett, and X. Xiao. Differential privacy in data publication and analysis. In *SIGMOD*, 2012.

[39] C. Zeng, J. F. Naughton, and J.-Y. Cai. On differentially private frequent itemset mining. *VLDB*, 2012.

[40] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. In *SIGMOD*, 2014.