Privacy preserving query model using inverse laplacian differential technique

K Sandhya Rani Kundra, J Hyma, P V G D Prasad Reddy and K Venkata Rao

1Assistant Professor, Department of Information Technology, GVPCOE (A), Vishakhapatnam. 
2Associate Professor, Department of computer science engineering, ANITS, Vishakhapatnam. 
3,4Professor, Department of Computer Science & Systems Engineering, Andhra University, Vishakhapatnam. 

E-mail: sandhyaranikks38@gmail.com

Abstract. Digitalization of data has proven its competent edge in the present digital era, leading to aggregation of mass amounts of data. This segregated data is made expendable by discovering interesting patterns made possible by “Data Mining”. Setting apart its perks, it possesses a discernible intimidation of outpouring personal information of data owners to anonymous resources. As always privacy is a prime issue of concern. On the other hand trading absolute privacy for loss of certainty of a query may not produce efficient patterns like in in-depth analysis of the source data sets. To maintain a balance among the above mentioned areas of interest “Privacy Preserving Data Mining (PPDM) using various privacy techniques is employed. The proposed work potentiates the working of privacy protection technique namely differential privacy using privacy parameter $\epsilon$ with heterogeneous adaptable method. In addition, this framework provides a basis for working with the inverse laplacian data distortion for better privacy protection in an interactive query model.

1. Introduction

Since there has been comprehensive amount of data expansion, there is an immense need to create methods by virtue of which one can work on the data-sets to gain information. Data Mining is one such technique of extracting required data in form of knowledge and can also be used in numerous ways [1] [2]. There are enormous applications raised by this area like weather forecasting, business, education, health sector etc. [3]. It is also evident that huge expansion of any technology leads to anomalies in real life situation. And thus there is a demand to make sure that whoever so uses such tools cannot use it in a wrong way.

More specifically, information gained by data mining can be used against the data owners [4] and leave them vulnerable to threats. Such mechanism can work against an individual, a firm or a community which can crop into public disclosure of sensitive data, business loss or prejudice against a community. This kind of obligation to privacy in such big data sets can be due to several reasons being legal reasons, personal choice, business concerns or medical reasons etc. Also intervening into private property is against the individual’s rights. As a field, data mining has introduced new concepts and algorithms [5] but it needs to be analyzed for privacy reasons and security breach [6].

To overcome such confidentiality issues new set of techniques which performs data mining along with a level of security so called privacy preserving techniques have been introduced [7] [8]. Among those large set of techniques an interactive model of privacy protection [9] [10] is chosen for the proposed
work. In this interactive framework, when every time a user poses a query on the data-set applies a mechanism to generate an output which does not impart private data to the person working on the database. In particularly a privacy technique which can alter the output by a small factor which makes it difficult to analyze the original data while the data-set remains intact. One such method is Differential Privacy [11].

This paper aims at exploring different techniques of providing privacy in an interactive model of the user data processing without hindering its performance and calculating a better way to get the job done. It will also speculate at all the significant points and methods of privacy which have endured over the past few years and then tender a final method taking all aspects into consideration.

In this paper sections are organized as follows, Section 2 emphasizes the notion of Differential privacy and $\varepsilon$-Differential Privacy Technique, Section 3 introduces Heterogeneous Differential Privacy. The proposed model along with the flow diagram described in section 4 followed by an algorithm in section 5. Result evaluation and Conclusion are mentioned in section 6 followed by Reference section in 7.

2. Differential privacy

Differential Privacy [11][12] is a technique exercised on Statistical Databases where the user is allowed to work using queries and the aim is that chances of diagnosing the actual records is belittle. One of the basic implementations of Differential Privacy technique is achieving $\varepsilon$-Differential Privacy [13].

2.1 $\varepsilon$-Differential Privacy Technique

This technique works by adding aptly chosen random noise to the result of a query asked by generating a noisy but, close to the answer, solution of the query asked. In the work published earlier [13] on Differential Privacy it has been stated to enact $\varepsilon$-Differential Privacy by adding a random noise whose magnitude is chosen on bases of query posed. The amount of noise added depends on the maximum change a single entity can have on the result.

