Differential Privacy in Data Publication and Analysis

Yin Yang[1,5] Zhenjie Zhang[1] Gerome Miklau[2] Marianne Winslett[1,3] Xiaokui Xiao[4]

[1]Advaned Digital Sciences Center
Illinois at Singapore Pte.
{yin.yang,zhenjie}@adsc.com.sg

[2]Computer Science Dept.
University of Massachusetts, Amherst
miklau@cs.umass.edu

[3]Dept. of Computer Science
Univ. of Illinois at Urbana Champaign
winslett@illinois.edu

[4]School of Computer Engineering
Nanyang Technological University
xkxiao@ntu.edu.sg

[5]Coordinated Science Laboratory
Univ. of Illinois at Urbana Champaign

ABSTRACT

Data privacy has been an important research topic in the security, theory and database communities in the last few decades. However, many existing studies have restrictive assumptions regarding the adversary's prior knowledge, meaning that they preserve individuals' privacy only when the adversary has rather limited background information about the sensitive data, or only uses certain kinds of attacks. Recently, *differential privacy* has emerged as a new paradigm for privacy protection with very conservative assumptions about the adversary's prior knowledge. Since its proposal, differential privacy had been gaining attention in many fields of computer science, and is considered among the most promising paradigms for privacy-preserving data publication and analysis. In this tutorial, we will motivate its introduction as a replacement for other paradigms, present the basics of the differential privacy model from a database perspective, describe the state of the art in differential privacy research, explain the limitations and shortcomings of differential privacy, and discuss open problems for future research.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Statistical Databases; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms

Security

Keywords

Differential privacy, privacy-preserving data publication, query processing, data analysis

1. INTRODUCTION

In data privacy research, the general goal is to develop mechanisms and protocols to publish data and analysis results without revealing sensitive information. Unfortunately, privacy preservation is generally a difficult task, since an adversary often can infer sensitive information, typically by exploiting background knowledge. In such scenarios, it seems nearly impossible to have a general solution for privacy protection that covers the full spectrum of possibilities with respect to the information the adversary may possess. Most of the existing studies on data privacy thus rely on specific assumptions about the prior knowledge of the adversary, leading to rather limited privacy protection.

In the past five years, *differential privacy* has emerged as a new paradigm that provides a more robust privacy guarantee, regardless of the adversary's prior knowledge [11]. In particular, given any two databases that differ on exactly one record r, a data analysis algorithm that satisfies differential privacy will output randomized results with almost identical probability distributions. Therefore, no matter how much the adversary knows about the other records in the database, or how many other analysis results she sees, the adversary will be unable to guess whether r is present in the database with high confidence. The precise meaning of "almost identical" is determined by a privacy parameter ϵ, set by the data owner. Smaller values of ϵ mean a stronger privacy guarantee.

The strong privacy guarantee of differential privacy comes at the price of noise added to the results of queries and analyses. One major line of research effort is devoted to reducing the amount of error that must be added to query and analysis results, while still satisfying differential privacy. Different types of analyses and data seem to require different error reduction techniques. A second major line of investigation seeks to extend the useful life of data for differentially private analyses, as each new query or analysis chips away at the total privacy budget represented by ϵ.

The last five years have seen over a hundred papers on differential privacy in the theory, security, database, machine learning, and statistics communities, with perhaps a dozen of these papers appearing in the top database venues. With differential privacy emerging as a preferred model for privacy research in the database community, a tutorial on this topic is timely.

2. TUTORIAL OUTLINE

Table 1 provides a detailed outline of this 3-hour tutorial. The tutorial contains five parts: (i) an introduction to privacy issues arising in real-world applications, with examples taken from recommendation systems, biomedical data analysis, and social network analysis; (ii) the basic idea, underlying theoretical foundation, and limits of differential privacy; (iii) existing techniques for high-accuracy differentially private query processing; (iv) differentially private algorithms for high-accuracy data mining; and finally, (v) interesting future directions and open problems in differential privacy.

