Analysis of Differential Privacy and study of its variation with privacy budget using real-world data.

Rakshit Chauhan
Department of Computer Science and Engineering, Delhi Technological University

Pooja Narula
Department of Computer Science and Engineering, Delhi Technological University

Shaurya Shekhar
Department of Computer Science and Engineering, Delhi Technological University

Manoj Kumar
Department of Computer Science and Engineering, Delhi Technological University

Abstract- The privacy of data is one of the most essential topics in the world of privacy and security. All the advancement in technology is rendered non-essential if the data cannot be secured and be protected from various organisations. Every person or organisation desires privacy, be it technological or otherwise. It is getting more and more complicated to preserve one's privacy and maintain the confidentiality of data due to numerous policies put forth by almost every organisation to exist. Differential privacy, which makes very careful assumptions about the adversary's past knowledge, has recently become a powerful element for privacy protection. Differential privacy has received attention ever since it was initially proposed and put forth, and it is now considered to be among the most potential concepts for privacy-preserving data release and analysis in many computer science and information technology domains. In this paper, we discuss the motivation for its introduction as a tool to replace other privacy methods, use databases and datasets for the comparison of data before and after using differential privacy on it. We also explore deep into differential privacy and replicate an attack on the data by an outsider to see if the data secured by differential privacy really is protected. Comparison of the probability distributions of data before and after applying differential privacy and the variation of these distributions is also reflected in this report.

Index Terms- Differential privacy, data privacy, smart noise, data analysis, privacy preserving

I. INTRODUCTION

Data privacy refers to the protection of sensitive or confidential information contained in datasets from unauthorized access, use, or disclosure. Names, addresses, Social Security numbers, and financial details are all part of this information, as well as private company information like secrets of trade and client information. Data privacy is important for several reasons. First and foremost, it helps to protect the privacy rights of individuals by preventing unauthorized access to their personal information. This can help to prevent identity theft, financial fraud, and other forms of harm that can result from the misuse of personal information. [1]

In addition to protecting individuals, data privacy is also important for businesses and organizations. Data breaches and other privacy violations can result in significant financial and reputational damage, as well as legal liability. By implementing strong data privacy policies and practices, businesses can protect themselves from these risks and ensure that they are in compliance with relevant laws and regulations.

Data privacy is also important for maintaining trust and transparency in the digital age. As more and more personal and sensitive information is collected and stored online, individuals are becoming increasingly concerned about the security and privacy of their data. By prioritizing data privacy and using better methods of practicing privacy and its policies, organizations can demonstrate their commitment to protecting the privacy of their customers and stakeholders.

Overall, data privacy is an essential component of responsible data management and is critical for protecting the rights and interests of individuals and organizations alike. By prioritizing data privacy and taking steps to safeguard sensitive information, businesses and organizations can build trust and demonstrate their commitment to responsible data stewardship.

Differential privacy [2] is a privacy-preserving technique that can help to protect the privacy of individuals whose data is contained in datasets. It accomplishes this by carefully introducing noise into the data while concealing any information about specific data points, preserving the data's statistical features. By doing so, differential privacy helps to protect the privacy of individuals by ensuring that their data is indistinguishable from other data points in the dataset.

There are several ways in which differential privacy helps to improve data privacy:

1. Protection against re-identification attacks: One of the main risks to data privacy is re-identification, where an attacker is able to link an individual to their data within a dataset. Differential privacy helps to protect against this type of attack by making it more difficult to identify individual data points within the dataset.
2. Protection against inference attacks: Another risk to data privacy is inference attacks, where an attacker is able to infer sensitive information about an individual from their data within a dataset. Differential privacy helps to protect against this type of attack by ensuring that the data released from the dataset is statistically similar irrespective of whether a specific individual's data is included or excluded.
3. Protection against membership attacks: Membership attacks involve an attacker trying to determine whether a specific individual's data is contained in a dataset. Differential privacy helps to protect against this type of attack by ensuring that the statistical properties of the data are consistent regardless of whether a specific individual's data is included or excluded.

Overall, differential privacy is an important tool for protecting the privacy of individuals whose data is contained in datasets. By incorporating differential privacy techniques into data analysis and sharing processes, organizations can help to ensure that sensitive information remains protected while still enabling meaningful analysis of the data.