Definition:- $f : D \rightarrow R^d$, the L1-sensitivity of $f$ is-

$$\Delta f = \max_{D1, D2} \| f(D1) - f(D2) \|$$

For all $D1, D2$ differing in at most one element.

Here, $\Delta f$ is the sensitivity of the function $f$.

There are divergent noise adding mechanisms such as Laplacian mechanism, exponential mechanism and posterior sampling used to achieve differential privacy. Laplacian and Inverse-Laplacian are proven to be the two mechanisms that add controlled noise to the data- set [14][15]. Here, the proposed model deals with the Inverse-Laplacian Mechanism and thus to draw a basic idea of differential privacy. As mentioned earlier in [14] the other modes of noise, such as the Gaussian Noise [18], can also be employed, but they may require a slight relaxation of the definition of differential privacy[18], hence the proposed work goes with the Inverse-Laplacian noise mechanism.

2.2 Inverse-Laplacian Noise

According to the work published in [14] on Laplacian Distribution, noise can be expressed as probability density function. Many differentially private methods add controlled noise to functions with low sensitivity [18]. The Inverse-Laplace mechanism [18] adds a noise from the Inverse-Laplace distribution [18], which can also be expressed as in equation 1.

$$\text{noise}(y) \propto \exp(-|y|/\lambda) \ldots \ldots \text{Equation 1}$$

Which has mean zero and standard deviation $\lambda$. Now in this case the output function of $A$ [18] is defined as a real valued function called as the transcript output by $A$[18] and is given in equation 2.

$$T_A(x) = f(x) + Y \ldots \ldots \text{Equation 2}[18]$$

Where $Y \sim \text{Lap}^{-1}(\lambda)$ and $f$ is the original real valued query or function we planned to execute on the database. Now clearly $T_A(x)$ can be considered to be a continuous random variable [18] given in equation 3 and equation 4 [15].

$$\frac{pdf(T_{A,D1}(x)=t)}{pdf(T_{A,D2}(x)=t)} = \frac{\text{noise}(t-f(D1))}{\text{noise}(t-f(D2))} \ldots \ldots \text{Equation 3}[18]$$
\[ \text{Lap}^{-1}(u, m, b_x) = m - b_x S \cdot sgn(u) \cdot \ln\left(1 - 2 \cdot (u)\right) \] 

\[ \frac{\Delta(f)}{\lambda} \text{ being the privacy factor } \epsilon \text{ which is at most } e^{\frac{|f(D_1) - f(D_2)|}{\lambda}} \leq e^{\frac{\Delta(f)}{\lambda}} \text{ [18]. Thus } T \text{ follows a differentially private mechanism (as can be seen from the definition above). It is a derived fact that in order to have } A \text{ as the } \epsilon \cdot \text{ differential private algorithm [18] we need to have } \lambda = \frac{1}{\epsilon}. \]

Final Value

\[ T_A(x) = \text{Original value}\left(f(x)\right) + \text{Lap}^{-1}(u, m, b_x) \]

Where \( \text{Lap}^{-1} \) – Inverse Laplacian Distribution,

\( u \) – Uniform (0,1)

\( m \) – Mean

\( b_x \) – Scaling Parameter \( \frac{\Delta(f)}{\epsilon} \), \( \Delta f \) is global sensitivity and \( \epsilon \) is the privacy budget.

even cloud merchant can’t get to the information either.

3. Heterogeneous Differential privacy

This section curtly familiarizes the concept of Heterogeneous Differential privacy, need of this type of mechanism and how to attain Heterogeneous Differential Privacy.

3.1 Need for Heterogeneous Differential Privacy

One of the major drawbacks of Homogeneous Noise addition is that it adds a fixed noise to each and every data set which tallies with the query, if every entity has same amount of noise added, if even one of the entries are known, the noise can be calculated easily by the adversary during data extraction. This leads to violation of privacy which is our prime interest.

