A Case Study on Differential Privacy

Author: Bihil SELESHI, Samrawit ASSEFFA

Supervisor: Lili JIANG

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in Computing Science

Department of Computing Science
Abstract

Throughout the ages, human beings prefer to keep most things secret and brand this overall state with the title of privacy. Like most significant terms, privacy tends to create controversy regarding the extent of its flexible boundaries, since various technological advancements are slowly leaching away the power people have over their own information. Even as cell phone brands release new upgrades, the ways in which information is communicated has drastically increased, in turn facilitating the techniques in which people’s privacy can be tampered with. Therefore, questioning the methodology by which people can maintain their privacy in the twenty-first century is a validated action undoubtedly conducted by the multitude of the world’s population.

Admittedly, data is everywhere. The world has become an explosion of information, and it should not come as a surprise, especially in a time when data storage is cheap and accessible. Various institutions use this data to conduct research, track the behavior of users, recommend products or maintain national security. As a result, corporations’ need for information is growing by the minute. Companies need to know as much as possible about their customers. Nonetheless, how can this be achieved without compromising the privacy of individuals? How can companies provide great features and maintain great privacy?

These questions can be answered by a current, anticipated research topic in the field of data privacy: differential privacy. Differential privacy is a branch of statistics that aims to attain the widest range of data while achieving a robust, significant and mathematically accurate definition of privacy.

Thus, the objective of this thesis will be describing and analyzing the concept of differential privacy and its properties that lead to the betterment of data privacy. Hence, we will try to study the basic state-of-the-art methods, the model and the challenges of differential privacy.

After analyzing the state-of-the-art differential privacy methods, this thesis will focus on an actual case study that is concerned with two types of different datasets which are experimented with one of the methods of differential privacy methods. We design a basic framework that tries to achieves differential privacy guarantee and evaluate the results regarding the level of privacy achieved.
Acknowledgements

We would like to express our deep, sincere gratitude to our advisor Lili for her relentless, patience, motivation, unending support and most importantly for many insightful conversations we had during the development of ideas in this thesis.
# Contents

Abstract iii  
Acknowledgements v  
1 Introduction 1  
2 Preliminary Knowledge on Privacy Preservation 5  
\[2.1\] Privacy Information .................................................. 5  
\[2.1.1\] Personally identifiable information ......................... 5  
\[2.1.2\] Privacy Preservation Models ................................. 6  
\[2.1.2.1\] Anonymization ................................................. 6  
\[2.1.2.2\] K-Anonymity .................................................. 6  
\[2.1.2.3\] Homogeneity Attack ......................................... 7  
\[2.1.2.4\] Background Knowledge Attack ............................ 8  
\[2.1.2.5\] l-Diversity ................................................... 8  
\[2.1.2.6\] t-Closeness ................................................... 9  
\[2.2\] Privacy Failures .................................................... 9  
\[2.2.1\] Insurance Commission (GIC) ............................... 10  
\[2.2.2\] Search Log ..................................................... 10  
\[2.2.3\] Netflix Prize .................................................. 11  
\[2.3\] Differential Privacy .............................................. 11  
3 Investigation on Differential Privacy 13  
\[3.1\] The Need for Differential Privacy ......................... 13  
\[3.2\] Definition of Differential Privacy ............................ 14  
\[3.2.1\] Queries and Database ....................................... 14  
\[3.2.1.1\] Counting Queries .......................................... 14  
\[3.2.2\] What is Privacy ............................................... 14  
\[3.3\] Properties of Differential Privacy .......................... 19  
\[3.3.1\] The Sensitivity ............................................... 19  
\[3.3.1.1\] The Global Sensitivity ................................. 19
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1.2 The Local Sensitivity</td>
<td>19</td>
</tr>
<tr>
<td>3.3.2 The Privacy Budget</td>
<td>20</td>
</tr>
<tr>
<td>3.3.2.1 Sequential composability</td>
<td>20</td>
</tr>
<tr>
<td>3.3.2.2 Parallel composability</td>
<td>21</td>
</tr>
<tr>
<td>3.4 Mechanisms of Differential Privacy</td>
<td>22</td>
</tr>
<tr>
<td>3.4.1 Laplace Mechanism</td>
<td>22</td>
</tr>
<tr>
<td>3.4.2 The Exponential Mechanism</td>
<td>23</td>
</tr>
<tr>
<td>3.4.3 The Median Mechanism</td>
<td>24</td>
</tr>
<tr>
<td>3.5 Data Release on Differential Privacy</td>
<td>26</td>
</tr>
<tr>
<td>3.5.1 Interactive Data Release</td>
<td>26</td>
</tr>
<tr>
<td>3.5.2 Non-Interactive Data Release</td>
<td>26</td>
</tr>
<tr>
<td>3.6 Challenges in Differential Privacy</td>
<td>27</td>
</tr>
<tr>
<td>3.6.1 Calculating Global Sensitivity</td>
<td>28</td>
</tr>
<tr>
<td>3.6.2 Setting of privacy parameter $\epsilon$</td>
<td>28</td>
</tr>
<tr>
<td>3.6.3 Uncertainty of outcome</td>
<td>29</td>
</tr>
<tr>
<td>4 Related Work</td>
<td>31</td>
</tr>
<tr>
<td>4.1 Early works on Statistical Disclosure Control Methods</td>
<td>31</td>
</tr>
<tr>
<td>4.1.1 Conceptual Technique</td>
<td>31</td>
</tr>
<tr>
<td>4.1.2 Query Restriction Technique</td>
<td>32</td>
</tr>
<tr>
<td>4.1.3 Data Perturbation (Input Perturbation)</td>
<td>33</td>
</tr>
<tr>
<td>4.1.4 Output Perturbation</td>
<td>33</td>
</tr>
<tr>
<td>4.2 Frameworks on Differential Privacy based Data Analysis</td>
<td>33</td>
</tr>
<tr>
<td>4.2.1 Privacy Integrated Queries (PINQ)</td>
<td>34</td>
</tr>
<tr>
<td>4.2.2 Airavat</td>
<td>34</td>
</tr>
<tr>
<td>4.2.3 Fuzz</td>
<td>35</td>
</tr>
<tr>
<td>4.2.4 GUPT</td>
<td>35</td>
</tr>
<tr>
<td>4.3 Work on Differential Privacy based Query Analysis</td>
<td>35</td>
</tr>
<tr>
<td>4.3.1 (Online) Setting</td>
<td>35</td>
</tr>
<tr>
<td>4.3.2 Non-Interactive (Offline) Setting</td>
<td>36</td>
</tr>
<tr>
<td>4.3.3 Histogram and Contingency Table</td>
<td>37</td>
</tr>
<tr>
<td>5 Experimental Framework Setup</td>
<td>39</td>
</tr>
<tr>
<td>5.1 Scenarios Description</td>
<td>39</td>
</tr>
<tr>
<td>5.2 Experimental Framework Overview</td>
<td>39</td>
</tr>
<tr>
<td>5.2.1 User Interface</td>
<td>40</td>
</tr>
<tr>
<td>5.2.2 Privacy Preservation Mechanism</td>
<td>40</td>
</tr>
</tbody>
</table>
5.2.3 Datastore ................................................. 41
5.2.4 Programming Language ................................. 41
5.3 Algorithm Description ....................................... 41
5.3.1 Algorithm for Laplace Mechanism ....................... 42

6 Experimental Study on Differential Privacy 43
6.1 Adult Dataset ................................................... 43
6.1.1 Dataset Description ......................................... 43
6.1.2 Experimental Design ........................................ 44
6.1.3 Results and Discussion ...................................... 47
6.2 Student Alcohol Consumption Dataset ...................... 49
6.2.1 Dataset Description ......................................... 49
6.2.2 Experimental Design ........................................ 51
6.2.3 Results and Discussion ...................................... 58

7 Conclusion and Future Work 65

Bibliography 67
List of Figures

3.1 Laplace Distribution with various distribution ......................... 22
3.2 Interactive and Non-Interactive data release ......................... 27
5.1 Architecture of the system ............................................. 40
6.1 Graph for the Number of years in Education by Income and Race . 46
6.2 Students Alcohol consumption and Health status .................... 55
6.3 Alcohol consumption and Health status with noise .................. 55
6.4 Alcohol consumption and absence days .............................. 56
6.5 Alcohol consumption and Absence days with noise .................. 56
6.6 Alcohol consumption and Final grade ............................... 57
6.7 Alcohol consumption and Final grade with noise .................... 57
6.8 Alcohol consumption and Final grade with high noise ............... 58
6.9 Result of simple linear regression of health status on alcohol con-
sumption ............................................................... 59
6.10 Result of simple Linear Regression of health status with added noise on alcohol consumption ................................. 59
6.11 Result of simple Linear Regression of school attendance on alcohol consumption .................................................... 60
6.12 Result of simple Linear Regression of school attendance with added noise on alcohol consumption ............................. 61
6.13 Result of simple Linear Regression of school performance on alcohol consumption ....................................................... 62
6.14 Result of simple Linear Regression of school performance with added noise on alcohol consumption (ϵ=0.1) ....................... 62
6.15 Result of simple Linear Regression of school performance with very high noise on alcohol consumption (ϵ=0.0001) ................. 63
List of Tables

2.1 Anonymity table of medical data .......................... 7

6.1 Adult dataset attributes ................................. 44
6.2 Mean query result for adult dataset ......................... 47
6.3 Result for Laplace mechanism for counting query ............ 48
6.4 Result for Laplace Mechanism for Mean query ............... 49
6.5 The pre-processed student related variables ................. 50
# Listings

<table>
<thead>
<tr>
<th>Listing</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listing 6.1</td>
<td>Counting query for adult dataset</td>
<td>45</td>
</tr>
<tr>
<td>Listing 6.2</td>
<td>Mean query for adult dataset</td>
<td>46</td>
</tr>
<tr>
<td>Listing 6.3</td>
<td>Connection string</td>
<td>51</td>
</tr>
<tr>
<td>Listing 6.4</td>
<td>Getting the data from the database</td>
<td>52</td>
</tr>
<tr>
<td>Listing 6.5</td>
<td>Apply Laplace Mechanism to the data</td>
<td>52</td>
</tr>
<tr>
<td>Listing 6.6</td>
<td>Query for number of students absence days</td>
<td>52</td>
</tr>
<tr>
<td>Listing 6.7</td>
<td>Apply Laplace Mechanism on the data</td>
<td>52</td>
</tr>
<tr>
<td>Listing 6.8</td>
<td>Query for the effect of Alcohol consumption</td>
<td>53</td>
</tr>
<tr>
<td>Listing 6.9</td>
<td>Apply Laplace Mechanism</td>
<td>53</td>
</tr>
<tr>
<td>Listing 6.10</td>
<td>Apply Laplace Mechanism</td>
<td>53</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

In the past decades, personal data records are collected increasingly by governments, health centers, social networks, organizations, companies, and individuals for data analysis and other different purposes. Hence, this received data has created opportunities for researchers, companies, organizations and decision makers. For example, medical records can be used to track the spread of disease, prevent epidemics, discover hidden links between illnesses, disease prevention, and early detecting and controlling of disease, etc. [40]. On the other hand, the collected data might be sold, exchanged or shared. For example, organizations can deliver their customers information to third parties like advertising agents.

Sharing and exchanging this large set of data has been a key critical tool for accomplishing specific requirements of the researchers and the organizations. For example, e-commerce, an information exchange created from plenteous activities including searching, browsing, and Internet shopping, which could improve productivity and also in the medical field, there is a huge amount medical record frameworks for the purpose of exchanging vital information. While releasing and sharing the data create new opportunities and help researchers and individuals with better data analysis, it is crucial protecting the privacy of each person’s information in the dataset. If anybody can be explicitly distinguishable from the released data, their private data will potentially be compromised. Therefore, before releasing the dataset, the data curator must secure data privacy preservation in a way that the individual’s identity contained in the information cannot be perceived.

In a setting where a trusted data creator or custodian owns a database consisting of rows of data with practical information about specific information, a privacy breach occurs when an adversary infer this particular information. The adversary preys on the background information and even if the data creator has published the anonymized version of the data this problem is commonly known as privacy-preserving data publishing.
Many anonymization techniques have been proposed to release the dataset securely [36, 1]. However, the dependence of the currently available privacy models on the background awareness of the adversary makes it hard to protect it from unpredicted auxiliary information. Because of their vulnerability to background knowledge attacks these techniques were not successful in preserving privacy [3, 30, 26, 1].

We will try to give a solution to the above-described problem with the approach of differential privacy [10]. Differential privacy is a concept that is currently growing rapidly and used in a wide range of aspects, such as starting from Apple’s iOS 10¹ to Google’s Chrome for Google Report Project. Differential privacy tries to guarantee the protection of this sensitive information about an individual irrespective of the background knowledge of the attacker. Thus, we will present the analysis of differential privacy and study its applicability on a chosen sample dataset. The main contributions of this thesis include:

- An extensive investigation on privacy preservation and differential privacy, including state-of-the-art privacy protection methodologies, differential privacy frameworks as well as their pros and cons.
- An experimental framework, which consists of back-end database deployment and algorithm implementation, to put differential privacy into practice and evaluation.
- Experimental study on differential privacy to evaluate various kinds of queries by applying one of the mechanisms in differential privacy on different datasets.

This thesis is organized as follows:

- In Chapter 2, we describe the pre-existing methodologies before the idea of differential privacy on privacy preservation; specifically, we present the basic understanding of the syntactic privacy models that use privacy preserving data publishing techniques such as k-anonymity, l-diversity, t-closeness and in the last section we will introduce differential privacy.

- In Chapter 3, we present the foundation of differential privacy that we depend on all the way throughout the chapters and continue defining differential privacy while showing known mechanisms and techniques for achieving differential privacy.

differential privacy. We then finalize the chapter by describing the significant challenges and the currently available frameworks of differential privacy.

- In Chapter 4, we present instances of previous works which aim to provide security to databases against information leakage and describe the current existing framework for differential privacy.

- In Chapter 5, we present the framework that we have applied to test out differential privacy and describe one of the algorithms that we have used in the experiment.

- In Chapter 6, we continue the investigation and present the process of achieving differential privacy by giving out the results and discuss the implication of differential privacy.

