**The Pennsylvania State University**
**The Graduate School**

# ATTACKING ANONYMIZATION AND CONSTRUCTING IMPROVED DIFFERENTIALLY PRIVATE CLASSIFIERS

A Thesis in
Electrical Engineering
by
Chandrasekhar Mothali



Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2011

The thesis of Chandrasekhar Mothali was reviewed and approved* by the following:

Daniel Kifer
Assistant Professor of Computer Science and Engineering
Thesis Advisor, Chair of Committee

David J. Miller
Professor of Electrical Engineering

Kenneth W. Jenkins
Professor of Electrical Engineering

Kultegyn Aydin
Professor of Electrical Engineering
Graduate Program Co-ordinator

*Signatures are on file in the Graduate School.

# Abstract

Privacy is an important issue when publishing data that involves individuals' sensitive information. As a consequence, *Data sanitization* which is the process of removing such sensitive information, gained a lot of importance. This thesis addresses issues concerning two sanitization schemes, namely *anatomy* and *differential privacy*.

A framework is developed for the purpose of attacking Anatomy sanitization scheme. This is done by trying to infer the sensitive values of individuals in data sets protected using Anatomy. A method for obtaining well calibrated probabilities of the sensitive value assignments is also proposed. Compared to the previous attacks, this attack algorithm is much simpler, quicker and also gives comparable accuracy levels.

Any sanitization scheme basically destroys some information present in the data set in the process of protecting the sensitive data. This reduces the utility or usefulness of the data set. The utility of differentially private statistics is affected due to the noise added in the process of sanitization. We present a novel empirical Bayes' estimate for such statistics which improves its utility. The statistics are then used to construct a differentially private classifier. We observe that the classifiers' accuracy is enhanced using the proposed estimate. What makes this technique unique is that the method for improving utility does not use any kind of auxiliary information.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

It is difficult to overstate my gratitude to my advisor, Dr. Daniel Kifer. Throughout my research, he has been an immense source of inspiration, encouragement and enthusiasm. His guidance has instilled in me an aptitude and temperament for research. I consider myself fortunate to have worked under his supervision. I would also like to thank Dr. David Miller for his time and insightful suggestions to improve the quality of this thesis and Dr. Kenneth Jenkins for agreeing to be on my thesis committee despite his extremely busy schedule.

I am indebted to my lab mates at the Machine Learning Lab, especially Bing-Rong Lin for the many enlightening discussions we had. I would also like to thank all my friends for helping me through tough times.

Words are inadequate to express my gratitude to my family. I would like to thank my father, for his unfailing support, advice and encouragement, my mother for her undying love and affection and my brother for the continuous encouragement, emotional support and also for always being there for me. I would be lost without them. Last but not the least, I would like to thank my grandparents, especially my grandfather for the love and support he gave me during his life.

Chapter 1

# Introduction

The importance of machine learning is rising because of the large amounts of data being collected either for business intelligence or research purposes. While this is good for machine learning, the data collected which is very useful can also lead to substantial privacy problems. The AOL data release scandal is a good example. AOL released a compressed text file containing twenty million search keywords for over $650,000$ users over a 3-month period, intended for research purposes[1]. The data did not identify any individual explicitly but contained personally identifiable information in many of the queries. An individual could be identified and matched to their account and search history by using these queries[1]. The New York Times was able to locate an individual from the released and anonymized search records by cross referencing them with phone book listings[2]. Consequently, the ethical implications of using such data for research is under debate. The attacks on the Netflix database[3] and the Massachusetts Group Insurance Commission medical encounter database[4] are other examples of situations where the privacy of individuals was compromised. As such, privacy preservation has become a very important aspect of data publishing.

Examples from [5] are used to illustrate different aspects of privacy in this thesis. Assume that a hospital wants to release patient's medical records in Table 1.1. referred to as *microdata*. The attribute *'Disease'* is sensitive, in the sense that no adversary should be able to infer the disease of a patient with significant confidence. Let us call the all the attributes except the sensitive attribute as the quasi-identifier(QI) attributes. Here, *Age*, *sex* and *Zip Code* are the QI attributes. Now, consider an adversary who knows that his friend Bob, who is 29 years old and lives in area code 90212 is present in the data base. Looking at the table, he can infer that Bob has Cancer. The adversary could infer Bob's disease because his QI attribute values(Gender=M, Age=29 & Zip Code=90212) were unique in the table.

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | M | 25 | 90210 | AIDS |
| 2 | F | 43 | 90211 | AIDS |
| 3 | M | 29 | 90212 | Cancer |
| 4 | M | 41 | 90213 | AIDS |
| 5 | F | 41 | 07620 | Cancer |
| 6 | F | 40 | 33109 | Cancer |
| 7 | F | 40 | 07620 | Flu |
| 8 | F | 24 | 33109 | None |
| 9 | M | 48 | 07620 | None |
| 10 | F | 40 | 07620 | Flu |
| 11 | M | 48 | 33109 | Flu |
| 12 | M | 49 | 33109 | None |

**Table 1.1.** Original Table consisting Micro-data

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | * | 25-49 | 9021* | AIDS |
| 2 | * | 25-49 | 9021* | AIDS |
| 3 | * | 25-49 | 9021* | Cancer |
| 4 | * | 25-49 | 9021* | AIDS |
| 5 | * | 25-49 | 0762* | Cancer |
| 6 | * | 25-49 | 0762* | Flu |
| 7 | * | 25-49 | 0762* | None |
| 8 | * | 25-49 | 0762* | Flu |
| 9 | * | 25-49 | 3310* | Cancer |
| 10 | * | 25-49 | 3310* | None |
| 11 | * | 25-49 | 3310* | Flu |
| 12 | * | 25-49 | 3310* | None |

**Table 1.2.** k-anonymous Generalization

## 1.1 Generalization

Generalization methods have been proposed which divide tuples into groups and transform their QI values into less specific forms, so that the tuples in the same group cannot be distinguished by their QI values. Let us refer to these groups as QI-groups as they are grouped based on the QI values of the tuples. Table 1.2 shows an example of one such generalization method applied on Table 1.1. In this table, the attribute *Gender* has been totally suppressed while attribute *age* has been coarsened to the range 25-49 and the attribute *Zip code* has been coarsened by suppressing the last digit. Tuples 1-4 form the QI-group 1, tuples 5-8 form QI-group 2 and tuples 9-12 form QI-group 3. We can see that in each QI-group, the non-sensitive attribute values of all the tuples are the same. Considering the adversary's knowledge, we can see that Bobs attributes fall in QI-group 1, in which there are 4 records. Thus, Bob's disease attribute is hidden among the disease attribute of these 4 records.

