Research Article
Journal of Research and Review in Science, 62-67
Volume 6, December 2019
DOI:10.36108/jrrslasu/9102/60(0190)

ORIGINAL RESEARCH

DIFFERENTIAL PRIVACY: A NON-STOCHASTIC APPROACH TO PRIVACY PARAMETER GENERATION

Adetunji Adewole[1], Olawunmi Olayiwola[1], Stanley Udeh[1]

[1]Department of Computer Science, Faculty of Science, University of Lagos, Nigeria

Correspondence
Adewole Adetunji Philip, Department of Computer Science, Faculty of Science, University of Lagos, Nigeria.
Email: padewole@unilag.edu.ng
Phone: +234(0)8033938277

Abstract:

Introduction: The need for information continually necessitates the gathering and analysis of data. Information is generated from data supplied by individuals and as much as individuals are willing to supply their data, they are much more concerned about their privacy. Anonymization was initially thought of to provide the needed privacy but was discovered that anonymization is prone to linkage attack. The privacy parameter plays a major role in ensuring the privacy of an individual in a database as well as in maintaining an accurate or near accurate information. A high value of this parameter may lead to inaccurate data and a low value of this parameter may lead to exposition of individual in a database.

Aims: The aim of this research is to deduce a non-stochastic method to generate the privacy parameter that proffers a reliable solution to data privacy preservation.

Materials and Methods: The University of California Irvine (UCI) adult dataset was used. It contains 32,562 instances (individual records). The dataset was originally used to predict individuals who earn above a certain amount based on a census data collected within a region. This research work investigated differential privacy as a better data privacy method by analyzing three differential privacy mechanisms; the Laplace mechanism, exponential mechanism, and the median mechanism.

Results: The model produced 0.69 as the epsilon value which is the privacy parameter. The generated value from this research shows an improved result that ensures privacy while maintaining accuracy of information that is based on the dataset used.

Conclusion: The result obtained shows that threat of re-identification of individuals from a particular survey based on some query by a potential attacker are completely eliminated.

To Keywords: Data privacy, Differential privacy, Laplace mechanism, Privacy parameter, Anonymization

All co-authors agreed to have their names listed as authors.

1. INTRODUCTION

The world has become an explosion of information, data is available in its abundance and this is due majorly to cheap data storage and accessibility. Organizations, governments, health centers, and individuals have extracted large volumes of personal data for data analysis and other different purposes. These data are being used for research, to track users' behavior, recommend products or for national security and has created opportunities for researchers, companies, organizations and decision makers. For example, medical records help to track the spread of disease, prevent epidemics, discover hidden links between illnesses, disease prevention, and early detection and controlling of disease, etc. Differential privacy proffers a reliable solution to data privacy preservation. Differential privacy is a branch of statistics that aims to attain the widest range of data while achieving a robust, significant and mathematically accurate definition of privacy [1].

The availability of data creates new opportunities and help researchers and individuals with better data analysis, but protecting the privacy of each person's information in the dataset is crucial. There is a great chance of having the data of individuals compromised if anybody can explicitly distinguish a person from released data. In order to ensure that individual's identity contained in the information remains uncompromised then the technique to ensure data privacy preservation must be secured enough.

For example, where a trusted data custodian owns a database consisting of data with practical information about a specific individual, a privacy breach may occur when an outsider can infer this particular information. Even if the data creator publishes the anonymized version of the data, the background information exposing the identity of the individual can be preyed-on. Differential privacy provides a better privacy option for a scenario as stated above. Differential privacy tries to guarantee the protection of sensitive information about an individual irrespective of the background knowledge of the attacker [2].