Another method to add noise is using random noise addition. But sometimes an unacceptable level of noise generation results during adding random noise mechanism.

Thus, there must be another level of privacy added to this methodology to overcome aforementioned drawbacks. The work on Heterogeneous Differential Privacy discussed in [16], acknowledges the fact that Privacy requirements are not homogeneous across users and items from the same user. This concept of people having varied preferences can help us create a way of adding heterogeneous noise for the posed query.

3.2 How to achieve Heterogeneous Differential Privacy

To achieve this, need to perceive the privacy claims of the Users or Entities. As a matter of fact Privacy requirements not only differ among individuals but also from attribute to attribute.

Thus, this work proposes a model where one can divide Data-Set into groups based on their privacy requirements. So that the chunk of noise added now differs amongst the sub-groups. This makes it unyielding to find the amount of noise added as it isn’t uniform throughout.

3.3 A Three Question Model

In this the main focus is to create a method such that it provides heterogeneous privacy to the Data-Set. The first step is the removal of Personal Identifiable Information (PII). The queries must be statistical queries and should not demand any PII data.

A three question model approach, where each question induces a level of privacy has been proposed. The questions being:-

Q1. Does the DB Admin thinks it is private?
Q2. How is each attribute correlated?
Q3. Does the User thinks it is private?

The first question gives us a 2-d vector namely domain privacy vector.
The second question gives us a 2-d vector where it finds an attribute’s relation to every other attribute by calculating the correlation using the Pearson correlation factor. Thirdly and lastly it checks whether the user thinks the data is private or not for him. This value is saved in a vector named privacy choice vector.

For each of these questions we generate a binary value. Combining the answers to these three questions we get the values ranging from 000 to 111 (having all possible combinations which lead up to 8 values). For these values we choose among those which are highly sensitive and which are not such as 000 will always have lower privacy preference and 111 should have highest privacy preference. A sample data categorization into various privacy classes is shown below.

4. Proposed model

After the above classification is done and the, amount of noise to be added is separately calculated for each of the category and then added to each entity of the respective category. Once the noise is added we generate the output of the query, as requested.
This section discusses about the proposed model. Its main goal is to set the privacy budget ‘ε’ initially and then distort the data by employing Differential privacy with Inverse Laplace distribution. The level of privacy depends on the ‘ε’ value called as privacy budget. Lower values of ε give a good amount of privacy to the data i.e., if ε is set to ‘0’ then it is absolute privacy and the privacy differs with the change in ‘ε’ value correspondingly.

To achieve the heterogeneity factor to the data a three question model discussed in the previous sections will be used so the entire data will have three different classes (high, medium, and low). Then it uses Inverse of Laplace distribution to calculate the noise for all the three classes where different amount of noise will be added to different classes. A noise threshold parameter control is also introduced further to achieve second level of data modification by taking the highest deviated data similarly. By this not only the model achieved employing heterogeneity but also increases the utility rate of the data by adding low noise to the data which does not need high privacy. The amount of noise to be added is separately calculated for each of the category and then added to each entity of the respective category. Once the noise is added then final privacy protected query result is generated. The flow diagram of given proposed model is pictured in figure [1].

![Flow diagram of proposed model](image)

**Figure 1. Flow diagram of proposed model**

## 5. Algorithm

**Algorithm Inverse Laplacian Data Distortion**

**Input:** Dataset D, Query function Q(), HDD privacy choice matrix  
**Output:** Perturbed Answer Q(D')  
**Start:**
Step-1: Divide the data into High, Medium, Low Privacy classes

Step 2: Calculate the Noise Parameter N_i
\[ N_i = \text{Lap}^{-1}(\Delta F_i / \epsilon_i) \] //where \ i \in \text{highmedlow}