Two notions, k-anonymity and l-diversity were proposed to measure the degree of privacy

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | M | 25-49 | ***** | AIDS |
| 2 | M | 25-49 | ***** | Flu |
| 3 | M | 25-49 | ***** | Cancer |
| 4 | M | 25-49 | ***** | None |
| 5 | F | 25-49 | ***** | Cancer |
| 6 | F | 25-49 | ***** | AIDS |
| 7 | F | 25-49 | ***** | None |
| 8 | F | 25-49 | ***** | Flu |
| 9 | * | 25-49 | ***** | Cancer |
| 10 | * | 25-49 | ***** | Aids |
| 11 | * | 25-49 | ***** | Flu |
| 12 | * | 25-49 | ***** | None |

**Table 1.3.** l-diverse generalization

preservation. A generalized table is k-anonymous[6, 7] if each QI-group involves at least k tuples(e.g. Table1.2 is 4-anonymous). However, in QI-group 1, in which Bob is present, AIDS occurs 3 times and Cancer once. So Bob's should have either AIDS or Cancer. This does not ensure Bob's privacy. *l-diversity*[8] on the other hand provides a stronger privacy protection. A table is $l$-diverse if, the most frequent sensitive value in a QI-group occurs no more than $l$ times the least frequent sensitive value in the QI-group. Table1.3 shows an example of a 4-diverse generalization. We can observe that all 4 disease values occur in each QI-group. However, the generalization process led to the attribute *Zip code* being suppressed totally.

Xiao et. al.[9] proposed a new anonymization method called Anatomy based on $l$-diversity which does not use generalization. Anatomy sanitization method releases a quasi-identifier table(which contains the quasi-identifier attributes) and a sensitive table(which contains the sensitive attributes) which separate QI values from the sensitive values. Table 1.4 shows an example of QI and sensitive tables obtained from Table 1.1.

| Tuple ID | Gender | Age | Zip Code | GID |
|---|---|---|---|---|
| 1 | M | 25 | 90210 | 1 |
| 2 | M | 29 | 90212 | 1 |
| 3 | F | 40 | 07620 | 1 |
| 4 | F | 24 | 33109 | 1 |
| 5 | F | 43 | 90211 | 2 |
| 6 | F | 41 | 07620 | 2 |
| 7 | M | 48 | 07620 | 2 |
| 8 | F | 40 | 07620 | 2 |
| 9 | M | 41 | 90213 | 3 |
| 10 | F | 40 | 33109 | 3 |
| 11 | M | 48 | 33109 | 3 |
| 12 | M | 49 | 33109 | 3 |

| GID | Disease |
|---|---|
| 1 | AIDS |
| 1 | Cancer |
| 1 | Flu |
| 1 | None |
| 2 | AIDS |
| 2 | Cancer |
| 2 | None |
| 2 | Flu |
| 3 | AIDS |
| 3 | Cancer |
| 3 | Flu |
| 3 | None |

**Table 1.4.** Quasi-identifier and Sensitive Tables

Construction of the anatomized tables is done as follows. First, the tuples of the *microdata*

are partitioned into several partitions or groups based on $l$-diversity, that is, the most frequent sensitive value in a group occurs no more than $l$ times the least frequent sensitive value in the group. Then the quasi-identifier table which includes all its exact QI values along with the group ID in a new column. Finally, the sensitive table is created which consists of the sensitive value (disease) of each tuple along with the group ID.

Now, if an adversary wants to predict the sensitive value of a person whose QI values are known, he has to randomly guess from the sensitive values present in the group to which the person belongs because the QIT does not indicate the sensitive value of any tuple. However, even using Anatomy, the privacy of individuals in a data set is not guaranteed. That is, the sensitive value of a record present in the anatomized tables can be predicted with probability more than intended. In the first part of this thesis, a Naive-Bayes' classifier is used to predict the sensitive value of of records in anatomized tables. A calibration method for the prediction probabilities obtained is also proposed.

## 1.2 Differential Privacy

Dwork proposed a more principled approach to privacy, in the form of Differential privacy[10] which had an extraordinary impact on the privacy community. The main idea behind differential privacy is that the presence or absence of a record in a database should not change in a significant way, the probability of obtaining a certain answer for a given query. In its simplest form, differential privacy works by computing the exact value of the statistic and then adding random noise to it. The random noise is chosen in such a way that the effect of any single individual on the statistic is masked. By adding random noise to the statistic, differential privacy distorts it from ground truth, thus reducing the usefulness or *utility* of the statistic. In the second part of the thesis, a method to improve the utility of differentially private statistics without the use of any auxiliary information is proposed.

## 1.3 Contributions

In this section, we summarize the contributions of this thesis. The work in this thesis is divided into two parts.

1. In the first part, an attack algorithm to attack the privacy of anatomy sanitized data sets is proposed.

2. In the second part, a method to improve the usefulness or utility of differentially private statistics is described. The method is then evaluated by constructing a classifier using these statistics.

## 1.4 Organization of the thesis

The rest of this Masters thesis is organized as follows. Chapter 2 gives an overview of related work in attacking anonymization methods. Chapters 3-5 describe the attack algorithm (first part of the thesis) and the results obtained. Chapters 6-8 constitute the second part of the thesis. Finally, Chapter 9 presents the conclusions of the thesis.

Chapter 2

# Related Work

In this chapter, an overview of previous work in the areas related to this thesis is presented. First, we describe the work relating to anonymization methods. Later, we describe the previous work in improving utility of differential privacy.

## 2.1 Anonymization

The first methods of anonymization relied simply on hiding or suppressing the personal attributes. For example, table2.1 shows the data of different individuals. The social security number attribute 'SSN' is the personal information and is supposed to be hidden. Hence the suppression technique would remove the 'SSN' attribute and result in table2.2.

| SSN | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1111111 | M | 25 | 90210 | AIDS |
| 2222222 | F | 43 | 90211 | AIDS |
| 3333333 | M | 29 | 90212 | Cancer |
| 4444444 | M | 41 | 90213 | AIDS |
| 5555555 | F | 41 | 07620 | Cancer |
| 6666666 | F | 40 | 33109 | Cancer |
| 7777777 | F | 40 | 07620 | Flu |
| 8888888 | F | 24 | 33109 | None |
| 9999999 | M | 48 | 07620 | None |
| 1010101 | F | 40 | 07620 | Flu |
| 1212121 | M | 48 | 33109 | Flu |
| 3434343 | M | 49 | 33109 | None |

**Table 2.1.** Table with personally identifiable information(SSN)

A similar technique was used by the Group Insurance Commission (GIC) on the medical records of the state employees of Massachusetts by suppressing the personally identifiable information. Sweeney[6] linked this data with the voter registration list and identified the medical

| Gender | Age | Zip Code | Disease |
|---|---|---|---|
| M | 25 | 90210 | AIDS |
| F | 43 | 90211 | AIDS |
| M | 29 | 90212 | Cancer |
| M | 41 | 90213 | AIDS |
| F | 41 | 07620 | Cancer |
| F | 40 | 33109 | Cancer |
| F | 40 | 07620 | Flu |
| F | 24 | 33109 | None |
| M | 48 | 07620 | None |
| F | 40 | 07620 | Flu |
| M | 48 | 33109 | Flu |
| M | 49 | 33109 | None |

**Table 2.2.** Suppressed table

records of the governor of Massachusetts. Such a *linking attack* was possible because the governors data from the voter list mapped to a single record in the medical records. Sweeney observed that altering the released information to map many possible records can thwart this kind of attack. This was the motivation for the concept of $k$-anonymity[6].