II. THEORETICAL FRAMEWORK

Differential privacy is a concept in data privacy that provides a way to protect sensitive information while still allowing useful statistical analysis to be performed on the data. The idea behind differential privacy is to add random noise to the data in such a way that the statistical properties of the data are preserved, while the individual data points are protected. The fundamental theoretical foundation of differential privacy entails the definition of a privacy budget, which denotes the greatest amount of data an adversary can gather about a person by looking at the results of a computation. The privacy budget is often calculated using a factor known as epsilon, which reflects the most data an attacker may discover about a person by looking at a computation's result.

To achieve differential privacy, a randomized algorithm is used to compute a noisy version of the data that satisfies the privacy budget constraint. The randomized algorithm adds random noise to the data in a way that is calibrated to the privacy budget, so that the amount of noise added increases as the privacy budget decreases. The key property of differential privacy is that it provides strong privacy guarantees even in the presence of a powerful adversary who has access to auxiliary information. In particular, differential privacy ensures that the probability of learning any specific piece of information about an individual is only slightly increased by the presence of that individual's data in the dataset. There are several different mechanisms that can be used to achieve differential privacy, including randomized response [3], Laplace noise, and Gaussian noise [4]. These mechanisms are designed to add noise to the data in a way that is calibrated to the privacy budget, so that the level of noise added increases as the privacy budget decreases.

Overall, the theoretical framework of differential privacy provides a powerful tool for protecting sensitive information while still allowing useful statistical analysis to be performed on the data. By adding carefully calibrated noise to the data, differential privacy ensures that individual data points are protected while statistical properties of the data are preserved.

III. LITERATURE REVIEW

Practising Differential Privacy in Health Care: A Review [Dankar et Amam, 2013 TDP] [5]

Through the review paper, authors alluded to multiple interpretations of Differential Privacy, specifically referring to the epsilon as the "knowledge gain ratio from one data set over the other" to the power of the Euler's number. It further makes the assertion that the values used in the literature generally are 0.01, 0.1, ln(2), and ln(3) respectively. As a compendium of the various techniques used to conceal private data, the review paper highlights the various limitations, and contemporaneous models implementing differential privacy. An interesting element of the work done by Dankar et Amam is the inter-disciplinary approach to differential privacy and its limitations especially with respect to the legality and its efficacy as a privacy tools. Alternatives to the traditional differential privacy methods, such as (ε,τ)-Differential Privacy and (ε,δ)-Differential Privacy methods. Further the review work highlights mechanisms used for achieving the conditions for differential privacy. Intertwined with a specialized approach to the healthcare industry, the review examines the feasibility of the achieving total privacy and whether it's as necessary as posited.

Differential Privacy [Cynthia Dwork, ICALP 2006] [2]

In this landmark paper, Cynthia et al. brought to the table the "Dream" of differential privacy, as a useful notion to codify the meaning of privacy and to ensure analyses of the datasets aren't compromised in quality all the while protecting the individual privacy. Here Cynthia presents a simple method to ensuring Differential Privacy by the use of additive noise mechanisms, where the additive noise is bounded by the largest difference a particular entry could make on the result of a query function. Further the paper delves into the mathematical framework for Differential privacy and the core observations made therein, including the observation of adding symmetric exponential noise to each coordinate of the query function and the importance of choosing noise as a function of the sensitivity.

Differential Privacy: What is all the noise about? [Roxana Danger, arXiv 2022] [6]

In this paper, the author examines the concepts of Differential Privacy as it relates to ML, the use of GANs to generate synthetic data, which accords to the definition of Differential Privacy and various mechanism involved to achieve the same. The author then delves into the various DP techniques in ML and makes the mention of Ensemble-DP and DP-SGD, with the special mention of the PATE algorithm, which combines the elements of both the aforementioned techniques. Finally, the paper goes to explore the concerns for privacy in Federated Learning (FL), taking a closer look at the types of FL, and specifically focusing on the aspect of privacy as it relates to the Centralized and Decentralized FL (Using Blockchain for the Decentralized aspect)

IV. WORKFLOW

The project utilises two datasets, one for the application of differential privacy and is a census dataset while the second dataset is for the attack analysis in order to check the efficiency of differential privacy over varied value of epsilons. The Adult Census Dataset [7] has been formed with the data collected from the 1994 Census database. It contains 32561 instances, with fifteen attributes. The dataset which was used to demonstrate the attack was collected from the synthetic random data generator. The dataset contains 10000 instances and six attributes. In order to apply differential privacy on the adult census data, we have pre-processed the data by assigning unique numeric identification to every different option for the following attributes- 1) Age, 2) Gender, 3) Race, 4) Country.

