Utility And Privacy Guarantees of Differential Privacy

by

Shubham Srivastava

Under the Supervision of Dr Debajyoti Bera

Indraprastha Institute of Information Technology Delhi

May, 2016

Utility And Privacy Guarantees of Differential Privacy

by

Shubham Srivastava

Submitted
in partial fulfillment of the requirements for the degree of
Master of Technology

Indraprastha Institute of Information Technology Delhi
May, 2016

CERTIFICATE

This is to certify that the thesis titled "Utility And Privacy Guarantees of Differential Privacy" being submitted by Shubham Srivastava to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May, 2016

Debajyoti Bera
Department of Computer Science
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Debajyoti Bera for providing excellent guidance and being supportive throughout the span of this thesis. Without his patience, critiques and thoughtful insights this work would never have been completed. A very special thanks to IIITD for providing an excellent infrastructure, environment and a flexible curriculum to carry out my work at a suitable pace. I am deeply thankful to all the faculty and admin staff here for there extremely supportive attitude. I am thankful to all my friends for being so supportive and helping me out with the problems whenever I got stuck. Fun time with you people always made me feel fresh and enthusiastic. I dedicate this work to my parents, who have always been there for me and provided me support both emotionally and financially. A special thanks to my elder sister Shubhi who always kept me motivated and my younger brother Priyam for all his love.

ABSTRACT

The primary aim of any survey is to facilitate analysis of sensitive data to extract useful information, without jeopardizing privacy of the participants. Privacy models like k-anonymity do not guarantee privacy against attackers with background knowledge. "Differential privacy" is a model for data release which formalizes data privacy and makes no assumption on the attackers' background knowledge. It builds on the idea that adding carefully computed noise to certain data can make it safer from privacy perspective while retaining utility of the data. We have addressed two problems for data release in this thesis and proposed algorithms that address the trade-off between utility and privacy. Following are the major contributions of this thesis:

1. We studied a functional mechanism to achieve differential privacy in regression analysis. We have extended an existing algorithm to achieve differential privacy for a more general form of linear regression. We have proved that the mechanism preserves differential privacy and provides better utility when compared to direct perturbation technique.

2. We have analyzed two strategies of achieving differential privacy for publishing summary statistics — the compose-then-perturb approach and perturb-then-compose approach. We prove that the perturb-then-compose approach indeed satisfies differential privacy. We then try to find out which approach provides better balance between utility and privacy, for summary information that we want to publish, for a dataset.

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

The primary aim of any survey is to learn useful information or trends about a population. Data collected from surveys offers great opportunities for mining useful information, but there is also a threat to privacy because data in raw form may contain sensitive information about individuals. Privacy-preserving data publishing [5] addresses this problem. Data curators use various methods like removing columns containing Personally Identifiable Information (PII), k-anonymity [15], l-diversity [10], t-closeness [9], etc to preserve privacy of the individuals. But all these techniques are vulnerable to attacks in one way or the other. For example, privacy models like k-anonymity do not guarantee privacy against attackers with background knowledge. "Differential privacy" on the other hand is backed by a strong mathematical foundation, and makes no assumption on the attackers' background knowledge. It builds on the idea that adding carefully computed noise to certain data can make it safer from privacy perspective while retaining utility of the data. We have addressed two problems for data release in this thesis and proposed algorithms that address the trade-off between utility and privacy.

1.1 Thesis structure

In this Chapter, we give the motivation behind this thesis. Subsequently, we shortly describe the work done in the thesis.

Chapter 2 contains the background information. It details privacy problems in

database that allows statistical aggregate queries. Then , we explain the concept of differential privacy and utility and privacy guarantees.

Chapter 3 briefly explains the related work and prior research on differential privacy.

In Chapter 4, we explain how functional mechanism can be used to perform differentially private linear regression. We also extend the approach to a more general variant of the regression analysis.

In Chapter 5, we discuss two strategies for release of composite functions, maintaining differential privacy. We then compare utility of above strategies for releasing average grade in each bucket, from student grade database.

Finally, in Chapter 6, we conclude our work. We then also discuss possible future work and research problems to work on.

1.2 Motivation

With the advances in IT, information gathering has become pervasive. Most of the times, data is collected by government, corporations and companies without an individual's knowledge and consent. For example, e-commerce companies keep track not only of the items you buy, but also the items you browse, to generate a detailed profile of each individual. Using this profile information they later can advertise specific products to an individual and plan business strategies accordingly.

When data about individuals or entities are to be publicized, care must be taken to avoid privacy violations. Various examples of attacks against publicly released data can be found in the literature. Some of the most popular ones include de-anonymization attack against the Netflix Prize data set [13] and the identification of individuals from a de-identified data set of AOL search queries.

Statistical Disclosure Control [2] (SDC) aims to allow the release of data and at the same time preserve the privacy of individuals. These techniques mask or modify the original data or the statistics that are to be published. This modification reduces the risk of privacy breach at the cost of utility. Thus, there is a trade-off to find a balance between privacy and utility. Traditional approach to evaluate the effectiveness of such a solution involved running experiments and trying to re-identify records from the published data. Since then, privacy models like k-anonymity and ε-differential privacy have been proposed to provide formal privacy guarantees. These models introduce uncertainty in the outcome of attacks against the privacy of individuals. In k-anonymity privacy model, the information for each person contained in the published dataset cannot be distinguished

from at least k-1 individuals whose information is also present in the published dataset. Though the method provides guarantees, attackers with some background knowledge can still make inferences that compromise an individual's privacy. The ε-differential privacy, is a privacy model that perturbs the response to query answers by adding independent random noise following a probability distribution. Although ε-differential privacy provides strong privacy guarantee, it is not all good news because to achieve the desired level of privacy, a lot of noise might get added which inadvertently leads to poor utility. Thus, there is a need to have algorithms for data release that address the trade-off between utility and privacy.

1.3 Contributions

We study ε-differential privacy in both interactive and non-interactive setting in detail and also define formally the privacy and utility guarantees in case of ε-differential privacy. Following are the major contributions of this thesis:

1. We studied a functional mechanism to achieve differential privacy in regression analysis. We have extended an existing algorithm to achieve differential privacy for a more general form of linear regression. We have proved that the mechanism preserves differential privacy and provides better utility when compared to direct perturbation technique.

2. We have analyzed two strategies of achieving differential privacy for publishing summary statistics — the compose-then-perturb approach and perturb-then-compose approach. We prove that the perturb-then-compose approach indeed satisfies differential privacy. We then try to find out which approach provides better balance between utility and privacy, for summary information that we want to publish, for a dataset.

CHAPTER 2

BACKGROUND

Privacy is the ability of an individual to seclude themselves, or information about themselves, and thereby express themselves selectively. Privacy, in case of census problem or survey, is based on the intuition that one's privacy is protected to the extent that one blends in with the crowd. Privacy aware individuals reveal private information to a third party, only if they are guaranteed that their participation will not disclose specifics which were contributed by the individual.

2.1 Privacy In Databases

Databases can serve many social goals, such as identifying generic indicators for disease, strategic implementation of government development plans, fair allocation of critical resources, etc. The data privacy problem in databases is to learn and publish statistical information about a population while maintaining confidentiality of the participants. If this confidentiality is not guaranteed, the concerned population may refrain from taking participation in the survey.

A privacy breach may include one or more of the following:

- leaking of individual records,
- linking with public databases to re-identify individuals,
- allowing adversary to reconstruct a database with significant probability.

There are two different settings to be considered when talking about privacy in databases — interactive and non-interactive.

- In the interactive setting, the aim of the data collector, which is a trusted party, is to provide a query interface to the user where queries about the data can be posed and the user is provided a response for the same.
- In the non-interactive setting, the data collector publishes a sanitized version of the dataset collected. The sanitized version may contain only summary information about the database.