The next set of privacy techniques relied on generalization using $k$-anonymity[7]. According to $k$-anonymity, in simple terms, generalizations should be done in such a way that each set of quasi-identifier attributes should appear at least $k$ times. Numerous variations[11, 12, 13, 14, 15, 16, 7, 6] of k-anonymous generalizations were proposed. Among the notable techniques in this group were global and local recoding. Global recoding sanitizes by coarsening the domain of each non-sensitive attribute. Table2.4 shows the output of global recoding on table2.3. Notice that the zip code has been coarsened by replacing the last digit with a *, age has been coarsened into intervals of length 25, and the gender has been suppressed. Local recoding([17], [18], [19]) refers to a similar class of partition based sanitization schemes where the generalizations are more flexible. In local recoding, the domain of the non-sensitive attributes can be coarsened in a different way for each group. Table 2.5 shows an example of a table created using local recoding. Note that the age and zip code are coarsened in different ways in each group.

The motivation behind such schemes is that if an attacker knows the nonsensitive attributes of an individual in the table, the attacker cannot be certain which tuple belongs to that individual. Such an attacker would not be able to identify the tuple belonging to an individual in Tables 2.4 and 2.5 with resolution better than a group size of 4.

The $k$-anonymization model for generalization only says that each tuple of non-sensitive attributes should appear at least $k$ times, but does not impose any restriction on the sensitive attributes. If a group is such that all the sensitive values in it are the same, such a table does not ensure any privacy at all to the members of that group. It was obvious that there needed to be a restriction on the occurrence of sensitive attributes too. This led to the concept of $l$-diversity[8]. According to $l$-diversity, the most frequent sensitive value in a group occurs no more than $l$ times the least frequent sensitive value in the group.

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | M | 25 | 90210 | AIDS |
| 2 | F | 43 | 90211 | AIDS |
| 3 | M | 29 | 90212 | Cancer |
| 4 | M | 41 | 90213 | AIDS |
| 5 | F | 41 | 07620 | Cancer |
| 6 | F | 40 | 33109 | Cancer |
| 7 | F | 40 | 07620 | Flu |
| 8 | F | 24 | 33109 | None |
| 9 | M | 48 | 07620 | None |
| 10 | F | 40 | 07620 | Flu |
| 11 | M | 48 | 33109 | Flu |
| 12 | M | 49 | 33109 | None |

**Table 2.3.** Original Table consisting Micro-data

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | * | 25-49 | 9021* | AIDS |
| 2 | * | 25-49 | 9021* | AIDS |
| 3 | * | 25-49 | 9021* | Cancer |
| 4 | * | 25-49 | 9021* | AIDS |
| 5 | * | 25-49 | 0762* | Cancer |
| 6 | * | 25-49 | 0762* | Flu |
| 7 | * | 25-49 | 0762* | None |
| 8 | * | 25-49 | 0762* | Flu |
| 9 | * | 25-49 | 3310* | Cancer |
| 10 | * | 25-49 | 3310* | None |
| 11 | * | 25-49 | 3310* | Flu |
| 12 | * | 25-49 | 3310* | None |

**Table 2.4.** Global Recoding

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | * | 25-45 | 9021* | AIDS |
| 2 | * | 25-45 | 9021* | AIDS |
| 3 | * | 25-45 | 9021* | Cancer |
| 4 | * | 25-45 | 9021* | AIDS |
| 5 | * | 40-50 | 07620 | Cancer |
| 6 | * | 40-50 | 07620 | Flu |
| 7 | * | 40-50 | 07620 | None |
| 8 | * | 40-50 | 07620 | Flu |
| 9 | * | 20-50 | 33109 | Cancer |
| 10 | * | 20-50 | 33109 | None |
| 11 | * | 20-50 | 33109 | Flu |
| 12 | * | 20-50 | 33109 | None |

**Table 2.5.** Local Recoding

Xiao et. al.[9] proposed a anonymization method based on l-diversity called Anatomy. In their work, they point out that releasing the non-sensitive attributes and sensitive attributes separately would enable them to avoid the generalization on the QI values. They also showed that the privacy of such an anonymization method is greater. Chapter 3 talks more about Anatomy sanitization method.

Kifer[5] observed that even in Anatomy, there exist correlations between attributes which an adversary can use, to learn a set of beliefs. DeFinetti attack[5] proposed by Kifer builds a classifier which, given the quasi-identifier attribute values of a tuple in a group predicts the corresponding sensitive value. First, a permutation of sensitive values in each group is guessed which assigns a sensitive value to each tuple in the group. Such an assignment then imposes a conditional probability distribution on each quasi-identifier attribute value given the sensitive attribute value. These conditional distributions in turn define a likelihood distribution on the permutations of sensitive values in each group. The permutations in each group are then evolved by sampling new permutations and adopting depending on the relative likelihood compared to the current permutation. The method described in this thesis is inspired from the work of Kifer[5].

### 2.1.1 Calibration of Prediction Probabilities

Obtaining calibrated probabilities of prediction is a very important aspect of decision making. Different ways of calibrating the prediction probabilities[20, 21, 22] were proposed for different prediction methods. All the existing methods follow a similar procedure for calibration. First, a prediction rule is learned using the training data. The predicted probabilities given by the learned rule are then quantized into bins. The ground truth of the training data is then used to adjust the probabilities of prediction appropriately. However, such methods are not applicable to the case in hand. When predicting the sensitive value of a record in a table, there is no training data or 'ground truth'.

## 2.2 Differential Privacy

Dwork proposed a more principled approach to privacy, in the form of Differential privacy[10] which had an extraordinary impact on the privacy community. The main idea behind differential privacy is that the presence or absence of a record in a database should not change in a significant way, the probability of obtaining a certain answer for a given query. In its simplest form, differential privacy works by computing the exact value of the statistic and then adding random noise to it. The random noise is chosen in such a way that the effect of any single individual on the statistic is masked.

### 2.2.1 Utility

The utility or usefulness of a data set is reduced when protected using any privacy technique. This is because any privacy technique relies on hiding some information. Differential privacy

adds random noise to any statistic of the data set that needs to be released, effectively distorting the statistic from ground truth. The more a statistic is distorted, the lesser its utility. The idea of post-processing the output of a differentially private mechanism(to improve its utility) when consistency constraints on the outputs(such as constraints on marginals of contingency tables) exist has been studied extensively[23, 24, 25, 26, 27, 28, 29, 30]. Barak et. al. [23] proposed a method for making a set of statistics obtained from the differentially private contingency tables consistent over multiple released tables. Hay et. al. [24] showed that the accuracy of differentially private histogram queries can be improved when there exist consistency constraints. However, all of the methods use auxiliary information in the form of the constraints. As opposed to the previous methods, the method proposed in this thesis does not use any auxiliary information.

Chapter 3

# Anatomy Sanitization

Anatomy sanitization technique was suggested by Xiao and Tao as an improvement over the generalization techniques. As opposed to the generalization techniques, where the nonsensitive attributes are divided into ranges of values, in the Anatomy sanitization technique, the attribute values are preserved the way they are. Instead, the sensitive and nonsensitive attributes are separated from one other. Then the tuples present in the database are divided into groups such that in each group, the sensitive values are distinct. In this chapter, we formally describe the method of Anatomy. For a more detailed description, refer to [9]. We first introduce some basic notation that will be used in the rest of the thesis and then the concepts required to introduce Anatomy.