The roles and the responsibilities throughout the course of this thesis is as follows,

- Samrawit has the responsibility of doing the research on the preliminary study, related work and experimental study on differential privacy, and she has covered Chapter 1, Chapter 2, Chapter 4 and Chapter 6.2

- Bihil has the responsibility of doing the research on the properties, analysis, implementation and experimental part of the differential privacy and has covered Abstract, Chapter 3 Chapter 5, Chapter 6.1 and Chapter 7 part.
Chapter 2

Privacy Preservation and Failure

2.1 Privacy Information

A data privacy perception model by Dalenius in 1977 [9] articulated a desideratum that depicts the privacy goal of databases: anything that can be learned from the database about a particular individual should be determined without the access to the database. The idea behind this notion is making sure that the measurement of the adversary’s before and after beliefs of a particular data is small. However, this type of privacy cannot be achieved. Dwork demonstrated that such privacy assurance is inconceivable because of the existence of background knowledge. Thus, Dwork came up with a new perspective of privacy preservation: the risk to one’s privacy, or in general, any risk, such as the risk of being denied automobile insurance, should not substantially increase as a result of participating in a database [9, 26].

Formerly, many works have been done to protect privacy; we will discuss the syntactic privacy models that use privacy preserving data publishing such as k-anonymity, l-diversity, t-closeness, and about differential privacy on the next sections.

2.1.1 Personally identifiable information

The collected data from different sources stored in a database. Mostly the privacy preserving data publisher categorize the database in four primary fields including Explicit Identifier, Quasi Identifier, Sensitive Attributes, and Non-Sensitive Attributes.

Specific Identifiers is a set of attributes that contains information that can be used to identify individuals such as name and security number uniquely. Quasi-identifier
(QID) is a set of attributes such as zip code, gender, a birth date in which the combination of these attributes could potentially identify individuals while sensitive attributes contain sensitive personal information such as medical history and salary. Non-sensitive attributes include attributes that are not listed in the other fields [39]. Among these categories, personally identifiable information (PII) is information that can be utilized all alone or with other data to identify individuals. For example, social security number, name, and phone number [6]. Before releasing the data, the curator has to ensure that individual’s personally identifiable information will not be revealed while the data is still valuable. Many privacy preserving methods have been proposed earlier, and we will discuss some of them below.

2.1.2 Privacy Preservation Models

2.1.2.1 Anonymization

Data anonymization is the process of removing personally identifiable information from the data set to protect individuals’ privacy and to make it possible for data users and owners to share data securely for data analysis, decision-making, research and different purposes so that individuals whose information is in the data set stay anonymous. The curator (the person who collected the data) modify the data by removing the specific identifiers such as name, security number, address and phone number. Even if the specific identifiers are removed, the availability of individual’s background information (e.g. in the public voter list) makes it easier for the adversary to re-identify individuals by linking the released data making it very hard to publish data without disclosing privacy [3]. Once the data is released to the third party, it is hard for the owners to control the way the data is manipulated. Latanya Sweeney, an MIT graduate student in computer science, had shown that individual’s information in the anonymously published data could be re-identified by linking the released data to publicly available data by re-identifying governor William weld’s medical information [3].

2.1.2.2 K - Anonymity

To deal with the shortcomings of simple data anonymization, researchers have proposed various methods to preserve privacy. One of the most popular methods of privacy preservation is the K-anonymity. To counter record linkage using quasi-identifiers Samarati and Sweeney [36] proposed the idea of k-anonymity, its endeavor is to release data with a scientific guarantee that a particular individual’s
data cannot be uniquely distinguished while the data could be utilized in a sensible manner. A k-anonymized data set has the property that each person contained in the record is similar to at least another k-1 other records on the potentially identifying variables. k-anonymity is defined as the level of data protection on inference by linking. It prevents linking the released data to other information sources (background information). Meanwhile, k-anonymity does not guarantee privacy. Machanavajjhala et al. [25] used two attacks to show how k-anonymity does not guarantee privacy. Let us discuss the two attacks based on the table 2.1:

<table>
<thead>
<tr>
<th>Name</th>
<th>Age</th>
<th>Zip code</th>
<th>Religion</th>
<th>Nationality</th>
<th>Medical History</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>20 ⩽ Age ⩽ 30</td>
<td>110**</td>
<td>*</td>
<td>USA</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>*</td>
<td>20 ⩽ Age ⩽ 30</td>
<td>110**</td>
<td>*</td>
<td>USA</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>*</td>
<td>20  ⩽ Age ⩽ 30</td>
<td>110**</td>
<td>*</td>
<td>Norway</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>*</td>
<td>40  ⩽ Age ⩽ 45</td>
<td>120**</td>
<td>*</td>
<td>Norway</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>*</td>
<td>40  ⩽ Age ⩽ 45</td>
<td>120**</td>
<td>*</td>
<td>Norway</td>
<td>Pneumonia</td>
</tr>
<tr>
<td>*</td>
<td>⩾ 46</td>
<td>130**</td>
<td>*</td>
<td>Mexico</td>
<td>Vitamin D deficiency</td>
</tr>
<tr>
<td>*</td>
<td>⩾ 46</td>
<td>130**</td>
<td>*</td>
<td>Mexico</td>
<td>Vitamin D deficiency</td>
</tr>
<tr>
<td>*</td>
<td>⩾ 46</td>
<td>130**</td>
<td>*</td>
<td>Mexico</td>
<td>Alzheimer’s</td>
</tr>
</tbody>
</table>

2.1.2.3 Homogeneity Attack

Homogeneity attack showed that when there is little diversity in the sensitive attributes, the adversary can identify the value of the sensitive attribute for that group of k-records. For example, a politician who intends to be elected to a post in the governance of a state utilizes the medical history of his opponent in demonstrating to the populace that his opponent cannot or is not ready to deal with the obligations as an agent of the state due to his medical problems. He will have to search for his opponent’s medical information by utilizing the released data of the 3-anonymous table from the hospital. Despite the likelihood that the data is a 3-anonymized table. Since he has some information about his opponent, he can recognize what ailment his opponent has because when there is no much contrasts (there is little diversity) in the sensitive data. Case in point, he knows that the patient is 25 years old American who lives in the postal division 11003, so due to this current data, he realizes that his rival has Heart Disease.
2.1.2.4 Background Knowledge Attack

In this attack, the adversary uses background knowledge to make the attack successful, and we will show that k-anonymity does not guarantee privacy against background knowledge attacks. For instance, a woman whose colleague’s father is sick needs to comprehend the nature of the sickness. She knows that her co-worker dad is old and he is from Mexico so she can conclude that he is suffering from either vitamin D inadequacy or Alzheimer’s. Nonetheless, it is realized that Mexicans, for the most part, could not be influenced by vitamin D insufficiency malady. Alzheimer’s is a common neurological ailment in old people. Therefore, it is easy for her to conclude that her colleague’s father has Alzheimer’s. Using background knowledge, she distinguishes what malady her colleague’s dad has. Therefore, from the above examples, it can be seen that k-anonymity does not guarantee privacy preservation.

2.1.2.5 l-Diversity

k-anonymity protects privacy against the identification of records, however, it is not generally successful for protecting privacy against inference attacks of the sensitive attributes. To address this problem, Machanavajjhala et al. [26] proposed a new notion called l-diversity which requires that each tuple that shares identical quasi-identifiers has at least l-diverse well-represented values for the sensitive attribute. Machanavajjhala et al. stated: equivalence class is said to have l-diversity if it contains at least l well-represented values for the sensitive attribute S. A table with equivalence classes, all of which are l-diverse, is said to be an l-diverse table. In short, L-diversity is a model which promotes intra-group heterogeneity of sensitive attributes by at least "L" different values.

Though l-diversity was proposed to address the shortcomings of k-anonymity, such as attribute linkage, Li et al. [25] showed that l-diversity does not address the issue of attribute disclosure sufficiently. To demonstrate this, they presented two attacks: Skewness attack (when the dataset has skewed distribution; and similarity attack (when the equivalence class contains distinct but semantically similar sensitive attribute values). In the former scenario, l-diversity fails to prevent attribute disclosure because the distribution for the real population is different from the dataset. This makes the distribution of sensitive attribute within equivalence
class differ from the real population creating attribute disclosure. In the latter sce-
nario, the adversary estimates the value of a sensitive attribute by first linking it
with another sensitive attribute.

2.1.2.6 t-Closeness

To prevent the limitations of l-diversity [25] proposed a notion of privacy called
t-closeness. The formal definition of t-closeness given by Li et al. is stated below.

An equivalence class is said to have t-closeness if the distance between
the distribution of a sensitive attribute in this class and the distribution
of the attribute in the whole table is no more than a threshold $t$. A table
is said to have t-closeness if all equivalence classes have t-closeness.

An equivalence class is a set of data that have the same values for their quasi iden-
tifiers.

2.2 Privacy Failures

Having huge amount of collected data and multiple ways of sharing has led to a
rapidly increasing accumulation of personal data, thereby leading to privacy is-
issues such as exposure of sensitive data and mass harvest of personal information
by third parties [39]. Studies demonstrate that majority of the US population can
be uniquely distinguished by joining zip code, gender and date of birth showing
clearly information released containing these attributes cannot be seen as anony-
mous data [38, 15, 39]. If individuals can be particularly distinguished in the re-
leased data, their private data will potentially be disclosed. The collected data sets
contain private or sensitive information of individual’s and releasing this data sets
can lead to the disclosure of personally sensitive information.
Protecting the database from background knowledge attacks is very challenging.
The adversary’s background information which the data curator cannot foresee
makes privacy preservation models vulnerable. As examined in the preceding sec-
tion the privacy standards do not protect privacy. Below we try to discuss instances
of privacy protection failures which led to the identification of individuals or users.
2.2.1 Insurance Commission (GIC)

Latanya Sweeney, an MIT graduate student in computer science, had shown that individual’s information in the anonymously released data could be re-identified by linking the released data to publicly available data (e.g., voter registration list) and by using some background knowledge about the individual or event. She had demonstrated this by re-identifying the medical data of the governor William Weld’s data, the Governor of Massachusetts by using hospital’s data which was released to the researchers by the Massachusetts Group Insurance Commission (GIC) since it is accepted to be anonymous. It was publicly known that the governor had collapsed on May 18, 1996, during the event of the graduation ceremony when he was about to receive an honorary doctorate. Even though GIC had modified the released data by deleting specific identifiers, Sweeney used her background knowledge from media regarding the hospital’s name where the governor was admitted to and regarding his residential address to identify the specific hospital from the released GIC data. She also bought the voter list which contains many attributes including individual’s name, address, zip code, birthdate, and gender from the governor’s residential county so that she can match the quasi-identifiers such as zip code, gender, date of birth, with the GIC data to re-identify the governor’s medical data. Her work proved that though data publishers release data set after anonymizing it by having all the personal identifiers removed, the remaining data can be used to identify individuals by linking it to other data, such as publicly available data sets [3, 41].

2.2.2 Search Log

In August 2006, American online (AOL)\(^1\) research publicly released 20 million detailed search logs of many numbers of AOL users collected over a three months’ period for research purposes. The AOL did not remove any data which brought privacy concerns; they tried to anonymize it by replacing identifiers such as AOL username and IP address with unique identification numbers to allow researchers to relate the searches to that of the individual’s. However, based on all the searches made by a single user it was possible to identify individual’s name quickly, social security numbers, and even more sensitive information is disclosed. Following

\(^1\)Chronicle of AOL search query log release incident, [http://sifaka.cs.uiuc.edu/xshen/aol/aol_querylog.html](http://sifaka.cs.uiuc.edu/xshen/aol/aol_querylog.html)
that, within a couple of days there was an article in The New York Times revealing the identity of one of the searchers, and subsequently, many other people were identified. AOL acknowledged the mistake and removed the released data and apologized for releasing, however, they couldn’t control the information leakage because the data was redistributed by others.²

2.2.3 Netflix Prize

In October 2006 Netflix the world’s largest online DVD rental service publicly released a data set containing 100 million anonymized movie ratings, created by 500,000 subscribers of Netflix. The purpose of the release was to improve their movie recommendation system. At the time, they secured the customer’s privacy by removing all the personal information they released the data that only contains an anonymized user Id, ratings, and the dates the subscriber rated the movie. Narayanan and Shmatikov from the university of Texas at Austin [30] demonstrated that an adversary who knows a little bit about individuals subscription could identify the subscriber’s information in the data set. By using IMDB (The Internet Movie Database) [29] as the source of the background knowledge they identified the subscriber’s record, and they have revealed the customer’s personal sensitive information [30, 29].

Protecting the database from these attacks is very challenging especially while the adversary has sufficient background information at his/her disposal that is even not anticipated by the data curator. Hence, there is a need for a better privacy preserving technique to prevent the leakage of individual’s private information and give a much better guarantee to preserve privacy against the worst-case scenarios where the adversary has almost all the background information. Thus, In this thesis, we choose differential privacy as a privacy model.

2.3 Differential Privacy

Differential privacy is a powerful standard for data privacy proposed by Dwork [10]. It is based on the idea that the outcome of the statistical analysis is essentially equally likely independent of whether any individual joins or refrains from joining the database, i.e., one learn approximately the same thing either way [27]. It gives a guarantee that the possibility that the adversary that brings harm or good

to any set of participants is basically the same regardless of whether or not any individual is in or out of the dataset. To accomplish this, Differential privacy adds a random noise to the output of the query so that the difference to the results of the output make by the presence or the absence of a single person will be covered up. Differential privacy has been studied theoretically and proved that, It gives a rigorous guarantee of privacy even when the adversary has the worst-case background knowledge [9]. It neutralizes all linkage attacks and statistical attacks because it uses the property of the data access mechanism which is not reliant on the presence or absence of background knowledge.

Thus, because of its strong privacy guarantee against the worst case background knowledge attacks of the adversaries, differential privacy has been considered as a promising privacy preserving technique. Therefore, throughout this thesis, we will try to describe its properties and analyze a selected case study on it.
Chapter 3

Investigation on Differential Privacy

In this chapter, we present the foundation of differential privacy that we depend on all the way throughout this thesis. First, we will give a summary about the need for data privacy and define the concept of data domain and query that we are going to include. Then, we formally define differential privacy and present the known underlying mechanisms and techniques for achieving differential privacy. We then finalize the chapter by describing the significant challenges that exist with differential privacy.

3.1 The Need for Differential Privacy

In today’s information realm, an extensive sensitive personal information is being possessed by the daily services we tend to use, such as search engines, mobile services, on-line social activity and so on. This vast amount of statistical sensitive personal information can be of enormous social value such as enhancing economic utility, comprehending the spread of disease, allocating resources and so on. As described in the last chapter information is obtained in several ways, starting with an opportunistic data collection that is simply promised “privacy” to a legally compelled one which both should be equally treated since there is no logical reasoning to engaging in the actions that generate them in the first place.