The classical intuition for privacy in databases given by Dalenius [2] in 1977, was very much similar to the notion of semantic security of encryption given by Goldwasser and Micali [6] in 1982. The intuition was that privacy meant anything that can be learned about a respondent from the statistical database can be learned without access to it. Unfortunately, it was shown by Dwork in 2006, that such type of privacy cannot be achieved. Since, we are not able to give absolute guarantees of disclosure, we turn towards relative guarantees.

Privacy models like k-anonymization, t-closeness provide such guarantees. The concept of **k-anonymity** was first formulated by Latanya Sweeney in 2002. The aim of "k-anonymization" is to produce a release of data, given a person-specific database with attributes, with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. In k-anonymity, the goal is to make each record indistinguishable from at least k-1 other records.

The problem with k-anonymization and other related techniques is that they are vulnerable to background knowledge attacks. The problem of making assumptions on the background knowledge of adversary can be understood from this example.

During World War II, Germany believed that its secret codes for radio messages were indecipherable to the Allies. German military used an Enigma machine for battlefield, naval and diplomatic communications. Due to the close link between German and Polish engineering industries, Polish were able to reconstruct an Enigma machine and read the messages of armed forces of Germany. This information was later shared with British, who then using the meticulous work of code breakers cracked the secret of German wartime communication, and played a crucial role in the final defeat of Germany.

Similarly, assumptions on the background knowledge of an adversary can lead to severe consequences when considering privacy in statistical databases. Using the background knowledge that heart attacks occur at a reduced rate in Japanese patients

Machanavajjhala et al. [10], in 2007, narrowed the range of values for a sensitive attribute of patient's disease.

We now elaborate with the help of examples, privacy breach in statistical databases.

Example 1. Consider a non-interactive setting, in which a survey is conducted in a large corporate office, to learn smoking habits of employees. The aim of the survey is to publish number of smokers cubicle wise. Assume we have a database D_1 containing records of smoking habits of employees in cubicle 1, where each record consists of *Name* and *Smoking Habits*, as shown in 2.1. The *Smoking Habits* value $\in$ {0,1} where 0 indicates that the person smokes and 1 indicates he/she doesn't.

Name	Smoking Habits
Alice	1
Bob	1
Jimmy	1
Paul	1
Jeremy	0

Table 2.1: Dataset D_1 with "Smoking Habit Records" of employees in cubicle 1

Suppose, an adversary possesses additional information that Jeremy in cubicle 1 doesn't smoke. Now, when the number of smokers is published, the number would be 4 for cubicle 1. The adversary will thus learn that remaining four employees working in cubicle 1 are smokers. Thus, it is clear that individual information can be compromised even without querying for any particular individual. Here the smoking habits of Alice, Bob, Jimmy and Paul are revealed.

Example 2. Consider an interactive setting, in which a university database of students and their grade marks between (1 - 10) is available for statistical queries like mean, sum, count, etc. A sample database D_2 of 5 students is shown in 2.2. A query interface is provided to ask queries like average grade marks of students whose name start with some characters.
Consider an adversary possesses auxiliary information that there are only 5 students in the database and only Alex is a student whose name starts with 'A'. The adversary can then issue two queries Q_1 and Q_2 as follows:

Q_1 : Find the average grades of students in the database whose names start with characters 'A', 'B', 'C', or 'D'.

Name	Grade Marks
Alex	7.5
Bob	8
Cathy	9
Barry	8.5
Dan	9

Table 2.2: Dataset D_2 with Student Grade Records

Q_2 : Find the average grades of students in the database whose names start with characters 'B', 'C', or 'D'.
Using the responses from above queries, grade marks of Alex, that is, 7.5 is revealed without even querying specifically for Alex's information.

A solution to the problem of information disclosure can be found if we consider the effect of uncertainty on data. To counter the background knowledge attacks, we can instead guarantee that whether or not an individual participates in a survey, cannot be inferred from summary published or responses generated when querying the database. This is exactly what differential privacy does.

2.2 The Concept of Differential Privacy

The concept of differential privacy [3] was first given given by Cynthia Dwork, in 2006, in the context of statistical database, where a trusted party holds a dataset D containing sensitive information of individuals. Each row contains the data of a single individual and the goal is to simultaneously protect every individual row and permit statistical analysis of the database as a whole. Thus,

- In non-interactive setting, the sanitized version is usually obtained by techniques such as data perturbation, by adding random noise to each row of the database. Any PII (Personally Identifiable Information) such as names, mobile number, aadhar number, etc are also removed.

- In interactive setting, the query interface returns to the user a carefully perturbed response as answer to the query.

Differential privacy guarantees that for any two **neighboring databases**, that is, databases that differ in only one row, the trusted third party's distribution over potential outputs are statistically close.

Mathematically, let us consider the following setting in this regard :

- A sensitive database $D \in U$, where U is the set of all databases
- Neighboring databases : D, $D^i \in U$, such that, $\delta(D, D^i) = 1$, where D^i is defined for i = $\{1, 2, ..., n\}$ as a database D with i^{th} row removed, n being the number of rows in database, and δ is a discrete distance function defined over two databases to the set of natural numbers $\{0, 1, 2....\}$
- A Randomized mechanism R, the range of which is denoted by Range(R) and S represents any subset of Range(R).

A randomized mechanism R which satisfies ε - differential privacy guarantees that it behaves similarly on similar input databases, where ε is a privacy parameter. Thus, a differentially private randomized mechanism must satisfy equation 2.1 below.

$$\mathbf{exp}(-\varepsilon)\mathbf{Pr}[\mathbf{R}(\mathbf{D^i}) \in \mathbf{S}] \leq \mathbf{Pr}[\mathbf{R}(\mathbf{D}) \in \mathbf{S}] \leq \mathbf{exp}(\varepsilon)\mathbf{Pr}[\mathbf{R}(\mathbf{D^i}) \in \mathbf{S}] \tag{2.1}$$

One of the basic technique that lets us achieve our goal is to add small random Laplacian noise to the answer of user's queries. This mechanism involves adding random noise that conforms to the Laplace statistical distribution with mean 0, and a scale parameter b that controls the amount of noise added. A random variable x has a Laplace(0, b) distribution if its probability density function is given by equation 2.2.

$$f(x|0, b) = \frac{1}{2b} exp\left(-\frac{|x|}{b}\right) \tag{2.2}$$

The scale parameter of Laplace mechanism depends on the privacy parameter ε as well as on the nature of query itself. The parameter that covers the nature of the query is called sensitivity, denoted by Δ. Sensitivity should be chosen in a way that it protects even the most different individual in the database. To achieve differential privacy using Laplace mechanism we take scale parameter b given by $\frac{\Delta}{\varepsilon}$. The For any real-valued query function, we perturb the query output by adding random noise with Laplacian distribution given by Lap$(0, \frac{\Delta}{\varepsilon})$. If we have a random variable Y drawn from the uniform distribution in the interval (-1/2,1/2), the random variable X, given by equation 2.3 follows Laplace distribution.

$$X = -\frac{\Delta}{\varepsilon} sgn(Y)\, ln(1 - 2|Y|) \tag{2.3}$$

We now see how Laplace mechanism achieves differential privacy by considering the example 2 related to university database of student grades, for some suitable value of

privacy parameter ε. The output of a differentially private mechanism to queries Q_1 and Q_2 can also be modeled as a single average grade query running on two datasets, one with and other without Alex's entry. We call the one with Alex's entry as D^O and the one without that entry as D^1, as shown in Table 2.3 and Table 2.4 respectively.

