International Journal of Latest Trends in Engineering and Technology
Vol.(16)Issue(1), pp.005-012
DOI: http://dx.doi.org/10.21172/1.163.02
e-ISSN:2278-621X

SUMMARY STATISTICS OF SENSITIVE DATA USING DIFFERENTIAL PRIVACY

Jaseem C K[1]

Abstract- The privacy of the ever-growing data in the modern world is a rising concern. This data has the potential to inform many useful insights while it is at stake because of the privacy issue. Recent decades have witnessed a number of cases in which the data were either stolen or personal information was identified from the statistical data. Solving this problem of privacy-preserving data analysis might encourage the flow of more data to be used. Differential Privacy is considered to be one of the state-of-the-art concepts that can help us achieve this goal. It is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. The project focuses on extracting the summary statistics of student alcohol consumption using differential privacy. The IBM differential privacy library is used to implement this and gain insights. The output is the summary statistics with a conclusive inference.

Key Words: Differential Privacy, Summary Statistics, Reconstruction Attack, Data Analysis

I. INTRODUCTION

The advent of the 21st century has seen an enormous collection of sensitive data from individuals on a daily basis. This piling up of sensitive and valuable data increased by the uprise of IoT devices. Recently, people tend to prefer IoT devices more. These produce millions of sensitive data every minute. The growing trend of internet consumption further boosts this behavior. The privacy of this ever-growing data is a rising concern in the modern world. There has been a number of incidents depicting a breach of privacy, which in turn have affected the people concerned. But these incidents help us rethink the way privacy was perceived and the ways in which these problems can be tackled. Contributions from two decades prior are clearly evident in the field of privacy-preserving statistical analysis. These include the development of general and robust definitions of privacy, the introduction of a meaningful measure of privacy loss, the design of basic privacy-preserving computational building blocks, and an investigation of the limits of what can be achieved by the statistical analysis while preserving privacy. Many cyber attacks were initiated that uses statistical data to know more about a person involved. These were threats to Personal Identifiable Information and privacy in general.

Detecting and preventing complex differencing attacks is an immense challenge in itself. There are currently a variety of types of differencing attacks, and it is not clear whether a future attack type might be conceived, which creates unforeseen privacy risk. Such attacks may use independent and uncoordinated releases, which can be difficult to anticipate. Hence traditional Statistical Disclosure Control(SDC) falls short in defending against these attacks. Given a dataset containing sensitive information, the goal is to release statistics about the dataset to the public. These statistics may be fixed in advance or may be chosen by the person who queries the dataset. The goal of privacy-preserving data analysis is to protect the privacy of the individual records in the dataset by defending any attack strategy designed to compromise privacy. This problem of privacy-preserving data analysis has a long history spanning multiple disciplines. The need for a robust and meaningful definition of privacy along with a computationally rich class of algorithms that satisfy this definition has never been more relevant than now. Differential privacy is such a formal mathematical model of privacy. It requires that the output of analysis should

[1] *BTech Student, Computer Science and Engineering, NSS College of Engineering, Palakkad, Kerala, India*

reveal almost no information specific to any individual within the dataset[1]. It has proven to defend against many attacks and hence already seen significant adoption in many fields.

This paper is organized as follows. Section II presents the background. Section III introduces Differential Privacy and it's theoretical aspects. Section IV discusses the implementation of the project and section V presents the outputs and results. Concluding remarks are given in section VI.

II. BACKGROUND

Data collection, analysis, and inference from data have been practiced from time immemorial. And it has become more and more important with time. In the current world, data inference is of utmost importance. Almost all companies run their data analysis. By the emergence of smart mobiles and IoT devices, data start to pile up each second. All these data have value to many sectors. The conclusion that can be attained from this data can be very valuable and some times even life-saving. But most of the time these data can be very sensitive and can include confidential and private information that the data holder does not wish to reveal. Obstructing the analysis of these data that has the potential to give much critical information can be very devastating. The data could have been used by tech companies to understand their user behavior or health centers to efficiently diagnose or researchers for forecasting and gaining insights on certain problems.