The loss of data privacy is imminent since the guarantee of data privacy incorporates controlling access to information, controlling the flow or purpose of information. Hence, as seen in the last chapter several methods try to preserve privacy which is insufficient enough to give the desired data privacy. Thus, the need for differential privacy arises with the hope of a better and robust data privacy.
3.2 Definition of Differential Privacy

Here we introduce some notation and the definition of differential privacy.

3.2.1 Queries and Database

Let $D = (d_1, \ldots, d_n) \in D$ be an input database or dataset (we will use these terms interchangeably) in which $d_i$ represents a row or an individual, and where $D$ is the space of all such databases of $n$ elements. Informally, any particular record is inconsiderate to a differentially private output. Therefore, the computing when seen from the point of view of any particular data is as if from a dataset that does not include it. Thus, we come to the concept of neighboring datasets.

**Definition 3.2.1.** The distance between two adjacent or neighboring datasets $D, D'$ with the $l_1$ norm is:

$$||D - D'||_1 = \sum_{i=1}^{|D|} |D_i - D'_i| \leq 1$$

3.2.1.1 Counting Queries

Counting queries are the essential class of queries that give some basic statistics on a database. They have a form "What portion of the database satisfy the property $q$?". Formally, it could be determined by a boolean predicate $q : X \rightarrow \{0, 1\}$. We use,

$$q(D) = \frac{1}{n} \sum_{i=1}^{n} q(x_i)$$

to denote the evaluation of the counting query on a database.

3.2.2 What is Privacy

The definition of privacy has a broad different aspects of understanding, and it might be a little profound to nail down precisely. Some have tried to define it from philosophical point of view, like Warren and Brandeis [43] seen it as "right to be let alone". While Dalenius [8] in 1977 tried to describe it as " anything that can be learned about a respondent from the statistical database should be learnable without access to the database". Some are more specialized and have extraordinarily less succinct quotes portraying their intent. Every definition endeavors to represent and inspire a particular kind of privacy, keeping in mind this is an astounding
academic interest, it does not regularly translate into immediately helpful guidance for the concerned data subject.

It is evident that the objective of privacy-preserving data analysis is to release information without giving up the privacy of any individuals whose data contribute to the database. Having data that is sensible enough to be used as utility and providing a secured privacy are the two compromises that pose conflicting objectives of differential privacy.

Thus, differential privacy is yet another privacy definition which is more unmistakably actionable and directly concerned with a specific concern the data subject have, what will be the result of in participating in a particular database, but it is also clearly both philosophical and technical. Differential privacy ensures that nothing will happen with access to your data, given that it could not occur without access to your data. Moreover, it makes a stronger guarantee [11] of privacy even when the adversary has arbitrary external knowledge. The possibility that ‘any particular thing’ happens with the access to your information is at most a multiple ‘X’ of chance it would happen without your data, where ‘multiple’ X determines how much privacy is guaranteed.

Before we get into the definition of differential privacy, we will start by defining randomized algorithm since the property of differential privacy is related to it.

**Definition 3.2.2.** (Randomized Algorithms[11]) A randomized algorithm $M$ with domain $A$ and discrete range $B$ is associated with a mapping $M : A \rightarrow (B)$. On input $a \in A$, the algorithm $M$ outputs $M(a) = b$ with probability $(M(a))_b$ for each $b \in B$. The probability space is over the coin flips of the algorithm $M$.

From the definition we could see that randomized algorithm is a deterministic algorithm, that takes two inputs: the data set and the string of random bits. As we are going to see the definition of differential privacy which is the probability of the randomness of the internal algorithm that holds the data set fixed. Hence, the important concept of the above definition where the probability space is over the coin flips of the algorithm $M$ implies that it is the source of randomness.
Definition 3.2.3. \((\epsilon\text{-differential privacy})\) [10] Let \(\epsilon > 0\). Define a randomized function \(\mathcal{M}\) to be \((\epsilon\text{-differentially private})\) if for all neighboring input datasets \(D_1\) and \(D_2\) differing on at most one element, and \(\forall S \subseteq \text{Range}(\mathcal{M})\), we have

\[
\frac{Pr[\mathcal{M}(D_1) \in S]}{Pr[\mathcal{M}(D_2) \in S]} \leq e^\epsilon
\]

where the probability is taken over the coin tosses of \(\mathcal{M}\).

The definition also implies a lower bound: since we can interchange \(D_1\) and \(D_2\) mutually

\[
Pr[\mathcal{M}(D_1) \in S] \geq e^{-\epsilon}.Pr[\mathcal{M}(D_2) \in S]
\]

That is, the probability of an output in \(S\) on a dataset \(D_1\) is at least \(e^{-\epsilon}\) times the probability of output in \(S\) on a neighboring dataset \(D_2\)\footnote{If some outputs are impossible on one input dataset, they must be impossible on all inputs, for example, if \(Pr[\mathcal{M}(D_1) \in S] = 0\) for some \(D_1\) and \(S\), \(Pr[\mathcal{M}(D_2) \in S] = 0\) for all dataset \(D_2\).}

A common weakening of \(\epsilon\text{-differential privacy}\) is the following notion of approximate privacy [19]

Definition 3.2.4. \(((\epsilon, \delta)\text{-differential privacy})\) [19]

Define a randomized function \(\mathcal{M}\) to be \(((\epsilon, \delta)\text{-differentially private})\) if for all neighboring input datasets \(D_1\) and \(D_2\) differing on at most one element, and \(\forall S \subseteq \text{Range}(\mathcal{M})\), we have

\[
Pr[\mathcal{M}(D_1) \in Y] \leq e^{\epsilon}.Pr[\mathcal{M}(D_2) \in Y] + \delta
\]

In this case the two parameters \(\epsilon\) and \(\delta\) control the level of privacy.

The strongest version of differential privacy, in which \(\delta = 0\), is known as \textit{pure differential privacy} while the more general case where \(\delta > 0\) is known as \textit{approximate differential privacy}, and is well less understood [14]. For instance, suppose that \(y\) is an output that discloses user \(z\)'s data where the dataset \(D_2\) does not contain \(z\)'s information under the assumption where \(Pr[\mathcal{M}(D_2) \in S] = 0\). While, \(\epsilon\text{-differential privacy}\) \(\mathcal{M}\) can never output \(z\) on any dataset \(((\epsilon, \delta)\text{-differential privacy})\), \(\mathcal{M}\) may output \(z\) with probability up to \(\delta\).

It is important to note that \(\epsilon\text{-differential privacy}\) is a property of the randomized algorithm \(\mathcal{M}\) but not of the dataset. The input dataset \(D_1\) generate a probability distribution on its range (\(\mathcal{R}(\mathcal{M})\)). Thus, the ratio of two probabilistic density functions comparable to the distribution generated on \(\mathcal{R}(\mathcal{M})\) by \(D_1\) and \(D_2\) concerning
the parameter $\epsilon$ as a boundedness condition is differential privacy. The privacy parameter $\epsilon$, when it is closer to zero, the closer the distribution are, and the higher the level of the privacy.

From the above definitions of differential privacy, it is quite evident what privacy means. Thus, the information acquired regarding a participant by the output of some algorithm is no more than the information we can acquire about that participant without the access to the output which we informally call as a pure semantic security. In fact, it is clarified in the semantically-flavoured explanation [9] which states regardless of external knowledge, an adversary with access to the sanitized database draws the same conclusions whether or not my data is included in the original database. Unluckily, the presence of arbitrary external information makes such privacy definition impossible. We could see this in the illustrated example [22]:

Consider a clinical study that explores the relationship between smoking and lung disease. A health insurance company who had any priori understanding of that relationship might dramatically alter its “beliefs” (as encoded by insurance premiums) to account for the results of the study. The study would cause the company to raise premiums for smokers and lower them for non-smokers, regardless of whether they participated in the study. In this case, the conclusions drawn by the company about the riskiness of any one individual (say Alice) are strongly affected by the results of the study. This occurs regardless of whether Alice’s data are included in the study.

From this understanding, rather than having a pure semantic privacy, we should target to a more proper definition of privacy that shows differential privacy obtains a relaxed version it. It is known that differential privacy states whether or not being in the output of an algorithm that describes a single user’s data, the amount of data that is learned by the adversary is practically the same. Thus, an assurance must be formed so that partaking in the statistical database will not utterly change the outcome functions run on the database for the sake of the user.

Supposing that the user participates in the database, we now formalize the argument that we have given. Thus, to bound statistical difference mathematically between the before presumptions/beliefs $b_1, b_2$, where $b_1$ is a before on the values in the database given an output $y = \mathcal{M}(D_1)$ where $D_1$ includes the user data, and $b_2$ is the after on the database given the same output $y = \mathcal{M}(D_2)$, where $D_2$ does not include the user’s data.
Differential privacy implies (relaxed) semantic privacy. [22]
To understand the concept of semantic privacy we need to define the statistical difference. If \( P \) and \( Q \) are two distributions on the same discrete probability space, the statistical difference between \( P \) and \( Q \) is defined as:

\[
SD(P, Q) = \max_{S \subseteq D} |P[S] - Q[S]|
\]

Our expectation is a randomized algorithm \( \mathcal{A} \) that satisfies \( \epsilon \)-differential privacy. Let \( b(D) \) denote the prior belief of the adversary on databases \( D \in \mathcal{D}^n \) and \( b(D|y) \) denote the posterior belief on databases, given an output \( y \in \mathcal{Y} \). Let \( b'(D|y) \) imply the posterior belief of the adversary where we use different randomized algorithm \( \mathcal{M}'(D) = \mathcal{M}'(D_{-n}) \) that \( D_{-n} \) is a database is where we keep the \( n - 1 \) values of \( D \) while the \( n \)-th value is changed by with some inconsistent \( d_n \in D \).

We now could argue that for all \( D \in \mathcal{D}^n \) and for all \( y \in \mathcal{Y} \)

\[
SD(b(D|y), b'(D|y)) \leq e^{2\epsilon} - 1
\]

**Theorem 3.2.5.** \( \epsilon \)-differential privacy implies semantic security.
Let \( \mathcal{M} \) be an \( \epsilon \)-differentially private algorithm. For all \( D \in \mathcal{D}^n \) and \( y \in \mathcal{Y} \), we have

\[
SD(b(D|y), b'(D|y)) \leq e^{2\epsilon} - 1
\]

**Proof.** By Bayes rule [33], we know that

\[
b(D|y) = \frac{\mu(y|D)b(D)}{\sum_{E \in \mathcal{D}^n} \mu(y|E)b(E)}
\]

This yields

\[
b(D|y) - b'(D|y) = \frac{\mu(y|D)b(D)}{\sum_{E \in \mathcal{D}^n} \mu(y|E)b(E)} - \frac{\mu'(y|D)b(D)}{\sum_{E \in \mathcal{D}^n} \mu'(y|E)b(E)}
\]

Then form the definition of differential privacy inequalities, we get

\[
|b(D|y) - b'(D|y)| \leq e^{2\epsilon} - 1
\]

\( \square \)
3.3 Properties of Differential Privacy

3.3.1 The Sensitivity

Sensitivity parametrizes the amount how much noise perturbation required in the differential privacy mechanism. Currently, the global and local sensitivity are being mainly used in differential privacy.

3.3.1.1 The Global Sensitivity

It shows how much the maximal differences is between the query results of the neighboring databases that are going to be used in one of the differentially private mechanisms. The formal definition:

**Definition 3.3.1. (Global Sensitivity) [12]**

For \( f : D^n \rightarrow \mathbb{R}^k \), and use the \( l_1 \) norm on \( \mathbb{R}^k \) (denoted \( ||.||_1 \), or simply \( ||.|| \)) as a distance metric on outcomes of \( f \). Then, the global sensitivity of \( f \) is

\[
GS(f) = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1
\]

In the process of releasing data while using queries such as count or sum that has low sensitivity work well with global sensitivity. We could take the case of count query as an example which has \( GS(f) = 1 \) that is smaller than the true answer. However, when it comes to queries like median, average the global sensitivity is much higher.

3.3.1.2 The Local Sensitivity

The extent of noise included by the Laplace mechanism rely upon \( GS(F) \) and the privacy parameter \( \epsilon \), but not on the database \( D \). The process of adding noise to the most of the functions applied to yield a much higher noise not resonating the function’s general insensitivity to individual’s input. Thus, Nissim [33] proposed a local sensitivity that satisfies differential privacy by adjusting the difference between query results on the neighboring databases [12]. The formal definition:

**Definition 3.3.2. (Local Sensitivity) [31]**

For \( f : D^n \rightarrow \mathbb{R}^k \) and \( x \in D^n \), the local sensitivity of \( f \)

\[
LS(f) = \max_{D_2} ||f(D_1) - f(D_2)||_1
\]
In here, we have to observe that the global sensitivity from the definition 3.3.1 is

\[ GS(f) = \max_{D_1} LS(f)(D_1) \]

which creates less noise for queries with higher sensitivity. For queries such as count or range, the local sensitivity is identical to the global sensitivity.

From the above definition, one could observe that on many inputs, every differentially private algorithm must add a noise at least as large as the local sensitivity. However, finding algorithms whose error matches the local sensitivity is not straightforward: an algorithm that releases \( f \) with noise magnitude proportional to \( LS(f) \) on input \( D_1 \) is not, in general, differentially private [31], since the noise magnitude itself can leak information.

### 3.3.2 The Privacy Budget

Absolute data privacy guarantee that holds despite the number of computations carried out on it is the best scenario that a data creator may hope to achieve for. However, accomplishing a significant meaning of privacy cannot be done with absolute privacy guarantee. The concept of privacy budget comes into action due to the need for restricting the number of queries and composability differential privacy [28].

In definition 3.2.3 the privacy level of the mechanism \( A \) is controlled by \( \epsilon \) which is defined as privacy budget [34]. Currently, the sequential composition and the parallel composition are mostly used in the design of mechanisms.

#### 3.3.2.1 Sequential composability

The possibility of computing the results of an independent differentially private \( k \) algorithms in sequence on a dataset, without giving up privacy could be achieved in sequential composability while privacy budget is added up for each step.

