International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Case Study of Differentially Private in Big Data Publishing

Ibraima Djaló

School of Information Technology and Engineering, Tianjin University of Technology and Education, Tianjin, China

Abstract: ***Privacy preserving data publishing is the main concern in current days, because the data being published through internet has been increasing day by day. The big challenge of data distribution is balancing privacy protection and data quality, which are typically considered to be a couple of contradictory factors. It is especially useful for the data owner to publish data, which preserves privacy-sensitive information. The mostcommonly used privacy protection method is differential privacy (DP) protection. However, the use of DP algorithm is not easy for non-professionals. In this research work, several examples of DP were presented by using Laplace mechanisms (LM), and exponential mechanisms (EM). The rule is created by analyzing data sets based on the calculation of support and differential privacy confidence. All experiment done using python language.***

Keywords: Case Study of Differential Private in Big Data Publishing

1. Introduction

The problem of privacy-preserving data publishing (PPDP) has become increasingly important in recent years. We often encounter situations where a data owner wants to make definite data available without revealing private and confidential information. For example, this is the case for revealing detailed information about citizens (census), patients (health data), investors (financial data), and so on. The canonical case in PPDP is when the database can be modeled as a table, where each row can contain information about an individual (for example, status details or employment information).

Following many attempts to formally define the requirements of privacy, the current state-of-the-art solution is to seek the differential privacy guarantee [1]. Informally, this model requires that what can be learned from the released data is (approximately) the same, whether or not any particular individual was included in the input database. This model offers strong privacy protection and does not make any limiting assumptions about the power of the notional adversary: it remains a strong model even in the face of an adversary with much background knowledge and reasoning power. In this paper, we present the definitions and privacy techniques that monitor an important role in this work. For a comprehensive view of the field, here is a textbook treatment in and research can be found in [2]. The actual task of the data provider is to develop methods and tools for publishing data in more antagonistic environment, so that the data will be available to the needed people. Differential privacy as a privacy preserving mechanism [14, 15] for big data. Differential privacy (DP) was considered to manage protection dangers to avert undesirable re-distinguishing proof and other security dangers to people whose individual data is available in big datasets, while giving helpful access to information and we quantified the risk of differential privacy under temporal correlations by formalizing, analyzing and calculating the loss of privacy against opponents who have varying degrees of temporal correlation. In this paper opens up interesting directions for future research and combines our methods with previous studies that neglected the effect of limiting the temporal privacy leak.

2. Related Work

In an era of big data analysis and personal computing, collecting individual information is increasingly central to decision making across different domains. Meanwhile, the increase of privacy concerns prevents researchers form making full use of data. Past privacy breach report by Fung et al [16], Narayanan and Shmatikov [17**],** have shown that various ad-hoc approaches failed to anonymize public records “Linkage Attacks” (to identify personal records by linking different databases). The concept of differential privacy formalizes the idea that a “privacy” mechanism should not reveal whether any individual is included in the input or not, much less what their data are. It quantifies the privacy “cost” of an algorithm such that researchers can develop mechanisms which achieve a good trade-off between privacy and utility. Such requirements of privacy are of growing interest in the computer science and statistics communities due to the impact on individual privacy by real world data analytics.

Dwork et al. [18] proposed the first differentially private mechanism, the Laplace mechanism, that is based on output perturbation through adding noise. The immediate follow-up work focused on the constructions of differential privacy preserving methods which have good utility by reducing the amount of noise injected Nissim et al**.** [19], McSherry and Talwar [20**]** proposed the exponential mechanism that releases a response with probability exponential in a utility function describing the usefulness of each response, with the best response having maximal utility. Other generic privatizing mechanisms include Gaussian Dwork and Roth [21], Bernstein Aldá and Rubinstein and more. Chaudhuri and Monteleone [22], Chaudhuri et al. [2011] proposed an approach that can be employed for privatizing regularized empirical risk minimization by adding a random term to the primal objectives.