Every individual leaves an extensive trail of potentially sensitive data in the modern world. The scale of this data collection is colossal. Data is also being disseminated more widely. The demand for open data is increasing with the growing adoption of analytics and data science. The demand for data-driven decision making requires to enable privacy and utility simultaneously. Even the non-sensitive open datasets are capable of raising the level of background knowledge which can aid in privacy attacks on other datasets[1]. The public is now more aware of privacy concerns and data misuse and hence new privacy laws and strict regulatory measures are taken against such issues. Privacy attacks are becoming more powerful by greater computing power and increasingly sophisticated techniques. There were several classes of innovative privacy attacks, that have been performed in recent years against which modern privacy techniques

1. Linkage attacks–

A linkage attack attempts to re-identify individuals in an anonymized dataset by combining that data with another dataset. In this attack, quasi-identifiers such as age, gender, and postcode are used in combination to determine the identity of a de-identified record, by linking to another dataset. This type of attack is particularly serious given the wealth of rich auxiliary data is easily available.

- Latanya Sweeney demonstrated that medical data stripped of direct identifiers could be re-identified by linkage with voter registration data available publicly. The k-anonymity model is used in this work by the author[2].
- The Netflix prize dataset revealed to be vulnerable to linkage attacks based on background knowledge from IMDb[3].

2. Differencing attacks and other composition attacks–

Present-day attempts at data privacy have shifted away from releasing de-identified microdata to releasing aggregate statistics instead. But this can still pose privacy risks and requires privacy techniques to avoid being disclosed. One of the simplest forms of attack on aggregate statistics is a differencing attack. It uses background knowledge about an individual person to learn sensitive information about that person using multiple statistics in which the target's data was included.

- Matthews et al. demonstrated the prevalence of differencing attack vulnerabilities in practice[4]. They reviewed multiple US state-level web-based data query systems allowing interactive queries of public health data in the US to give flexible tabular outputs. Despite being designed with SDC methods in place, many systems were found vulnerable to differencing attacks.
- There have been multiple attacks on Facebook users using microtargeted advertising, despite Facebook's internal SDC methodologies[5].

3. Reconstruction attacks–

A reconstruction attack is any method for reconstructing a private dataset partially from public aggregate information. In a reconstruction attack, an attacker is able to use statistics released about a sensitive dataset to infer with high accuracy a significant portion of the dataset itself. This casts serious doubts on the ability of traditional statistical disclosure control (SDC) methods to protect systems that release aggregate statistics.The phenomenon of Fundamental Law of Information Recovery by Dwork and Roth formulated as "overly accurate answers to too many

questions will destroy privacy in a spectacular way" shows that in order to preserve even a very weak notion of individual privacy, the statistics need to be sufficiently distorted.

- The US Census reported vulnerabilities to reconstruction attacks in its 2000 and 2010 Census data releases[6].

4. Membership inference attacks–

Membership inference attack is another attack on aggregate statistics that focus on merely determining whether or not someone is in a dataset known. It is the process of determining whether a sample comes from the training dataset of a trained ML model or not.

- A membership inference attack on the results of Genome-Wide Association Studies (GWAS) has been demonstrated[7], this type of membership inference proves to work even when the statistics are noisy. Membership inference attacks have also been performed on ML models, whereby an adversary determines whether the data of an individual was used to train a model[8].

There exist many such attacks that can possibly leak the data and gain personal information from the statistical data. Traditional SDC methods were devised for a data ecosystem very different from that of the current scenario. Computing power is increasing, and richer data is publicly available. Simultaneously, new attack methodologies are being developed and existing attacks are becoming stronger. This highlights the need for a fresh and rigorous look at privacy. As traditional SDC falls short in defending against these attacks, there requires a new method for keeping the data safe from such attacks. Approaches that focus on defending against currently-identified threats are not suited to defend against unidentified threats. Auxiliary knowledge is capable of making attacks on aggregate statistics easier, just as it strengthens attacks on row-level data. These methodologies that rely on a specific attack is in danger of not being sufficiently future-proof. At the same time, too high accuracy may result in a privacy breach. One solution is to measure and evaluate the cumulative risk of outputting the results of statistical analyses on private data. This measurement of privacy risk can be used to guide choices about the number and type of statistics to release and it's accuracy. This approach is taken by differential privacy.