Suppose there are \( k \) algorithms \( M_i(D; x_i) \), where the \( x_i \) represents some auxiliary input while each of the \( M_i \)'s are \( \epsilon \)-differentially private for any auxiliary input \( x_i \). Consider a sequence of computations \( \{x_1 = M_1(D), x_2 = M_2(D; x_1), x_3 = M_3(D; z_1, z_2), \ldots\} \) and \( M(D) = x_k \)

**Theorem 3.3.3.** (Sequential composability \( A \) is \( \|\epsilon \)-differentially private [27]).

Let \( k_i(D) \), for some \( i \in I \), be computations over \( D \) providing \( \epsilon_i \)-differential privacy. The sequence of computations \( (k_i(D))_{i \in I} \) provides \( (\sum_{i \in I} \epsilon_i) \)-differential privacy.
3.3. Properties of Differential Privacy

Proof. Let $D_1, D_2$ be two neighboring databases. Then

\[
Pr[\mathcal{M}(D)_1 = x_k] = Pr[\mathcal{M}_1(D_1) = x_1]Pr[\mathcal{M}_2(D_1; x_1) = x_2] \ldots Pr[\mathcal{M}_k(D_1; x_1, \ldots, x_{k-1})]\]

\[
\leq e^{k\epsilon} \prod_{i=1}^{k} Pr[\mathcal{M}_i(D_2; x_1, \ldots, x_{i-1}) = x_i]
\]

\[
= e^{k\epsilon}Pr[\mathcal{M}(D)_2 = x_k]
\]

3.3.2.2 Parallel composability

In situations where we have a sequence of queries made on non-intersecting sets, we could apply parallel composability. The largest privacy budget would guarantee the ultimate privacy in this composability. We consider a situation where we have $k$ disjoint subsets, $D_i$ from a partitioned database $D$ and suppose we have $k$ algorithms $\mathcal{M}(D_i; x_i)$ which each are differentially private.

**Theorem 3.3.4.** (Parallel composibility. $\mathcal{A}$ is $\|\epsilon$-differentially private [27]).

Let $k_i(D)$, for some $i \in I$, be computations over $D_i$ providing $\epsilon$-differential privacy. If each $D_i$ contains data on a set of subjects disjoint from the sets of subjects of $D_j$ for all $j \neq i$, then $(k_i(D))_{i \in I}$ provides $\epsilon_i$-differential privacy.

Proof. Let $D_1, D_2$ be two neighboring databases. Assume that the $j$-th partition contains the differing element. Then

\[
Pr[\mathcal{M}(D)_1 = x_k] = \prod_{i=1}^{k} Pr[\mathcal{M}_i(D_1; x_1, \ldots, x_{i-1}) = x_i]
\]

\[
\leq e^{\epsilon}Pr[\mathcal{M}_j(D_2; x_1, \ldots, x_j) = x_j] \prod_{i \neq j}^{k} Pr[\mathcal{M}_j(D_1; x_1, \ldots, x_{i-1})]
\]

\[
= e^{\epsilon}Pr[\mathcal{M}(D)_2 = x_k]
\]
3.4 Mechanisms of Differential Privacy

3.4.1 Laplace Mechanism

In this section we introduce one of the most basic mechanism in differential privacy. The Laplace mechanism involves adding random noise that adjusts to the Laplace distribution with mean 0 and scales $\frac{GS(f)}{\epsilon}$ and add independently to each query response, thus making sure that every query is perturbed appropriately.

To analyze Laplace Mechanism we first need to define Laplace distribution.

**Definition 3.4.1.** Laplace distribution [11]

Laplace distribution is characterized by location $\theta$ (any real number) and scale $\lambda$ (has to be greater than 0) parameters with the following probability density function:

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

![Figure 3.1: Laplace Distribution with various distribution](https://en.wikipedia.org/wiki/Laplace_distribution)

**Definition 3.4.2.** The Laplace Mechanism [12]

Given any function $f : \mathcal{D}^n \rightarrow \mathbb{R}^k$, the Laplace mechanism is defined as:

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + (Y_1, \ldots, Y_k)$$

where $Y_i$ are i.i.d. are random variables from the definition 3.4.1

---

3.4. Mechanisms of Differential Privacy

**Theorem 3.4.3.** Let \( f : \mathcal{D}^n \rightarrow \mathbb{R}^k \) be a real valued query of sensitivity 1. Then the mechanism \( M = f(D) + \frac{GS(f)}{\epsilon} \) satisfies \((\epsilon, 0)\)-differential privacy where \( GS(f) \) was defined in Definition 3.3.1.

**Proof.** Let \( D_1, D_2 \) be neighboring databases at any point \( x \in \mathbb{R} \). It is enough to compare the probability density \( f \) of \( M(D_1) \) with the \( g \) of \( M(D_2) \). Hence, without loss of generality we could assume that \( GS(D_1) = 0 \). Put \( c = f(D_2) \) and note that \( |c| \leq \Delta = \Delta(q) \). Hence, may assume that

\[
\frac{f(x)}{g(x)} = \frac{\exp(-\epsilon|x|/GS(f))}{\exp(-\epsilon|x-c|/GS(f))} \leq \frac{\exp(\epsilon|x|/GS(f))}{\exp(-\epsilon|c|/GS(f))} \leq \exp(\epsilon)
\]

where the first inequality comes from the triangle inequality and the second inequality comes from the definition of global sensitivity 3.3.1.

3.4.2 The Exponential Mechanism

A real-valued noise is being added to the actual answer when it comes to Laplace Mechanism. Nonetheless, it is known that all queries functions cannot return numerical values to their output all the time. Hence, McSherry and Talwar [28] proposed a more general method that can be applied to answer non-numeric queries in a differential manner.

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

**Definition 3.4.4.** (The Exponential Mechanism) [28]

Given a database \( D \in \mathcal{X}^n \) and a quality function \( q \) with respect to \( D \), the global sensitivity of \( s \) be \( GS_q = \max_{D_1, D_2, r \in \mathcal{R}} |q(D_1, r) - q(D_2, r)| \) and query range \( \mathcal{R} \), the exponential mechanism \( M_E(D, q, \mathcal{R}) \) gives the output \( r \in \mathcal{R} \) based on the probability:

\[
Pr[M_E(D, q, \mathcal{R}) = r] \propto \exp(\frac{\epsilon q(D, r)}{2GS_q}).
\]
When the range of exponential mechanism is super-polynomially large in the natural parameters of the problem, it can define a complex distribution over a large arbitrary domain and so it may not be possible to implement it efficiently.

**Theorem 3.4.5.** The exponential mechanism maintains $\epsilon$-differential privacy. [28]

**Proof.** We compare the ratio of the probability that the mechanism outputs for some elements $r \in \mathcal{R}$ on two neighboring databases $D_1$ and $D_2$ (i.e., $||D_1 - D_2|| \leq 1$). Let

$$N_{D_1} = \sum_{r \in \mathcal{R}} \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right)$$

be the normalizing constant for the database $D_1$. Then we have

$$\frac{Pr[\mathcal{M}_E(D_1, q, \mathcal{R}) = r]}{Pr[\mathcal{M}_E(D_2, q, \mathcal{R}) = r]} = \frac{N_{D_2} \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right) \exp \left( -\frac{\epsilon q(D_2, r)}{2GS_q} \right)}{N_{D_1} \exp \left( \frac{\epsilon q(D_1, r) - q(D_2, r)}{2GS_q} \right)} \leq \frac{N_{D_2}}{N_{D_1}} \exp \left( \frac{\epsilon}{2} \right)$$

where the last inequality follows from the definition of global sensitivity. Now we consider the ratio of the normalizing constants.

$$\frac{N_{D_2}}{N_{D_1}} = \frac{\sum_{r \in \mathcal{R}} \exp \left( \frac{\epsilon q(D_2, r)}{2GS_q} \right)}{\sum_{r \in \mathcal{R}} \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right)} \leq \frac{\sum_{r \in \mathcal{R}} \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right) \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right)}{\sum_{r \in \mathcal{R}} \exp \left( \frac{\epsilon q(D_1, r)}{2GS_q} \right)} = \exp \left( \frac{\epsilon}{2} \right)$$

So we conclude that

$$\frac{Pr[\mathcal{M}_E(D_1, q) = r]}{Pr[\mathcal{M}_E(D_2, q) = r]} \leq \exp(\epsilon)$$

where the inequality follows from the definition of global sensitivity. A similar argument would show that this ratio is also at least $\exp(\epsilon)$. So we conclude that the exponential mechanism is $\epsilon$-differentially private \(\square\)

### 3.4.3 The Median Mechanism

The median mechanism is an interactive differentially private mechanism that answers arbitrary predicate queries $f_1, \ldots, f_k$ that arrive on the fly without the future knowledge queries, where $k$ could be large or even super-polynomial. It performs...
much better than the other mechanisms (e.g. Laplace Mechanism) when it comes to answering more queries exponentially and gives fixed constraints. Theoretically, the mechanism is suitable for defining and identifying the equivalence of queries in the interactive setting. [34]

Catagorizing queries as "hard" and "easy" with low privacy cost is the core concept of the Median mechanism. The number of "hard" queries is bounded to $O(\log k \cdot \log |X|)$ due to a Vapnik-Chervonekis(VC) dimension argument [42] and the constant factored reduction of databases every time we answer a "hard" query. A query is considered "easy" from consonant answers maintained from the set of databases. The median mechanism can be explained in the following steps [26]:

1. Initialize $C_0 = \{\text{database of size } m \text{ over } X\}$

2. For each query $f_1, f_2, \ldots, f_k$ in turn:
   
   (a) Define $r_i$ and let $\hat{r}_i = r_i + Lap\left(\frac{2}{e\alpha}\right)$.
   
   (b) Let $t_i = \frac{3}{4} + j\gamma$, where $j \in \{0, 1, \ldots, \frac{1}{\gamma} \frac{3}{20}\}$ is chosen with probability proportional to $2^{-j}$
   
   (c) If $\hat{r}_i < t_i$, set $a_i$ to be the median value of $f_i$ on $C_{i-1}$.
   
   (d) If $\hat{r}_i < t_i$, set $a_i$ to be $f_i(D) + Lap\left(\frac{1}{e\alpha}\right)$.
   
   (e) If $\hat{r}_i < t_i$, set $C_i$ to the databases $S$ of $C_{i-1}$ with $|f_i(S) - a_i| \leq \frac{\epsilon}{50}$; otherwise $C_i = C_{i-1}$.
   
   (f) If $\hat{r}_j < t_j$ for more than $20m \log |X|$ values of $j \leq i$, then halt and report failure.

The mechanism makes use of several additional parameters. Roth [42] sets them to

$$m = \frac{160000 \ln k \ln \frac{1}{\epsilon}}{\epsilon^2}$$

$$\alpha' = \frac{\alpha}{720m \ln |X|} = \Theta\left(\frac{\alpha \epsilon^2}{\log |X| \log k \log \frac{1}{\epsilon}}\right)$$

$$\gamma = \frac{4}{\alpha' \epsilon n} \frac{2k}{\alpha} = \Theta\left(\frac{\log |X| \log^2 k \log \frac{1}{\epsilon}}{\alpha \epsilon^3 n}\right)$$
The $\alpha$ can be seen as privacy cost as a function of the number of queries. The value of $a_i$ of the median mechanism is defined as

$$r_i = \sum_{S \in C_{i-1}} \exp(-\epsilon^{-1}|f_i(D) - f_i(S)|) / |C_{i-1}|$$

### 3.5 Data Release on Differential Privacy

Delivering aggregate statistics on a dataset without exposing any of the individual’s record is the primary goal of data release. In the process of this private data release, there exists the data curator which is the trusted party that has a database $D$ with sensitive information and privacy parameters $\epsilon > 0$. The other party included in this process is the user/data analyst which is the seen as the entrusted one that lay-out a sequence of queries $q_1, \ldots, q_k$. The interactive and non-interactive as seen in the figure 3.2 are the two settings that are being applied in the mechanisms depending upon the arrangement of answers for the query sets.

#### 3.5.1 Interactive Data Release

In this setting, to answer the queries from the users, an interactive differential privacy interface is being inserted between the users and the dataset for the sake of privacy. Moreover, each of the queries that are being answered in this setting depends upon the privacy budget. Thus, the curator and the user interact in $x$ rounds. In round $z \in \{1, \ldots, x\}$, the curator gives an answer $a_z$ for a chosen query $q_z$ from the collection of queries $Q$ and this query may be selected on the dependency from the previous interaction $q_1, a_1, \ldots, q_{z-1}, a_{z-1}$.

For an interactive mechanism, even if we know all the queries in advance, we can still obtain a privacy preserving answers by running the interactive mechanism on each of them, which makes this setting preferable comparing it to the non-interactive one.

#### 3.5.2 Non-Interactive Data Release

In this setting, a set of queries $Q$ can be answered in a batch which contributes a higher flexibility for data analysis when it is compared to the interactive environment. A much higher noise is being added to the answers given for the queries to reassure differential privacy, despite this, there is a significant decline in the usage
of the data. Thus, the major issue in this setting is keeping the trade-off between the usage and privacy by answering more with a limited privacy budget.

![Diagram of Interactive and Non-Interactive data release]

**Figure 3.2:** Interactive and Non-Interactive data release

### 3.6 Challenges in Differential Privacy

Differential privacy is a strong concept of privacy. However, the notion still has practical challenges and limitations. From the definition of differential privacy, we have seen that an individual has a limited impact on the published dataset whenever we apply one of the mechanisms of differential privacy. The dominant theory in here is that the person’s privacy is not breached as long as the information gained from the dataset does not include that individual, although this particular information could be used to know about private information about the individual. This assumption implies that without disregard to the notion of differential privacy specific types of background information can be utilized with the differentially private outcome to learn accurate information about the person, thus implying that the assurance of differential privacy is not absolute confidentiality [14].
3.6.1 Calculating Global Sensitivity

To get a sensible constrained global sensitivity for a query, one must include all aspects of the domain with all conceivable tuples. For example, if we consider a hypothetical scenario given by Lee and Clifton [23], Purdue University has put together a “short list” of alumni as possible commencement speakers. A local newspaper is writing a story on the value Indiana taxpayers get from Purdue and would like to know if these distinguished alumni are locals or world travelers. Purdue does not want to reveal the list (to avoid embarrassing those that are not selected, for example), but is willing to show the average distance people on the list have traveled from Purdue in their lifetimes. Hence, having the assumption that by applying differential privacy mechanism to add noise to the resulting average will protect individuals, but how much noise is needed? Outliers in the data, such as Purdue’s Apollo astronaut alumni (who have been nearly 400,000km from campus) result is in need of a significant amount of noise. In this case calculating the global sensitivity will result in changing the appropriate value of $\epsilon$, which makes the result of having less utility. Simple queries with low sensitivity, for example, count queries have a weak effect on the utility of the data. However, high sensitivity queries are involved in numerous applications. In general, having privacy and applicable noise while computing a global sensitivity is a challenging task.