## 3.1 Notation

Let $T = t_1, t_2, ..., t_N$ be a table consisting of N records or tuples with non-sensitive attributes $R_1, ..., R_{|R|}$ and sensitive attribute $S$. Let $t.R_j$ represent the $j^{th}$ nonsensitive attribute and $t.S$ represent the sensitive attribute of tuple $t$.

## 3.2 Anatomy

As with any generalization technique, Anatomy needs the data to be partitioned first. A partition or QI-group basically consists of several subsets of table T, such that each tuple belongs to exactly one subset. The partitions need to be l-diverse, that is, the most frequent sensitive value in a group should occur no more than l times the least frequent sensitive value. Formally, an l-diverse partition is defined as follows.

**Definition** l-diverse partition [8] A partition with m QI-groups is l-diverse, if each QI-group $QI_j (1 \leq j \leq m)$ satisfies the following condition. Let $v$ be the most frequent sensitive value in

$QI_j$, and $c_j(v)$, the number of tuples $t \in QI_j$ with $t.S = v$;then

$$c_j(v)/|QI_j| \leq 1/l$$

where $|QI_j|$ is the size (the number of tuples) of $QI_j$.

The formal definition of Anatomy follows.

**Definition (Anatomy [9])** Given an l-diverse partition with m QI-groups, anatomy produces a quasi-identifier table (QIT) and a sensitive table (ST) as follows. The QIT has schema

$$(R_1, R_2, ..., R_{|R|}, Group - ID).$$

For each QI-group $QI_j (1 \leq j \leq m)$ and each tuple $t \in QI_j$, QIT has a tuple of the form:

$$(t.R_1, t.R_2, ..., t.R_{|R|}, j).$$

The ST has the schema

$$(Group - ID, S, Count).$$

For each QI-group $QI_j (1 \leq j \leq m)$ and each distinct S value $v$ in $QI_j$, the ST has a record of the form:

$$(j, v, c_j(v))$$

where $c_j(v)$ is the number of tuples $t \in QI_j$ with $t.S = v$. Apart from the tuples, the QIT or ST do not contain any other data.

The QIT and ST are collectively referred to as the anatomized tables. Xiao et. al.[9] also showed that the anatomized tables capture the correlation in T more accurately than generalized tables. For example, consider the data of individuals in table 3.1. The sensitive values *AIDS, Cancer, Flu, None* each occur thrice in the table. Based on 4-diverse partitions, Anatomy produces QIT and ST tables as shown in Table 3.2. Observe that the each sensitive value occurs once in each group. In each group, $count(AIDS) = count(Cancer) = count(Flu) = count(None) = 1$. Since the size of each group is 4, these groups are 4-diverse.

### 3.2.1 Privacy Preservation

Xiao et. al.[9] illustrate the privacy preservation of anatomy sanitization in the following way.

Consider any tuple $t \in T$, which is contained in QI group $QI_j$ (in the underlying l-diverse partition for some $j \in [1, m]$. For example, consider tuple 5 of the QI-table given in table 3.2. The adversary who attempts to find out $t.S$ of tuple 5, can obtain $j = 2$ from the QIT which, however, does not have S data. Hence, the adversary can only conjecture that $t.S$ equals one of the S values pertinent to $QI_2$ in the ST. Without any other information, the adversary assumes that every tuple in $QI_2$ has an equal chance to carry any S value relevant to $QI_2$. Hence the

| Tuple ID | Gender | Age | Zip Code | Disease |
|---|---|---|---|---|
| 1 | M | 25 | 90210 | AIDS |
| 2 | F | 43 | 90211 | AIDS |
| 3 | M | 29 | 90212 | Cancer |
| 4 | M | 41 | 90213 | AIDS |
| 5 | F | 41 | 07620 | Cancer |
| 6 | F | 40 | 33109 | Cancer |
| 7 | F | 40 | 07620 | Flu |
| 8 | F | 24 | 33109 | None |
| 9 | M | 48 | 07620 | None |
| 10 | F | 40 | 07620 | Flu |
| 11 | M | 48 | 33109 | Flu |
| 12 | M | 49 | 33109 | None |

**Table 3.1.** Original Table consisting Micro-data

| Tuple ID | Gender | Age | Zip Code | GID |
|---|---|---|---|---|
| 1 | M | 25 | 90210 | 1 |
| 2 | M | 29 | 90212 | 1 |
| 3 | F | 40 | 07620 | 1 |
| 4 | F | 24 | 33109 | 1 |
| 5 | F | 43 | 90211 | 2 |
| 6 | F | 41 | 07620 | 2 |
| 7 | M | 48 | 07620 | 2 |
| 8 | F | 40 | 07620 | 2 |
| 9 | M | 41 | 90213 | 3 |
| 10 | F | 40 | 33109 | 3 |
| 11 | M | 48 | 33109 | 3 |
| 12 | M | 49 | 33109 | 3 |

| GID | Disease |
|---|---|
| 1 | AIDS |
| 1 | Cancer |
| 1 | Flu |
| 1 | None |
| 2 | AIDS |
| 2 | Cancer |
| 2 | None |
| 2 | Flu |
| 3 | AIDS |
| 3 | Cancer |
| 3 | Flu |
| 3 | None |

**Table 3.2.** Quasi-identifier and Sensitive Tables

probability of a tuple in a group having a sensitive value is the fraction of tuples in the group having that sensitive value. Hence the adversary concludes that tuple 5 can have AIDS, Cancer, Flu or None with probability 1/4 each. Since anatomy uses l-diverse partitions, this probability is always bound above by the value $1/l$.

## 3.3 Vulnerability of Anatomy

Anatomy anonymization method, as pointed out in [9], does capture the correlations between the attributes better than generalization. However, the privacy guarantees given by Xiao et. al.[9] is flawed. It fails to take into account the learning ability of an adversary. This was pointed out by Kifer[5]. The following examples illustrates the vulnerability of anatomy.

Consider Table 3.3. It shows the anatomized tables which contain the one non-sensitive attribute and one sensitive attribute. The non-sensitive attribute records whether or not an individual smokes and the sensitive attribute is the individuals disease. The group size is 2.

From the motivation of anatomy sanitization scheme, a person in any group would be equally likely to have one of the two sensitive values present in that group. For example, tuple 5 in group 3 would be equally likely to have cancer or no disease. However, an attacker who is willing to learn may make an observation that whenever a smoker is in a group, then cancer is one of the diseases that appear in the group (this occurs in groups 1,3,4, and 6). On the other hand, groups that do not have any smokers do not contain cancer(true for groups 2 and 5). Thus the attacker may reason that the appearance pattern of smoking and cancer in this table is evidence of a correlation between smoking and cancer. Therefore, even though tuple 6, a non-smoker, appears in a group where one individual has cancer and the other is healthy, tuple 6 should be more probable to have no disease. This is because tuple 5, a smoker in the same group, is more likely to have cancer according t the correlation exhibited by the table.

| Tuple ID | Smoker? | GID |
|---|---|---|
| 1 | y | 1 |
| 2 | y | 1 |
| 3 | n | 2 |
| 4 | n | 2 |
| 5 | y | 3 |
| 6 | n | 3 |
| 7 | y | 4 |
| 8 | y | 4 |
| 9 | n | 5 |
| 10 | n | 5 |
| 11 | y | 6 |
| 12 | n | 6 |

| GID | Disease |
|---|---|
| 1 | Cancer |
| 1 | Flu |
| 2 | Flu |
| 2 | None |
| 3 | Cancer |
| 3 | None |
| 4 | Flu |
| 4 | None |
| 5 | Flu |
| 5 | None |
| 6 | Cancer |
| 6 | None |

**Table 3.3.** Quasi-identifier and Sensitive Tables

The vulnerability of anatomy basically arises due to an assumption made by Xiao et. al. [9], which is that the tuples in different groups are independent. The above example shows that the correlations between attributes of tuples in different groups imposes a probability distribution on the sensitive value of a tuple. Such a distribution can be used by an adversary to predict the sensitive value of a tuple with probability more than that intended by anatomy. In the following chapter, we see how these probability distribution can be learned from anatomized tables using a Naive-Bayes model.