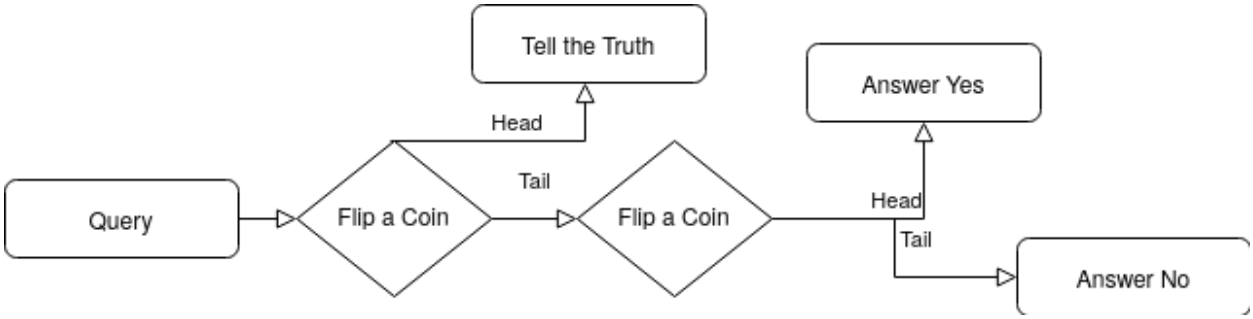
Differential privacy is a privacy model for limiting statistical disclosure and controlling privacy risk. It is a definition, or a standard, that specifies a particular requirement that data release methods may or may not satisfy. If a data release method satisfies the requirement, then it would protect an individual's information essentially as if his or her information were not used in the analysis at all. There are many differentially private data release techniques that involve releasing aggregate statistics perturbed with random noise. Noise is added during the computation of the release in a way to provide privacy while maximizing the accuracy of results. The volume of available data is increasing inexorably and attacks such as differencing and reconstruction attacks pose a real threat. Thus a data analysis method that does not compromise privacy and is capable of generating genuine insights is required. Differential privacy was framed as a means by which to help address these needs.

III. Differential Privacy

A major motivating factor for the adoption of differential privacy has been the discovery of many new privacy attacks. These attacks aim to reverse data privacy protection mechanisms to expose sensitive information about individuals in a dataset. Techniques that are used for these attacks can target aggregate data like summary counts, histograms, or average statistics. Easy access to rich datasets, more computing power, and novel methods have made these techniques a growing threat to the confidentiality of aggregate data releases. The primary concern of Differential Privacy (DP) is to assure that a data subject is not affected by their entry or participation in a database while maximizing utility or data accuracy for the queries. That is DP describes a promise, whereby a data subject will not be affected, adversely or otherwise, by allowing the person's data to be used in any study or analysis[9].

Differential privacy can be considered as a definition of privacy tailored to the problem of privacy-preserving data analysis. Data cannot be fully anonymized and still remain useful. The richer the data, the more useful it is. Differential privacy addresses this paradox learning useful information about a population while learning nothing about an individual. This ensures that the same conclusions are extracted independent of whether any individual opts in or out of the data set. It ensures that any sequence of outputs is "essentially" equally likely to occur, independent of the presence or absence of any individual[10]. The probabilities are taken over random choices made by the privacy mechanism, and the term "essentially" is captured by a parameter, ε. A smaller ε will yield better privacy. There can be many DP algorithms for achieving a computation task T in an ε-differentially private manner for a given value of ε. Some algorithms will have better accuracy than others.

A DP algorithm for removing private information in the data, whereby you can perform the analysis on the data can be depicted by the flip of coins[9]. Thus, for each entry, the curator will apply this algorithm:



Hence, each person is protected with "plausible deniability", because a person is plausible to deny the answer by the randomness of flipping a coin. The formal mathematical definition of DP is given as follows. A randomized function K gives ε-differential privacy if for all data sets D and D′ differing on at most one element, and all S □ Range(K),

$$\Pr[K(D) \square S] <= \exp(\varepsilon) * \Pr[k(D') \square S] \qquad (1)$$

The probability is taken is over the coin tosses of K. Epsilon (ε) is a measure of privacy loss at a differential change in data. The smaller the value, the more protected it is.