3.6.2 Setting of privacy parameter $\epsilon$

The question on how to set the privacy parameter $\epsilon$ has been present since the introduction of differential privacy. However, setting the right value of $\epsilon$ has not been adequately addressed. The way in which $\epsilon$ influences the ability to identify an individual is not clear, although differential privacy has apparently stated that if it is hard to determine if a person is incorporated into a database, it then is defiantly hard to know that individual’s record. In the usual sense, the parameter $\epsilon$ in $\epsilon$-differential privacy does not show what has been revealed about the person; it rather limits the outcome an individual has on the result. The influence of $\epsilon$ that has in determining an individual is less clear for queries that try to retrieve more general properties of data. However, for queries that ask specific information, e.g. “Is Mr. X in the database?” $\epsilon$ directly relates to the exposure of the information. Lee and Clifton showed [23] for a given setting of $\epsilon$, the ability that an adversary can have on a particular person in a database would
3.6. Challenges in Differential Privacy

vary depending on the query, values in the data, and even on values that are not in the data. However, while for changes of values both in the data and outside the dataset $\epsilon$-differential privacy adjusts as seen in the calculation of the query’s global sensitivity, but this is not a direct measure of what is disclosed about an individual. The improper value of $\epsilon$ causes a privacy breach, even for the same values of $\epsilon$ the level of protection that $\epsilon$-differential privacy mechanism provides is different based on the values of the domain attributes and the type of queries.

3.6.3 Uncertainty of outcome

The results obtained from a differentially private mechanism can differ enormously which makes the quality unreliable to use. Whenever applying Laplace mechanism, there might be a significant difference in the answer. An example provided in [37] a differentially private query for the mean income of a single U.S. county, with $\epsilon = 0.25$ (resp. $\epsilon = 1.0$), deviates from the true value by 10,000$ or less only 3% (resp. 12%) of the time! This can be extremely misleading, given that the true value is 16,708$. 

Chapter 4

Related Work

As discussed above, differential privacy requires that the outcome of computation remains insensitive to any given individual record in the dataset. The need to avoid unauthorized disclosures of information on records has led to the development of different analysis techniques which improve the privacy preservation of records. We will discuss, afterward, instances of previous works which aimed to provide security to databases against information leakage.

4.1 Early works on Statistical Disclosure Control Methods

A statistical database (SDB) system allows only the retrieval of aggregate statistics, such as frequency and sample mean. The methods which are developed to provide security to SDBs fall under four general techniques: conceptual, query restriction, data perturbation, and output perturbation [35].

4.1.1 Conceptual Technique

Two models which fall under this general framework had been proposed in early 1980’s. The conceptual model, which was described by Chin and Ozsoyoglu, serves as a platform for investigation of security issues at the conceptual data model level. In this model, only the population (entities with common attributes), and its statistics are accessible for the user. Merging and intersecting populations cannot be done by any data manipulation language, such as relational algebra. Within the conceptual model, the concept of atomic (A-) populations (i.e., the smallest sub-populations which cannot be decomposed any further and which contain either null or at least two entities) is used to deny information request on a singular entity. The lattice model describes SDB information represented in high-dimensional
Chapter 4. Related Work

The cell suppression technique employed by the conceptual model for single entries may lead to the disclosure of redundant information. In the lattice design, several aggregate tables are produced by aggregating the high dimensional table along each dimension, corresponding to each attribute. This process leads to a creation of coarser aggregate tables which form a lattice. Allowing access only to aggregate tables gives a more robust security to the database.

4.1.2 Query Restriction Technique

General query restriction techniques which have been developed to restrict queries include query-set-size control, query-set-overlap control, cell suppression, and partitioning. The query-set-size control method permits the release of statistics only if the size of the query set $|C|$ satisfies the condition: $K \leq |C| \leq L - K$, where $L$ is the size of the database (the number of entities represented in the database) and $K$ is a parameter set by the curator (with the condition $0 \leq K \leq \frac{L}{2}$). The drawback of this technique is that the privacy which is provided by this technique can be compromised by the usage of tracker even when $K$ is close to $\frac{L}{2}$.

The Query-Set-Overlap Control mechanism responds only to query set when the size of the intersection between this query set and preceding queries of a given user is less than some parameter. This technique has its drawbacks because it does not fend off attacks from multiple users, it does not allow statistics release for both a group and subgroups (example, all students, and students taking a particular course) - which limits the utility of the database and finally because user profile should be updated continuously.

The partitioning mechanism is based on the concept of clustering individual entities in several mutually exclusive subsets (similar to the idea of an atomic population) by partitioning the values of attributes. As in the case of atomic population, one of the problems which arise from partitioning is the emergence of subsets with the only single entity. Merging distinct attributes may solve this issue, but it may also lead to information loss. Cell suppression technique is used to remove from the released table all cells which lead to a release of confidential information (all cells with sensitive information, and all cells with non-confidential information but which can potentially result in the release of confidential information). Finally, auditing involves keeping the trail of all queries entered by unique users and always checking if privacy is compromised with the entry of the new query.
for this reason, is not economical regarding CPU time and storage as log keeps on adding several queries.

4.1.3 Data Perturbation (Input Perturbation)

It is a technique which is used to introduce noise in the primary statistical database before making it available for users is considered relatively active in privacy protecting of statistical databases when compared to the previously discussed mechanisms. There are two types of data perturbation techniques which are discussed in this thesis. The probability distribution approach considers the database as a sample from a given population with a particular probability distribution. Thus, the technique takes another sample from the same population (thus with similar probability distribution) and replaces the original statistical database with it. The probability distribution, instead of another sample, can also be used to replace the original statistical database. The value falsification approach perturbs the values of the attributes within the statistical database using multiplicative or additive noise, or other randomized processes.

4.1.4 Output Perturbation

This technique is different from data perturbation because with output perturbation noise is injected to each query result instead of the database itself.

4.2 Frameworks on Differential Privacy based Data Analysis

While early works on differential privacy focused on manually proving whether a particular algorithm is differentially private, i.e., the responses to specific queries do not lead to information leakage. Lately, several systems, which can mechanically perform differentially private data analysis (without expert intervention), have been developed. These systems allow untrusted users, with no expertise in privacy, to write algorithms and run statistical analyses without being occupied with privacy requirement beyond the defined privacy policy. In the interactive setting of data analysis, in which the user can only access the data through the interface and obtains only aggregate information, the systems provide the functionality of differential privacy to their users. We will discuss some of these works in the following paragraphs.
4.2.1 Privacy Integrated Queries (PINQ)

The first mechanism, which was proposed by McSherry [17] with the aim of protecting $\epsilon$-differential privacy, had its basis on the concept of designing agents tracking the privacy budget consumption by the query (at run-time) and canceling the computation when the budget is exhausted. This mechanism was used by McSherry to implement PINQ as a capability-based system. The implementation is based on LINQ\(^1\) declarative query language, which is SQL-like language, embedded in C\#. Data providers can use PINQ to wrap LINQ data sources in protected objects with encoded differential privacy bound $\epsilon$. The wrapper uses Laplace noise and the exponential mechanism to enforce differential privacy. The analyst can access a database through a query interface exposed by the thin privacy preserving layer. The access layer implements differential privacy by adding carefully calibrated noise to each query. PINQ’s restrictive language which does not allow database transformation operations (example, Select, Where, Join, Group by, unless they are followed by aggregation operation) and run-time checks ensure the total amount of noise respects the encoded privacy bound $\epsilon$. Despite PINQ’s restrictive language, the analyst can use this interface to use aggregate operations such as count (NoisyCount), sum (NoisySum) and average (NoisyAvg).

4.2.2 Airavat

Airavat combines an approach similar to PINQ and uses Mandatory Access Control (MAC) in a distributed, Java-based MapReduce framework. This usage can accelerate the process of large datasets. Airavat implements a simple model consisting of a query which contains a sequence of not necessarily trusted codes, chained micro-queries, called “mappers” and a subset from among fixed list of macro-queries within the system, called “reducers” which are part of the trusted base. The mappers are responsible for updating privacy budget and determining whether to continue or abort the analysis based on the adequacy of the privacy budget. When the analyst submits a query, they must also declare the anticipated numerical range of its outputs, action similar to stating the sensitivity level. The reported sensitivity level is important for the calculation by Airavat of the amount of noise which must be added to the reducers outputs (after applying aggregation function) to achieve $\epsilon$-differential privacy.

4.2.3 Fuzz

PINQ and Airavat assume that the adversary can see only the results of his/her query, which ignores the fact that the adversary can guess with a high level of certainty some attributes by strategically observing CPU activity, execution time and global variables. Fuzz addresses this shortcoming by isolating the physical machine and allowing users to communicate with the database over the network only [7]. Fuzz uses a new type system to statically infer the privacy cost of arbitrary queries written in a special programming language, and it uses a primitive called predictable transactions to prevent information leakage through execution time side channel. Fuzz splits each query to a set of micro-queries. Each micro-query is expected to be returned within the specified time, deadline. Otherwise, it is aborted, and a default value is returned as a result of micro-query execution. If the micro-query execution takes less time, and afterward the system waits and returns the result after that particular time. Using this approach each query takes the same predictable amount of time for all databases of the same size.

4.2.4 GUPT

GUPT [34] is a platform which uses a new model of data sensitivity which decreases the privacy requirement of data over time. It uses the aging model of data sensitivity which enables the description of privacy budget regarding the accuracy of the final output. This model enables GUPT to select an optimal size that reduces the perturbation added for differential privacy. GUPT also automatically distributes privacy budget to each query according to the accuracy requirements. It uses a sample and aggregate differential privacy framework. The data re-sampling method used by GUPT minimizes the error which is caused by the data partitioning scheme. Finally, the GUPT platform is safe under side-channel attacks such as it attacks, privacy budget attacks, and state attacks.

4.3 Work on Differential Privacy based Query Analysis

4.3.1 (Online) Setting

In this setting, the data analyst submits queries to the administrator in an interactive way, based on the observed answers to previous queries, and the queries are answered immediately with no knowledge of future queries. Under interactive setting, maintaining privacy and accuracy at the same time is difficult if a large
number of queries are submitted. Since the early work of Dinur and Nissim [18], by which they applied polynomial reconstruction algorithm to SDBs to show large perturbation is necessary to maintain privacy, several types of research have been conducted to address this issue.

Roth and Roughgarden [13] introduced the median privacy mechanism which improves upon independent Laplace mechanism and answers exponentially more interactive counting queries. A basic implementation of the median mechanism meant it is inefficient and sampling from a set of super-polynomial size is needed. More efficient implementation, on the other hand, means weaker utility. Classification of queries as “hard” and “easy” (with hard queries defined as queries the answers to which completely determine the answers to all the other queries) without exhausting the privacy budget is the motivation for the development of this mechanism.

Private multiplicative weights mechanism, whose goal is using a privacy preserving multiplicative weights mechanism, was later developed by Hardt and Rothblum [16]. The main result is achieving a running time only linear in N (for each of the k queries), while the error scales roughly as $\frac{1}{\sqrt{n \log k}}$. Moreover, the proposed mechanism makes partial progress for side-stepping previous negative results in the work of Dwork et al. [4] by relaxing the utility notion. Hardt and Rothblum [16] considered accuracy guarantees for the class of pseudo-smooth databases (i.e., underlying distributions that do not put too much weight on any particular data item) with sublinear running time. In later work, Gupta et al. [45], through a simple modular analysis, had given improved accuracy bounds for linear queries in private multiplicative weights mechanism.

4.3.2 Non-Interactive (Offline) Setting

In this setting, the curator sanitizes the data before publishing “more secure” or “anonymized” version of the DB (example, histograms, summary tables) or synthetic DB (with the same distribution as the original DB) once and for all, and has no role after releasing.

In the last few years, DP has transitioned from conceptual level to application. It has been applied to several real-world data which few of them we discuss below.
4.3. Work on Differential Privacy based Query Analysis

4.3.3 Histogram and Contingency Table

Using histogram is an effective way of statistical summarization of attributes. It aggregates data points into intervals (groups) and represents each group with non-overlapping bins which correspond to the exact count of the data points. Chawla et al. [20] were first to introduce histogram sanitization after a formal privacy definition was agreed. They proposed two sanitization mechanisms: recursive histogram sanitization, in which bins are partitioned recursively into smaller cells until no region contains $2t$ or more real data points; and density-based input perturbation, noise from spherically symmetric distribution (e.g., a Gaussian distribution) is added to data points. Recently, Xu et al. [44], being motivated by their observation that the quality of a histogram structure depends on the balancing of information loss and noise scales, proposed two algorithms for differential privacy compliant histogram computation: NoiseFirst (which determines the histogram structure after injecting random noise) and StructureFirst (which injects random Laplace noise to each count after determining the optimal histogram structure on the original count sequence).

Afterward, they adapted DP-histograms to answer arbitrary range-count queries. The researchers used real-world datasets from IPUM’s census record of Brazil, search logs from Google Trends and AOL search logs, NetTrace (an IP level network trace data) and Social Network files to experiment their proposals. The experimental results show that compared to other proposed methods [12, 20, 24], NoiseFirst usually returns more accurate results for range count queries with short ranges, especially for unit-length queries providing histogram with better visualization of the data distribution. On the other hand, large-length queries are better handled by StructureFirst.

A contingency table is a table which summarizes the frequency (count) of attributes in the dataset. These counts are called marginals and can be used to compute correlations between attributes within the dataset. The calculated correlations are used to reduce the amount of noise required for privacy protection. Barak et al. [2] have proposed methods to release a set of consistent marginals of a contingency table preserving all the privacy and accuracy of the original data. Their approach can be viewed as a general approach for synthetic data production. They utilize a Fourier transformation of the data to estimate low-order marginals with counts which are non-negative integers, and their sum is consistent with a set of marginals. Xiao has devised a differentially private method, privilege (based on Haar wavelet transformation), which optimizes range count queries (count queries where the predicate
on each attribute is a range) by reducing the magnitude of noise needed to guarantee $\epsilon$-differential privacy to publish multidimensional frequency matrix. Privelet preserves privacy by modifying the frequency matrix $M$ of the input data. The wavelet transform is first applied to this model to produce another matrix $C$ to which a polylogarithmic noise is then injected. They had considered a scenario where there is overlap between count queries (i.e., where there is at least one tuple satisfying multiple queries). Under their approach, queries with smaller answers are injected with less noise while queries with bigger answers are injected with more noise. For their experiment, Xiao et al. [46] used the census data from both Brazil and the US.