The workflow can be broadly categorised into two parts- Applying differential privacy to various attributes of the dataset, and launching an attack or trying to read the dataset after adding SmartNoise.

A. Applying differential privacy to various attributes of the census dataset

Firstly, the dataset has been pre-processed, that is, the required attributes have been assigned integral records through Excel functions. The resultant data is then read as the csv file via the Python Pandas. For different attributes, the categories array has been created of the unique numerical values the particular attribute stores. The differential privacy has been implemented for various attributes using the SmartNoise library of the OpenDP framework of Python. For each of the attributes, a dp_histogram has been created which takes the data and results into an array of categories distribution, with each index of the array storing the noisy value for that particular value of the attribute. Thereafter, the true distribution array is formed for various attributes using the actual data in the Excel and comparison has been made between the true distribution and the noisy distribution using Python Pandas library for creating pie charts. The above process of comparison using pie charts has been carried out for different values of epsilon (0.01, 0.1, ln2, ln3) to observe the variety of deviation from true distribution.

B. Launching an attack on the dataset value after the smart noise has been added.

Random data is generated for the purpose of launching an attack and a person of interest is chosen, in our case it is the 10th record of the dataset. We print the true income value of the person of interest and calculate the true mean of income of the dataset by calculating the sum of income of all the records and dividing it with the number of records.

V. OUTCOME & ANALYSIS

Thereafter, the modified data is created by adding noise via OpenDP framework to each record and through this data, we access the income attribute of the person of interest in order to observe the deviation in the values. We also find the new mean of modified data in order to notice the deviation in mean values. The mentioned process of modifying the data is repeated for different values of epsilon- 0.01, 0.1, 1, ln2, ln3 and the income value is recorded and the mean value is calculated.

Upon categorising the various attributes into different groups and applying differential privacy on the same using different values of epsilons, we were able to generate the results in the form of pie charts of the distribution of the attribute data. Each attribute has five pie charts distribution corresponding to them- based on true values, for epsilon with value 0.01, 0.1, ln 2 and ln 3. The results for the age attributes are shown as follows-

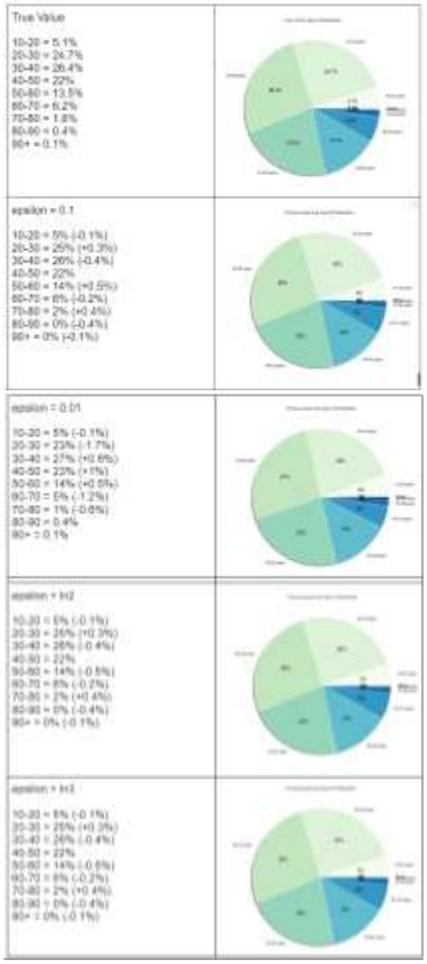
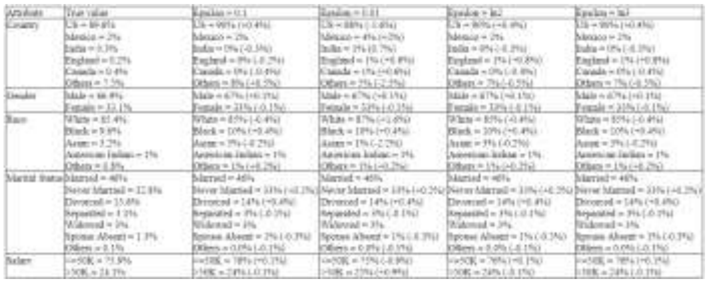


Figure 1: Distribution of various attributes

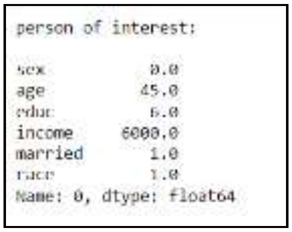
Similar comparisons have been done for other attributes namely salary, gender, country, race and marital status. The table for the same is below as follows.