Step 2.1: For each data attribute \( d_i \in D \)
If (class (\( d_i \)) == high)
Then \( N_h = \text{Lap}^{-1}(\Delta F_h / \epsilon_i) \) //for all i \in high
Else If (class (\( d_i \)) == Med)
Then \( N_m = \text{Lap}^{-1}(\Delta F_m / \epsilon_i) \) //for all i \in medium
Else
\( N_l = \text{Lap}^{-1}(\Delta F_l / \epsilon_i) \) //for all i \in low

Step 4: Calculate Noise Threshold
\[ N_T = \sum_{i \in \text{highmedlow}} Max(\text{noise}) \]

Step 3: \( Q(D') = Q(D) + N_i + N_T \) //\forall i \in \text{highmedlow}

Step:  

6. Result analysis
This section valuates the ability of the proposed solution to acquire efficient data utility while preventing the individual privacy. First, various sample statistical queries posed are mentioned and the query results obtained in different methods like Homo-differential privacy, Heterogeneous-Differential Privacy are clearly given in table [1] and table [2]. To perform result evaluation Amazon Sales Database and Adult Data set [17] are used.

Dataset and Sample Query Description
D1: Amazon Sales Attributes present: 5 [Date, Amzn closing price, Amzn price return, KO closing price, KO price return]. No of records: 2518.
Query1: What is the average closing price for Amzn where daily percent return is greater than 1.00?
Query2: What is the minimum closing price for KO where daily percent return is between 0 to 1
Query3: What is the max closing price for Amzn
Query4: What is the average daily return price for Amzn
Query5: What is the minimum closing price for KO where daily percent return is less than 1.00?
Query6: What is the sum of all KO closing prices
Query7: What is the sum of all the Amzn closing prices where percent return is greater than 3
Query8: What is the maximum of all the KO percent return where KO closing price is greater than 20?

D2: Database Income Attributes present: 9 [Salary, Commision, Age, Elevel, Car, Zipcode, Hvalue, Hyears, Loan] No of records: 32519
Query1: What is the average salary for people without commission (i.e. commission = 0)
Query2: What is the maximum salary for people in the age limit 40 to 60?
Query3: What is the minimum salary for people who have only one car?
Query4: What is the avg loan for people living in zipcode 530009?
Query5: What is the sum of havalues where hyears is greater than 20?
Query6: What is the average age of people without loan (i.e. loan=0)
Query7: What is the maximum no of cars owned by an individual?
Query8: What is the average Elevel

| Table 1. Query evaluation with various privacy techniques |
|---------------------------------|
| Sales Data Set    | Original | Homo-Diff Privacy | Hetero-Diff. Privacy |
|-------------------|----------|-------------------|----------------------|

6
The query results of homogeneous and heterogeneous differential privacy techniques for the given statistical queries are given in table [1] and [2]. From the obtained results given in figure [2] and figure [3] it is noticed that the amount of deviation from the original result is slightly changed in a normal differential privacy technique, where as in heterogeneous differential technique this amount of deviation is high than compared to the previous results. In addition to this additional privacy protection, it is also evident that for any adversary, intruding into the original data and targeting a single row or a subset is made difficult than earlier.

![Comparison Plot of Statistical Queries](image.png)

**Figure 2.** Query result comparison on sales data

In the normal differential privacy technique as it is adding same amount of noise to all the records, gives an easy way to interpret all other remaining original values when adversary has information of any one single record presence.
Figure 3 Comparison of mean absolute deviation

Table 2. Query evaluation with various privacy techniques on income data set

| Query  | Original | Homo-Diff Privacy | Hetero-Diff. Privacy |
|--------|----------|--------------------|----------------------|
| Query1 | 112415.05| 112416.09          | 112419.1             |
| Query2 | 74991    | 74992.05           | 74995.11             |
| Query3 | 21703    | 21704.42           | 21708.61             |
| Query4 | 100000   | 100001.06          | 100004.2             |
| Query5 | 4535     | 4589.74            | 4675.93              |
| Query6 | 62.54    | 63.66              | 68.91                |
| Query7 | 20       | 21.01              | 25.04                |
| Query8 | 2.95     | 3.96               | 7.97                 |
| Mean   | 39216.19 | 39225.12           | 39238.1              |
| Mean Absolute Error | nil | 8.93 | 21.91375 |

Figure 4. MAE Evaluation on a sample query sequence
The proposed work analyzed the trade-off between accuracy and privacy and explored how the given solution is advantageous in protecting privacy. The randomly chosen statistical queries and the corresponding error deviation are represented in figure [4].