Hay et al. [32] propose an approach based on hierarchical sums and least squares for achieving $\epsilon$-differential privacy while ensuring a polylogarithmic noise variance in range-count query responses. Given a uni-dimensional frequency matrix $M$, Hay et al.’s algorithm add Laplace noise directly to the replies of range count queries on $M$ which it computed beforehand. Then it produces a noisy frequency matrix $M^*$ based on these noisy responses. The researchers have tested their method on three datasets: SearchLog, NetTrace and Social Network data. Though both Private and Hay et al.’s method have comparable utility guarantee, unlike the Privelet algorithm which is applicable also to simple and multi-dimensional queries, Hay et al.’s algorithm are devised for uni-dimensional datasets. Recently, Li et al. [5] generalized both the Privelet and Hay et al.’s algorithms by introducing two stage process, the matrix mechanism, for answering a workload of linear counting queries Separate set of queries (strategy queries) are used as a query proxy to the DB. Noisy responses are produced for these strategy queries using the Laplace mechanism, and then noisy responses to the workload queries are derived from these responses. The researchers used this approach to exploit stronger noise distribution correlation which preserves differential privacy but increases accuracy.
Chapter 5

Experimental Framework Setup

In this chapter, we address a detailed description of the architecture that we have used to perform an experiment on the application of differential privacy as it will be a test bed and guidance for the next chapter. Moreover, we present the scenarios that we thought would fit best to see the actual effect of differential privacy. Later in the chapter, we present choices of the programming language that we have used.

5.1 Scenarios Description

Currently, there are several ways to implement differential privacy with various kinds of settings. Thus, some assumption has to be revealed, and we have considered in this thesis to follow a basic model or architecture where a secured server is connected to a data store that is efficient enough to present differential privacy mechanisms which are going to be shown in the next section.

The actual scenario we have assumed is when a data owner puts datasets in a secured system for the purpose of some data analyst to use the information for a particular purpose while providing data privacy by utilizing differential privacy methods.

In the case of mechanisms of differential privacy, we have applied one of the primary methods of differential privacy that are essential in protecting sensitive data. The method is Laplace mechanism where the algorithms used is described in the next section.

5.2 Experimental Framework Overview

The proposed model aims to approve the thesis’s scenario and verify that one of the existing mechanisms of differential privacy increase the privacy level of the
Chapter 5. Experimental Framework Setup

database. Thus, as we have presented above, we have used an underlying architecture in implementing differential privacy as seen in the figure 5.1 and most importantly we have followed the interactive model of differential privacy. Each of the major parts and their respective purpose in this environment is presented in the next sections.

**Figure 5.1: Architecture of the system**

5.2.1 User Interface

It is an abstract notion of the system where a data analyst or user will connect to the dataset/database through the user interface and request data using a query and receive data back with the noise added to the true data by the methods of differential privacy.

5.2.2 Privacy Preservation Mechanism

In this part of the framework, the Laplace Mechanism and the methods accompanying it to show the applicability of differential privacy is implemented. It is known that at this point the one using this framework need not know how to enforce the privacy requirement nor be an expert in privacy since one of the properties of the interactive model as described in section 3.4.1 manage this.

Hence, in this part of the framework, it receives a query requesting for data from the data analyst/user where it brings the raw data from the data store. It then enforces differential privacy by adding noise depending upon the global sensitivity
5.3. Algorithm Description

to each query. Depending upon the mechanism of differential privacy, which in this case is Laplace mechanism as described in sections 3.4.1 and 3.3.1, the magnitude of the noise added is chosen to mask the influence of any particular record on the outcome.

5.2.3 Datastore

This part of the framework is where all the sensitive datasets are stored in PostgreSQL database.

5.2.4 Programming Language

The selection of programming language is on the basis upon for the reason of processing a large dataset with the smallest amount of time and having the capability of supporting mathematical operations. Thus, for those grounds we have used Python that is implemented utilizing the interactive environment called IPython Notebook \(^1\). Furthermore, for managing a large dataset we have used Pandas \(^2\), and Numpy\(^3\) and SciPy\(^4\) libraries are used.

5.3 Algorithm Description

For the purpose of showing how to apply differential privacy, we used one of the primary method, Laplace mechanism. We will describe the algorithm that we used in implementing the mechanism in the next section.

\(^1\)IPython The Jupyter Notebook, https://ipython.org/notebook.html
\(^2\)pandas: Python Data Analysis Library, http://pandas.pydata.org/
\(^3\)Numpy, http://www.numpy.org/
Chapter 5. Experimental Framework Setup

5.3.1 Algorithm for Laplace Mechanism

Algorithm 1 Laplace Mechanism’s algorithm

1: function LAPLACE(D, Q : \mathbb{N}^{|x|} \rightarrow \mathbb{R}^k, \epsilon) \triangleright \text{the laplace based on the dataset, query and the epsilon value}
2: \triangle = GS(Q) \triangleright \text{Calculate the global sensitivity}
3: \textbf{for} i \leftarrow 1, k \textbf{do}
4: \quad y_i \sim \text{Lap}(\frac{\triangle}{\epsilon}) \triangleright \text{Get the noise based on the } \epsilon \text{ and sensitivity from Laplace distribution}
5: \textbf{end for}
6: \textbf{return} Q(D) + (y_1, \ldots, y_k) \triangleright \text{the noise added plus true value}
7: end function

A controlled noise is added to the functions with low sensitivity for many differentially private methods [12]. In the algorithm 1 above it adds the perturbation/noise depending upon the value of the sensitivity it gets form the Global sensitivity of the query and passing it to the Laplace distribution [11] where its density function \( \frac{1}{2\lambda} e^{-\frac{|X - \mu|}{\lambda}} \) while \( \mu \) is a mean and \( \lambda(>0) \) is a scale factor. It is known that for the query function \( Q \) and the dataset \( D \), a randomized mechanism that returns \( Q(D) + (y_1, \ldots, y_k) \) as a response where \( y_1, \ldots, y_k \) is drawn independent and identically distributed from \( \text{Lap}(\triangle) \), gives \( \epsilon \)-differential privacy [12].
Chapter 6

Experimental Study on Differential Privacy

In this chapter, we present the dataset that has been used and the analysis of the implementation that we have applied. We have used two types of datasets that we have implemented with differential privacy. Each of the dataset description and analysis is described in the next section.

6.1 Adult Dataset

6.1.1 Dataset Description

We have used one of the famous UCI’s adult dataset\(^1\) which was acquired from US Census data (1994) and was donated in 1996. It contains more than 30,000 instances (customer records) with the following 15 attributes (columns):

### Table 6.1: Adult dataset attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fnlgwt</td>
<td>Numerical</td>
<td>Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool</td>
</tr>
<tr>
<td>Education</td>
<td>Nominal</td>
<td>Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse</td>
</tr>
<tr>
<td>Relationship</td>
<td>Nominal</td>
<td>Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried</td>
</tr>
<tr>
<td>Race</td>
<td>Nominal</td>
<td>White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black</td>
</tr>
<tr>
<td>Sex</td>
<td>Nominal</td>
<td>Female, Male</td>
</tr>
<tr>
<td>Capital-gain</td>
<td>Numerical</td>
<td>United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&amp;Tobago, Peru, Hong, Holand-Netherlands</td>
</tr>
<tr>
<td>Capital-loss</td>
<td>Numerical</td>
<td>United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&amp;Tobago, Peru, Hong, Holand-Netherlands</td>
</tr>
<tr>
<td>Hours-per-week</td>
<td>Numerical</td>
<td>United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&amp;Tobago, Peru, Hong, Holand-Netherlands</td>
</tr>
<tr>
<td>Income</td>
<td>Nominal</td>
<td>&lt;=50K, &gt;50K</td>
</tr>
</tbody>
</table>

### 6.1.2 Experimental Design

The whole experiment was conducted on a Linux OS based system that has an Intel i5-4200 CPU processing capacity with a 4GB RAM while the other system was a Windows based system with an AMD CPU processing capacity with 4GB RAM. To experiment with differential privacy, we have used two different datasets with a different aspect of scenarios. The scenarios are assumed to be dependent upon the
utilization of the dataset in a way it makes sense when applying the experiment. Thus, we have given into consideration to implement different kinds of queries and see the result and behavior of them when using the methods in $\epsilon$-differential privacy.

The first scenario we considered is being a data analyst and want to find out the number of years a person have spent in an education by income and race using the adult dataset. The first step in our experimental design, in this case, would be applying the required counting query to get the data:

```sql
SELECT
    education_num,
    COUNT(CASE WHEN race = 'White' THEN 'White' END) as "NoWhite",
    COUNT(CASE WHEN race = 'Black' THEN 'Black' END) as "NoBlack",
    COUNT(CASE WHEN race = 'Asian-Pac-Islander' THEN 'Asian-Pac-Islander' END) as "NoAsianPacIslander",
    COUNT(CASE WHEN race = 'Amer-Indian-Eskimo' THEN 'Amer-Indian-Eskimo' END) as "NoAmerIndianEskimo",
    COUNT(CASE WHEN race = 'Other' THEN 'Other' END) as "NoOther"
FROM
    adulti
WHERE
    salary_group = '<=50K'
GROUP BY
    education_num, salary_group
ORDER BY
    education_num;
```

Listing 6.1: Counting query for adult dataset

which would have give the data analyst the following data as shown in the graph 6.1 but before that happens one of the methods differential privacy is applied to it in order to preserve the privacy. The results and the behaviour of the differential privacy is discussed in the next part.
The second scenario we considered under this dataset is understanding the characteristics of the mean query under aggregate function when using it differential privacy. Thus, in here also as being a data analyst, we want to find out the average working hour per work for each job class that existed in the dataset. To get the data, we apply the appropriate mean query:

```sql
SELECT workclass, AVG(hours_per_week) as AVGHoursPerWeek
FROM adulti
GROUP BY workclass
```

which would deliver the data as shown below before applying it to one of the methods in differential privacy.
6.1. Adult Dataset

<table>
<thead>
<tr>
<th>work_class</th>
<th>hours/week</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 State-gov</td>
<td>39.031587</td>
</tr>
<tr>
<td>1 Federal-gov</td>
<td>41.379167</td>
</tr>
<tr>
<td>2 Private</td>
<td>40.267096</td>
</tr>
<tr>
<td>3 Local-gov</td>
<td>40.982800</td>
</tr>
<tr>
<td>4 Self-emp-inc</td>
<td>48.818100</td>
</tr>
<tr>
<td>5 Self-emp-not-inc</td>
<td>44.421881</td>
</tr>
<tr>
<td>6 NA</td>
<td>31.919390</td>
</tr>
<tr>
<td>7 Never-worked</td>
<td>28.428571</td>
</tr>
<tr>
<td>8 Without-pay</td>
<td>32.714286</td>
</tr>
</tbody>
</table>

Table 6.2: Mean query result for adult dataset

6.1.3 Results and Discussion

Since the individuals’ privacy is the primary concern of differential privacy, the count and mean types of queries are considered to examine the level of privacy and as well as the utility of the data provided by the proposed model. Thus, mainly as shown in sql 6.1 is a counting query where it manages to ask about the behavior of races concerning education and income which have a very sensitive information. In the other type of query, as seen in sql 6.2 is a mean query that asks information about the average working time that also needs to be protected using differential privacy method. The main reason why we have considered to include the mean query is that the way of calculating the sensitivity of the query affects the characteristics of differential privacy that we are going to see in the next section.

In the Laplace mechanism, the major step in achieving differential privacy is generating a noise to be added from the Laplace distribution. Therefore, since the value of \( \epsilon \) govern the amount of the noise created in the experiment, for the sake of accuracy it was repeated 50, 100, 150 and 250 times over each dataset for the different values of \( \epsilon \). The values taken into consideration for the parameter \( \epsilon \) is from the understanding of the definition of differential privacy [10, 23, 11, 21] where the level of privacy is higher when the value of \( \epsilon \) is close to the value of zero and as stated in [10] the value of \( \epsilon \) is assumed to be small as 0.01, 0.1 or even in some cases, \( \ln 2 \) or \( \ln 3 \). Although, in our case the random choice for the values \( \epsilon \) will give the needed privacy but calculating in order to choose the right value of \( \epsilon \) is considered under the future work. Thus, the values taken into consideration for the parameter \( \epsilon \) are
1, 0.5, 0.01, and 0.001. The choices of $\epsilon$ contribute to the test’s performance and the characteristics it has the utility of the data.

For the case of the counting query, the first step in getting the noise from the Laplace distribution is calculating the amount of sensitivity. From the definition of global sensitivity in section 3.3.1 we know that it is the maximum difference between the two neighboring databases and in this case, it is the effect a record have on being included in a dataset or not which is at most one.