The DP technique takes as input the raw data, finds the answers from the original input data, and then introduces distortion based on a variety of factors. Noise is added through the Laplace distribution[9]. It is similar to the normal distribution/bell-curve. Many other mechanisms can also be used in place of the Laplace distribution. while the Laplacian Mechanism works for any function with a real number as an output, the Exponential Mechanism can be used in functions without real number as output. Ensuring privacy is crucial for numerous applications such as maintaining the integrity of sensitive information, eliminating the opportunity for adversaries to track users, etc. DP is appealing for a variety of applications because it guarantees these privacy needs.

III. Project Implementation

1. Dataset Collection

Student Alcohol Consumption data is collected from the website kaggle.com[11]. The data were obtained in a survey of about 674 students of math and Portuguese language courses in secondary school. It contains sensitive data of students such as alcohol intake in workdays and weekends, current health status, quality of family relations, etc. There are two data files of students each in math and Portuguese language course. The files are in comma-separated value(CSV) format.

2. Data Preprocessing

The two data files on student alcohol consumption of math and Portuguese course are merged together for data analysis. The description of columns is matched to the abbreviate column names.

3. Data Analysis

The IBM differential Privacy library 'diffprivlib'[12] is installed and imported. It is a general-purpose, open-source Python library for differential privacy. Its purpose is to allow experimentation, simulation, and implementation of differentially private aggregates and models. This contains a number of mechanisms, tools, and models. This library is used to calculate mean, variance, standard deviation, and histograms of certain variables of the project data set. The epsilon (ε) used in this project is 0.1. Summary statistics of the age of students in total and of both genders are generated. Then, the alcohol consumption of male and female students are evaluated. Comparison and plotting of the data depicting the alcohol consumption of students during weekends and workdays are carried out. Consumption behavior of rural and urban students are also examined. The relation of the current health status of students and their consumption levels are also inspected in addition. Consumption behavior is also compared with the quality of family

relationships. All this data analysis is done using the differential privacy library and hence it defends the possible statistical attacks.

III. Results and Outputs

After collecting, preprocessing, and analyzing the data, insightful statistics, and inferences are obtained. The summary statistics and alcohol consumption behavior of students of both genders, from different residential backgrounds and of variant family situations are evaluated. Alcohol consumption levels were made in ranges from 1 to 5.

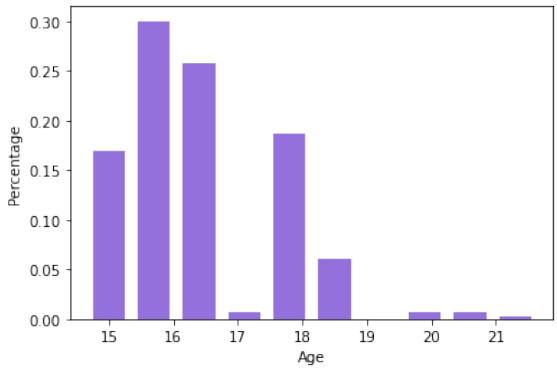
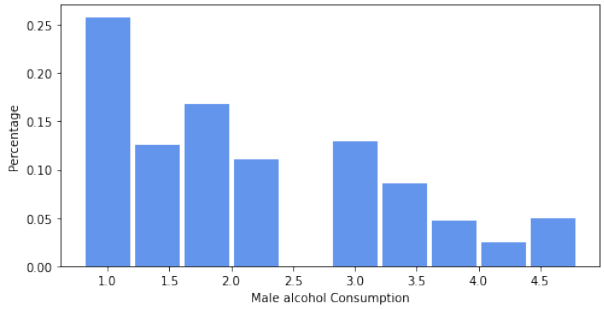


Figure 1. Histogram of age of students in the Dataset

(a)



(b)