2. MATERIAL AND METHODS

2.1 Mechanisms of Differential Privacy

2.1.1 Laplace Mechanism

Laplace mechanism is known as one of the most basic mechanisms in differential privacy [1]. The mechanism involves adding random noise that adjusts to the Laplace distribution with mean 0 and scales $\frac{GS(f)}{\varepsilon}$ and adds independently to each query response, thus making sure that every query is perturbed appropriately where *GS(f)* is known as the global sensitivity of *f*; which is a measure of the difference between the query results of the neighboring databases used in the differential privacy mechanism.
It suffices here to say that to analyze the Laplace mechanism we first need to define the Laplace distribution. The Laplace distribution is hereby defined in Equation 1:

$$f(x|\theta,\lambda) = \frac{1}{2\lambda} exp\left(-\frac{|x-\theta|}{\lambda}\right) \quad \text{.....Equation 1}$$

Laplace distribution is characterized by location θ (any real number) and scale λ (has to be greater than 0) parameters with the probability density function (Equation 1).

The Laplace mechanism is defined thus:
For a given function $f: \mathcal{D}^n \longrightarrow \mathbb{R}^k$ the Laplace mechanism is defined as:

$$\mathcal{ML}(x, f(.), \varepsilon) = f(x) + (Y1, \ldots, Yk) \quad \text{...Equation 2}$$

where Y_i are random variables from the definition in Equation 1.

Figure 1 figuratively explains the working principles of the Laplace mechanism.

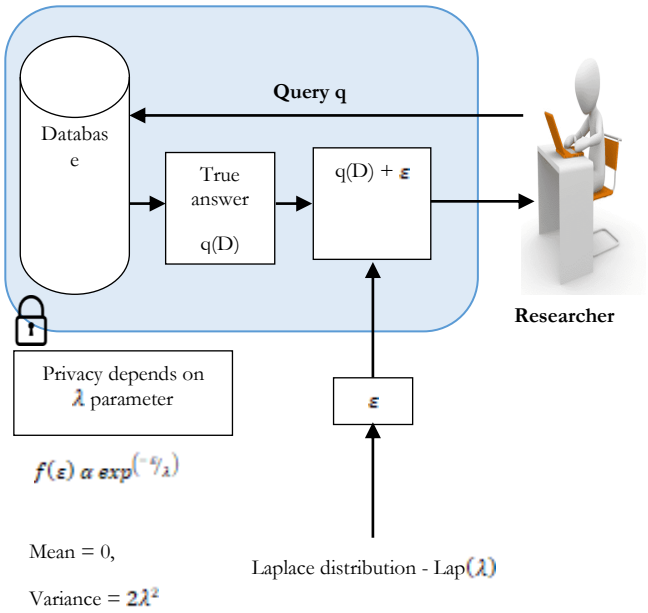

Fig 1: The working principles of Laplace mechanism

2.1.2 Exponential Mechanism

Given a quality function $q : \mathcal{D}\ X\ \mathcal{R} \longrightarrow \mathbb{R}$ the exponential mechanism selects an output from d with n elements from domain $\mathcal{D}$ and an arbitrary range $\mathcal{R}$, based on the score which represents the quality of *r* in *d*. The final output would be close to the ideal choice on *q* since the mechanism appoints exponentially higher probabilities of being selected to the higher outputs.

The formal definition for the exponential mechanism according to [3] is seen in Equation 3 and Equation 4:

Given a database $D \in \mathcal{X}^n$ and a quality function *q* with respect to *D*, the global sensitivity of s be

$$GS_q = \max_{D_1, D_2, r \in \mathcal{R}} |q(D_1, r) - q(D_2, r)|$$

....Equation 3

and query range $\mathcal{R}$, the exponential mechanism $M_E(D, q, \mathcal{R})$ gives the output $r \in R$ based on the probability:

$$Pr[M_E(D, q, \mathcal{R}) = r] \propto exp(\frac{\varepsilon q(D,r)}{2GSq})$$ ....Equation 4

From the equations above, we could say that the exponential mechanism is useful for functions that do not return a real number as well as when perturbation leads to invalid outputs.

