Privacy Preserving Data Publishing: A Classification Perspective

A N K Zaman
School of Computer Science
University of Guelph
Guelph, ON, CANADA

Charlie Obimbo
School of Computer Science
University of Guelph
Guelph, ON, CANADA

***Abstract*—The concept of privacy is expressed as release of information in a controlled way. Privacy could also be defined as privacy decides what type of personal information should be released and which group or person can access and use it. Privacy Preserving Data Publishing (PPDP) is a way to allow one to share anonymous data to ensure protection against identity disclosure of an individual. Data anonymization is a technique for PPDP, which makes sure the published data, is practically useful for processing (mining) while preserving individuals sensitive information. Most works reported in literature on privacy preserving data publishing for classification task handle numerical data. However, most real life data contains both numerical and non-numerical data. Another shortcoming is that use of distributed model called Secure Multiparty Computation (SMC). For this research, a centralized model is used for independent data publication by a single data owner. The key challenge for PPDP is to ensure privacy as well as to keep the data usable for research. Differential privacy is a technique that ensures the highest level of privacy for a record owner while providing actual information of the data set. The aim of this research is to develop a framework that satisfies differential privacy standards and to ensure maximum data usability for a classification tasks such as patient data classification in terms of blood pressure.**

***Keywords*—privacy preserving data publishing; differential privacy**

I. INTRODUCTION

Increase in large data repositories in the recent past by Corporations and Governments have given credence to developing information-based decision-making systems through data-mining. For example, all California based, licensed hospitals have to submit person-specific data (Zip, Date of Birth, admission and release dates, principal language spoken etc.) of all discharged patients to the California Health Facilities Commission to make that data available for interested parties (e.g., insurance, researchers) to promote Equitable Healthcare Accessibility for California [1]. In 2004, the Information Technology Advisory Committee of the President of the United States published a report with the title Revolutionizing Health Care through Information Technology [2], to impose emphasis to implement a nationwide electronic medical record system to promote and to encourage medical knowledge sharing throughout the computer-assisted clinical decision support system. Publishing data is beneficial in many other areas too. For example, in 2006 Netflix (online DVD Rental Company) published 500,000 movie ratings data set from subscribers to encourage research to improve the movie recommendation accuracy on the basis of personal movie preferences [3]. From Oct 2012 Canada and US governments started a pilot project called ”Entry/Exit pilot project” [4] to share travellers (citizens and permanent residents of both countries) biographic data of people who cross the US/Canada border among The Canada Border Services Agency (CBSA) and the Department of Homeland Security (DHS). This is an example of data sharing between two governments.

Table I presents a raw data about patients, where, every row belongs to a single patient. After applying, generalization, anonymized data is published in Table II.

TABLE I: RAW DATA ABOUT PATIENT

Record ID	Sex	Age	Disease Code	Class
1	Female	33	15, 16, 31, 32	N
2	Female	60	15, 31	Y
3	Female	37	16	Y
4	Female	35	15, 16	N
5	Male	16	15	N
6	Male	36	16, 31	Y
7	Female	46	15, 16, 31, 32	N
8	Male	27	16, 31, 32	Y

TABLE II: ANONYMIZED DATA TABLE FOR PUBLICATION

Age	Sex	Disease Code	Class	Count
[15-60)	Female	1*	Y	2
[15-60)	Female	1*	N	3
[15-60)	Male	1*	Y	2
[15-60)	Male	1*	N	1
[15-60)	Female	1*, 3*	Y	1
[15-60)	Female	1*, 3*	N	2
[15-60)	Male	1*, 3*	Y	2
[15-60)	Male	1*, 3*	N	0

The taxonomy tree used for generalization Table I is given in Figure 1. The taxonomy tree is presenting two attributes age and disease code. The attribute age could be divided into two different calsses as 15 to 30 and 30 to 60. In a similar way, four different disease codes are generalized as 1* and 3*. In Table I, although there is no identifiable information (e.g. name or phone number) about any patient, still the privacy of patient is vulnerable due to background knowledge of a malicious user of the data set. For example, if a malicious user knows that the disease code 32 belongs to a Male patient, then it is easy to identify the record #8, as it is the only Male