Section	Topic
Motivation	Overview of types of privacy concerns
	Case study: Netflix data [33, 29]
	Case study: genomic data (GWAS) [19, 39]
	Case study: social network data [2]
	Limits of k-anonymity [38], l-diversity [28], t-closeness [26]
Basics	Differential privacy definition [11]
	Running example COUNT query
	Laplace mechanism [10]
	Budgets and multiple queries
	Budgets and dynamic data
	SUM and other aggregates
	Small case studies where DP works [21]
	Where DP fails: SARS
	Where DP fails: highly correlated data [20]
	Categorical data: exponential mechanism [31]
	Other mechanisms [12, 40, 17]
	(ϵ, δ)-differential privacy [34]
Query Processing	Noise reduction goals
	(Non)-interactive processing
	Privlets for range queries [41]
	Universal histograms for range queries [18]
	Noise or dirty data?
	Enforcing consistency [18]
	Multi-query optimization [23, 24]
	Multi-dimensional query processing [42, 8]
	Histogram construction [43]
	Time series aggregation [36]
	Compressive sensing [27]
	Non-interactive methods [7, 13]
Data Mining	Decision tree construction [15, 32]
	Frequent pattern mining [4]
	Data cube publication [9]
	Clustering [34]
	Regression analysis [5, 6, 30, 22]
	Classification [37]
Open Problems	Privacy budget management
	Arbitrary updates
	Multiparty computation
	Complex data
	Complex relational queries

Table 1: Tutorial outline

2.1 Motivation for Differential Privacy

Several major privacy breaches have occured in the past few years. In 2006, AOL released the search records of more than 650,000 users, collected over a 3-month period [1]. Although AOL had removed all personal identifiers from this data, the search records themselves contain personally identifiable information, such as the names of people and places. Consequently, New York Times reporters were able to re-identify an individual, by cross referencing these records with phonebook listings [1]. Similarly, the Netflix Prize (`http://netflixprize.com`) was a well known competition for collaborative filtering techniques, using training and evaluation datasets taken from Netflix's movie recommendation records. Although the released data did not contain any personal identifiers, researchers found that recommenders' movie ratings are quite uniquely identifying, to the extent that many recommenders could be re-identified by any who knew a few of the recommenders' favorite and/or least favorite movies. Further some recommenders can be re-identified by combining the released data with the comments retrieved from IMDB (`http://www.imdb.com`) [33].

Another well-known incident concerns the privacy issues in biomedical studies, where one can apply an attack model. We will present one such attack model originally designed for forensic purposes. Homer et al. [19] showed that it is possible to verify with high confidence whether an individual has participated in a genome-wide association study (GWAS). Soon after the publication of [19], the U.S. National Institute of Health (NIH) removed all aggregate results from its dbGaP online database, a key resource for biomedical research with a genetic component. Even today, the aggregate results from most studies are not freely available on dbGaP. Instead a researcher who wants to see aggregate results from dbGaP studies must first obtain IRB approval [14], an onerous process that used to be required only for access to the underlying raw data in dbGaP. Further, Wang et al. [39] developed a stronger attack, which is so powerful that it is able to re-identify patients from the results routinely published in GWAS research papers.

The Netflix and dbGaP examples illustrate that privacy risks are real. Yet society can clearly benefit from continued publication of some form of these datasets. Unfortunately, existing privacy models often fail to eliminate such risks, even with popular anonymization-based privacy models, e.g., k-anonymity [38] and l-diversity [28].

2.2 Differential Privacy Basics

The most commonly-used definition of differential privacy is ϵ-differential privacy, which guarantees that any individual tuple has negligible influence on the published statistical results, in a probabilistic sense. Specifically, a randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if and only if for any two databases DB, DB' that differ in exactly one record, and any possible output $\mathcal{O}$ of $\mathcal{A}$, the ratio between the probability that $\mathcal{A}$ outputs $\mathcal{O}$ on DB and the probability that $\mathcal{A}$ outputs $\mathcal{O}$ on DB' is bounded by a constant. Formally, we have

$$\frac{Prob(\mathcal{A}(DB) = \mathcal{O})}{Prob(\mathcal{A}(DB') = \mathcal{O})} \leq e^{\epsilon},$$

where ϵ is a constant specified by the user, and e is the base of the natural logarithms. Intuitively, given the output $\mathcal{O}$ of $\mathcal{A}$, it is hard for the adversary to infer whether the original

data is DB or DB', if parameter ϵ is sufficiently small. Similarly, ϵ-differential privacy also provides any individual with *plausible deniability* that her/his record was in the database.