Name	Grade Marks
Alex	7.5
Bob	8
Cathy	9
Barry	8.5
Dan	9

Table 2.3: Student Grade Dataset D^O

Name	Grade Marks
Bob	8
Cathy	9
Barry	8.5
Dan	9

Table 2.4: Neighboring Dataset D^1

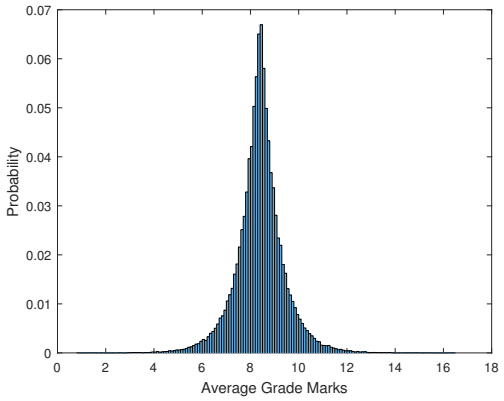

Figure 2.1: Probability distribution of response on original student grade database

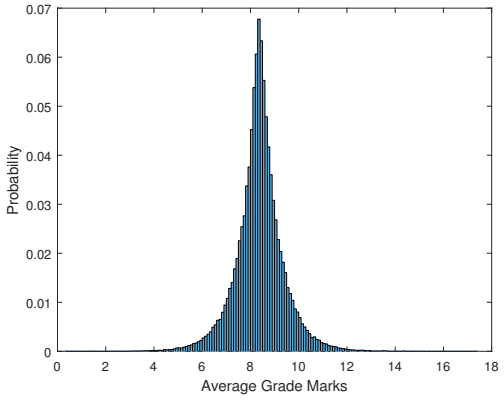

Figure 2.2: Probability distribution of response on a neighboring database

The probability of output distribution of differentially private mechanism on Dataset D^O and D^1 is shown in Figure 2.1 and Figure 2.2 respectively. In Figure 2.1, Pr (Average Marks = 8.4) $\approx$ 0.069 while in Figure 2.2, Pr (Average Marks = 8.375) $\approx$ 0.069. Also the probability distribution of both the datasets are very similar. Thus, an adversary trying to distinguish between the two databases will not be able to do so and hence the adversary will not be able to extract grade marks of Alex.

The task of preserving differential privacy stand alone isn't hard as the trusted third party can add so much noise to the users' queries so that it renders the data collected and stored useless. Table 2.5 gives an example where a lot of noise gets added to the query response on dataset D^O and D^1, derived from the example of university database of student grades. The reason for large noise addition may be attributed to high sensitivity of query or improper choice of privacy parameter. Obviously, the usefulness of such analysis is very low.

Dataset	Query Response
D^O	17.3691
D^1	3.2814

Table 2.5: Query response for Student Grade example dataset when large noise gets added

Similarly the task of preserving utility stand alone isn't difficult too. To answer the user queries in this case, just compute on the data collected and return the answers without any alteration. Of course, there is no privacy preserved in this case and all the participants of the survey are vulnerable. The interesting question thus is "how to preserve differential privacy and provide good utility guarantees at the same time?". The answer is utility and privacy trade-off.

Laplace mechanism can't be used for query functions which are not real-valued. For such cases, we can achieve differential privacy using exponential mechanism [12]. Preserving ε-differential privacy can be a very strong condition in many applications. For example, in case of output distribution on D^O and D^1 for university database of student grade, probability of occurrence of some range of outputs is 0 for D^O. As per the definition of definition of differential privacy given by equation 2.1, probability of occurrence of these ranges of outputs should be 0 for D^1. But, this cannot be easily achieved. A weaker notion of(ε, δ)-differential privacy, given by equation 2.4 is used for these cases.

$$exp(-\varepsilon)Pr[R(D^i) \in S] + \delta \leq Pr[R(D) \in S] \leq exp(\varepsilon)Pr[R(D^i) \in S] + \delta \tag{2.4}$$

2.3 Privacy and Utility Guarantees

Differentially private mechanisms must make confidential data available for accurate data analysis. The strategy to enforce ε-differential privacy depends on the nature of statistical analysis that is to be performed on the data collected. Once we are aware of the analysis that is to be done or the queries that will be issued, we find the sensitivity of the function, come up with a privacy budget ε and add noise using some mechanism. In this work, we deal with real-valued queries, so we use Laplacian mechanism [14] for noise addition. The value for privacy budget can be taken as a small constant like 2 or 3, or by experiments values from 0.01 to 100 can be tried. To choose the privacy budget, an economic model has been proposed by J. Hsu et al.[7]. If we know the type of numeric queries the user poses beforehand and the actual answer to it is, say Q(D^O), then we have sensitivity given by,

$$\Delta = max_{D^O, D^i}(Q(D^O) - Q(D^i))$$

where, D^i for i from 1 to n. is a neighboring database of D^O.

To re-iterate the fact, when or where to add the noise generated from the distribution depends on the setting (interactive or non-interactive), mechanism employed to preserve differential privacy and on the statistical queries being performed.

The privacy guarantees are given by the definition of differential privacy itself. To do an analysis of the utility of the mechanism employed, we first proceed with the more straight forward case of non-interactive setting [8]. In this case, since the trusted third party releases a perturbed database, if we run queries on it, the answers should be reasonably close to answer the original database would have given. So we consider the following setting:

- D^O is the original database that has the collected records from individuals,
- D' is the perturbed database released by the trusted party,
- Q is the query, which gives response based on the database provided as input,

Now, for very small quantities β and γ, we can write the utility requirements as,

$$\Pr[\,|\,R(D^O) - R(D')\,|\, = \beta] \geq 1 - \gamma$$

In the case of interactive setting we need to keep in mind that, generally no perturbed dataset is released but we add noise while answering the queries itself. Now,

- D^O denotes the database of collected records as before,
- Q(D^O) denotes the actual answer to the query posed by the user, that is, without considering any setting for differential privacy,
- R(D^O) denotes the noise perturbed answer to the user's query, that is, actually returned by the deferentially private mechanism

similar to previous case, for small quantities β and γ we can write utility requirement as,

$$\Pr[\,|\,R(D^O) - Q(D^O)\,|\, = \beta] \geq 1 - \gamma$$

The measure of utility can be estimated by the values of β and γ. Closer these values to zero more is the usefulness.
We now see how the privacy and utility of a differentially private mechanism can be evaluated. Consider that the actual results of count query on a statistical database is $C^0.C^1,C^2.....C^n$, where C^0 is the answer to count query on original database and C^i is the result of count query on original database with i^{th} row removed. The differentially private release mechanism instead answers $D^0,D^1,D^2.....D^n$, for each count query, as depicted in Figure 2.3

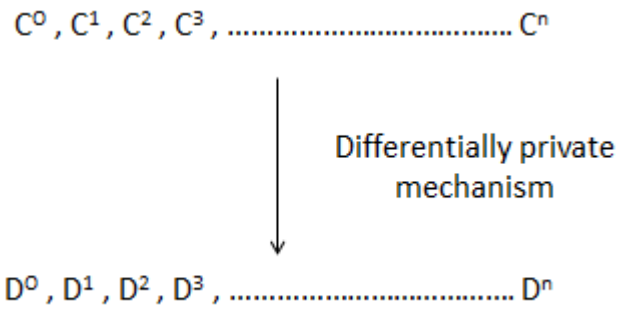

Figure 2.3: Unperturbed and perturbed responses to count query

The privacy requirement can be met if, $D^0 \approx D^i$ for all values of i from 1 to the number of rows.
The utility of the mechanism has two requirements as mentioned below:

- $\mathbb{E}[D^i] = C^i$
- $\mathbb{V}ar[D^i]$ should be small

CHAPTER 3

RELATED WORK

Ever since the concept of differential privacy was introduced by Cynthia Dwork [3, 4], a lot of work has been done by various researchers in the field. A major portion of work focused on achieving differential privacy for analysis tasks that can be performed in data mining. A lot of work has also been done on how to publish differentially private summary of datasets like health records, diseases outbursts in various community. Apart from these, methods are being devised to perform social network analysis and graph analysis maintaining differential privacy.