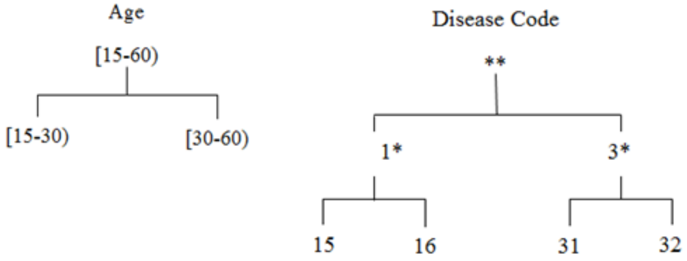

Fig. 1: Taxonomy tree for Attributes Age and Disease Code

has disease 32. On the other hand, after location that patient, the malicious user can also know that the targeted Male has diseases 16 and 31.

The rest of the paper is organized as follows: Section II surveys the related recent published work. Section III sates the problem statement. Section IV discusses the proposed system and experimental setup. Section V mentions the contributions of this research. Section VI presents the pseudocode of the proposed algorithm. Section VII concludes this paper.

II. Related Works

Researchers have proposed many algorithms for Privacy Preserving Data Mining (PPDM) and PPDP, however, not much is found in literature that addresses the privacy preservation to achieve the goal of classification [5]. The following section will discuss recent works on data anonymization for classification.

Iyengar [6] first wrote his paper on privacy of data and classification. He proposed usage based metrics (general loss metric, LM and Classification metric, CM) and showed by applying generalization and/or suppression, the anonymized data is still usable for classification tasks.

A bottom-up anonymization method was proposed by Wang et al. [7], that is only able to handle categorical data for the purpose of the data classification task. Later, the same authors introduced another method called TDS (top-down specialization method) [8] and then another improvement of TDS called TDR [9] (Top-Down Refinement) which is capable to handle both categorical and numerical values for data anonymization.

Lefevre et al. [10] proposed an algorithm called Mondrian and its improved version named as InfoGain Mondrian [11] to address various anonymized data processing including classification. InfoGain Mondrian showed better performance than TDS algorithm, and it is considered as one of the benchmark algorithms for anonymized data classification task. k-anonymous decision trees [12] based algorithms was proposed by Friedman et al. in 2008. Sharkey et al. [13] also proposed a method that generated pseudo data by following the decision tree model.

Kisilevich et al. [14] presented a multi-dimensional hybrid approach called compensation which achieved privacy by utilizing suppression and swapping techniques. The authors investigated data anonymization for data classification by satisfying k-anonymization. They claimed that their work resulted in better classification accuracy on anonymized data. If suppression techniques are applied, then one of the major drawbacks is that sparse data results in high information loss [15].

Li et al. [16] proposed and demonstrated the k-anonymity based algorithm. They utilized global attribute generalization and local value suppression techniques to produce anonymized data for classification. Their algorithms showed better classification performance compared to InfoGain Mondrian [11].

Table III presents some recent works published on data anonymization and classification. Still most published works are using k-anonymity based algorithms.

TABLE III: CLASSIFICATION MODEL USED BY DIFFERENT PRIVACY PRESERVED ALGORITHMS

	K-Anonymity	$\in$-differential privacy	
[17]	Y		Hierarchical Conditional Entropy-based Top-Down Refinement (HCE-TDR)
[18]	Y		SVM Classifier
[14]	Y		Decision tree
[19]		Y	Decision tree

Fung et al. [5] presented different existing anonymization based algorithms in their paper. It is seen that most of the algorithms can handle only two attack models. So, more efficient algorithms are needed to ensure the privacy of a dataset donor and/or owner.