Attribute	True value	Epsilon = 0.1	Epsilon = 0.01	Epsilon = ln2	Epsilon = ln3
Country	US = 89.6% Mexico = 2% India = 0.3% England = 0.2% Canada = 0.4% Others = 7.5%	US = 90% (+0.4%) Mexico = 2% India = 0% (-0.3%) England = 0% (-0.2%) Canada = 0% (-0.4%) Others = 8% (+0.5%)	US = 88% (-1.6%) Mexico = 4% (+2%) India = 1% (0.7%) England = 1% (+0.8%) Canada = 1% (+0.6%) Others = 5% (-2.5%)	US = 90% (+0.4%) Mexico = 2% India = 0% (-0.3%) England = 1% (+0.8%) Canada = 0% (-0.4%) Others = 7% (-0.5%)	US = 90% (+0.4%) Mexico = 2% India = 0% (-0.3%) England = 1% (+0.8%) Canada = 0% (-0.4%) Others = 7% (-0.5%)
Gender	Male = 66.9% Female = 33.1%	Male = 67% (+0.1%) Female = 33% (-0.1%)	Male = 67% (+0.1%) Female = 33% (-0.1%)	Male = 67% (+0.1%) Female = 33% (-0.1%)	Male = 67% (+0.1%) Female = 33% (-0.1%)
Race	White = 85.4% Black = 9.6% Asian = 3.2% American Indian = 1% Others = 0.8%	White = 85% (-0.4%) Black = 10% (+0.4%) Asian = 3% (-0.2%) American Indian = 1% Others = 1% (+0.2%)	White = 87% (+1.6%) Black = 10% (+0.4%) Asian = 1% (-2.2%) American Indian = 1% Others = 1% (+0.2%)	White = 85% (-0.4%) Black = 10% (+0.4%) Asian = 3% (-0.2%) American Indian = 1% Others = 1% (+0.2%)	White = 85% (-0.4%) Black = 10% (+0.4%) Asian = 3% (-0.2%) American Indian = 1% Others = 1% (+0.2%)
Marital Status	Married = 46% Never Married = 32.8% Divorced = 13.6% Separated = 3.1% Widowed = 3% Spouse Absent = 1.3% Others = 0.1%	Married = 46% Never Married = 33% (+0.2%) Divorced = 14% (+0.4%) Separated = 3% (-0.1%) Widowed = 3% Spouse Absent = 2% (-0.3%) Others = 0.0% (-0.1%)	Married = 46% Never Married = 33% (+0.2%) Divorced = 14% (+0.4%) Separated = 3% (-0.1%) Widowed = 3% Spouse Absent = 1% (-0.3%) Others = 0.0% (-0.1%)	Married = 46% Never Married = 33% (+0.2%) Divorced = 14% (+0.4%) Separated = 3% (-0.1%) Widowed = 3% Spouse Absent = 1% (-0.3%) Others = 0.0% (-0.1%)	Married = 46% Never Married = 33% (+0.2%) Divorced = 14% (+0.4%) Separated = 3% (-0.1%) Widowed = 3% Spouse Absent = 1% (-0.3%) Others = 0.0% (-0.1%)
Salary	<=50K = 75.9% >50K = 24.1%	<=50K = 76% (+0.1%) >50K = 24% (-0.1%)	<=50K = 75% (-0.9%) >50K = 25% (+0.9%)	<=50K = 76% (+0.1%) >50K = 24% (-0.1%)	<=50K = 76% (+0.1%) >50K = 24% (-0.1%)

Table 1:Compiled Result for Country, Gender, Race, Marital Status and Salary

The second part of the project was to launch an attack on the data after the noise has been added to it. The attack has been launched using five different values of epsilon, i.e 1, 0.01, 0.1, ln 2 and ln 3, to compare and contrast the variance in the actual value and the deviated value. The person of interest that is the record on which the attack has been launched is the tenth record of the dataset.