Dwork et al. showed that ε-differential privacy can be achieved by using Laplace mechanisms. Sarathy et al. [14] used this mechanism to achieve differential privacy for numeric data. This method works for aggregate queries and for cases where output is real number. Differentially private release of histograms was then proposed by Xu et al. [17]. For queries whose output space is discrete, McSherry et al. [12] proposed the exponential mechanism. Using exponential mechanism various interesting problems were then solved. The task of privacy preserving regression analysis was first undertaken by Chaudhari et al. [1]. The technique worked for linear regression but could not be generalized to other regression tasks. Zhang et al. [18] then proposed the functional mechanism to achieve differentially private regression. The mechanism proposed was generic and it could be applied to any optimization problem. Our work extends the application of functional mechanism to achieve the most generic differentially private linear regression model of the form $y = \omega * x + c$, instead of the less generic one, that is,

$y = \omega * x$.

The problem of perturbing data for differential privacy, and optimizing their effect on utility has already been discussed in literature.Some strategies for achieving differential privacy provide more utility compared to others. Wang et al.[16] proposed a divide-and-conquer approach for function computation in private network analysis. They proposed to equally split the privacy budget among unit function computations. In line with the above approach, we have proposed perturb-then-compose method, which for some privacy budget allocations outperforms the direct approach of noise addition.

CHAPTER 4

REGRESSION ANALYSIS UNDER DIFFERENTIAL PRIVACY

The basic idea behind Regression Analysis is to use data to identify relationships among variables and use these relationships to make predictions. Regression analysis techniques involves solving an optimization problem. Starting with the original data, it is hard to decide on the amount of noise needed to make the optimization results differentially private. We can instead, perturb the objective function of the optimization problem itself. This is the main idea behind the functional mechanism of enforcing ε-differential privacy. Privacy preserving regression model should guarantee protection against attempt of an adversary to infer whether an individual was included in the training set used to learn the regression model.

4.1 Prior Related Work

Only a few methods have been used to preserve privacy in case of regression analysis using differential privacy. One way to do the same is to generate synthetic data in a differentially private way using original sensitive data. This synthetic data can be then used to generate the regression model. However, to generate the synthetic data we need to add Laplacian noise to the original dataset. Unfortunately, this method injects a large amount of noise and so is unable to produce accurate regression results.

Other technique to preserve differential privacy is to perturb the coefficients of the

actual model that we get when applying regression analysis to the original dataset.This technique can then be used to generate newer synthetic data which can be published or can be used for further analysis. Adding Laplace noise directly to the coefficients although preserves privacy, but due to this direct noise addition, the accuracy of model drops. Zhang et al. proposed a functional mechanism for regression analysis under differential privacy. They apply functional mechanism on a simpler form of linear regression and propose an algorithm for the same. We have extended the functional mechanism to a more generic form of linear regression.

4.2 Functional Mechanism for Differentially Private Regression Analysis

Consider D is a database with n records r_1, r_2,....., r_n and d+1 attributes X_1, X_2,....., X_d, Y. A record in a database can be denoted as r_i = (x_{i1}, x_{i2},....,x_{id},y_i).

Our aim is to generate a linear regression model that predicts the value of attribute Y of a record based on the value of attributes X_1, X_2,....., X_d for the same record. Thus, the function f that we obtain as a regression model takes as input (x_{i1}, x_{i2},....,x_{id}) and outputs a prediction of y_i that is as accurate as possible.

Consider ω as the model parameter, which is a d-dimensional vector where the j-th number in the vector (j ϵ 1,2,.....,d) is the weight of x_{ij} in the function f. Additionally, depending on the type of regression model we can have more model parameters. zhang et al. [18] in their work applied functional mechanism to a simpler version of linear regression with ω as the only model parameter and an objective function given by $\omega^* = argmin_\omega \sum_{i=1}^{n} (y_i - x_i^T.\omega)^2$.

Here we consider the most general form of linear regression with model parameters ω and α. The cost function cf helps to evaluate if the model parameters leads to an accurate model. It takes as input r_i, ω and α and outputs a score that measures the distance between original and predicted values of y_i. Let the optimal model parameters be ω^* and α^*. Also consider sum square error of the predicted Y values as the cost function cf. Then,

$$(\omega^*,\alpha^*) = argmin_{\omega,\alpha} \sum_{i=1}^{n} cf(r_i,\omega,\alpha)$$
$$\Rightarrow (\omega^*,\alpha^*) = argmin_{\omega,\alpha} \sum_{i=1}^{n} (y_i - x_i^T.\omega - \alpha)^2$$

The linear regression on D thus gives us a model function $\mathbf{f = x_i^T.\omega^* + \alpha^*}$.

Direct publication of model parameters ω*, α* violates ε-differential privacy since it reveals information about the Database D. This issue can be addressed by adding noise to the model parameters using Laplace mechanism but the sensitivity analysis of ω*, α* is not straight forward due to no direct relationship between the model parameters and the Database D. We instead use functional mechanism to achieve ε-differential privacy and perturb the cost function. We then release the model parameters that minimize the perturbed cost function. To obtain the perturbed cost function we add Laplacian noise to the coefficients of the cost function equation for the database D.

To calculate sensitivity Δ, we first write the polynomial representation of $\text{cf}_D(\omega,\alpha)$. ω and α are vector that contains d values, (ω_1, ω_2, ω_3.....,ω_d) and (α_1, α_2, α_3,....,α_d). Let ϕ denote some product of powers of ω_1, ω_2, ..., ω_d, α_1, α_2,...,α_d. The powers depend on the value of d, for which the regression is being performed. Let Φ_j denote the set of all the products of (ω_1, ω_2, ω_3.....,ω_d) and (α_1, α_2, α_3,....,α_d) with degree j.

For any record r_i, we can write cf(r_i, ω, α) now in terms of ϕ, for all products of model parameters possible. So, for any J $\in$ [0,∞], we can write

$$\text{cf}(r_i, \omega, \alpha) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi r_i} \phi$$

where $\lambda_{\phi r_i} \in$ R denotes the coefficient of ϕ in the polynomial. Using the same method we can express $\text{cf}_D(\omega,\alpha)$ as a polynomial.

We then perturb $\text{cf}_D(\omega,\alpha)$ by injecting Laplace noise into the polynomial coefficients, and then find the parameters $\overline{\omega}$ and $\overline{\alpha}$ that minimizes the perturbed function $\overline{cf_D}$ (ω,α).

Now, let us consider D and D^k be any two neighboring databases. We denote the objective functions of the regression analysis on D and D^k as $\text{cf}_D(\omega,\alpha)$ and $\text{cf}_{D^k}(\omega,\alpha)$, respectively. These are represented as follows:

$$\text{cf}_D(\omega,\alpha) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \sum_{r_i \in D} \lambda_{\phi r_i} \phi,$$
$$\text{cf}_{D^k}(\omega,\alpha) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_d} \sum_{r_i \in D^k} \lambda_{\phi r_i} \phi$$

We now write down a result that we use further to prove that our mechanism is differentially private. The proof of the result is omitted here, but can be found in section 4.1 of [18]. We then give the algorithm **Functional Mechanism** that takes as input Database D, the objective function and the privacy budget and outputs the model parameters ω* and α* that should be published to preserve privacy. The algorithm and the proof of it being differentially private, follows the same steps as the one given in [18], with some minor changes. We include these here for the sake of completeness.

Result 1: Let D and D^k be neighboring databases and $cf_D(\omega,\alpha)$ and $cf_{D^k}(\omega,\alpha)$, as shown above be the objective functions of regression analysis on D and D^k respectively. Then,

$$\sum_{j=0}^{J}\sum_{\phi\in\Phi_j}\left|\left|\sum_{r_i\in D}\lambda_{\phi r_i}-\sum_{r_i\in D^k}\lambda_{\phi r_i}\right|\right|_1 \leq 2max_r\sum_{j=0}^{J}\sum_{\phi\in\Phi_j}||\lambda_{\phi r}||_1.$$

where r_i is an arbitrary record.