III. Problem Statement

The key challenge for PPDP is to ensure individuals privacy as well as to keep the data usable for research. The aim of this research is to develop a framework that satisfies differential privacy standards and to ensure maximum data usability to deal with the classification task for knowledge miners. The core benefit of this work is to ensure the ease of availability of high quality data to promote collaborative scientific research to achieve new findings.

IV. Proposed system and experimental design

The objective of this research work is to develop a stable PPDP system by addressing specific research issues for publishing anonymized data. One of the primary goals is to publish useful data set to satisfy certain research needs, e.g., classification. The following sections will discuss research questions and the proposed system to be developed:

A. *Privacy Constraint*

One of the main objectives of the proposed system is to preserve individual's privacy. k-anonymization based algorithms are susceptible to attacks that may use individual "contributor's" background and link them to the repository, to expose which tuples belong to the given individual. They are, therefore, vulnerable to record-linkage and attribute-linkage attacks. In literature, it is also found that many privacy preserving models also suffer from table linkage and probabilistic attacks. In the proposed system, differential privacy ($\in$-differential privacy) privacy will be used that is capable to protect date published from the above mentioned attacks.

Differential privacy is a new algorithm that provides a strong privacy guarantee. Partition-based [20] [21] privacy models ensure privacy by imposing syntactic constraints on the output. For example, the output is required to be indistinguishable among k records, or the sensitive value to be well represented in every equivalence group. Instead, differential privacy makes sure that a malicious user will not be able to get any information about a targeted person, despite of whether a data set contains that persons record or not. Informally, a differentially private output is insensitive to any particular record. Therefore, while preserving the privacy of an individual, the output of the differential privacy method is computed as if from a data set that does not contain her record.

Differential privacy also prevents linking a victims sensitive information from an adversary has capturing may be interested in quasi-identifiers.

1) Definition: $\in$-differential privacy: Let us consider two data sets $DB1$ and $DB2$ that differ only in one element. For both data sets $DB1$ and $DB2$, a certain query response Rs should be the same as well as satisfy the following probability distribution Pr:

$$\frac{Pr(An(DB1) = Rs)}{Pr(An(DB2) = Rs)} \leq e^{\epsilon} \tag{1}$$

where, An presents an anonymization algorithm. The parameter $\epsilon > 0$ is chosen by the data publisher. Stronger privacy guarantee could be achieved by choosing a lower value of ϵ. The values could be 0.01, 0.1, or may be $\ln 2$ or $\ln 3$ [22]. If it is a very small ϵ then

$$e^{\epsilon} \approx 1 + \epsilon \tag{2}$$

To process numeric and non-numeric data with differential privacy model, following techniques will be needed.

B. Laplace Mechanism

Dwork et al. [23] proposed the Laplace mechanism to add noise for numerical values and ensure differential privacy. The Laplace mechanism takes a database DB as input and consists of a function f and the privacy parameter λ. The privacy parameter λ specifies how much noise should be added to produce the privacy preserved output. The mechanism first computes the true output $f(DB)$, and then perturbs the noisy output. A Laplace distribution having probability density function

$$pdf(\frac{x}{\lambda}) = \frac{1}{2\lambda} e^- |x|/\lambda \tag{3}$$

generates noise, where, x is a random variable; its variance is $2\lambda^2$ and mean is 0. The sensitivity of the noise is defined by the following formula:

$$\hat{f}(DB) = f(DB) + lap(\lambda) \tag{4}$$

where, $lap(\lambda)$ is sampled from Laplace distribution. The mechanism

$$\hat{f}(DB) = f(DB) + lap(\frac{1}{\epsilon}) \tag{5}$$

ensures ϵ-differential privacy.