Figure 2. (a) Histogram of Female Alcohol Consumption Level (b) Histogram of Male Alcohol Consumption Level

Male students are identified as consuming more alcohol as compared to their students of the female gender. It was also observed that high alcohol consumption is seen more on weekend days. Rural and urban students' alcohol

consumption was almost alike but rural students were more in high consumption. The current health status of the students did not seem to vary with alcohol consumption. The summary statistics of the health status of students with all the ranges of alcohol consumption are similar. In general, students of this dataset have a good level of relationship with their family. Slight variation can be seen in the students of high alcoholic behavior. Students with high alcohol consumption tend to have a low quality of family relationships while students with low alcohol consumption behavior perform well in family relations.

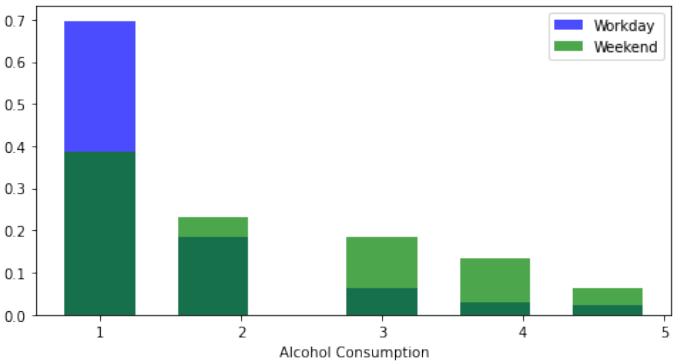


Figure 3. Histogram of Alcohol Consumption in Workdays and Weekends

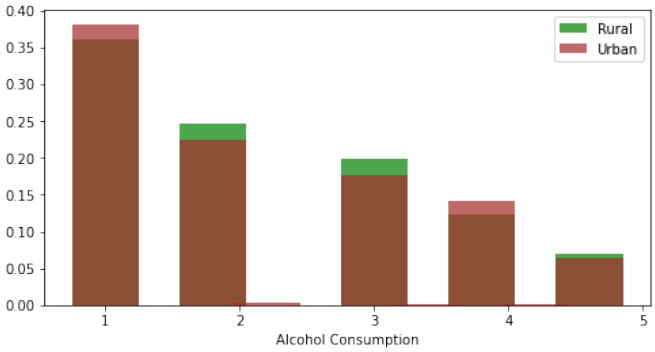


Figure 4. Histogram of Alcohol Consumption by Rural and Urban Students

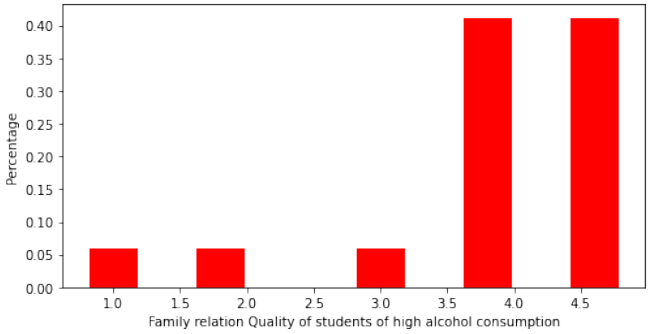


Figure 5. Histogram of Quality of Family Relation by students of high Alcohol Consumption

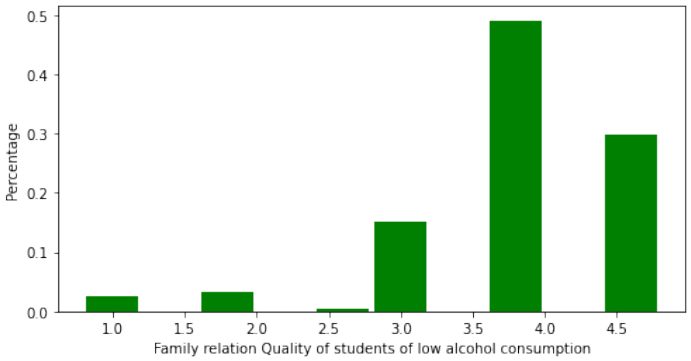


Figure 6. Histogram of Quality of Family Relation by students of low Alcohol Consumption