2.1.3 Median mechanism

The median mechanism is an interactive differentially private mechanism that answers arbitrary predicate queries $f_1, \ldots, f_k$ that arrive on the fly without the future knowledge queries, where *k* could be large or even super-polynomial. It performs much better than the other mechanisms (for example, Laplace Mechanism) when it comes to answering more queries exponentially and gives fixed constraints. Theoretically, the mechanism is suitable for defining and identifying the equivalence of queries in the interactive setting [4]. The median mechanism is defined by Equation 5,

$$r_i = \frac{\sum_{S \in C_{i=1}} exp\left(-\epsilon^{-1} |f_i(D) - f_i(S)|\right)}{|C_{i-1}|}$$ .....Equation 5

Where *D* is the domain database, *S* is a subset of *D* and C_i are elements of the domain.

2.1.4 The Privacy loss parameter

Choosing a value for ε can be thought of as tuning the level of privacy protection required. This choice also affects the utility or accuracy that can be obtained from the analysis. A smaller value of ε results in a smaller deviation between the real-world analysis and each opt-out scenario and is therefore associated with stronger privacy protection but less accuracy. For example, when ε is set to zero, the real-world differentially private analysis mimics the opt-out scenario of each individual perfectly. However, an analysis that perfectly mimics the opt-out scenario of each individual would require ignoring all information from the input and accordingly could not provide any meaningful output. Yet when ε is set to a small number such as 0.1, the deviation between the real-world computation and each individual's opt-out scenario will be small, providing strong privacy protection while also enabling an analyst to derive useful statistics based on the data.

Figure 2 shows the difference between two databases where an entity is excluded and included in the database and the eventual privacy exposure.

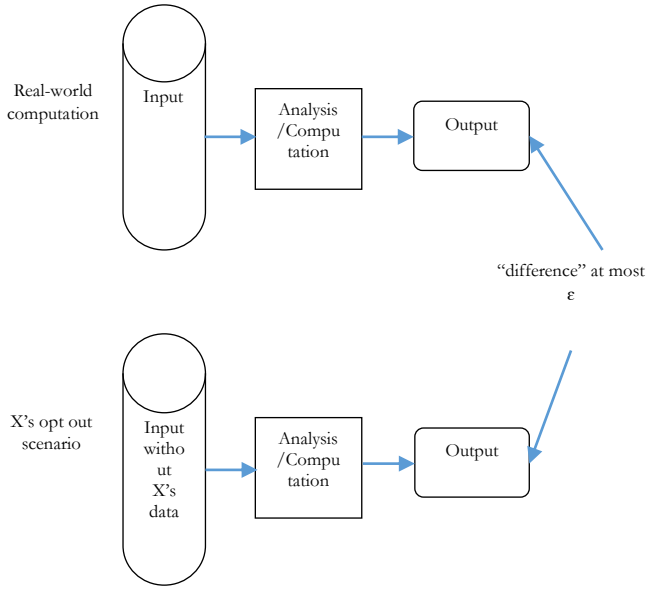

Fig 2: Maximum deviation between included and excluded individual data from a database

2.2 Setting of privacy parameter ε

The question on how to set the privacy parameter ε has been present since the introduction of differential privacy. However, setting the right value of ε has not been adequately addressed. The way in which ε influences the ability to identify an individual is not clear, although differential privacy has apparently stated that if it is hard to determine if a person is incorporated into a database, it is then definitely hard to know that individual's record. In the usual sense, the parameter ε in ε-differential privacy does not show what has been revealed about the person; it rather limits the outcome an individual has on the result.

The influence that ε has in determining an individual is less clear for queries that try to retrieve more general properties of data. However, for queries that ask specific information, for example, "Is Mr. X in the database?" ε directly relates to the exposure of the information.