7. Conclusion and future work
This method synthesizes data distortion with heterogeneous differential privacy to provide us an auxiliary level of security. This method assorts the dataset entities into three classes each time a query is posed. This method can be used to further sub-divide the entities into more significant groups on the basis of questions asked and by altering the amount of noise added on the basis of answers. Also this method can be adopted for increasing the level of privacy by doing some other alterations to the “Three Question Model”, however a proper care has to be taken for not to denying the data usage. As a future work data modification with various privacy budgets for each individual privacy class can be verified.

References
[1] Chen M, Han J and Yu P 1996 Data mining: An Overview from a database Prospective, IEEE Trans 8, pp 866-83.
[2] Bharat Bhushan Agarwal and Sumit Prakash Tayal 2009 Data Mining and Data Warehousing Laxmi Publications Ltd.
[3] Bharati M Ramageri Data Mining Techniques and Applications, Indian Journal of Computer Science and Engineering 1 4301-305.
[4] Nikita Jain and Vishal Srivastava 2013 Data Mining Techniques: A Survey Paper IJRET, 2.
[5] Sadiq Hussain 2017 Survey on Current Trends and Techniques of Data Mining Research London Journal of Research in Computer Science and Technology 17 Issue 1.
[6] Agrawal R and Srikan R 2000 Privacy-preserving data mining In Proc SIGMOD00 pp 439-50.
[7] Li H S 2010 Study of privacy preserving data mining In Proceedings of the International Symposium on Intelligent Information Technology and Security.
[8] Jian Wang, Yongcheng Lou, Yen Zhao and Jiain Le 2009 A Survey on Privacy Preserving Data Mining, International Workshop pp 111-14.
[9] Khushbu Agrawal and Vandan Tewari 2017 Analysis of Privacy Preserving mechanisms for Outsourced Data Mining 27-29.
[10] Noman Mohammed, Rui Chen, Benjamin C M Fung and Philip S Yu 2011 Differentially Private Data Release for Data Mining San Diego California USA.
[11] Arik Friedman and Assaf Schuster 2010 Data Mining with Differential Privacy, Washington, DC, USA.
[12] Qi Li, Yuqiang Li, Guicai Zeng and Aihua Liu 2017 Differential Privacy Data Publishing MethodBased on Cell Merging IEEE 978-1- 5090-4429- 0/17.
[13] Cynthia Dwork and Adam Smith 2009 Differential Privacy for Statistics: What we Know and What we Want to Learn Journal of Privacy and Confidentiality 1 pp 135-54.
[14] Dwork C, McSherry F, Nissim K and Smith A 2006 Calibrating Noise to Sensitivity in Private Data Analysis.
[15] Rathindra Sarathy and Krish Muralidhar 2009 Differential privacy for numeric data in proceedings Joint UNECE/Eurostat work session on statistical data confidentiality.
[16] Mohammad Alaggan, ebastien Gambs and Anne-Marie Kermarrec 2016 Heterogeneous Differential Privacy Journal of Privacy and Confidentiality 7 p 127-58.
[17] Dua D and Karra Taniskidou E 2017 UCI MachineLearningRepository [http://archive.ics.uci.edu/ml]. Irvine CA: University of California School of Information and Computer Science.
[18] Differential Privacy (From Wikipedia, the free encyclopedia)
https://en.wikipedia.org/wiki/Differential_privacy