Table -1 Summary Statistics of dataset[11]

	Mean	Variance	Standard Deviation
Age of Students	**16.7431**	**1.1351**	**1.7900**
Alcohol Consumption of Students (in range of 1-10)	**3.6986**	**9.6689**	**2.0496**
Alcohol Consumption of Female Students (in range of 1-10)	**3.3818**	**4.6119**	**2.1962**
Alcohol Consumption of Male Students (in range of 1-10)	**4.2436**	**11.3504**	**1.6439**
Alcohol Consumption of Students in workdays (in range of 1-5)	**1.4863**	**0.1945**	**0.9260**
Alcohol Consumption of Students in weekends (in range of 1-5)	**2.2288**	**1.7318**	**1.2603**
Current Health Status of Students (in range of 1-5)	**3.5401**	**5.0934**	**1.4324**
Quality of family relations of students (in range of 1-5)	**3.9736**	**0.3051**	**0.5637**

IV.CONCLUSION

The project successfully generates summary statistics of the dataset without compromising privacy. This is achieved using differential privacy methods. The statistics can defend itself from attacks such as differential and reconstruction attacks that can disclose the privacy of individuals. Differential privacy is still a relatively young field of research and users are still learning how to bring it effectively into practice. The privacy-utility trade-off problem is difficult to tackle when used in practical applications. However, data privacy researchers continue to find improvements to differentially private algorithms, with state of the art approaches being developed for a wide variety of analyses. It is better for the organizations that release sensitive data to assess the impact of traditional and new privacy attacks, and evaluate whether differential privacy is a suitable and beneficial method of defense beforehand. Differential privacy is particularly well-suited to the use case of releasing statistics about national populations or similar data sets to the public because differentially private algorithms perform best in use cases with pre-determined statistics and large sample sizes.

REFERENCES

[1] Hector Page, Charlie Cabot, Kobbi Nissim, “Differential privacy: an introduction for statistical agencies”, Privitar, December 2018.

[2] Sweeney, L. (2002). “K-Anonymity: A Model for Protecting Privacy.” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 05 (October 2002): 557–70

[3] Narayanan, A. and Shmatikov, V. (2008). “Robust De-Anonymization of Large Sparse Datasets.” In Proceedings of the 2008 IEEE Symposium on Security and Privacy, 111–125. SP ’08. Washington, DC, USA: IEEE C

[4] Matthews, G., Harel, O. and Aseltine, R. (2017). “A Review of Statistical Disclosure Control Techniques Employed by Web-Based Data Query Systems.” Journal of Public Health Management and Practice 23, no. 4 (2017): e1–4.

[5] Korolova, A. (2010). “Privacy Violations Using Microtargeted Ads: A Case Study.” In 2010 IEEE International Conference on Data Mining Workshops, 474–82, 2010.

[6] Garfinkel, S. (2018). “Modernizing the Disclosure Avoidance System for the 2020 Census”. Accessed May 8 2020.

[7] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J. et al. (2008). “Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays.” PloS Genetics 4, no. 8 (August 29, 2008).

[8] Shokri, R., M. Stronati, C. Song, and V. Shmatikov. (2017).“Membership Inference Attacks Against Machine Learning Models.” In 2017 IEEE Symposium on Security and Privacy (SP), 3–18, 2017.

[9] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy", Foundations and Trends in Theoretical Computer Science Vol. 9, Nos. 3–4 (2014) 211-407 2014.

[10] Cynthia Dwork and Adam Smith, "Differential Privacy for Statistics: What we Know and What we Want to Learn," Journal of Privacy and Confidentiality (2009), Number 2, pp. 135–154, 2009.

[11] UCI Machine Learning, Student Alcohol Consumption, 2016 (https://www.kaggle.com/uciml/student-alcohol-consumption). Accessed May 10 2020.

[12] Naoise Holohan, Stefano Braghin, P´ol Aonghusa, and Killian Lev-acher. "Diffprivlib: The ibm differential privacy library", arXiv:1907.02444, 2019.