C. Exponential Mechanism

McSherry and Talwar [24] proposed an exponential mechanism to handle non-numeric data. This mechanism takes as input, a data set DB that encompass an output range, τ, privacy parameter, ϵ and a utility function

$$u : (DB \times \tau) \rightarrow R$$

to produce an output, $t \in \tau$, having real value score where a better utility is proportional to higher score. A probability distribution is introduced by this mechanism which samples an output over the range τ. The sensitivity of the function is defined by

$$\Delta u = max_{\Delta_{(t,DB,D\hat{B})}} |u(DB, t) - (D\hat{B}, t)| \tag{6}$$

The probability associated with every output is proportional to

$$e^{\frac{\epsilon u(DB,t)}{2\Delta u}} \tag{7}$$

$i.e.$ the higher score should be chosen exponentially with an output.

D. Anonymization

Ideas of interactive and non-interactive [19] anonymization techniques are as follows. The non-interactive approach is chosen for this research work. In literature, differential privacy method is widely used in an interactive framework [25] [23] [26] [27]. In case of a non-interactive framework, anonymized data set is published by the owner for public use after changing the raw data to an anonymous form. In this research the non-interactive framework is adopted. This is due to the fact that this approach has a number of advantages over its counterpart (interactive approach). In an interactive framework, noise is added to every query response to ensure privacy. To ensure privacy, a database owner has a privacy constraint to answer queries with a certain limit before he/she has to increase the noise level to a point that the answer is no longer useful. Thus, the data set can only support a fixed number of queries for a given privacy budget. This is a critical problem when there are a large number of data miners that want to access the data set, because each user (data miner) can only allow to ask a small number of queries. Even for a small number of users, it is not possible to explore the data for testing various hypotheses.

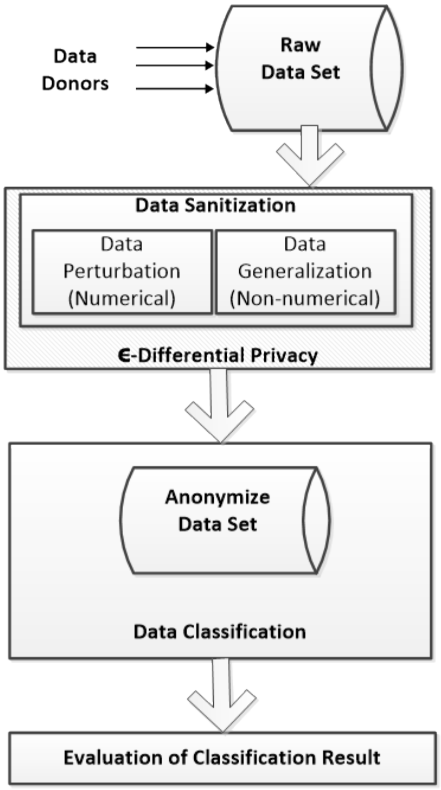
Data Donors
Raw Data Set
Data Sanitization
Data Perturbation (Numerical)
Data Generalization (Non-numerical)
ϵ-Differential Privacy
Anonymize Data Set
Data Classification
Evaluation of Classification Result

Fig. 2: Data Flow Diagram of the Proposed System

E. Generalization

Definition: Let

$$DB = r_1, r_2, ..., r_n$$

be a set of records, where every record r_i represent the information of an individual with attributes

$$A = A_1, A_2, ..., A_d$$

It is assumed that each attribute A_i has a finite domain, denoted by $\Omega(A_i)$. The domain of DB is defined as

$$\Omega(DB) = \Omega(A_1) \times \Omega(A_2) \times ... \times \Omega(A_d)$$

To anonymize a data set DB, the process of generalization takes place by substituting an original value of an attribute with a more general form of a value. The exact general value is chosen according to the attribute partition.

Figure 2 represents the data flow diagram of the proposed system. In the first step, the raw data is collected from the data donors', for example, in case of a medical data, patients of a hospital would be the data donors. After collecting the raw data, a sensitization algorithm is applied on the data to preserve individual's privacy. Finally, the sanitized data is released for the research community for further processing for knowledge mining.