The result of the first query 6.1 is seen in the table for all the $\epsilon$ parameter values

<table>
<thead>
<tr>
<th>$\epsilon=0.001$</th>
<th>$\epsilon=0.01$</th>
<th>$\epsilon=0.1$</th>
<th>$\epsilon=0.5$</th>
<th>$\epsilon=1$</th>
<th>$\epsilon=2$</th>
<th>True Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.812493</td>
<td>35.390516</td>
<td>37.326686</td>
<td>38.004313</td>
<td>37.949046</td>
<td>38.142612</td>
<td>38</td>
</tr>
<tr>
<td>114.25916</td>
<td>129.886678</td>
<td>128.211722</td>
<td>128.904062</td>
<td>128.987817</td>
<td>129.9923</td>
<td>129</td>
</tr>
<tr>
<td>376.422304</td>
<td>264.810698</td>
<td>265.716605</td>
<td>266.990090</td>
<td>267.004823</td>
<td>266.913432</td>
<td>267</td>
</tr>
<tr>
<td>508.410383</td>
<td>521.427331</td>
<td>514.27641</td>
<td>515.432698</td>
<td>515.11026</td>
<td>515.017971</td>
<td>515</td>
</tr>
<tr>
<td>318.298624</td>
<td>367.727981</td>
<td>380.454155</td>
<td>381.233755</td>
<td>381.025898</td>
<td>380.948868</td>
<td>381</td>
</tr>
<tr>
<td>633.457919</td>
<td>710.750242</td>
<td>708.279636</td>
<td>707.957029</td>
<td>708.001879</td>
<td>707.936147</td>
<td>708</td>
</tr>
<tr>
<td>905.903116</td>
<td>912.219137</td>
<td>927.73722</td>
<td>926.672365</td>
<td>927.043154</td>
<td>927.001282</td>
<td>927</td>
</tr>
<tr>
<td>229.255766</td>
<td>310.303729</td>
<td>307.250924</td>
<td>308.205609</td>
<td>308.102301</td>
<td>308.059411</td>
<td>308</td>
</tr>
<tr>
<td>7361.250424</td>
<td>7363.948766</td>
<td>7365.089796</td>
<td>7361.968016</td>
<td>7362.020765</td>
<td>7361.984653</td>
<td>7362</td>
</tr>
<tr>
<td>4766.829785</td>
<td>4986.021341</td>
<td>4951.782225</td>
<td>4951.46874</td>
<td>4951.640962</td>
<td>4951.942688</td>
<td>4952</td>
</tr>
<tr>
<td>999.507002</td>
<td>889.21688</td>
<td>874.203274</td>
<td>873.741552</td>
<td>874.00838</td>
<td>874.015033</td>
<td>874</td>
</tr>
<tr>
<td>668.459507</td>
<td>680.438207</td>
<td>678.677963</td>
<td>679.927983</td>
<td>679.971105</td>
<td>680.02492</td>
<td>680</td>
</tr>
<tr>
<td>2745.336556</td>
<td>2675.991294</td>
<td>2667.150378</td>
<td>2666.96552</td>
<td>2666.995558</td>
<td>2666.956198</td>
<td>2667</td>
</tr>
<tr>
<td>643.886931</td>
<td>672.841862</td>
<td>665.517948</td>
<td>666.025606</td>
<td>666.140248</td>
<td>666.002405</td>
<td>666</td>
</tr>
<tr>
<td>93.347772</td>
<td>141.364006</td>
<td>131.312386</td>
<td>131.979095</td>
<td>132.031384</td>
<td>132.039191</td>
<td>132</td>
</tr>
<tr>
<td>114.494777</td>
<td>89.435127</td>
<td>92.744101</td>
<td>93.117938</td>
<td>93.166598</td>
<td>92.96753</td>
<td>93</td>
</tr>
</tbody>
</table>

**Table 6.3: Result for Laplace mechanism for counting query**

From the result above, we see the effect $\epsilon$ has on the quality of the result from this mechanism. This is due to differential privacy is achieved since Laplace mechanism requires generating amounts of noise from Laplace distributions. Therefore, as the values of $\epsilon$ becomes smaller the amounts of generated noise become larger, and smaller while the values of it become larger.

When it comes to the scenario of a mean query, the process of achieving differential privacy is the same as of the counting query, but the major difference is calculating sensitivity. Unlike the counting query, the sensitivity of the mean query is the maximum average difference of the two neighboring databases which have a variable amount other than one. This value will lead to not having a complete sensible
noise addition that could be used in the dataset. We could clearly see the result in the table 6.4

<table>
<thead>
<tr>
<th>True Value</th>
<th>$\epsilon=2$</th>
<th>$\epsilon=1$</th>
<th>$\epsilon=0.5$</th>
<th>$\epsilon=0.1$</th>
<th>$\epsilon=0.01$</th>
<th>$\epsilon=0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.379167</td>
<td>41.290438</td>
<td>41.125853</td>
<td>41.26991</td>
<td>42.599621</td>
<td>25.580234</td>
<td>293.379021</td>
</tr>
<tr>
<td>40.267096</td>
<td>40.324735</td>
<td>40.353773</td>
<td>40.33514</td>
<td>41.172319</td>
<td>30.570535</td>
<td>60.370953</td>
</tr>
<tr>
<td>40.9828</td>
<td>41.125241</td>
<td>40.803465</td>
<td>41.515622</td>
<td>37.188855</td>
<td>55.853787</td>
<td>-121.15732</td>
</tr>
<tr>
<td>48.8181</td>
<td>48.783353</td>
<td>48.97514</td>
<td>48.997429</td>
<td>45.874634</td>
<td>51.523733</td>
<td>-124.031842</td>
</tr>
<tr>
<td>44.421881</td>
<td>44.452711</td>
<td>44.177682</td>
<td>43.37441</td>
<td>43.694063</td>
<td>63.597372</td>
<td>-41.977862</td>
</tr>
<tr>
<td>31.91939</td>
<td>32.051182</td>
<td>31.890406</td>
<td>31.856859</td>
<td>32.275897</td>
<td>85.031996</td>
<td>362.218115</td>
</tr>
<tr>
<td>32.714286</td>
<td>32.846824</td>
<td>32.713454</td>
<td>32.852096</td>
<td>29.470895</td>
<td>47.369689</td>
<td>119.513105</td>
</tr>
</tbody>
</table>

TABLE 6.4: Result for Laplace Mechanism for Mean query

From the above results, we could observe that the Laplace mechanism has different characteristics towards the different types queries that run against the dataset. From this, the first thing to understand is that the noise that is going to be added depends upon the kind of the dataset of which $\epsilon$ plays a major part in. The other thing that we could see from this mechanism is that the trade offs between utility and privacy are also influenced by the value of $\epsilon$

### 6.2 Student Alcohol Consumption Dataset

#### 6.2.1 Dataset Description

We use student’s alcohol consumption dataset from UCI Machine Learning Repository: datasets for our study. The data contains Alcohol consumption of secondary level students including student’s grades, demographic, social and school-related features. The data was first gathered and analyzed by Paulo Cortez and Alice Silva, University of Minho, Portugal (Portuguese student on two courses Mathematics and Portuguese). It was collected by using school reports and questionnaires from two public schools, from the Alentejo region of Portugal amid the 2005 - 2006 school year. Students are evaluated three times, G1 - first-period grade, G2 - second-period grade and G3 - final grade. The grading scale is 20 points grading scale from 0 to the lowest to 20 the highest and these grades are related to the course subject, Math or Portuguese [32, 5].
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description(Domain)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>student sex (binary: female or male)</td>
</tr>
<tr>
<td>Age</td>
<td>student age (numeric: from 15 to 22)</td>
</tr>
<tr>
<td>School</td>
<td>student school (binary: Gabriel Pereira or Mousinho da Silveira)</td>
</tr>
<tr>
<td>Address</td>
<td>student home address type (binary: urban or rural)</td>
</tr>
<tr>
<td>Pstatus</td>
<td>parent cohabitation status (binary: living together or apart)</td>
</tr>
<tr>
<td>Medu</td>
<td>mother education (numeric: from 0 to 4a)</td>
</tr>
<tr>
<td>Mjob</td>
<td>mother job (nominalb)</td>
</tr>
<tr>
<td>Fedu</td>
<td>father education (numeric: from 0 to 4a)</td>
</tr>
<tr>
<td>Fjob</td>
<td>father job (nominalb)</td>
</tr>
<tr>
<td>Guardian</td>
<td>student guardian (nominal: mother, father or other)</td>
</tr>
<tr>
<td>Famsize</td>
<td>family size (binary: ≤ 3 or &gt;3)</td>
</tr>
<tr>
<td>Famrel</td>
<td>quality of family relationships (numeric: from 1 – very bad to 5 – excellent)</td>
</tr>
<tr>
<td>Reason</td>
<td>reason to choose this school (nominal: close to home, school reputation, course)</td>
</tr>
<tr>
<td>Traveltime</td>
<td>home to school travel time (numeric: 1 – &lt;15 min., 2 – 15 to 30 min., 3 – 30)</td>
</tr>
<tr>
<td>Studytime</td>
<td>weekly study time (numeric: 1 – &lt;2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4)</td>
</tr>
<tr>
<td>Failures</td>
<td>number of past class failures (numeric: n if 1 ≤ n &lt;3, else 4)</td>
</tr>
<tr>
<td>Schoolsup</td>
<td>extra educational school support (binary: yes or no)</td>
</tr>
<tr>
<td>Famsup</td>
<td>family educational support (binary: yes or no)</td>
</tr>
<tr>
<td>Activities</td>
<td>extra-curricular activities (binary: yes or no)</td>
</tr>
<tr>
<td>Paidclass</td>
<td>extra paid classes (binary: yes or no)</td>
</tr>
<tr>
<td>Internet</td>
<td>Internet access at home (binary: yes or no)</td>
</tr>
<tr>
<td>Nursery</td>
<td>attended nursery school (binary: yes or no)</td>
</tr>
<tr>
<td>Higher</td>
<td>wants to take higher education (binary: yes or no)</td>
</tr>
<tr>
<td>Romantic</td>
<td>with a romantic relationship (binary: yes or no)</td>
</tr>
<tr>
<td>Freetime</td>
<td>free time after school (numeric: from 1 – very low to 5 – very high)</td>
</tr>
<tr>
<td>Goout</td>
<td>going out with friends (numeric: from 1 – very low to 5 – very high)</td>
</tr>
<tr>
<td>Walc</td>
<td>weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)</td>
</tr>
<tr>
<td>Dalc</td>
<td>workday alcohol consumption (numeric: from 1 – very low to 5 – very high)</td>
</tr>
<tr>
<td>Absences</td>
<td>number of school absences (numeric: from 0 to 93)</td>
</tr>
<tr>
<td>G1</td>
<td>first period grade (numeric: from 0 to 20)</td>
</tr>
<tr>
<td>G2</td>
<td>second period grade (numeric: from 0 to 20)</td>
</tr>
<tr>
<td>G3</td>
<td>final grade (numeric: from 0 to 20)</td>
</tr>
</tbody>
</table>

**Table 6.5: The pre-processed student related variables**

1. 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.

2. teacher, health care related, civil services (e.g. administrative or police), at home or other.
6.2. Student Alcohol Consumption Dataset

6.2.2 Experimental Design

Our analysis intends to investigate whether student’s alcohol consumption has effect on their health status, education performance and school attendance. The student alcohol consumption dataset contains two parts of alcohol consumption of students, Dalc (alcohol consumption) and Walc (weekend alcohol consumption). We combine the two parts to calculate the total amount of alcohol taking by a student in an entire week using the following formula.

\[ A_{lc} = \frac{(W_{alc} \times 2) + (D_{alc} \times 5)}{7} \]

We imported the CSV file of student’s alcohol consumption dataset to PostgreSQL database and connected the database to IPython using the following code:

```python
try:
    conn = pg.connect("dbname='Student' user='postgres' host='localhost' password='samri'")
except:
    print("I am unable to connect to the database")
```

Listing 6.3: Connection string

After connecting IPython (Jupyter Notebook) to the database (PostgreSQL), we used the SQL counting queries on IPython to find the total alcohol consumption taken by individual students in a week and their health status, school attendance and their education performance (Final grade score). After getting the results from the SQL query, to protect the privacy of the data of the actual value, we added noise to the data using the Laplace mechanism.

We tried to do the analysis based on both limited observations and all 649 observations within the dataset. However, the regression analysis on limited observations resulted in outputs which are entirely different from the regression result from the entire dataset. For example, the regression analysis of school attendance on alcohol consumption showed alcohol consumption was not a predictor for school attendance when the observations used limited while it showed otherwise when the complete data was used for the regression analysis. So, we opted to use the entire dataset to avoid bias which led to our qualification of observations for the regression analysis.
Chapter 6. Experimental Study on Differential Privacy

sql3 = '''SELECT (((walc * 2) + (dalc * 5))/7) AS AlcoholCons, health AS HealthStatus FROM adulti ORDER BY AlcoholCons;'''
dataframe3 = psql.read_sql_query(sql3, conn)
dataframe3

### Listing 6.4: Getting the data from the database

The results for this query is shown on the figure 6.2. After getting the true value of this dataset, we used Laplace mechanism to add noise to the true value to make $\epsilon$ differentially private data. The code used to manipulate the data is shown in the listing 6.5, and the results are shown on the figure 6.3.

```
dataframe3Copy = dataframe3
data03=add_laplace_noise(dataframe3.healthstatus, len(dataframe3.healthstatus), 0.001, 1)
dataframe3Copy.healthstatus = data03
dataframe3Copy
```

### Listing 6.5: Apply Laplace Mechanism to the data

To investigate the effect of Alcohol consumption on school attendance we use the query as shown in the listing 6.6. The result for this query is shown on the figure 6.4.

Number of students absence days from school and their Alcohol consumption from 1 - very low to 5 - very high

sql2 = '''SELECT (((walc * 2) + (dalc * 5))/7) AS AlcoholConsumption, absences AS AbsenceDays FROM adulti ORDER BY (((walc * 2) + (dalc * 5))/7), absences;'''
dataframe2 = psql.read_sql(sql2, conn)
dataframe2

```
dataframe2Copy = dataframe2
data01=add_laplace_noise(dataframe2.absencedays, len(dataframe2.absencedays), 0.001, 1)
dataframe2Copy.absencedays = data01
dataframe2Copy
```

### Listing 6.6: Query for number of students absence days

The code used to manipulate the data is given below in the listing 6.7 and the result is shown on the figure 6.5.
To investigate the effect of Alcohol consumption on education performance we use the following query as shown in the listing 6.8, the result is shown on the figure 6.6.

```sql
sql4 = '''SELECT (((walc * 2) + (dalc * 5))/7) AS AlcoholConsumption, G3 AS FinalGrade
FROM adultos
ORDER BY AlcoholConsumption, FinalGrade ;'''
dataframe4 = psql.read_sql_query(sql4, conn)
dataframe4
```

**Listing 6.8: Query for the effect of Alcohol consumption**

We manipulate the true value we get from this query using the following code as shown in the listing 6.9 and the result is shown on the figure 6.7.

```python
dataframe4Copy = dataframe4
data04 = add_laplace_noise(dataframe4.finalgrade, len(dataframe4.finalgrade), 0.1, 1)
dataframe4Copy.finalgrade = data04
dataframe4Copy
```

**Listing 6.9: Apply Laplace Mechanism**

We have performed the experiment for different values of epsilon (\( \epsilon = 0.1 \) and \( \epsilon = 0.0001 \)) on the final grade attribute to study the level of protection \( \epsilon \)-differentially private mechanism gives for the same attribute when the value of the privacy parameter is different.