Rubinstein et al. [23**]** proposed a set of privacy preserving classification methods using support vector machines with

an output perturbation approach. Other learning algorithms including principal component analysis [Chaudhuri et al., [24], the functional mechanism [Zhang et al., [25] and tress [Jagannathan et al., [26] have also been adapted to maintain differential privacy.Kifer and Machanavajjhala[27] proved a no free lunch theorem, which defines non privacy as a game, to argue that it's not possible to provide privacy and utility without making assumptions about how the data are generated.

3. Preliminaries to Differential Privacy

Differential privacy is a relatively new notion of privacy and it is also one of the most popular privacy concepts, like most computer science theories, differential privacy has its roots in mathematics. Its algorithm foundation was selected by Dwork and Roth [7] in 2014 and more, but their mathematical foundations have been used before [8, 9]**.** This chapter provides basic background information, summary of these notes, definition of some key terms. From the basic concept of privacy, this section presents the notion, the implementation mechanism, and two important features of differential privacy, giving an example. In the end, the two noise generation mechanisms can be used in more detail.

3.1 What's Privacy?

In large data sets of sensitive personal information are becoming increasingly common, are no longer the domain only of census agencies for example: Hospitals, clinics, Social network, insurance companies and search engines have vast amounts of confidential data. These organizations face legal, financial, and moral pressures to make their data available to the public, but also face legal, financial, and moral pressures to protect the identities of individuals in their data set. Formal guarantees that violate privacy and usefulness are therefore of utmost importance.

One of most important parts of a privacy preservation technique is the definition when an individual's privacy is violated. Striking a balance between actually providing information to the user they can work with and privacy, making sure people's confidential information doesn't leak has always been hard to get right. Rocking too much to one side negatively affects one part, while perhaps helping the other; working openly with confidential data may be ethically questionable, but it can lead to better results.

3.2 Definition Differential privacy

Differential privacy is the dominant standard for privacy. A random algorithm that satisfies differential privacy offers protection to individuals by ensuring that their output is insensitive to changes caused by data from any individual entering or leaving the dataset. An algorithm can be made differentially private by applying one of several general-purpose mechanisms to randomize the calculation appropriately, for example by adding calibrated noise to the sensitivity of the quantity being computed, where the sensitivity captures how much depends on the quantity data from any individual [3]. Because of the obvious importance of protecting individual privacy by extracting population level inferences from data, differentially private algorithms have been developed for a wide range of machine learning tasks [4, 5].

An essential part of differential privacy is randomization. The intuition behind this is that an adversary can deterministic algorithm. This is done by executing the query on two neighboring datasets and observing the result. Because data sets differ in only one entry, the adversary can easily determine the value of that entry, violating privacy in the process.

3.3 Mechanism of Differential Privacy

Data privacy is an important factor that data owners must take into consideration when collecting, storing and publishing user data. This extends to publishing statistics on user data. In recent years, differential privacy has emerged as a popular privacy framework, thanks to its robust mathematical privacy guarantees. The three primary noise mechanisms in differential privacy are Laplace Distribution, Laplace Mechanism and Exponential Mechanism. The magnitude of noise alludes to privacy and overall sensitivity that will focus on two noise generation mechanism that can achieve privacy and differential: the Laplace mechanism, as previously discussed by different authors, which collects noise values, a Laplace distribution parameterized by the sensitivity of the query function.

3.4 Case Study of DP

a) The Data set

In this section, we present the Dataset that was used and the implementation analysis we applied in research developed in this thesis will consist of an implementation of the aforementioned differentially privacy mechanisms, after which an analysis of their performance will be made. The optimality of the three noise generation mechanisms, a work environment must first be set up. Python was used as the main programming language, because of its high level of extendibility and ease of use. To keep everything up-to-data the most recent stable release of Python was used. Differential privacy wouldn't exist if there wasn't data that needed to be privatized. A classic data set that is used for classification in machine learning is the "Adult Data Set" from the UCI Machine Learning Repository [12]. We have used one of the famous UCI's Adult dataset [13] which was acquired more than 30.000 instances (customer records) with the following 15 attributes (columns). According to the website, its purpose is to predict whether an individual's income exceeds $50k per year based on a collection of their attributes. There are 15 attributes, and they're listed in table 1, together with their possible values. The data was extracted from the 1994 Census database by Barry Becker.