The earliest and most widely-adopted approach for enforcing ϵ-differential privacy is the *Laplace mechanism* [10], which works by injecting random noise following a Laplace distribution into the output of the original, deterministic algorithm $\mathcal{G}$ to obtain its randomized version $\mathcal{A}$. One major limitation of the Laplace mechanism is that it requires the output of $\mathcal{G}$ to be real numbers. This motivates the *exponential mechanism* [31], which handles integer as well as non-numeric outputs by sampling from the output space of $\mathcal{G}$ in a randomized manner, instead of injecting noise directly. This mechanism is often employed in existing work for designing differentially private algorithms that involve complex operations on the input data (e.g., partitioning of datasets [7]). In addition, the geometric mechanism [16] is optimized specifically for algorithms with integer outputs, and achieves higher results accuracy than both the Laplace and the exponential mechanisms for such data.

There also exist some variations of ϵ-differential privacy. Nissim et al. [34] proposes (ϵ, δ)-differential privacy, which is a relaxed version of ϵ-differential privacy that allows privacy breaches to occur with a very small probability controlled by parameter δ. This relaxed notion is adopted in existing work to tackle scenarios where enforcing ϵ-DP would lead to unacceptably low data utility [21]. In addition, Dwork et al. [12] extend ϵ-DP for the scenarios where (i) the input data change with time and (ii) the output of a randomized algorithm $\mathcal{A}$ needs to be re-computed upon changes in the input. Finally, Kifer and Machanavajjhala [20] discuss the limitations of differential privacy, suggesting that a stronger privacy protection scheme is required for datasets with highly correlated records.

2.3 Differentially Private Query Processing

Besides the generic ϵ-DP methods described in the previous section, previous work has developed optimized solutions for specific classes of query workloads. In particular, Xu et al. [43] investigate how the counts in one-dimensional histograms can be released in a differentially private manner. Barak et al. [3] propose a solution for releasing *marginals*, each of which contains the counts pertinent to a projection of the original dataset onto a subset of its attributes. Hay et al. [18] and Xiao et al. [41, 42] initiate studies on the optimization for arbitrary count queries with a range selection on each attribute of a dataset. Hay et al.'s approach is designed for one-dimensional data, and it achieves superior data utility by exploiting the correlations between different queries. Xiao et al.'s method is based on *wavelet transforms* and it achieves the same asymptotic bounds (in terms of data utility) as Hay et al.'s solution; in addition, Xiao et al.'s method supports multi-dimensional data.

Cormode et al. [7] propose a solution that not only achieves comparable performance to Hay et al. and Xiao et al.'s methods but also incurs a smaller time overhead. Cormode et al. [8] and Peng et al. [35] present techniques for constructing differentially private indices (e.g., quad-trees), and show that the indices can be utilized to provide higher utility than Hay et al. and Xiao et al.'s methods on multi-dimensional datasets with large domains. Li et al. [23, 25] generalize Hay et al. and Xiao et al.'s approaches and develop solutions for optimizing arbitrary linear counting queries (whose selection predicates are not necessarily continuous ranges). Li et al. [27] employ compressive sensing techniques to improve the accuracy of point queries on sparse data. While the above techniques focus on minimizing the *absolute errors* of count queries, Xiao et al. [40] propose a solution that optimizes the *relative errors* instead.

In addition to count queries, previous work has also investigated more complex types of query workloads. For example, Rastogi and Nath [36] study the publication of time-series in a distributed setting. Dwork et al. [13] present a general solution that supports any queries that map the dataset to real numbers.

2.4 Differentially Private Knowledge Discovery

Privacy-preserving data mining and machine learning has attracted extensive research efforts recently, since sensitive information is often involved in knowledge discovery applications, e.g. disease causality studies and customer behavior analysis. A variety of mining and learning problems have been re-investigated under the context of differential privacy.