V. Contributions

The following sections will discuss the important contributions of this research.

A. Securing Data Donors Privacy

By surveying the literature it is found that k-anonymy and various extension are susceptible to different attacks such as attribute linkage attack, background knowledge attack, table linkage attack and probabilistic attack. Differential privacy provides the strongest privacy guarantee and a differentially private output is insensitive to any particular record. Differential privacy model is able to protect all above mentioned attacks. In this research, differential privacy will be used along with generalization.

B. Handling High Dimensionality of Data Set

Measuring and Collecting information about an individual is becoming easier due to the improved technology. As a result, the number of attributes is rising and the size of the domain is growing exponentially. To handle that high dimensional data set, this research is going to implement the idea of multiple releases of anonymized data instead of publishing the whole data in a single time. A data set with different attributes could be utilized with different interest groups for their own research needs. Suppose there is a Table T contains data donors personal data, for example, (Employment Status, Gender, Age, Race, Disease, income). An interested group (for example a health insurance company) for the mentioned Table T, interested to classify data and wants to model the relation between disease and gender, age, income. Another interested group (for example a non-government organization (NGO) that works for different social services) may be interested to cluster data with attributes employment status, age, race. One solution is to publish the attributes in a single release Employment status, Gender, age, race, income for both interested groups; however, the problem with such release is that none of the group needs all released attributes. On the other hand, publishing more attributes together makes data donors vulnerable to malicious users. If the required information for different analysis is separate then publishing data for both cases at once may not necessary for those cases. Alternatively, publishing anonymized data based on the data recipients need is a better way to address the specific need of an analysis. Publishing multiple views of data, may be a more efficient way to handle high-dimensional data sets.

C. Re-usability of Data

Another goal of this research is to increase data re-usability through multiple publications of anonymous data. By the course of time, every day, data is collected and stored. So, multiple publishing of anonymized data gives an opportunity for data re-usability. For example, say the data owner has two sets of health care data for the years 1995-2004 and 2005-2014. One can publish the entire data set in an anonymous form in a single time. However, if any researcher wants data from the range 2004-2009, then the data owner could publish the anonymous data for the desired range instead of giving two different data sets with lots of redundant information.

D. Minimizing Redundancy in Published Anonymized Data

In literature, all the existing non-interactive privacy preserving models publish data once and made the data available for the interested parties. One of the major drawbacks of this paradigm is data redundancy. For this research, purpose-based (e.g. based on time span or based on specific attributes etc.) releases of anonymized data are aimed to address the classification task to avoid data redundancy.

VI. Pseudocode for the proposed algorithm

The following section presents the pseudocode for the proposed system:

```
1) START
2) Read the raw data set DB
3) Create purpose-based tailored data set TDB
     $a$). Based on certain time span [reflects section V(C)] (if NO go to b)
     $b$). Based on selection of attributes [reflects section V(B)]
4) Follow taxonomy tree for TBD
5) Apply generalization and ensure differential privacy:
     $a$). Apply Exponential Mechanism [case of non-numeric data]
     $b$). Apply Laplace Mechanism [case of numeric data]
6) Generate generalized privacy preserve data set GTDB.
7) Apply classification technique
8) Evaluation of classification accuracy.
9) END.
```

VII. Conclusion

In this paper the idea of privacy preserving data publishing is discussed for data classification purpose. The goal of this work is to implement a practical privacy preserving framework to keep privacy of an individual while keeping the anonymized data useful for the researcher. The core benefit of this work is to promote data sharing for knowledge mining. Differential privacy along with generalization creates a strong privacy guarantee and data utility for data miners.

References

[1] R. S. B. David and C. Schoenfelder, "California inpatient data reporting manual, medical information reporting for california," Office of Statewide Health Planning and Development, Tech. Rep., 09 2013. [Online]. Available: http://www.oshpd.ca.gov/hid/mircal/

[2] P. I. T. A. Committee, "Revolutionizing Health Care Through Information Technology," www.nitrd.gov/pitac/meetings/2004/, June 2004.