Table 1: Adult Dataset

Attribute	Type	Values
Age	Numerical	
Workclass	Nominal	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlgwt	Numerical	Final sampling weight
Education	Nominal	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th,

		Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-num	Numerical	
Marital-Status	Numerical	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Nominal	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Nominal	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	Nominal	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex	Nominal	Female, Male
Capital-gain	Numerical	
Capital-loss	Numerical	
Hours-per-week	Numerical	
Native-Country	Nominal	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad &Tobago, Peru, Hong, Holland and Netherlands
Income	Nominal	<=50K, >50K

b) Case Study of Laplace Distribution

The ***Laplace distribution***, one of the earliest known probability distributions, is a continuous probability distribution named after the French mathematician Pierre Simon Laplace. Like the normal distribution, this distribution is unimodal (one peak) and symmetrical. However, it has a sharper peak than normal distribution. The Laplace distribution is the distribution of the difference of two independent random variables with identical exponential distributions (**Leemis, n. d**) [10]. It's often used to model phenomena with heavy tails or when data has a higher peak than the normal distribution. This distribution is the result of two exponential distributions, one positive and one negative; It's sometimes called the double exponential distribution, because it looks like two exponential distributions spliced together back-to-back.

The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function:

$$\text{Lap}(x|b) = \frac{1}{2b}\exp\left(-\frac{|x|}{b}\right)$$

The formula for the Laplace Distribution contains two parameters that have an influence on the dispersion:

- This value corresponds to the x-coordinate of the center of the graph's peak.
- b, also known as the scale parameter. Its value must be greater than zero, and it affects the area under the curve is centered around the peak. See Figure 1 for the impact the scale parameter has on the graph. The red line shows the Laplacian Distribution when beta=0,5, the green line for beta=1, the blue line for beta=2, the light blue line for beta=3 and the red line for beta=4.

The variance of this distribution isσ^2 =2b^2. We will sometimes write Lap(b) to denote the Laplace distribution with scale b, and will sometimes abuse notation and write Lap(b) simply to denote a random variable X ~ Lap(b). Laplace distribution is characterized by location 0 (any real number) and scale b (has to be greater than 0) parameters with the following probability density function:

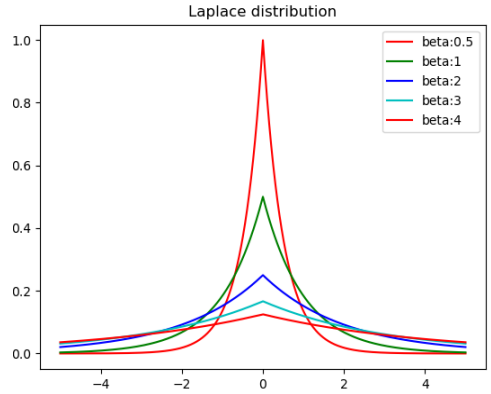

Figure 1: Laplace Distribution for beta=4, beta=1, beta=2, beta=3, and beta=0,5

c) The Laplace Mechanism

Laplace mechanism is the workhorse of differential privacy, frequently utilized in applications on numerical data. Its strength lies in its mathematical and computational simplicity, in contrast to other mechanism such as the Exponential mechanism. In spite of its popularity however, the Laplace mechanism lacks consistency in its output. Consider, for example, adding noise from the Laplace mechanism to a count query; negative results hold no meaning, yet are a valid output of the mechanism, occurring especially frequently for low numbered counts. As the name suggests the Laplace mechanism will just computation function, and perturb each coordinate with noise drawn from the Laplace mechanism distribution. The scale of the will be adjusted to the sensitivity of the function (divided by ε). Laplace mechanism used when the output is numerical given a dataset and the function.

3.5 Experiment Results

To check efficiency of differential privacy, we use different dataset with a different aspect of parameters through the python interface, all possible database functions and embedded data mean that once the performance of noise generation engines, they need to be implemented.