Decision trees are commonly used to build classification models, due to their simplicity and effectiveness. To construct a differentially private decision tree, Friedman and Schuster [15] employ a carefully designed boundary selection mechanism to ensure that the partitioning at every node in the tree satisfies differential privacy. Mohammed et al. [32] attempt to tackle the problem from another angle, by merging the records before the beginning of the decision tree construction algorithm. Association rule and frequent pattern mining is another popular data mining technique, which has proven its value in the analysis of commodity transactions. To avoid possible privacy leakage when publishing association rules and frequent patterns, Bhaskar et al. [4] present a differentially private approach which applies the exponential mechanism to the identification of top frequent patterns. McSherry and Mironov [30] address the privacy issue within the Netflix prize dataset and discuss the possibility of tackling the problem in [33] using differential privacy techniques. Ding et al. [9] solve the privacy problem in a data warehouse, i.e., data cube publication.

In the machine learning literature, Chaudhuri et al. [5, 6] design a differentially private regression method by injecting noise into the objective function instead of the raw data. Their method is applicable to regression problems in which the objective function satisfies a property called strong convexity. A similar technique was independently proposed by Rubinstein et al. [37], which focuses on classification methods using the support vector machine with kernels satisfying l-Lipschitz continuity. In [22], Lei presents a simple solution to regression problems under differential privacy, which generates a differentially private histogram after partitioning the data space with a grid. His analysis shows that the regression accuracy is well preserved when the granularity of the grid is appropriately selected.

2.5 Open Problems

Although the notion of differential privacy has attracted attention in quite a few areas, many open problems remain, especially in data management tasks involving large-scale sensitive personal information. In the following, we list some open problems that we believe are important and deserving of additional attention from researchers.

First, the physical meaning of the privacy budget ϵ is unclear for real database users. While ϵ was originally derived from the mathematical / probabilistic domain, it is difficult to quantitatively measure the strength of the privacy protection provided by differential privacy with a specific value of ϵ from a practitioner's point of view. This makes it hard for ordinary users to select the appropriate value for ϵ for privacy protection, while maximizing the utility of the analysis results produced by the different privacy mechanism. Second, most existing studies focus on query processing on *static* databases. It is more difficult to design differential privacy protocols to handle arbitrary updates. Third, current differential privacy techniques assume a central database with a single owner. When the database is distributed or owned by different parties, e.g., by different Internet service providers (ISPs), the problem of statistical data sharing becomes the key bottleneck for collaborative analysis tasks, e.g. detection of botnets across ISPs. It is thus interesting to investigate the possibility of using differential privacy to solve the problem. Finally, most existing studies on differential privacy assume a simple data model, such as statistics on single/multiple dimensional numerical spaces. To extend the applicability of differential privacy, novel mechanisms are required for more complicated data domains (e.g. graphs and strings) or complex query plans (e.g. recursive SQL queries).

3. PRESENTERS

Yin Yang is a research scientist at the Advanced Digital Sciences Center (ADSC) in Singapore, and a principal research affiliate in the Coordinated Science Lab at the University of Illinois at Urbana-Champaign (UIUC). His research interests include database security and privacy, keyword search, and spatio-temporal databases. Yin shares responsibility for the query processing and open problems sections of the tutorial.

Zhenjie Zhang is a research scientist at ADSC, where he joined ADSC's differential privacy project in 2010. He has published several papers on differential privacy at top venues in databases and computer security. His research interests include multimedia indexing, privacy-preserving data analysis and computational advertising. Zhenjie is responsible for the data mining section of the tutorial.

Gerome Miklau is an associate professor in the Department of Computer Science, University of Massachusetts at Amherst. He has published extensively in the area of data privacy, including several high-impact papers on differential privacy. He received the SIGMOD Dissertation Award in 2006 and an NSF CAREER Award in 2007. Gerome shares responsibility for the query processing and open problems sections of the tutorial.

Marianne Winslett is a professor in the Department of Computer Science at UIUC, and the director of ADSC. She is the principal investigator of a large A*STAR-funded project on differential privacy for biomedical data. She is an ACM Fellow and has received best-paper prizes for work on information security from USENIX Security, VLDB, and Storage Security and Survivability. Marianne is responsible for the motivation section of the tutorial.

Xiaokui Xiao is an assistant professor at the School of Computer Engineering, Nanyang Technological University (NTU). He has published extensively on database privacy, including k-anonymity, l-diversity, and more recently, differential privacy. He is a co-PI of the differential privacy project mentioned above, a recipient of the Young Scientist Award in Physical/Mathematical Science from the Hong Kong Institution of Science, and was designated a Nanyang Assistant Professor. Xiaokui is responsible for the basics section of the tutorial.