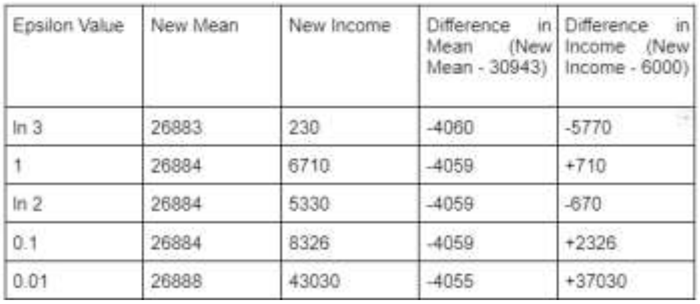


```
person of interest:

sex           0.0
age          45.0
educ          6.0
income     6000.0
married       1.0
race          1.0
Name: 0, dtype: float64
```

Figure 2: A sample person of interest

The attack has been launched on the income attribute of the person of interest. The true value of the income is 6000 and the overall mean of the incomes of the dataset is 30943.45. These values have been compared to the results after applying the differential privacy, which have been accumulated in the table mentioned below.



Epsilon Value	New Mean	New Income	Difference in Mean (New Mean - 30943)	Difference in Income (New Income - 6000)
ln 3	26883	230	-4060	-5770
1	26884	6710	-4059	+710
ln 2	26884	5330	-4059	-670
0.1	26884	8326	-4059	+2326
0.01	26888	43030	-4055	+37030

Table 2: Comparison between the Mean and Individual values

As it can be deduced from the table that as the value of epsilon deviates farther away from 1, the new calculated value of the income inculcates more noise and therefore, the actual value is difficult to find after the data has been treated with the differential privacy. As the epsilon approaches 1, the deviation in the income’s value reduces. It is also noticeable that even though the true value of an attribute, here income, varies greatly with the change in epsilon, the variation in the mean value of the attribute over the whole data stays the same for all cases and doesn’t deviate much from the actual figure.

To sum up, this quality of the differential privacy helps in adding the smart noise through which the actual information of each record is not available but the crux of the overall data is maintained which can be utilised by the receiver of the data.

VI. LIMITATIONS

Differential privacy is an approach of safeguarding people's privacy when gathering, analysing, and exchanging data. By introducing random noise to the data to mask individual data points, it tries to offer strong privacy assurances. Yet, while applying differential privacy in practise, it is important to take into account a number of its drawbacks [8]. The main drawbacks of differential privacy are as follows:

1. Utility versus privacy trade-off: One of the main obstacles to differential privacy is the utility versus privacy trade-off. The level of privacy protection is determined by the quantity of noise added to the data, which also has an impact on the precision of the data analysis. While too little noise can jeopardise people's privacy, too much noise can render the data unusable for study. To prevent the privacy protection offered by differential privacy from detracting from the usefulness of the data, this trade-off must be carefully weighed.

2. Limited protection against inference attacks: Differential privacy is intended to defend against direct attacks that seek to identify specific data points, but it offers only limited defence against inference assaults. The protection it offers against inference attacks, which utilise statistical methods to infer private information about specific people, is, however, somewhat restricted. When the data collection is tiny or the attacker has access to additional information not present in the data set, inference assaults can be especially potent.

3. Data perturbation sensitivity: Differential privacy is sensitive to the type of data perturbation. To effectively preserve privacy, the quantity of noise that is introduced to the data must be precisely adjusted. Yet, even little data modifications might significantly affect how much privacy is protected by differential privacy.

4. Limitations of the Laplace mechanism: The Laplace mechanism is an amazing tool to achieve differential privacy, but it comes with its own cons. It uses a Laplace distribution to choose data, adding random noise to the data. The Laplace mechanism, however, has a number of drawbacks. For instance, it makes the assumption that the data is continuous and quantitative, which may not be true for all data sets. Choosing a good noise parameter can also prove to be a strenuous task while using the Laplace mechanism.

5. Difficulty of implementing differential privacy: Achieving differential privacy can be difficult, especially for firms without a strong background in data privacy and security. A thorough understanding of statistical procedures, data analysis methodologies, and privacy-preserving algorithms is necessary for the design and implementation of a differentially private system. To ensure that differential privacy offers the necessary level of privacy protection, the implementation must also be rigorously evaluated and tested.