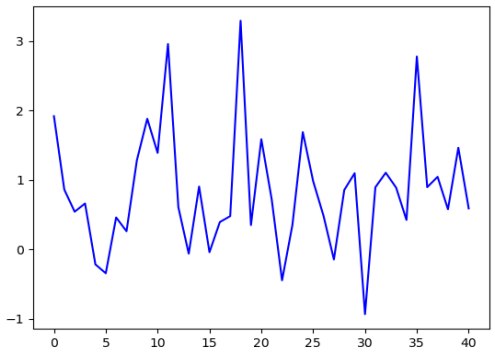

Figure 2: Experiment result the Laplace Mechanism

Figure 2 show the Laplace mechanism noise in graph.

Table 2: Experiment result the Laplace Noise

-0.3097794	2.03273239	1.56019272	0.07567007	0.76932996	0.96425559
1.51525539	1.00447818	1.30787974	0.67466642	0.33390887	1.37060093
1.53759992	0.1376869	0.59715679	0.29051631	0.31716623	0.4312335
1.83667553	0.36331844	0.76430748	1.43983438	0.48956341	0.17716806
2.84580121	1.01343489	1.9967973	2.50348004	1.73311782	0.07243189
1.60215755	0.82685951	2.64759557	0.16493273	1.81557373	0.63907356
1.84780331	1.0471751	1.76875144	1.80872075	0.72570086	0.54683782

Table 2 represent the values of noise for different values.

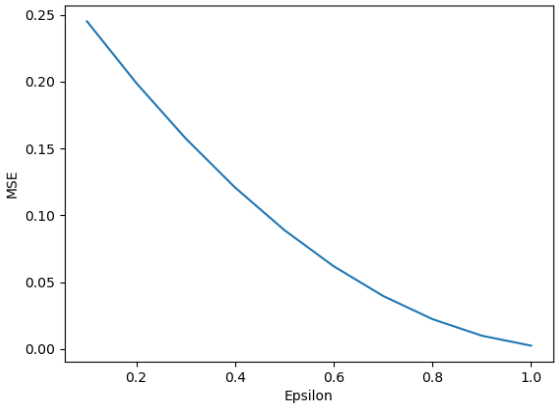

Figure 3: Experiment result The Laplace function Epsilon (Average Mean Square)

In figure 4 we represent the Laplace epsilon function in graph. X-axis represent the epsilon function and Y-axis represent the mean square epsilon value.

d) Case Study of Exponential Mechanism

The Exponential mechanism is a fundamental tool of differential privacy (DP) due to its strong privacy guarantees and flexibility. We study its extension to setting with summaries based on infinite dimensional output such as with functional data analysis, shape analysis, and nonparametric statistics. We show that the mechanism must be designed with respect to a specific base measure over the output space such as Gaussian process. One of the earliest mechanisms designed to satisfy ε-DP, is the Exponential mechanism, introduced by McSherry &Talwar (2007). It's uses an objective function, used for a (non-private) statistical or machine learning analysis, making it especially easy to link DP with existing inferential tool. A simple proof for proposition can be found in McSherry &Talwar (2007) [6]. Proposition (Exponential Mechanism: McSherry &Talwar (2007).

Experiment Results

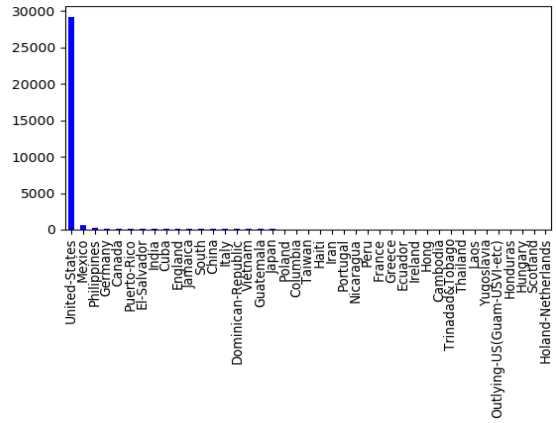

Figure 4: Experiment result the Laplace Exponential

Figure 4 show the result for Laplace exponential mechanism and for this we use country for distribution.