Notation	Description
D	Database of n records
$r_i=(x_i,y_i)$	i^{th} record in D
d	number of entries in vector x_i
(ω,α)	the model parameter vectors for generic linear regression model
$cf(r_i,\omega,\alpha)$	the cost function of the linear regression model that evaluates whether model parameters (ω,α) leads to an accurate prediction for a record r_i
$cf_D(\omega,\alpha)$	$\sum_{r_i\in D} cf(r_i,\omega,\alpha)$
$(\omega*,\alpha*)$	$(\omega*,\alpha*) = argmin_{\omega,\alpha}\ cf_D(\omega,\alpha)$
ϕ	a product of one or more values in (ω,α)
Φ_j	the set of all possible $\phi(\omega,\alpha)$ of degree j
$\lambda_{\phi r_i}$	the polynomial coefficient of ϕ in $cf(r_i,\omega,\alpha)$
$\overline{cf_D}(\omega,\alpha)$	noisy version of $cf_D(\omega,\alpha)$
$(\overline{\omega},\overline{\alpha})$	$(\overline{\omega},\overline{\alpha}) = argmin_{\omega,\alpha}\ \overline{cf_D}(\omega,\alpha)$

Table 4.1: Table of notations

Theorem 1: Algorithm 1 satisfies ϵ-differential privacy.

Proof: Without loss of generality, let us consider that D and D^n are two neighbor databases that differ in the last record. Suppose r_n (r'_n) be the last record in D (D^n). Calculation of sensitivity Δ is done on the first line of Algorithm 1, and $\overline{cf_D}$ (ω,α) gets output from Line 9. We need to find the ratio of probability to achieve the same perturbed objective function $\overline{cf_D}$ (ω,α) using databases D and D^n.

$$\frac{Pr(\overline{cf}(\omega,\alpha)|D)}{Pr(\overline{cf}(\omega,\alpha)|D^n)} = \frac{\prod_{j=0}^{J}\prod_{\phi\in\Phi_j}\exp\left(\frac{\varepsilon*\left|\left|\sum_{r_i\in D}\lambda_{\phi r_i}-\lambda_\phi\right|\right|_1}{\Delta}\right)}{\prod_{j=0}^{J}\prod_{\phi\in\Phi_j}\exp\left(\frac{\varepsilon*\left|\left|\sum_{r'_i\in D^n}\lambda_{\phi r'_i}-\lambda_\phi\right|\right|_1}{\Delta}\right)}$$

$$\leq \prod_{j=0}^{J} \prod_{\phi \in \Phi_j} \exp\left(\frac{\varepsilon}{\Delta} * \left\| \sum_{r_i \in D} \lambda_{\phi r_i} - \sum_{r'_i \in D^n} \lambda_{\phi r'_i} \right\|_1 \right)$$

$$= \prod_{j=0}^{J} \prod_{\phi \in \Phi_j} \exp\left(\frac{\varepsilon}{\Delta} * \left|\left| \lambda_{\phi x_n} - \lambda_{\phi x'_n} \right|\right|_1 \right)$$

$$= \exp\left(\frac{\varepsilon}{\Delta} * \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \left|\left| \lambda_{\phi r_n} - \lambda_{\phi r'_n} \right|\right|_1 \right)$$

Now by Using Result 1, we can write

$$\leq \exp\left(\frac{\varepsilon}{\Delta} * 2max_r \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \left|\left| \lambda_{\phi r} \right|\right|_1 \right)$$

$$= \exp(\varepsilon)$$

Thus, we have proved that Algorithm 1 is differentially private.

Algorithm 1 Functional Mechanism (Database D, objective function $cf_D(\omega, \alpha)$, privacy budget ε)

1. Set $\Delta = 2max_r \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \left|\left| \lambda_{\phi r} \right|\right|_1$
2. for each 0 ≤ j ≤ J do
3. for each $\phi \in \Phi_j$ do
4. set $\lambda_\phi = \sum_{r_i \in D} \lambda_{\phi r_i}$ + Lap $(\frac{\Delta}{\varepsilon})$
5. end for
6. end for
7. Let $\overline{cf_D}\ (\omega,\alpha) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \lambda_\phi\ \phi(\omega, \alpha)$
8. Compute $(\overline{\omega}, \overline{\alpha})$ = arg $min_{\omega,\alpha}\ \overline{cf_D}\ (\omega,\alpha)$
9. Return $(\overline{\omega}, \overline{\alpha})$

4.3 Applying Functional Mechanism to achieve Differentially Private Linear Regression

Linear regression finds the linear relationship between the input attributes that fits the input data most. For the sake of simplicity we now consider the case when d=1, that is,

we have only two attributes in total, x and y. We are thus finding a prediction of y using the x values.

To add Laplacian noise to the cost function so that we obtain a perturbed cost function we first need to find the sensitivity Δ. To keep the analysis simple, we take record r_i = (x_i,y_i) in the database D with $x_i \leq 1$ and $y_k \ \epsilon$ [-1,1]. We now calculate sensitivity Δ using line 1 of Algorithm 1.

$$\Delta = 2max_{r_i} \sum_{j=0}^{J} \sum_{\phi \epsilon \Phi_j} \left|\left|\lambda_{\phi r}\right|\right|_1$$

$$\leq 2max_{r_i}(y^2 + x^2 + 2yx + 2x + 2y)$$

$$= 16$$

(As per the conditions on value of attributes in record)

Following the Algorithm 1 we then add Lap ($\frac{16}{\varepsilon}$) noise to each coefficient. Thus, we obtain the perturbed cost function $\overline{cf_D}$ (ω,α). We then differentiate $\overline{cf_D}$ (ω,α) partially with respect to ω and α and equate them to zero.

$$\frac{\partial}{\partial \omega} \overline{cf_D}\ (\omega,\alpha) = 0$$
$$\frac{\partial}{\partial \alpha} \overline{cf_D}\ (\omega,\alpha) = 0$$

Solving the above two equations we get the value of optimized model parameters $\overline{\omega}$ and $\overline{\alpha}$. Finally, we get to release the ε-differentially private linear regression analysis equation $\overline{f}$ given by,

$$\overline{f} = x.\overline{\omega} + \overline{\alpha}$$

Now we perform an experiment to illustrate the process. For example, let us assume we have a two-dimensional database D with six records as shown in Table 4.2. Also let us take an arbitrary value of privacy budget ε. Here, to avoid adding too much noise to the coefficients of cost function and showing the utility of mechanism, we choose ε = 8.

x	y
0.3	0.50
0.4	0.35
1.0	0.9
0.6	0.75
0.8	0.9
0.25	0.2

Table 4.2: Two dimensional database with 6 points

The objective function for linear regression is,

$$cf_D(\omega,\alpha) = 2.3125\omega^2 + 6.7\omega\alpha - 4.5\omega + 6\alpha^2 - 6.8\alpha + 2.275$$

Optimal values of the model parameters are ω^* = 0.7951 and α^* = 0.1228. So, linear regression without any privacy mechanism would output regression line given by,

$$f = 0.7951x + 0.1228$$

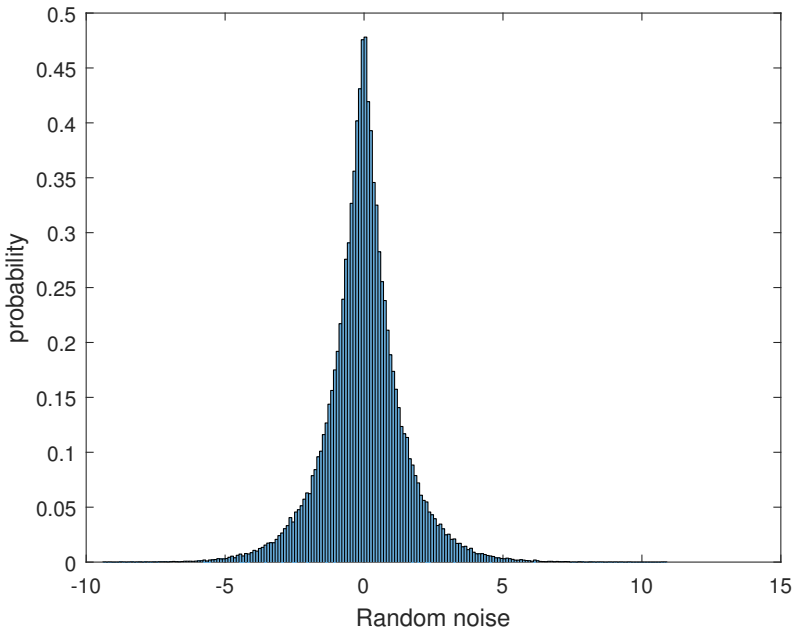

Figure 4.1: Laplace distribution for noise generation in direct perturbation method