Chapter 4

# Attacking Privacy Using Naive-Bayes Classifier Trained using EM Algorithm

In this chapter, a Naive-Bayes classifier is described for the purpose of extracting sensitive information from anatomy sanitized tables. The parameters of the Naive-Bayes classifier will be trained using the Expectation-Maximization Algorithm.

## 4.1 Notation

The notation used in the following sections is described in this section. Let a data set contain N records or tuples. $t$ denotes a tuple in the data. $R_1, ..., R_{|R|}$ denote the domain of the non-sensitive attributes and $S$ denote the domain of the sensitive attribute. $t.R_j$ and $t.S$ denote the value of the $j^{th}$ non-sensitive attribute and the sensitive attribute of the tuple $t$ respectively. We are required to find the permutation over the sensitive values which is most likely. Using the Naive Bayes formulation, the nonsensitive attributes are conditionally independent given the sensitive value S. Hence, the Naive Bayes parameters are the probabilities $P(t.S = s_i)$ and $P(t.R_j = r_{jk}|t.S = s_i)$, where $r_{jk}$ represents the $k^{th}$ value the non-sensitive attribute $R_j$ can take and $s_i$ is the $i^{th}$ value the sensitive attribute $S$ can take. We will use $n$ as the tuple index.

## 4.2 Expectation Maximization

We introduce the latent variable matrix $\mathbf{z}$ consisting row vectors $\bar{z}_n$ with elements $z_{ni}$. $z_{ni}$ is the indicator variable for the sensitive attribute $s_i$ for tuple $t_n$, i.e., $z_{ni}$ is 1 if sensitive value $s_i$ is assigned to the tuple $t_n$. Let $T$ indicate the tuples present in the quasi-identifier table(without

the sensitive attribute). Let $W$ represent the Naive Bayes parameters.

The likelihood function is given by,

$$P(T, Z, W) = P(T, Z|W)P(W)$$

The term $P(W)$ represents the prior distribution of the Naive Bayes parameters. Taking natural logarithm on both the sides, we get,

$$\Rightarrow \ln P(T, Z, W) \quad = \quad \ln P(T, Z|W) + \ln P(W) \tag{4.1}$$

Since the nonsensitive attributes are conditionally independent given the sensitive attribute(Naive Bayes assumption), the likelihood of a single tuple is given by

$$P(t_n|W) = P(t_n.S) \prod_{j=1}^{|R|} P(t_n.R_j|t_n.S)$$

Using the latent variables $z_{ni}$, this can be written as follows.

$$P(t_n, z_n|W) = \prod_{i=1}^{|S|} \left( P(t_n.S = s_i) \prod_{j=1}^{|R|} P(t_n.R_j|t_n.S = s_i) \right)^{z_{ni}}$$

Since the tuples are independent of each other given the Naive Bayes paramters, we have

$$\begin{aligned} P(T, Z|W) &= \prod_{n=1}^{N} \prod_{i=1}^{|S|} \left( P(t_n.S = s_i) \prod_{j=1}^{|R|} P(t_n.R_j|t_n.S = s_i) \right)^{z_{ni}} \\ \Rightarrow \ln P(T, Z|W) &= \sum_{n=1}^{N} \sum_{i=1}^{|S|} z_{ni} \left( \ln P(t_n.S = s_i) + \sum_{j=1}^{|R|} \ln P(t_n.R_j|t_n.S = s_i) \right) \\ &= \sum_{n=1}^{N} \sum_{i=1}^{|S|} z_{ni} \left( \ln P(t_n.S = s_i) + \sum_{j=1}^{|R|} \sum_{k=1}^{|R_j|} \delta_{njk} \ln P(t.R_j = r_{jk}|t.S = s_i) \right) \end{aligned}$$

Let $P(t.S = s_i) = \psi(i)$ and $P(t_n.R_j|t_n.S = s_i) = w_{ijk}$.

$$\Rightarrow \ln P(T, Z|W) = \sum_{n=1}^{N} \sum_{i=1}^{|S|} z_{ni} \left( \ln \psi_i + \sum_{j=1}^{|R|} \sum_{k=1}^{|R_j|} \delta_{njk} \ln(w_{ijk}) \right) \tag{4.2}$$

where $\delta_{njk}$ is an indicator as to whether the $j^{th}$ nonsensitive attribute of the $n^{th}$ tuple takes the value $r_{jk}$, that is

$$\delta_{njk} = \begin{cases} 0 & \text{if } t_n.R_j \neq r_{jk}; \\ 1 & \text{if } t_n.R_j = r_{jk}; \end{cases}$$

The second term in equation 4.1 is given by

$$\begin{aligned} P(W) &= P(\psi_1, ..., \psi_{|S|}) \prod_i \prod_j P(w_{ij1}, ..., w_{ij|R_j|}) \\ \ln P(W) &= \ln P(\psi_1, ..., \psi_{|S|}) + \textstyle\sum_i \sum_j \ln P(w_{ij1}, ..., w_{ij|R_j|}) \end{aligned} \tag{4.3}$$

$$\tag{4.4}$$

Therefore, from 4.1, 4.2 and 4.3, we have,

$$\begin{aligned} \ln P(T, Z, W) &= \sum_{n=1}^{N} \sum_{i=1}^{|S|} z_{ni} \left( ln\psi_i + \sum_{j=1}^{|R|} \sum_{k=1}^{|R_j|} \delta_{njk} ln w_{ijk} \right) \\ &\quad + \ln P(\psi_1, ..., \psi_{|S|}) + \sum_i \sum_j \ln P(w_{ij1}, ..., w_{ij|R_j|}) \end{aligned}$$

Dirichlet priors are used for the Naive Bayes parameters randomize the initialization. Hence, we have,

$$P(\psi_1, ..., \psi_{|S|}) = Dir(1+\alpha, ..., 1+\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{|S|} \psi_i^{\alpha}$$

$$P(w_{ij1}, ..., w_{ij|R_j|}) = Dir(1+\beta, ..., 1+\beta) = \frac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{|R_j|} w_{ijk}^{\beta}$$

Therefore,

$$\frac{\partial \ln P(W)}{\partial \psi_i} = \frac{\alpha}{\psi_i}$$

$$\frac{\partial \ln P(W)}{\partial w_{ijk}} = \frac{\beta}{w_{ijk}}$$

Therefore from 4.1, the expected value of the log-likelihood under the posterior distribution of the latent variables $Z$ is given by

$$E_z[\ln P(T, Z, W)] = E_z[\ln P(T, Z|W)] + \ln E_z[P(W)] \tag{4.5}$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{|S|} E_z[z_{ni}] \left( \ln \psi_i + \sum_{j=1}^{|R|} \sum_{k=1}^{|R_j|} \delta_{njk} \ln (w_{ijk}) \right) + \ln P(W) \tag{4.6}$$

$$\tag{4.7}$$

Now, we are required to find $E_z[z_{ni}]$. We consider two possible cases here.