In academics, the range of ε is usually between 0.01 to 1. However, industry implementations so far have set ε in the range of 1 to 10 (Apple set ε to 2, 4, and 8 in different applications [5]; Google uses ln(3)≈1.1 [6]; Census OnTheMap used 8.99 [7]). Using high ε's weakens the worst-case guarantee of differential privacy, but leaves in place other desirable properties of differential privacy such as composability, transparency, and quantifiability of privacy risk. Some works have been done using auction theory to set ε, but this is based on the assumption that individuals must be compensated for their privacy loss, which is often not true [8].

2.2.1 Laplace noise and differential privacy

From [9], it was noted that since the Laplace probability density function is symmetrical, hence, there is a need to consider the two possible directions in which erring can occur. The aim to describe the privacy protection level through a mathematical statement of the following form is given in Equation 6:

$$P[\hat{c} - wc < c < \hat{c} + wc] = p \quad \text{.....Equation 6}$$

Where c and $\hat{c}$ are the true value of a query and the noise-added value respectively, w and p are a measure respectively of the confidence interval width and the confidence level. For a fixed p, the larger the w the higher the level of protection privacy, since the recipient of the query output finds it more difficult to estimate the true value.

It can be verified further from applying the Laplace density function to arrive at Equation 7.

$$\varepsilon = -\frac{ln(1-p)}{wc} \quad \text{.....Equation 7}$$

The Equation 7, according to [9] measures the privacy parameter that will be used to add noise to the original query result before releasing the result to the data-seeker.

2.3 Proposed model

So far, research has shown that the level of exposure an individual has concerning re-identification from a database depends on how much information can be derived from a database. The number of columns specifies the various fields of the database. The number of rows in the database specifies the size of the database in terms of the number of individuals present in the database.

From Equation 1, we can modify the Laplace distribution as a function of the variables of the dataset. Thus we have,

$$f(x) = \frac{1}{2\varepsilon} exp(-\varepsilon x) \quad \text{.....Equation 8}$$

where x is the number of variables in the dataset and ε is the privacy parameter. Considering the entire dataset, the overall effect of adding differential privacy noise to the output will amount to the sum of all noisy output generated.

Therefore, applying Equation 7 gives:

$$\varepsilon = -\frac{ln(|1-N|)}{x} \quad \text{.....Equation 9}$$

where x and N are the number of variables and entries in the dataset respectively.

Studies have shown that the more we know about a database the more the inference that can be drawn from it [10]. The proposed algorithm to determine the value of epsilon was implemented using the Laplace algorithm shown below:

Algorithm 1:
Calculating required epsilon value by using Epsilon Generation Model (EGM)

1: load dataset D
2: function *EGM(p,q)* – sums the number of attributes *(p)* and number of entries *(q)* from the dataset
3: return *p, q* from *count(D)* – *p* = total number of attributes in dataset, *q* = total number of entries in dataset
4: $\varepsilon = \ln\left(\frac{|1-q|}{p}\right)$ - Calculate the epsilon value to be added
5: return ε - the privacy parameter to be used in the next stage
6: end function

Algorithm 2:
Laplace mechanism's algorithm

1: function LAPLACE$(D, Q : \mathbb{N}^{|x|} \longrightarrow \mathbb{R}^k, \varepsilon)$) - the Laplace based on the dataset, query and the epsilon value
2: $\Delta = GS(Q)$ - Calculate the global sensitivity
3: for $i \leftarrow k$ do
4: $y_i \backsim Lap\left(\frac{\Delta}{\varepsilon}\right)$ - Get the noise based on the ε and sensitivity from Laplace distribution
5: end for
6: return $Q(D) + (y_1, \ldots, y_k)$ - the noise added plus true value
7: end function

3. RESULTS AND DISCUSSION

The proposed model aims to improve the privacy level of the output of a query from a data-seeker to a database. Thus, as it has been presented above, an underlying architecture was used in implementing differential privacy as seen in Figure 3 and most importantly an interactive model of differential privacy has been followed. Each section of the design and their purpose are depicted in figure 3:

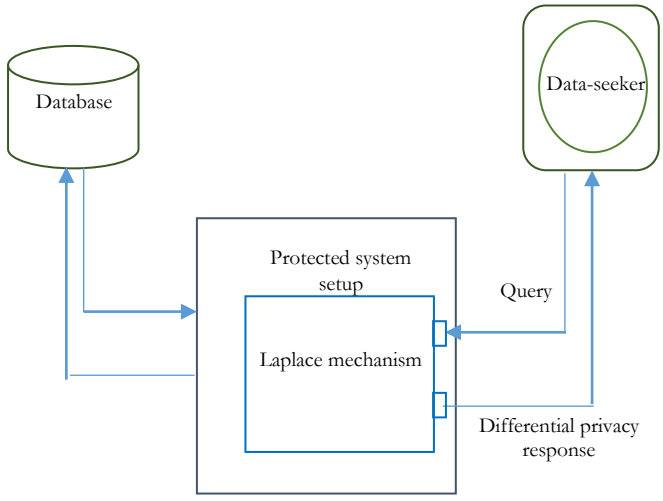

Fig 3: System design

3.1 Description of dataset

The University of California Irvine (UCI) adult dataset which was acquired from US Census data (1994) and was donated in 1996 [11] was used. It contains 32,562 instances (individual records). The dataset was originally used to predict individuals who earn above a certain amount based on a census data collected within a region.

3.2 Implementation design

The whole experiment was conducted on a Windows OS based system with Intel i7-2620m CPU processing capacity and 16GB RAM. To experiment with differential privacy, the researchers implemented a dataset with different scenarios. The scenarios are assumed to be dependent upon the utilization of the dataset in a way it makes sense when applying the experiment.

Here, the researchers considered the characteristics of the mean query of this dataset using aggregate function when implementing its differential privacy. An attempt was made to find out the average working hour per week for each job class that existed in the dataset. To get the data, the appropriate mean query was applied in Query 1 and Table 1 delivers the data before applying it to the Laplace mechanism for differential privacy.

```
SELECT
workclass, AVG(hours_per_week) as AVGHoursPerWeek
FROM staff
```

Query 1: getting the mean query

Table 1: Results from getting the mean query

SN	Work class	AVGHoursPerWeek
0	Federal-gov	41.38
1	Local-gov	40.98
2	Never-worked	28.43
3	Private	40.27
4	Self-emp-inc	48.82
5	Self-emp-not-inc	44.42
6	State-gov	39.03
7	Without-pay	32.71

3.3 Program design

The programming language used to implement this research was the Java programming language; which is an object-oriented language that has rich and robust features to help us implement this design.

NetBeans 8.2 IDE was used to create a graphical interface to display results as seen in Figure 4. Figure 4 shows the implementation process that generated the values as seen in Table 2.

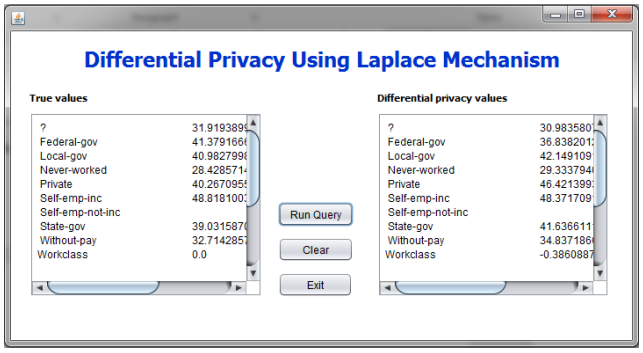

Fig 4: Result display after applying the Laplace mechanism

The table 2 displays the result of running the query based on various epsilon values.