By directly adding Laplace noise to the model parameters using distribution given by Figure 4.1, we get regression line given by,

$$f' = 0.4636x + 0.7078$$

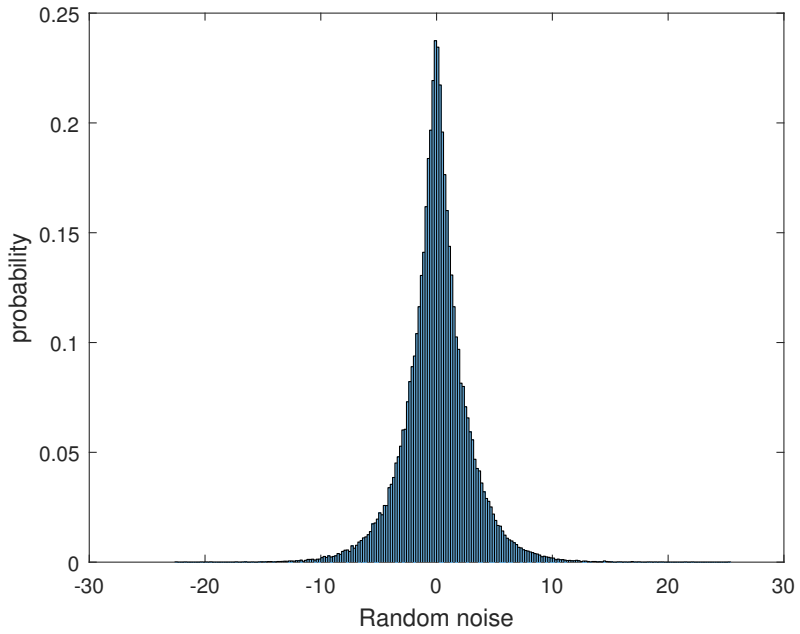

Figure 4.2: Laplace distribution for noise generation in functional mechanism

Now we apply Algorithm 1 by setting $\Delta = 16$. The Laplace distribution used to generate the noise to be added is shown in Figure 4.2.

We then generate the noisy objective function given by,

$$\overline{cf_D}(\omega,\alpha) = 2.4396\omega^2 + 6.802\omega\alpha - 4.5731\omega + 6.2961\alpha^2 - 6.9984\alpha + 2.1652$$

Optimal values of the model parameters that will be published to maintain differential privacy will be $\overline{\omega} = 0.6579$ and $\overline{\alpha} = 0.2004$. Thus, the result of the differentially private linear regression using functional mechanism is given by,

$$\overline{f} = 0.6579x + 0.2004$$

The results of original linear regression, differentially private linear regression using direct perturbation and using functional mechanism along with data points of the database D are plotted in Figure 4.3. It can be easily seen that the accuracy of differentially private linear regression using functional mechanism is more when compared to the direct perturbation technique.

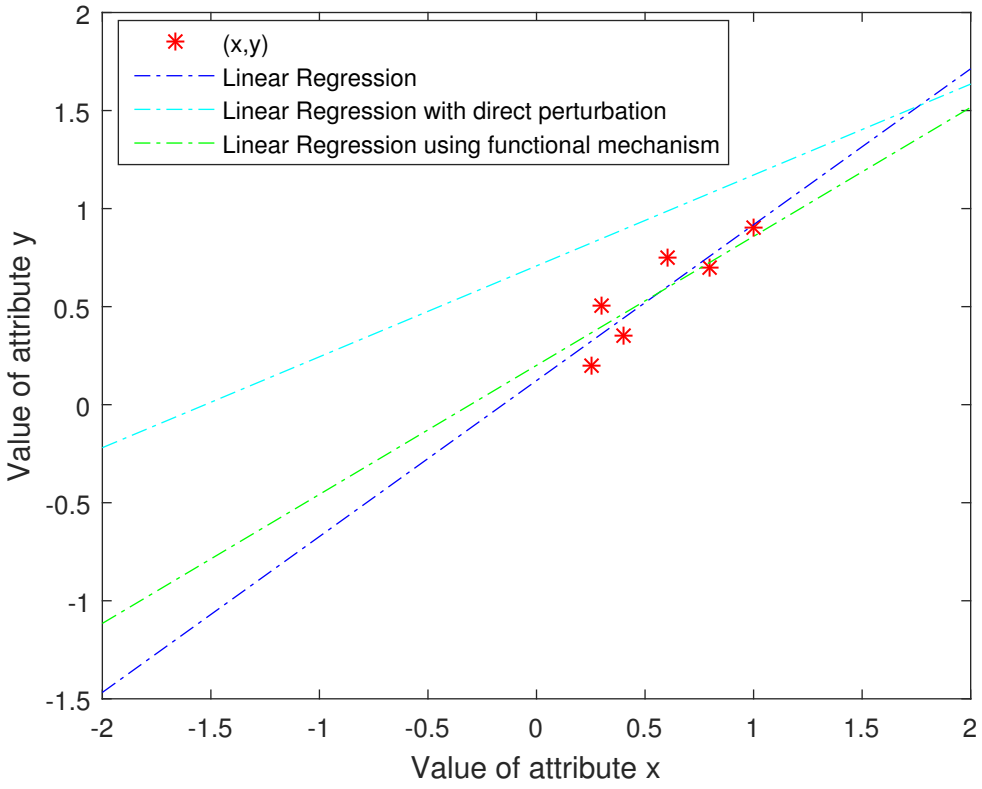

Figure 4.3: Result comparison of different methods of Linear Regression for Database D in Table 4.2

4.4 Conclusion and Future Work

Functional mechanism is a general approach for differentially private regression analysis. Other techniques to achieve privacy in regression tasks add too much noise to the results. The utility of functional mechanism approach is best when the coefficients of the perturbed objective function are approximately preserved, that is, the perturbed coefficients are very close to the coefficients of the original objective function. The generic linear regression model gives a line of the form "$y = w * x + c$", which fits the given data most. Previous work for the differentially private linear regression using functional mechanism focused on the simpler model "$y = w * x$". Here, we extended the functional mechanism to generic linear regression and compared it's result to direct perturbation technique. It was found that for a dataset, the results of functional mechanism are much more accurate.

The functional mechanism approach to differential privacy can be applied to many

other optimization problems also. There are still some aspects for functional mechanism that demand further research

- The utility of the functional mechanism depends on the sensitivity value. More analysis needs to be done to find if sensitivity of the objective function for regression tasks can be lowered or not. If we are able to get a lower value for sensitivity, we will be able to provide more accurate models.
- There is no proper mechanism to choose privacy budget ε for functional mechanism. More research in finding a suitable value for privacy budget in functional mechanism would prevent disclosure of any sensitive information.

CHAPTER 5

RELEASE OF COMPOSITE FUNCTIONS MAINTAINING DIFFERENTIAL PRIVACY

Enabling accurate analysis of data while preserving differential privacy is a challenging task due to the high sensitivity of query functions. Various interesting functions like clustering coefficient in cluster analysis or average, are different from traditional aggregate functions, in a way that these involve computations of two or more aggregate functions. Here, we call these functions as composite functions. In this chapter, we first give two strategies to release composite functions, that preserve ε-differential privacy. We then compare utility of above strategies for releasing average grade in each bucket, for a sample student grade database.