**Case1: Considering the effect of permutations in the group** In this case, the probability of a tuple having a particular sensitive value is dependent on the other tuples in the group.

In this case, we have,

$$\begin{aligned} E_z[z_{ni}] &= P(z_{ni}=1|t_n,W) \\ &= \frac{\sum_{\pi\in\pi^i} P(\pi)}{\sum_i \sum_{\pi\in\pi:\pi_n=s_i} P(\pi)} \end{aligned}$$

where $\pi$ is a permutation of sensitive values present in the group in which $t_n$ is present, $\pi^i$ is the set of all permutations which assign the sensitive value $s_i$ to the tuple $t_n$. Let $\pi_m$ represent the sensitive value that is assigned to tuple $t_m$ in the group by the permutation $\pi$ of sensitive values present in that group. Hence, the above quantity is given by,

$$\begin{aligned} E_z[z_{ni}] &= \frac{\sum_{\pi\in\pi^i} \prod_{m=1}^{g_size} P(t_m.S=\pi_m)P(t_m.R|t_m.S=\pi_m)}{\sum_{i=1}^{|S|}\sum_{\pi\in\pi^i} \prod_{m=1}^{g_size} P(t_m.S=\pi_m)P(t_m.R|t_m.S=\pi_m)} \\ &= \frac{\sum_{\pi\in\pi^i} \prod_{m=1}^{g_size} \psi_{\pi_m}^{old} \prod_{j=1}^{|R|}\sum_k \delta_{njk} w_{\pi_{mjk}}^{old}}{\sum_{i=1}^{|S|}\sum_{\pi\in\pi^i} \prod_{m=1}^{g_{size}} \psi_{\pi_m}^{old} \prod_{j=1}^{|R|}\sum_k \delta_{njk} w_{\pi_{mjk}}^{old}} \\ &= \gamma_{ni} \end{aligned}$$

**Case2: Without considering the effect of permutations in the group** In this case, the probability of a tuple having a particular sensitive value is not dependent on the other tuples in the group. We shall call this case as the approximate algorithm. Such an approximation is considered because the E-step becomes smaller which *might* result in smaller running time for the algorithm. We have,

$$\begin{aligned} E_z[z_{ni}] &= P(z_{ni}=1|t_n,W) \\ &= \frac{P(t_n.S=s_i)\prod_{j=1}^{|R|} P(t_n.R_j|t_n.S=s_i)}{\sum_q \left(P(t_n.S=s_q)\prod_{j=1}^{|R|} P(t_n.R_j|t_n.S=s_q)\right)} \\ &= \frac{\psi_i^{old}}{\sum_q \left(\psi_q^{old}\prod_{j=1}^{|R|}\sum_{k=1}^{|R_j|}\delta_{njk} w_{qjk}^{old}\right)} \\ &= \gamma_{ni} \end{aligned}$$

Using $E_z[z_{ni}] = \gamma_{ni}$ in 4.6, we get,

$$E_z[\ln P(T,Z,W)] = \sum_{n=1}^{N}\sum_{i=1}^{|S|}\gamma_{ni}\left(\ln\psi_i + \sum_{j=1}^{|R|}\sum_{k=1}^{|R_j|}\delta_{njk}\ln w_{ijk}\right) + \ln P(\psi_1,...,\psi_{|S|}) + \sum_i\sum_j \ln P(w_{ij1},...,w_{ij|R_j|}) \tag{4.8}$$

The loss function is then given by,

$$\begin{aligned} Q(W, W^{old}) &= E_z[\ln P(T, Z, W)] \\ &= \sum_{n=1}^{N}\sum_{i=1}^{|S|}\gamma_{ni}\left(\ln\psi_i + \sum_{j=1}^{|R|}\sum_{k=1}^{|R_j|}\delta_{njk}\ln w_{ijk}\right) \quad (4.9) \\ &\quad + \ln P(\psi_1, ..., \psi_{|S|}) + \sum_i\sum_j \ln P(w_{ij1}, ..., w_{ij|R_j|}) \quad (4.10) \\ & \quad (4.11) \end{aligned}$$

We are required to minimize the loss function $Q(W, W^{old})$. We also need to consider the constraints on $\psi_{ni}$ and $w_{ijk}$.

### 4.2.1 Constraints on parameters:

1. The constraint on $\psi_{ni}$ says that each tuple will take one of the sensitive values.

$$\Rightarrow \sum_{i=1}^{|S|}\psi_i = 1$$

2. The constraint on $w_{ijk}$ says that each non-sensitive attribute $R_j$ will take one of the possible values $r_{jk}$.

$$\Rightarrow \sum_{k=1}^{|R_j|} w_{ijk} = 1$$

Therefore, the Lagrange function required to be minimized is given by,

$$\Lambda(\psi_i, w_{ijk}) = Q(W, W^{old}) + \eta\left(\sum_{i=1}^{|S|}\psi_i - 1\right) + \psi\left(\sum_{k=1}^{|R_j|} w_{ijk} - 1\right) \quad (4.12)$$

Differentiating 4.12 with respect to $\psi_{ni}$ and equating to zero, we have,

$$\frac{\partial\Lambda}{\partial\psi_i} = \frac{\sum_{n=1}^{N}\gamma_{ni} + \alpha}{\psi_i} + \eta = 0$$

$$\Rightarrow \psi_i = \frac{\sum_{n=1}^{N}\gamma_{ni} + \alpha}{\sum_{n=1}^{N}\sum_{i=1}^{|S|}\gamma_{ni} + |S|\alpha}$$

Similarly, differentiating with respect to $w_{ijk}$ and equating to zero, we have,

$$\frac{\partial\Lambda}{\partial w_{ijk}} = \frac{\sum_{n=1}^{N}\gamma_{ni}\delta_{njk} + \beta}{w_{ijk}} + \psi = 0$$

$$\Rightarrow w_{ijk} = \frac{\sum_{n=1}^{N} \gamma_{ni}\delta_{njk} + \beta}{\sum_{k=1}^{|R_j|}\left(\sum_{n=1}^{N} \gamma_{ni}\delta_{njk} + \beta\right)}$$

## 4.3 Assigning Sensitive Values

At the end of the EM cycles, we have the values $\gamma_{ni}$ which represents the probability of $t_n$ having the sensitive value $s_i$ given the nonsensitive attributes of $t_n$. The sensitive values present in a group need to be assigned to the tuples in the group using these probabilities. Specifically, for every group, we need to find the permutation of sensitive values that is most probable. Let $g$ represent the number of tuples in a group or the group size. Let $t_1, ..., t_g$ and $s_1, ..., s_g$ be the tuples and sensitive values present in the group respectively. The probability of a permutation $\pi$ of the sensitive values is given by

$$Pr(\pi) = \frac{\prod_{i=1}^{g} \gamma_{i\pi_i}}{\sum_{\pi}\prod_{i=1}^{g} \gamma_{i\pi_i}}$$

where $\pi_i$ denotes the $i^{th}$ sensitive value in the permutation $\pi$ of sensitive values. To maximize the above probability, we need to maximize the numerator, which is $\prod_{i=1}^{g} \gamma_{i\pi_i}$, with respect to $\pi$.