4. Conclusion

This research consists of studying the data privacy on the experiment and the discussion of the application of differential privacy mechanisms. We show that differential privacy protects data by adding a random noise taken from a Laplace distribution, and from the result obtained, we show that the mechanism used under differential privacy provides data privacy, for data set based on different values. Based on the result obtained, the Laplace mechanism is one of the solid approaches to obtain differential privacy, depending on the type of data set and the privacy parameter used. However, it has limited applicability to apply to data sets are not normally of a specific type, such as requests that are not numeric. To test this, a dataset was consulted in a python environment and noise was added with the Laplace distribution and the ladder. After finding the staircase distribution much slower, the noise calculation function was written as the Laplace sampling function. This led to an improvement in performance. So much that the ladder

mechanism was faster than the Laplace mechanism by a factor of approximately. This corresponded to the claims of the manufacturers of the ladder distribution, solidifying their grades.

References

[1] Dwork, Cynthia. "Differential privacy." *Encyclopedia of Cryptography and Security* (2011): 338-340.
[2] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211-407.
[3] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In *Theory of cryptography conference*, pp. 265-284. Springer, Berlin, Heidelberg, 2006.
[4] Chaudhuri, Kamalika, and Claire Monteleone. "Privacy-preserving logistic regression." In *Advances in neural information processing systems*, pp. 289-296. 2009.
[5] Bassily, Raef, Adam Smith, and Abhradeep Thakurta. "Private empirical risk minimization: Efficient algorithms and tight error bounds." In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464-473. IEEE, 2014.
[6] Dwork, Cynthia. "Differential privacy." *Encyclopedia of Cryptography and Security* (2011): 338-340.
[7] Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211-407.
[8] Thorve, Swapna, Lindah Kotut, and Mary Semaan. "Privacy Preserving Smart Meter Data." (2018).
[9] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In *Theory of cryptography conference*, pp. 265-284. Springer, Berlin, Heidelberg, 2006.
[10] Leemis, L.(n.d.). Laplace Exponential. Retrieved January 10, 2018 from http://www.math.wm.edu/~leemis
[11] Laplace Distributionhttps://blog.csdn.net/article/details
[12] httpps://archive.ics.uci.edu/ml/dataset/adult
[13] Lichman, M. "UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California, School of Information and Computer Science." *URL: http://archive. ics. uci. edu/ml* (2013).
[14] Jain P, Gyanchandani M, Khare Direndrapratap singh N, Rajesh L. A survey on big data privacy using Hadoop architecture. Int J Computer Sci Network Security (IJCSNS). 2017;
[15] Al-Zobbi M, Shahrestani S, Ruan C. Improving MapReduce privacy by implementing multi-dimensional sensitivity-based anonymization. J Big Data. 2017;
[16] Benjamin Fung, Ke Wang, Rui Chen, and Philip S Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4):14, 2010.
[17] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparsen datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008
[18] Cynthia Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.
[19] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Thirty-ninth AnnualACM Symposium on Theory of Computing*, pages 75–84. ACM, 2007
[20] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103, 2007
[21] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Rheingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. Technical Report 1411.2664, arXiv, 2014
[22] Francesco Aldà and Benjamin I. P. Rubinstein. The Bernstein mechanism: Function release under differential privacy. In *Proceedings of the 31st AAAI Conference onArtificial Intelligence (AAAI'2017)*, 2017
[23] Kamalika Chaudhuri, Claire Monteleone, and Anand D Sarwat. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12 (Mar):1069–1109, 2011.
[24] Benjamin I P Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: privacy-preserving mechanisms for SVM learning. *Journalof Privacy and Confidentiality*, 4(1):4, 2012
[25] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winglet. Functional mechanism: regression analysis under differential privacy. *Proc. VLDBEndowment*, 5(11):1364–1375, 2012
[26] G Jagannathan, K Pillaipakkamnatt, and R N Wright. A practical differentially private random decision tree classifier. In *IEEE International Conference on DataMining Workshops*, pages 114–121, dec 2009
[27] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Managementof data*, pages 193–204. ACM, 2011.