Table 2: Results comparison of applying the Laplace mechanism for mean query

True Value	ε=2	ε=1	ε=0.5	ε=0.1	EGM (ε-value) = 0.693
41.38	41.28	40.87	42.41	63.60	48.42
40.98	40.56	41.01	40.99	41.79	38.96
28.43	28.86	22.66	29.89	25.52	30.32
40.27	39.86	39.70	38.59	55.52	40.43
48.82	48.87	49.41	47.24	58.90	54.62
44.42	43.99	45.03	51.56	39.07	47.89
39.03	38.15	37.95	41.37	41.46	40.16
32.71	32.94	34.57	32.06	30.85	32.28

From the results, it could be observed that the Laplace mechanism has different characteristics towards the different values of epsilon for the same query that run against the dataset. It is a worthy point to note that the noise to be added depends on the dataset of which ε is the privacy parameter that determines the level of noise to be added. The research also reveals that the

tradeoffs between utility and privacy are also influenced by the value of ε.

The value of the generated epsilon and eventual outcomes are display in the last column of Table 2. From the dataset number of columns and a total number of entries, it could be easily deduced from the proposed model that the value of epsilon is 0.693. Using this value for epsilon in the Laplace mechanism gives the final figures as the output to be released to the researcher.

Based on the results, it could be concluded that the proposed model created a better result that ensures privacy without jeopardizing accuracy in the process.

4. CONCLUSION

Differential privacy as a way of preserving individual privacy while ensuring accuracy of analysis has shown to be a very useful tool over the years both in research and industry. The threat of re-identification of individuals from a particular survey based on some query by a potential attacker are completely eliminated by the approach used in this study. The right choice for the privacy parameter has been a research question all the while. This research considered these factors when choosing what the privacy parameter value would be.

ACKNOWLEDGEMENTS

The authors acknowledge the invaluable inputs of the anonymous reviewers whose contributions gave rise to this paper.

COMPETING INTERESTS

We declare that there are no competing interests with anyone on this research work and the manuscript.

AUTHORS' CONTRIBUTIONS

Adewole Adetunji designed the study and wrote the first draft of the manuscript. Olawunmi Olayiwola managed the literature searches. Stanley Udeh wrote the protocol and managed the implementation process. All authors read and approved the final manuscript.

REFERENCES

1. Asseffa, S., & Seleshi, B. (2017). A Case Study on Differential Privacy (Dissertation). Retrieved from DiVA: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-136789.
2. Ashwin, M., Michael, H., & Xi, H. (2017). Tutorial: Differential Privacy in the Wild. In M. Ashwin, H. Michael, & H. Xi, Tutorial: Differential Privacy in the Wild (p. 50).
3. Frank, M., & Kunal, T. (2007). Mechanism Design via Differential Privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (pp. 94–103). Washington, DC, USA: IEEE Computer Society.
4. Roth, A., & Roughgarden, T. (2010). Interactive Privacy via the Median Mechanism. Proceedings of the Forty-second ACM Symposium on Theory of Computing (pp. 765–774). Cambridge, Massachusetts, USA: ACM.
5. Apple. (2018, October 10). Learning with Privacy at Scale. Retrieved from Apple Machine Learning Journal: https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.
6. Erlingsson, U., Pihur, V., & Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. 2014 ACM SIGSAC Conference on Computer and Communications Security (pp. 1054–1067). New York, NY, USA: ACM.
7. OnTheMap. (2008, November 21-22). Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-Use Data. 3rd IAB Workshop on Confidentiality and Disclosure.
8. Ghosh, A., & Roth, A. (2015, May 1). Selling Privacy at Auction. Retrieved from Games and Economic Behavior 91: https://doi.org/10.1016/j.geb.2013.06.013.
9. Maurizio, N., & Giuseppe, D. (2015). Differential privacy: An estimation theory-based method for choosing epsilon. arXiv preprint arXiv:1510.00917.
10. Vijay, A., & John, H. (2000). Research Advances in Database and Information Systems Security. Thirteenth working conference on Database Security. Seattle, Washington: Kluwer Academic Publishers.
11. Becker, R. K. UCI Machine Learning Repository. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/adult