$$\begin{aligned}
\pi^* &= \arg\max_{\pi} \prod_{i=1}^{g} \gamma_{i\pi_i} \\
&= \arg\max_{\pi} \sum_{i=1}^{g} \log \gamma_{i\pi_i} \\
&= \arg\min_{\pi} \sum_{i=1}^{g} -\log \gamma_{i\pi_i}
\end{aligned}$$

The Hungarian algorithm [31] can be used to find the permutation $\pi^*$ that minimizes the above quantity since it is an assignment problem. The full attack algorithm is given in Table 4.1. The attack algorithm using the approximate EM algorithm is given in Table 4.2

## 4.4 Interpretation of the update equations

1. The update equation for the parameters $\psi_i$, which is the probability of sensitive value $i$ is as follows.
$$\psi_i^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} + \alpha}{N + |S|\alpha}$$
It is the fraction of tuples that effectively have the sensitive value $i$.

2. The update equation for the parameters $w_{ijk}$, which is the probability that given a tuple

**Table 4.1.** Attack Algorithm

| | |
|---|---|
| 1: | Initialize the parameters $\psi_i$ and $w_{ijk}$. |
| 2: | **E-Step:**Calculate the posterior probabilities $\gamma_{ni}$ of $z_{ni} = 1$ once we have observed $t.R$. $\gamma_{ni} = \frac{\sum_{\pi \in \pi^i} \prod_{m=1}^{g_{size}} \psi_{\pi_m}^{old} \prod_{j=1}^{|R|} \sum_k \delta_{njk} w_{\pi_{mjk}}^{old}}{\sum_{i=1}^{|S|} \sum_{\pi \in \pi^i} \prod_{m=1}^{g_{size}} \psi_{\pi_m}^{old} \prod_{j=1}^{|R|} \sum_k \delta_{njk} w_{\pi_{mjk}}^{old}}$ where $\pi^i$ is the set of all permutations which assign the sensitive value $s_i$ to the tuple $t_n$. |
| 3: | **M-Step:**Re-evaluate the parameters from the posterior probabilities $\gamma_{ni}$. $\pi_i^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} + \alpha}{\sum_{n=1}^{N} \sum_{i=1}^{|S|} \gamma_{ni} + |S|\alpha} = \frac{\sum_{n=1}^{N} \gamma_{ni} + \alpha}{N + |S|\alpha}$ $w_{ijk}^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta}{\sum_{k=1}^{|R_j|} \left( \sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta \right)} = \frac{\sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta}{\sum_{k=1}^{|R_j|} \sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + |R_j|\beta}$ |
| 4: | count=count+1 If count = 20, assign sensitive values and find accuracy. |
| 5: | If accuracy has changed in the past 200 EM cycles, go back to step 2. |

**Table 4.2.** Approximate Attack Algorithm

| | |
|---|---|
| 1: | Initialize the parameters $\psi_i$ and $w_{ijk}$. |
| 2: | **E-Step:**Calculate the posterior probabilities $\gamma_{ni}$ of $z_{ni} = 1$ once we have observed $t.R$. $\gamma_{ni} = \frac{\pi_i^{old} \prod_j \sum_k \delta_{njk} w_{ijk}^{old}}{\sum_q \left( \pi_q^{old} \prod_{j=1}^{|R|} \sum_{k=1}^{|R_j|} \delta_{njk} w_{qjk}^{old} \right)}$ |
| 3: | **M-Step:**Re-evaluate the parameters from the posterior probabilities $\gamma_{ni}$. $\pi_i^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} + \alpha}{\sum_{n=1}^{N} \sum_{i=1}^{|S|} \gamma_{ni} + |S|\alpha} = \frac{\sum_{n=1}^{N} \gamma_{ni} + \alpha}{N + |S|\alpha}$ $w_{ijk}^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta}{\sum_{k=1}^{|R_j|} \left( \sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta \right)} = \frac{\sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta}{\sum_{k=1}^{|R_j|} \sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + |R_j|\beta}$ |
| 4: | count=count+1 If count = 20, assign sensitive values and find accuracy. |
| 5: | If accuracy has changed in the past 200 EM cycles, go back to step 2. |

has the sensitive value $i$, the value of its $j^{th}$ non-sensitive attribute is $r_{jk}$ is as follows

$$w_{ijk}^{new} = \frac{\sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + \beta}{\sum_{k=1}^{|R_j|} \sum_{n=1}^{N} \gamma_{ni} \delta_{njk} + |R_j|\beta}$$

It is the fraction of the effective number of tuples that have the sensitive value $i$ and the value of $j^{th}$ non-sensitive attribute being $r_{jk}$.

## 4.5 Experiments

In this section we present experiments demonstrating the effectiveness of the described attack algorithm against Anatomy. The attack code was implemented in MATLAB and run on a machine with Intel dual core 3.16GHz processor with 4GB main memory. The data used came from the Adult Dataset from the UCI Machine Learning Repository [32].

All the tuples with missing values were removed. This resulted in a data set with 30162 tuples. The attributes *workclass, relationship, gender, salary class* (whether it is above or below

| Group Size | $Measurement$ | $Acc(\%)$ | $Acc_{definetti}(\%)$ |
|---|---|---|---|
| 2 | Min | 70.73 | 76.6 |
| | Max | 72.35 | 78.5 |
| | Mean | 71.22 | 77.0 |
| | Baseline | 50.00 | 50.0 |
| 3 | Min | 53.72 | 56.8 |
| | Max | 54.57 | 57.9 |
| | Mean | 54.05 | 57.6 |
| | Baseline | 33.33 | 33.33 |
| 4 | Min | 43.33 | 40.0 |
| | Max | 43.87 | 40.8 |
| | Mean | 43.55 | 40.6 |
| | Baseline | 25.00 | 25.0 |

**Table 4.3.** Performance of Attack Algorithm

| Group Size | Measurement | Accuracy |
|---|---|---|
| 2 | Min | 33.66 |
| | Max | 66.72 |
| | Mean | 50.32 |
| | Baseline | 50.00 |
| 3 | Min | 23.63 |
| | Max | 45.91 |
| | Mean | 32.93 |
| | Baseline | 33.33 |
| 4 | Min | 17.15 |
| | Max | 36.67 |
| | Mean | 24.84 |
| | Baseline | 25.00 |

**Table 4.4.** Performance of Attack Algorithm using Approximate EM

\$50K), and *occupation* were retained. The occupation was treated as the sensitive attribute, and it had 14 distinct values.