5.1 Strategies to achieve Differential Privacy

Our aim is to release ratio of two measures from a database, preserving differential privacy, and at the same time we want the released value close to the actual value. Average is a composite function that is a ratio of two aggregate functions, sum and count. Mentioned below are the two strategies:

5.1.1 Compose-then-Perturb

In this method, differential privacy is achieved by adding carefully calculated Laplace noise after the value of composite function is calculated. First, we find

global sensitivity Δ for the composite function and then generate noise, from a Laplace distribution with scale parameter given by $\left(\frac{\Delta}{\varepsilon}\right)$, where ε is the user-specified privacy budget, which is then added directly to the value of composite function.

5.1.2 Perturb-then-Compose

In this method, the target function is decomposed into several unit functions, which are less complex and connected by basic mathematical operators. We then perturb each unit function computation with Laplace noise derived from its own sensitivity and chosen privacy budget. Finally, we use the perturbed output of each unit computations to compute the perturbed output of composite function.

Let f denote the value of target computation, and $f_1, f_2, ...f_n$, denotes the value of each of the n unit functions, involved in the target computation. Let sensitivity and chosen privacy budget of these n functions be $\Delta_1, \Delta_2 ...., \Delta_n$ and $\varepsilon_1, \varepsilon_2, ...\varepsilon_n$ respectively. We denote by f_i', the perturbed value of each unit computations and by f' the perturbed value of target computation.

To guarantee ε-differential privacy using this approach, we need to allocate privacy budget among all unit computations maintaining $\varepsilon = \sum_{i=1}^{n} \varepsilon_i$. One possible approach to do this is to distribute privacy budget equally among all unit computations, that is, $\varepsilon_i = \frac{1}{m}\varepsilon$.

For average function, the unit computations involved are sum and count. So, f_1 is computed value of sum and f_2 is computed value of count. These are the perturbed with $Lap\left(\frac{\Delta_1}{\varepsilon_1}\right)$ and $Lap\left(\frac{\Delta_2}{\varepsilon_2}\right)$ respectively. Thus we obtain f_1' and f_2' which we use to compute f'.

Proceeding forward, we first describe the dataset D, whose statistics is to be published. We then give proofs of various results that we use in our approach.

5.2 Dataset Description And Statistics Published

For our experiments, we consider the dataset of student grades. The grades of students are assumed to have any value between 4 - 10. Thus, if we consider an interval of 1, we will have 6 buckets of grades as 4 - 5, 5 - 6,..., 9 - 10. The aim is to publish average grade in each of these buckets maintaining differential privacy. The dataset with student

grades is shown in Table 5.1. A published summary that doesn't provide any privacy guarantee, but provides accurate results for the student grades dataset of Table 5.1 is shown in Table 5.2.

S.No	Name	Grade Points
1	Alice	9.6
2	Bob	9.4
3	Jimmy	9.1
4	Paul	8.7
5	Jeremy	8.8
6	Lynda	8.6
7	Ryan	8.2
8	Mike	7.9
9	Henry	7.4
10	Andy	7.5
11	Judith	7.2
12	Laurel	6.3
13	Jones	6.7
14	Rob	6.4
15	Arya	6.2
16	Jimmy	6.1
17	Emma	5.9
18	Emily	5.7
19	Kevin	5.6
20	Dean	5.3
21	Tim	5.2
22	Christine	4.8
23	Jill	4.9
24	Amanda	4.7

Table 5.1: Dataset with Student Grade records

S.No	Grade Range	Grade Sum	Count	Average Grade
1	4 - 5	14.4	3	4.8
2	5 - 6	27.7	5	5.54
3	6 - 7	31.7	5	6.34
4	7 - 8	30	4	7.5
5	8 - 9	34.3	4	8.575
6	9 - 10	28.1	3	9.367

Table 5.2: Summary of Student Grade records

5.3 Mathematical Results and Proofs

Result 1 The expected value of the quotient of two perturbed results with Laplace noise is equal to the expected value of the two original values without the perturbation.

$$\mathbb{E}\left(\frac{f_1 + Lap\left(0, \frac{\Delta_1}{\varepsilon_1}\right)}{f_2 + Lap\left(0, \frac{\Delta_2}{\varepsilon_2}\right)}\right) = \frac{f_1}{f_2}$$

Proof
Let $f_1' = f_1 + e_1$ and $f_2' = f_2 + e_2$ where $e_1 \sim Lap\left(0, \frac{\Delta_1}{\varepsilon_1}\right)$ and $e_2 \sim Lap\left(0, \frac{\Delta_2}{\varepsilon_2}\right)$.
Since, f_1' and f_2' are independent random variables,

$$\mathbb{E}\left(\frac{f_1'}{f_2'}\right) = \frac{\mathbb{E}(f_1')}{\mathbb{E}(f_2')}$$

Now, using linearity of expectations and that mean for Laplace distribution used is 0,

$$\frac{\mathbb{E}(f_1')}{\mathbb{E}(f_2')} = \frac{f_1 + \mathbb{E}(e_1)}{f_2 + \mathbb{E}(e_2)} = \frac{f_1}{f_2}$$

Result 2 The perturb-then-compose approach for a computation that can be expressed in the form $\frac{f_1}{f_2}$ guarantees ε-differential privacy when $\varepsilon = \varepsilon_1 + \varepsilon_2$, where ε_1 is the privacy budget allocated to f_1 and ε_2 is the privacy budget allocated to f_2.
Proof
Consider that randomized mechanism R_1 and R_2 are independent mechanisms, with privacy guarantees ε_1, ε_2 differential privacy respectively, give response to computation of f_1 and f_2, respectively on a dataset.

Using the sequential composability property [11] of differential privacy, that if there are n independent mechanisms: $R_1, R_2, ..., R_n$, whose privacy guarantees are $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ differential privacy, respectively, then any function of them is $(\sum_{i=0}^{n} \varepsilon_i)$-differentially private. So, the computation will be $(\varepsilon_1 + \varepsilon_2)$-differentially private.
But, we have $\varepsilon_1 + \varepsilon_2 = \varepsilon$, so it guarantees ε-differential privacy.

5.4 Utility Comparison of Proposed Strategies

To do a comparison of the two strategies, we first take an overall value of privacy budget that should be guaranteed by both the strategies. Suppose the overall privacy budget is 2.
For compose-then-perturb approach, consider the following parameter values for publishing the summary:

- Privacy budget $\varepsilon = 2$
- Sensitivity $\Delta = 10$

Table 5.3 shows the actual average grade and perturbed average grade using compose-then-perturb approach.

S.No	Grade Range	Actual Average Grade	Perturbed Average Grade
1	4 - 5	4.8	4.4550
2	5 - 6	5.54	5.0235
3	6 - 7	6.34	6.2314
4	7 - 8	7.5	7.4078
5	8 - 9	8.575	8.7093
6	9 - 10	9.367	9.6608

Table 5.3: Perturbed average grade calculated using compose-then-perturb approach

We now publish the perturbed average grade using perturb-then-compose approach for three different privacy budget allocations:

- Privacy budget for computation of sum function ε_1 = 1.5
- Sensitivity of numerator computations Δ_1 = 10
- Privacy budget for denominator computations ε_2 = 0.5
- Sensitivity of denominator computations Δ_2 = 1

Table 5.4 shows the actual average grade and perturbed average grade using compose-then-perturb approach, with ε_1=1.5 and ε_2=0.5.