4. ACKNOWLEDGEMENT

Marianne Winslett, Yin Yang, Xiaokui Xiao and Zhenjie Zhang are supported by SERC Grant No. 1021580074 from Singapore's A*STAR. In addition, Xiaokui Xiao is supported by Nanyang Technological University under SUG Grant M58020016 and AcRF Tier 1 Grant RG 35/09. Gerome Miklau is supported by the NSF through grants CNS-1012748, IIS-0964094, and IIS-0643681.

5. REFERENCES

[1] http://search-logs.com/aol/about.

[2] L. Backstrom, C. Dwork, and J. M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW*, pages 181–190, 2007.

[3] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282, 2007.

[4] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, pages 503–512, 2010.

[5] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, pages 289–296, 2008.

[6] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12:1069–1109, 2011.

[7] G. Cormode, C. M. Procopiuc, D. Srivastava, and T. T. L. Tran. Differentially private publication of sparse data. In *ICDT*, 2012.

[8] G. Cormode, M. Procopiuca, E. Shen, D. Srivastava, and T. Yu. Differentially private spatial decompositions. In *ICDE*, 2012.

[9] B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In *SIGMOD Conference*, pages 217–228, 2011.

[10] C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[12] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum. Differential privacy under continual observation. In *STOC*, pages 715–724, 2010.

[13] C. Dwork, G. N. Rothblum, and S. P. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60, 2010.

[14] S. E. Fienberg, A. B. Slavkovic, and C. Uhler. Privacy preserving gwas data sharing. In *ICDM Workshops*, pages 628–635, 2011.

[15] A. Friedman and A. Schuster. Data mining with differential privacy. In *KDD*, pages 493–502, 2010.

[16] A. Ghosh, T. Roughgarden, and M. Sundararajan. Universally utility-maximizing privacy mechanisms. In *STOC*, pages 351–360, 2009.

[17] Hardt and Rothblum. A multiplicative weights mechanism for privacy preserving data analysis. In *FOCS*, pages 61–70, 2010.

[18] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 3(1):1021–1032, 2010.

[19] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8), 2008.

[20] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD Conference*, pages 193–204, 2011.

[21] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW*, pages 171–180, 2009.

[22] J. Lei. Differentially private m-estimators. In *NIPS*, 2011.

[23] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, pages 123–134, 2010.

[24] C. Li and G. Miklau. Efficient batch query answering under differential privacy. *CoRR*, abs/1103.1367, 2011.

[25] C. Li and G. Miklau. An adaptive mechanism for accurate query answering under differential privacy. *PVLDB*, 2012.

[26] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, pages 106–115, 2007.

[27] Y. D. Li, Z. Zhang, M. Winslett, and Y. Yang. Compressive mechanism: Utilizing sparse representation in differential privacy. In *WPES*, pages 177–182, 2011.

[28] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *ICDE*, page 24, 2006.

[29] A. Machanavajjhala, A. Korolova, and A. D. Sarma. Personalized social recommendations – accurate or private? *PVLDB*, 4(7):440–450, 2011.

[30] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *KDD*, pages 627–636, 2009.

[31] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007.

[32] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu. Differentially private data release for data mining. In *KDD*, pages 493–501, 2011.

[33] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.

[34] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84, 2007.

[35] S. Peng, Y. Yang, Z. Zhang, M. Winslett, and Y. Yu. DP-Tree: Indexing multi-dimensional data under differential privacy. In *SIGMOD Conference*, 2012 (poster).

[36] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *SIGMOD Conference*, pages 735–746, 2010.

[37] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *Journal of Privacy and Confidentiality*, 2011.

[38] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS*, page 188, 1998.

[39] R. Wang, Y. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. In *ACM CCS*, 2009.

[40] X. Xiao, G. Bender, M. Hay, and J. Gehrke. iReduct: Differential privacy with reduced relative errors. In *SIGMOD*, pages 229–240, 2011.

[41] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. In *ICDE*, pages 225–236, 2010.

[42] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. *TKDE*, 23(8):1200–1214, 2011.

[43] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu. Differentially private histogram publication. In *ICDE*, 2012.