To compare the results, we find using different values of \( \epsilon \) we use \( \epsilon = 0.0001 \), for the final grade attribute the code we used is shown in the listing 6.10, and the result is shown on the figure 6.8.

```python
dataframe4Copy = dataframe4
data04 = add_laplace_noise(dataframe4.finalgrade, len(dataframe4.finalgrade), 0.0001, 1)
dataframe4Copy.finalgrade = data04
dataframe4Copy
```

**Listing 6.10: Apply Laplace Mechanism**

In our effort to draw the scatter plot of true value data and the noisy data on the same graph, we observed that the resulted graph is too dense because the y-axis scale is large the noisy values range from decimal values near 0 to large values.
This results in overlap of markers in the given range making the graph congested. Thus, we opted to represent the figures for the actual value data and the noisy data separately.

In our effort to draw the scatter plot graph for the true value data and noisy data on the same graph, we observed that the resulted graph is too dense because the y-axis scale is large since the noisy values range from decimal values near 0 to large values. This results in an overlap of markers in the given range which is making the graph congested. Thus, we opted to represent the figures for the actual value and noisy data separately.
6.2. Student Alcohol Consumption Dataset

**Figure 6.2:** Students Alcohol consumption and Health status

ε = 0.001

**Figure 6.3:** Alcohol consumption and Health status with noise
Chapter 6. Experimental Study on Differential Privacy

Figure 6.4: Alcohol consumption and absence days

$\epsilon = 0.001$

Figure 6.5: Alcohol consumption and Absence days with noise
6.2. Student Alcohol Consumption Dataset

Figure 6.6: Alcohol consumption and Final grade
\[ \epsilon = 0.1 \]

Figure 6.7: Alcohol consumption and Final grade with noise
6.2.3 Results and Discussion

In the implementation introduced above, we presented the distinctions between the actual value results and the differentially private results. In this section, we discuss the results we get from the implementation, i.e., the level of privacy protection guarantee that $\epsilon$-differential privacy mechanism gives and the utility of the data we get when the epsilon value increases and decreases. Before discussing these results, we will discuss the findings from the simple linear regression we performed using the Stata software version 13\(^2\) to find out if there is a correlation between the alcohol consumption and students performance, alcohol consumption and students health status, alcohol consumption and students absence days from school.

In the implementation introduced above, we presented the distinctions between the actual value results and the differentially private results. In this section, we discuss the results we get from the implementation, i.e., the level of privacy protection guarantee that $\epsilon$-differential privacy mechanism gives and the utility of the data we get when the epsilon value increases and decreases. Before discussing these results, we will discuss the findings from the simple linear regression we performed using the Stata software version 13\(^2\) to find out if there is a correlation between the alcohol consumption and students performance, alcohol consumption and students health status, alcohol consumption and students absence days from school.

6.2. Student Alcohol Consumption Dataset

In our school, we have performed simple linear regression using Stata software version 13. The results will be discussed below.

```
. regress Health_Status Alcohol_Consumption
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 649</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F( 1, 647) = 2.35</td>
</tr>
<tr>
<td>Model</td>
<td>4.90462072</td>
<td>1</td>
<td>4.90462072</td>
<td>Prob &gt; F = 0.1258</td>
</tr>
<tr>
<td>Residual</td>
<td>1350.49488</td>
<td>647</td>
<td>2.08732785</td>
<td>R-squared = 0.0036</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.0021</td>
</tr>
<tr>
<td>Total</td>
<td>1355.39908</td>
<td>648</td>
<td>2.09166526</td>
<td>Root MSE = 1.4448</td>
</tr>
</tbody>
</table>

```

| Health_Status | Coef.  | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---------------|--------|-----------|---|-----|---------------------|
| Alcohol_Consumption | .0950501 | .0639666 | 1.55 | 0.126 | -0.0275531  .2236534 |
| _cons         | 3.390267 | .1108184 | 30.59 | 0.000 | 3.172666  3.607875 |
```

**Figure 6.9:** Result of simple linear regression of health status on alcohol consumption

```
. regress Health_Status_With_Noise Alcohol_Consumption
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 649</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F( 1, 647) = 0.92</td>
</tr>
<tr>
<td>Model</td>
<td>7.30234533</td>
<td>1</td>
<td>7.30234533</td>
<td>Prob &gt; F = 0.3577</td>
</tr>
<tr>
<td>Residual</td>
<td>5180.8341</td>
<td>649</td>
<td>8.01993633</td>
<td>R-squared = 0.0014</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = -0.0001</td>
</tr>
<tr>
<td>Total</td>
<td>5198.12645</td>
<td>648</td>
<td>8.02585255</td>
<td>Root MSE = 2.8319</td>
</tr>
</tbody>
</table>

```

| Health_Status_Wit-e | Coef.  | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|---------------------|--------|-----------|---|-----|---------------------|
| Alcohol_Consumption | .1202956 | .12533 | 0.96 | 0.330 | -.1259072  .3664944 |
| _cons               | 3.38138 | .2272202 | 15.11 | 0.000 | 2.884808  3.707892 |
```

**Figure 6.10:** Result of simple Linear Regression of health status with added noise on alcohol consumption

As it can be seen from table 6.9, alcohol consumption by itself does not seem to predict the health status of a student (p-value=0.338) at a significant level of 0.05. Table 6.10 also shows that adding a noise with \(\epsilon=0.001\) does not change the result of the regression (whether the correlation is significant or not) albeit some shift in the value of the coefficient from 0.098 to 0.12. This model can be further investigated by adding other predictors (age, sex) but as we are performing the analysis just to evaluate the effect of adding noise, we will not perform such analysis.

```
regress Absence_days Alcohol_Consumption

Source | SS    | df | MS    | Number of obs = 649
-------|-------|----|-------|----------------------
Model  | 402.148003 | 1  | 402.148003 | F(1, 647) = 19.20
Residual | 13555.5982  | 647 | 20.9483744 | Prob > F = 0.0000
-------|-------|----|-------|----------------------
Total  | 13957.7412 | 648 | 21.5366423 | R-squared = 0.0288
       |       |    |       | Adj R-squared = 0.0273

Absence_days | Coef. | Std. Err. | t     | P>|t|     | [95% Conf. Interval]
--------------|-------|-----------|-------|--------|---------------------
Alcohol_Consumption | .8078452 | .2026378 | 4.00 | 0.000 | [.498838 1.216852]
_cons | 2.337968 | .3510692 | 6.66 | 0.000 | [1.640596 3.035341]
```

**Figure 6.11:** Result of simple Linear Regression of school attendance on alcohol consumption
6.2. Student Alcohol Consumption Dataset

We also investigated if there is a change in the result of the regression analysis when a noise is added with $\epsilon=0.001$. As we can see from the tables 6.11 and 6.12, the level of alcohol consumption significantly predicts the school attendance of a student ($p$-value <0.05) though the model itself describes only 2.9% of the variance (R-squared=0.0288) and 3.0% of the variance when noise is added to school attendance. This analysis shows that for a unit increase on a scale of alcohol consumption, we would expect a 0.89 unit increase in the absence record of a student when no noise is introduced to the outcome variable and a 0.99 unit increase in the absence record of a student when noise is added.

Finally, we run the regression analysis of school performance on alcohol consumption level by adding a varying degree of noise on the outcome variable (final grade). In the table 6.13 shows that increased alcohol consumption significantly predicts a decrease in the school performance of a student ($p$-value <0.05 and coefficient=-0.72). Adding a noise with a value of $\epsilon=0.1$ did not affect either this result or the statement that approximately 4% of the variance of school performance is accounted for by the given model in the table 6.14. However, adding a noise with $\epsilon=0.0001$ results in the different outcome as shown in the table 6.15. Here, we can see that the effect of alcohol consumption on school performance is not significant ($p=0.563$) and only 0.05% of the variance of final grade is accounted for by the given model, neither of results similar to the result obtained in the table 6.13.
Chapter 6. Experimental Study on Differential Privacy

. regress Final_grade Alcohol_Consumption

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 649</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F( 1, 647) = 26.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>267.195413</td>
<td>1</td>
<td>267.195413</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>6466.07315</td>
<td>647</td>
<td>10.0402865</td>
<td>R-squared = 0.0595</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6733.26656</td>
<td>648</td>
<td>10.4371396</td>
<td>Adj R-squared = 0.0380</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 6.13: Result of simple Linear Regression of school performance on alcohol consumption

. regress Final_grade_with_noise Alcohol_Consumption

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 649</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F( 1, 647) = 26.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>266.058839</td>
<td>1</td>
<td>266.058839</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>6467.89803</td>
<td>647</td>
<td>10.0431191</td>
<td>R-squared = 0.0393</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6733.95797</td>
<td>648</td>
<td>10.4302966</td>
<td>Adj R-squared = 0.0379</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Final_grade_with_noise</th>
<th>Coef.</th>
<th>Std. Err.</th>
<th>t</th>
<th>Prob &gt;</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol_Consumption</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>_cons</td>
<td>12.33835</td>
<td>.2430814</td>
<td>53.40</td>
<td>0.0000</td>
<td>12.50303  12.45768</td>
</tr>
</tbody>
</table>

FIGURE 6.14: Result of simple Linear Regression of school performance with added noise on alcohol consumption (\(\epsilon=0.1\))


6.2. Student Alcohol Consumption Dataset

\texttt{. regress Final\_grade\_very\_high\_noise Alcohol\_Consumption}

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 649</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>174.077539</td>
<td>1</td>
<td>174.077539</td>
<td>( F(1,647) = 0.34 )</td>
</tr>
<tr>
<td>Residual</td>
<td>335706.154</td>
<td>647</td>
<td>518.865771</td>
<td>( \text{Prob} &gt; F = 0.5626 )</td>
</tr>
<tr>
<td>Total</td>
<td>335880.231</td>
<td>648</td>
<td>518.83369</td>
<td>( \text{R-squared} = -0.0005 )</td>
</tr>
</tbody>
</table>

\underline{Final\_grade\_very\_high\_noise} | \underline{Coeff.} | \underline{Std. Err.} | \underline{t} | \underline{P>|t|} | \underline{[95\% Conf. Interval]}

| Alcohol\_Consumption | -0.5841397 | 1.000493 | -0.58 | 0.563 | -2.564454 | 1.396175 |
| const | 13.19057 | 1.74721 | 7.55 | 0.000 | 9.759686 | 16.62146 |

\textbf{FIGURE 6.15:} Result of simple Linear Regression of school performance with very high noise on alcohol consumption (\( \epsilon = 0.0001 \))

\( \epsilon \) is a privacy parameter on the level of privacy given. Choosing the proper value of \( \epsilon \) is very challenging. Even for the same value of \( \epsilon \), the privacy guarantee provided by differential privacy differs based on the type of query and the attributes on the domain.

To compare the results of the data analysis when the \( \epsilon \) value changes, we have performed an experiment on student’s final grade dataset (\( \epsilon = 0.1 \) and \( \epsilon = 0.0001 \)) to show the difference between the results we get from the differentially private mechanism.

Based on our experiment, we observed that when \( \epsilon \) decreases the utility of the data decreases and the privacy level increases, and when \( \epsilon \) increases the utility of the data increases and the privacy level decreases.

When the \( \epsilon \) value equals 0.0001, the results we get from the differentially private data release shows that the privacy level guarantee that \( \epsilon \)-differential privacy gives is very high, but the data has no utility, that is it leads to the wrong data analysis results (change in p-value). When we make a statistical data analysis as shown in the table 6.15, it shows that there is no correlation between student’s alcohol consumption and students performance (final grade results) which is the wrong conclusion.
For $\epsilon = 0.1$ the utility of the data increased, but the privacy level of the data decreased there is less difference between the outputs of the data compared to the real values of the dataset.

From our experiment using student’s alcohol consumption dataset, we observed that giving sufficiently small value for $\epsilon$ protects the privacy of the data from the adversary. As the $\epsilon$ value decreases the privacy level increases, it will be difficult for the adversary to guess the exact value of the query results, but the utility of the data decreases.
Chapter 7

Conclusion and Future Work

In this chapter we summarize the main themes in this thesis and gives a possible future directions for research. Differential privacy has become the de facto standard to guarantee privacy, and currently, it is one of the most anticipated research topics in data privacy used in a wide range of applications. The research presented in this thesis consists of four parts: the first part focuses on the study of methodologies that are used in the process of data privacy before the use of the differential privacy; the second part focuses on the study of differential privacy and its properties which continue with studies that are related to it; the third part concentrates on the framework and an algorithm used in the case study of differential privacy; the last part concentrates on the experiment and discussion of applying a mechanism of differential privacy on different types of data sets while using different kinds of queries. Theoretical and experimental results have led to the following conclusions of this thesis.

- As we have seen in chapter 3 differential privacy is equipped with a property and methodology that proves to be a strict privacy methodology with sufficient theory in support.

- We have proposed a simple framework for the assumption we came up with in order to show how data privacy is achieved using differential privacy. The framework compromises the basic building blocks that are necessary and moreover, it compromises one of the basic methods of differential privacy in order to show that data privacy is achieved as seen in chapter 6

- In chapter 6 we showed that differential privacy protects data by adding randomized noise taken from a Laplace distribution as seen in sections 6.1.2 and
6.2.2 to an actual value, and from the result we got it showed that the mechanism used under differential privacy give a data privacy for the two types of datasets based upon the different values of $\epsilon$ used.

- From the results we got, Laplace mechanism is one of a sound approach in achieving differential privacy, depending upon the type of the dataset and the privacy parameter used. However, it got limited applicability towards applying to datasets that are not typically of a specific type like requests that are not numerical.

Up to the current date there are dozens of inquiries that has been made in choosing the right value of $\epsilon$ which plays a significant role in differential privacy, however, there is still no general agreement towards selecting the right value. Therefore, future work should consider in tackling this issue towards coming up with a better general approach in choosing the right amount of $\epsilon$.

While the underlying mechanisms of differential privacy are efficient enough to give the data privacy needed they might not be flexible sufficiently to apply them in all aspect of real life scenarios which might hinder in achieving the maximum security and usability needed. Hence, it is an ideal to analyze and customize other types of mechanisms besides the basic ones as a future work.

Under data dependency differential privacy has a vulnerable assumption that can lead to a reduction in the expected privacy level when applied to real-world datasets that show natural dependence owing to various social, behavioral, and genetic relationships between users. This weak privacy assumption is observed under cases where attacks such as inference exist under differential privacy mechanism. Thus, future work should also consider in working towards a better mechanism that significantly improves the existing tools under this assumption.
Bibliography


[32] Fabio Pagnotta and HM Amran. “Using data mining to predict secondary school student alcohol consumption”. In: Department of Computer Science, University of Camerino ()