To conclude, differential privacy is a potent instrument for safeguarding people's privacy when gathering, analysing, and exchanging data. But, while putting it into effect, there are a number of restrictions that must be properly taken into account.

VII. SCOPE OF STUDY

The current research figures out the significance of applying differential privacy to the data and the impact of privacy budget values to the smart noise. The research can be further expanded to understand the importance and use cases of applying differential privacy over data which gets used in the training of various machine learning models.

Depending on whether the job at hand requires supervised learning or unsupervised learning, several methods can be used to include differential privacy into machine learning algorithms. The primary idea of focus is to randomize part of the mechanism's behaviour to provide privacy, where mechanism refers to a learning algorithm, but the differential privacy framework can be further expanded and applied to study any algorithm ensuring privacy.

The target is to train the various models with both the raw data as well as the differentially private data and subsequently compare the results of statistical analysis, which should ideally remain similar. In contrary to this, the comparison of results of data prediction of individual records must give wrong results for the differentially private data, thus protecting the individual's data and ensuring data privacy.

VIII. CONCLUSION

The "concept of differential privacy" has brought a revolution in the data privacy sector. This non-conventional technique of protecting data of individuals yet conveying the essence and the crux of data is significantly useful in the field of data analysis.

Through the project, we tried to understand the application of differential privacy on various data of individuals and compared the results of deviation for various epsilon values. As it can be noted from the results, the differential privacy adds the "smart noise" to the data, which barely disturbs the overall distribution of data into various categories, but at the same time prevents any kinds of breaching of individual records of the data through attacks.

Although the mean and the categorical distribution of the modified data remains the same for analytical studies by any third party or other organisations, the individual record values change highly, especially when the value of epsilon deviates far from 1, like in the case of 0.01 as epsilon values. This helps in simultaneous data protection as well as analysis.

The application of the project holds a special importance when the census data is required to be sent to various agencies or municipal corporations to analyse the data in order to understand the possible ideologies, needs and demands of the population. This understanding is required to prepare the propagandas for future projects in an area.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all of the individuals and organizations who supported us during the course of this research project. First and foremost, we would like to thank our supervisor, Prof. Manoj Kumar, for his guidance, support, and encouragement throughout the project. His insights and expertise were invaluable in helping us to design and carry out our research. In addition, we are grateful to Delhi Technological University for their support and assistance with various aspects of the research, including provision of required hardware and software. Their help was essential to the success of the project. Finally, we would like to thank our families and friends for their unwavering support and encouragement throughout the project. Their patience and understanding were critical to our ability to complete this work. Once again, we would like to express our deepest gratitude to all of those who supported us during this research project.

REFERENCES

[1] E. Aïmeur and D. Schőnfeld, "The ultimate invasion of privacy: Identity theft," in *Ninth Annual International Conference on Privacy, Security and Trust*, Montreal, 2011.

[2] C. Dwork, "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP*, Venice, Italy, 2006.

[3] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias.," *Journal of the American Statistical Association,* pp. 63-69, 1965.

[4] A. Nikolov, K. Talwar and L. Zhang, "The geometry of differential privacy: the sparse and approximate cases," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, Palo Alto, USA, 2013.

[5] F. K. Dankar and K. E. Imam, "Practicing Differential Privacy in Health Care: A Review," *Transactions on Data Privacy,* vol. 6, no. 1, p. 35–67, 2013.

[6] R. Danger, "Differential Privacy: What is all the noise about? " *arXiv preprint arXiv:2205.09453,* 2022.

[7] D. Dua and C. Graff, *Adult Census Income Dataset, UCI Machine Learning Repository,* Irvine, CA: the University of California, School of Information and Computer Science, 2019.

[8] K. Nissim, T. Steinke, A. Wood, M. Bun, M. Gaboardi, D. R. O'Brien and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience" 3 March 2017. [Online]. Available: https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_0.pdf.

AUTHORS

First Author – Rakshit Chauhan, Bachelor of Technology, Delhi Technological University,

Second Author –Pooja Narula, Bachelor of Technology, Delhi Technological University,

Third Author –Shaurya Shekhar, Bachelor of Technology, Delhi Technological University,

Fourth Author – Manoj Kumar, Professor , Delhi Technological University,