S.No	Grade Range	Perturbed Grade Sum	Perturbed Count	Actual Average	Perturbed Average Grade
1	4 - 5	14.5221	3.0021	4.8	4.837
2	5 - 6	25.1185	4.7448	5.54	5.294
3	6 - 7	36.0057	5.6488	6.34	6.374
4	7 - 8	29.861	3.8802	7.5	7.695
5	8 - 9	36.8639	4.4571	8.575	8.270
6	9 - 10	28.3952	3.1274	9.367	9.079

Table 5.4: Perturbed average grade calculated using compose-then-perturb approach with ε_1=1.5 and ε_2=0.5

Next, we publish the summary using the perturb-then-compose approach using different set of parameters as below:

- Privacy budget for numerator computations ε_1 = 0.1
- Sensitivity of numerator computations Δ_1 = 10

- Privacy budget for denominator computations ε_2 = 1.9
- Sensitivity of denominator computations Δ_2 = 1

S.No	Grade Range	Perturbed Grade Sum	Perturbed Count	Actual Average	Perturbed Average Grade
1	4 - 5	18.478	4.390	4.8	4.209
2	5 - 6	25.60	4.9905	5.54	5.129
3	6 - 7	44.5273	7.2402	6.34	6.15
4	7 - 8	35.325	4.458	7.5	7.924
5	8 - 9	27.4943	3.117	8.575	8.8208
6	9 - 10	43.8314	4.4708	9.367	9.804

Table 5.5: Perturbed average grade calculated using perturb-then-compose with ε_1 = 0.1 and ε_2 = 1.9

Now, We publish the summary using the perturb-then-compose approach with third set of parameters as below:

- Privacy budget for numerator computations ε_1 = 1
- Sensitivity of numerator computations Δ_1 = 10
- Privacy budget for denominator computations ε_2 = 1
- Sensitivity of denominator computations Δ_2 = 1

S.No	Grade Range	Perturbed Grade Sum	Perturbed Count	Actual Average	Perturbed Average Grade
1	4 - 5	17.327	3.5955	4.8	4.819
2	5 - 6	27.42	5.1300	5.54	5.345
3	6 - 7	46.128	7.44	6.34	6.20
4	7 - 8	36.309	4.4901	7.5	7.41
5	8 - 9	30.1735	3.4967	8.575	8.629
6	9 - 10	47.278	5.079	9.367	9.308

Table 5.6: Perturbed average grade calculated using perturb-then-compose with ε_1 = 1 and ε_2 = 1

We now compare our approaches of perturb-then-compose for 3 sets of parameters and the compose-then-perturb approach. Table 5.7 shows mean of the L2-norm for total privacy budget of 2.

S.No	Approach	Mean L2-norm error
1	Compose-then-Perturb (ε=2)	0.1027
2	Perturb-then-Compose (ε_1=1.5, ε_2=0.5)	0.0459
3	Perturb-then-Compose (ε_1=0.1, ε_2=1.9)	0.1642
4	Perturb-then-Compose (ε_1=1, ε_2=1)	0.07383

Table 5.7: Mean of squared error for differentially private approaches

From the error values, it can be seen that the perturb-then-compose approach with ε_1 = 1.5, and ε_2 = 0.5, is the approach that has the most utility. For given values of overall privacy budget and individual privacy budget, using above error values, we can decide which approach will have more utility. It can also be deduced that for different set of values of privacy parameter, a different strategy may provide more utility.

5.5 Conclusion and Future work

Of the few strategies that exist to preserve differential privacy while publishing some statistics, there is no definite answer to the question : "Which strategy is better?".

In this chapter, we studied about two approaches to preserve differential privacy when the query is a composite function. We have published range wise average as a summary statistics for a sample student grade database. In our experiments, we have published this summary using four approaches, one compose-then-perturb and other perturb-then-compose. Using mean of squared error, we also find out the best approach to publish the summary. Although the perturb-then-compose method is applied to publish average grades in each range, it can be applied to any function computation which involves calculating unit functions linked with basic mathematical operators.

There is a lot of scope for future work and research. Among them, a few are listed below:

- In this thesis, we have focused on only one composite function, that is, average and it involves two aggregate functions. It would be an interesting to find out, if the work can be extended to unit functions combined with more complex operators, like, log, square root, etc.

- To find the most optimal privacy budget distribution is a hard problem. It would be great to find out how to allocate privacy budget among the unit functions to get more accurate results

- There is no actual way to determine the amount of privacy guarantee a mechanism should provide. Explicitly finding the privacy budget for the task of ratio publishing, without leaking any private information, would also be a good research problem.

CHAPTER 6

CONCLUSION

Differential privacy is a strong privacy notion, but most of the times we need to add a lot of noise. Designing differentially private mechanisms with more utility is an important and challenging task.

In this thesis, we have extended the differentially private linear regression using functional mechanism, to the most generic expression of linear regression. We have given the extended algorithm and showed that it preserves differential privacy. Using a sample dataset, we have verified that the accuracy of linear regression under functional mechanism is more, when compared to direct perturbation techniques.

We then solved the problem of releasing a composite function for a given dataset. The composite function average, can be published by adding noise directly to response or by adding noise independently to numerators and denominators and then finding the response. We call these two approaches compose-then-perturb and perturb-then-compose respectively. Using both the strategies, we published average in each range, for a sample student grade database and compared utility using mean of square error values.

BIBLIOGRAPHY

[1] K. CHAUDHURI AND C. MONTELEONI, *Privacy-preserving logistic regression*, In Proceedings of the 20th Annual Conference on Neural Information Processing Systems, (2008).

[2] T. DALENIUS, *Towards a methodology for statistical disclosure control*, Statistik Tidskrift, (1977), pp. 429–444.

[3] C. DWORK, *Differential privacy*, in Automata, Languages and Programming, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, (2006), pp. vol. 4052, 1–12.

[4] C. DWORK AND A. ROTH, *The Algorithmic Foundations of Differential Privacy*, Foundations and Trends in Theoretical Computer Science, 2014.

[5] B. FUNG, K. WANG, R. CHEN, AND P. YU, *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys, (2010), pp. vol. 42, no. 4, 14:1–53.

[6] S. GOLDWASSER AND S. MICALI, *Probabilistic encryption & how to play mental poker keeping secret all partial information*, Annual ACM Symposium on Theory of Computing, (1982).

[7] J. HSU, M. GABOARDI, A. HAEBERLEN, S. KHANNA, A. NARAYAN, B. C. PIERCE, AND A. ROTH, *Differential privacy: An economic method for choosing epsilon*, In Proceedings of 27th IEEE Computer Security Foundations Symposium (CSF), (2014), pp. 486–503.

[8] D. LEONI, *Non-interactive differential privacy: a survey*, In Proc. of 1st Int. Workshop on Open Data, Nantes, France, (2012), pp. v.6 n.5, 301–312.

[9] N. LI, T. LI, AND S. VENKATASUBRAMANIAN, *t-closeness: Privacy beyond k-anonymity and l-diversity*, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), (2007).

[10] A. MACHANAVAJJHALA, D. KIFER, J. GEHRKE, AND M. VENKITASUBRAMANIAN, *L-diversity: Privacy beyond k-anonymity*, ACM Transactions on Knowledge Discovery from Data, (2007), pp. vol. 1, no. 1, 24.

[11] F. MCSHERRY, *Privacy integrated queries: an extensible platform for privacy-preserving data analysis*, Proceedings of the 35th SIGMOD international conference on Management of data, (2009), pp. 19–30.

[12] F. MCSHERRY AND K. TALWAR, *Mechanism design via differential privacy*, Proceedings of the 48th Annual Symposium of Foundations of Computer Science, (2007).

[13] NARAYANAN AND SHMATIKOV, *Robust de-anonymization of large sparse datasets*, In Proceedings of the 2008 IEEE Symposium on Security and Privacy, (2008), pp. 111–125.

[14] R. SARATHY AND K. MURALIDHAR, *Evaluating laplace noise addition to satisfy differential privacy for numeric data*, Transactions on Data Privacy, (2011), pp. 1–17.

[15] L. SWEENEY, *k-anonymity: A model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, (2002), pp. 557–570.

[16] Y. WANG, X. WU, J.ZHU, AND Y. XIANG, *On learning cluster coefficient of private networks*, ACM International Conference on Advances in Social Networks Analysis and Mining, (2012), pp. 395–402.

[17] J. XU, Z. ZHANG, X. XIAO, Y. YANG, AND G. YU, *Differentially private histogram publication*, In ICDE, (2012).

[18] J. ZHANG, Z. ZHANG, X. XIAO, Y. YANG, AND M. WINSLETT, *Functional mechanism: Regression analysis under differential privacy*, Proceedings of the VLDB Endowment (PVLDB) Vol. 5, No. 11, (2012), pp. 1364–1375.