[3] I. Netflix, "Netflix Prize," http://www.netflixprize.com//index, February 2013.

[4] C. B. S. A. (CBSA), "Entry/Exit pilot project," http://www.cbsa-asfc.gc.ca/media/release-communique/2012/2012-09-28b-eng.html, July 2014.

[5] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010. [Online]. Available: http://doi.acm.org/10.1145/1749603.1749605

[6] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 279–288. [Online]. Available: http://doi.acm.org/10.1145/775047.775089

[7] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proceedings of the Fourth IEEE International Conference on Data Mining*, ser. ICDM '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 249–256. [Online]. Available: http://dl.acm.org/citation.cfm?id=1032649.1033461

[8] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proceedings of the 21st International Conference on Data Engineering*, ser. ICDE '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 205–216. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2005.143

[9] C. M. Fung Benjamin, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 5, pp. 711–725, May 2007. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.1015

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proceedings of the 22Nd International Conference on Data Engineering*, ser. ICDE '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 25–. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.101

[11] L. Kristen, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization techniques for large-scale datasets," *ACM Trans. Database Syst.*, vol. 33, no. 3, pp. 17:1–17:47, Sep. 2008. [Online]. Available: http://doi.acm.org/10.1145/1386118.1386123

[12] A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1007/s00778-006-0039-5

[13] P. Sharkey, H. Tian, W. Zhang, and S. Xu, "Privacy-preserving data mining through knowledge model sharing," in *Proceedings of the 1st ACM SIGKDD International Conference on Privacy, Security, and Trust in KDD*, ser. PinKDD'07. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 97–115. [Online]. Available: http://dl.acm.org/citation.cfm?id=1793474.1793482

[14] S. Kisilevich, Y. Elovici, B. Shapira, and L. Rokach, "Protecting persons while protecting the people," C. S. Gal, P. B. Kantor, and M. E. Lesk, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. kACTUS 2: Privacy Preserving in Classification Tasks Using k-Anonymity, pp. 63–81. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10233-2-7

[15] J. Liu and K. Wang, "Anonymizing transaction data by integrating suppression and generalization," in *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, ser. PAKDD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 171–180. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13657-3-20

[16] J. Li, J. Liu, M. M. Baig, and R. C. Wong, "Information based data anonymization for classification utility," *Data Knowl. Eng.*, vol. 70, no. 12, pp. 1030–1045, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2011.07.001

[17] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing classification data using rough set theory," *Knowl.-Based Syst.*, vol. 43, pp. 82–94, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2013.01.007

[18] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29 2009 - April 2 2009, Shanghai, China*, 2009, pp. 429–440. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2009.19

[19] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 493–501. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020487

[20] R. C. wing Wong, J. Li, A. W. chee Fu, and K. Wang, "(, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *In ACM SIGKDD*, 2006, pp. 754–759.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, 2007, pp. 106–115. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2007.367856

[22] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, 2011. [Online]. Available: http://doi.acm.org/10.1145/1866739.1866758

[23] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878-14

[24] F. McSherry and K. Talwar, “Mechanism design via differential privacy,” in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, 2007, pp. 94–103. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/FOCS.2007.41

[25] I. Dinur and K. Nissim, “Revealing information while preserving privacy,” in *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS ’03. New York, NY, USA: ACM, 2003, pp. 202–210. [Online]. Available: http://doi.acm.org/10.1145/773153.773173

[26] A. Roth and T. Roughgarden, “Interactive privacy via the median mechanism,” in *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, ser. STOC ’10. New York, NY, USA: ACM, 2010, pp. 765–774. [Online]. Available: http://doi.acm.org/10.1145/1806689.1806794

[27] A. Friedman and A. Schuster, “Data mining with differential privacy,” in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’10. New York, NY, USA: ACM, 2010, pp. 493–502. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835868