Anatomized tables with 2 tuples per group, 3 tuples per group and 4 tuples per group were generated. Since 4 does not evenly divide 30162, an anatomized table with 4 tuples per group also has a few groups of 5. For each anatomized table, the attack algorithm was run 1000 times, each with a different initial starting point. The accuracy of the assignment was checked after every 10 EM cycles. A run was assumed to be converged if the accuracy was constant over 200 EM cycles. Tables 4.3 and 4.4 show the minimum, maximum and the mean and the baseline of the accuracies obtained for group sizes 2, 3 and 4. The baseline is simply $100/g\%$ since the in the worst case, a tuple has a probability $1/g$ of having one of the sensitive values present in the group. In table4.3, the results obtained for the algorithm which considers the permutations of sensitive values, with those obtained for the definetti attack described in [5].

### 4.5.1 Analysis

Table 4.3 shows the results obtained for the attack algorithm taking the permutations of sensitive values in each group into consideration. We observe that the minimum and the maximum values of accuracy obtained are close to each other showing that there is good convergence. In comparison with the attack algorithm described in [5], we see that the accuracies are lesser for group sizes 2 and 3. However, for group size 4, the accuracy obtained using the EM algorithm is better. Infact, as the group size increases, the attack using EM algorithm gets better compared to the definetti attack. It should also be noted that in the definetti attack in [5] was used to predict the sensitive value of only 1000 tuples at a time because of the high computational complexity. In contrast, the EM algorithm is used to predict the sensitive values of all the tuples present in the data set. This is possible because of the simple structure of the classifier used and the minimal computation effort involved. The performance of the algorithm on the anatomized tables of group sizes greater than 4 is yet to be examined.

Table 4.4 shows the results obtained for the attack algorithm without taking the permutations of sensitive values in each group into consideration. We observe that the minimum and the maximum values of accuracy are wide apart. This indicates poor convergence of the algorithm. Also, the mean accuracy is very close to the baseline showing that not much is being learned by the algorithm. Comparison with the definetti attack in this case is therefore omitted.

Chapter 5

# Calibration of Probabilities

In the previous chapter, we saw that a simple Naive-Bayes classifier can be used to predict the sensitive values in an anatomy sanitized database by using the correlations between the sensitive and non-sensitive attributes. When we talk about predicting a sensitive value, it is not always enough to just predict the most probable one. How probable the predicted value is, is also an important quantity. For example, if the sensitive attribute attribute is whether a person has cancer or not, and if we conclude that the person has cancer when the person actually does not, its very risky. As such, the probability of the person having cancer is the an important quantity. Unfortunately, a Naive-Bayes classifier gives overly optimistic probabilities of prediction, that is, the prediction probabilities are pushed towards 0 or 1. In this chapter, we look at a way of calibrating the probabilities of the sensitive value assignments. This will allow us choose a threshold of probability for predicting the sensitive value. For example, if the sensitive attribute is whether a person has cancer or not, we would want to predict that a person has cancer only if the prediction probability is at least, say 0.95. The threshold will depend on the type of sensitive attribute being predicted.

First, we look at ways of evaluating the prediction probabilities. Then, we evaluate the probabilities obtained from the Naive-Bayes classification experiments. We then present a method for calibration along with the justifications. The calibrated probabilities are then evaluated using the same evaluation methods and then compared with the uncalibrated probabilities.

## 5.1 Measures to Evaluate Prediction Probabilities

The previous section emphasizes the need for ways of evaluating the accuracy of the prediction probabilities. In this section, we consider two methods of evaluating prediction probabilities, namely Reliability Diagrams and Brier scores.

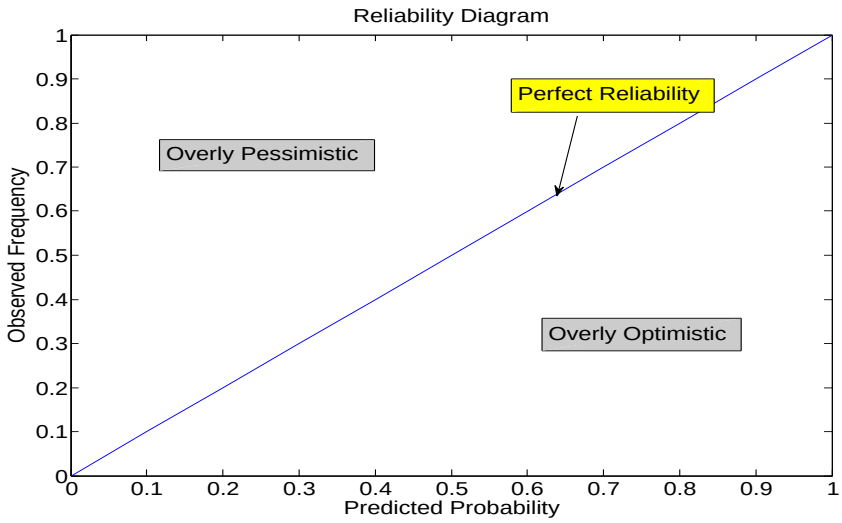



**Figure 5.1.** Reliability Diagram Illustration

### 5.1.1 Reliability Diagrams

Reliability diagrams [33] are simply the graphs of the observed frequency of an event plotted against the predicted probability of the event. This effectively tells how often a forecast probability actually occurred. A perfect prediction system will result in prediction with the probability of $x$ being consistent with the eventual outcome $x$ of the time. Hence when plotting the reliability diagram, comparisons are made with the diagonal. Figure5.1 illustrates the reliability diagram.

### 5.1.2 Brier Score

The Brier Score [34] is a score function that measures the accuracy of a set of probability assignments.

Consider a sequence of n events $E_1, ..., E_n$ assessed with probabilities $(p_1, ..., p_n)$ respectively. Let $(I_1, ..., I_n)$ indicate the occurrence of the n events.

$$I_j = \begin{cases} 1 & \text{if the event } E_j \text{ occurred;} \\ 0 & \text{otherwise;} \end{cases}$$

The Brier score(BS) of the probability assessments is defined as

The Brier score is given by

$$BS = \frac{1}{n}\sum_{j=1}^{n}(I_j - p_j)^2 \tag{5.1}$$

Now, quantize these n probabilities into k distinct values $p_i$ for $i = 1, ..., k$. Let $n_i$ indicate the number of events with associated probability $p_i$. Since the total number of events is $n$, we

have $n = \sum_{i=1}^{k} n_i$. Finally, let $r_i$ be the fraction of events with associated probability $p_i$ which are observed to occur. The calibration (CAL) score of the sequence of probability assessments is

$$CAL = \frac{1}{n} \sum_{i=1}^{k} n_i (r_i - p_i)^2 \tag{5.2}$$

The refinement (REF) score of the sequence of probability assessments is

$$REF = \frac{1}{n} \sum_{i=1}^{k} n_i r_i (1 - r_i) \tag{5.3}$$

It can easily be shown[35, 36] that $BS = CAL + REF$ for any sequence of observed events and predicted probabilities. For any group of events predicted with the same probability $p_i$, the calibration score is $(r_i - p_i)^2$. This measures the extent to which the observed frequency of occurrence among events in group $i$ differs from their common probability assessment. The refinement score for group $i$ is $r_i(1 - r_i)$. This measures the uniformity of occurrence within groups of events assessed with the same probability. The contribution to refinement from any group equals zero if the frequency of occurrence, $r_i$, is either zero or one. It reaches its maximum level of 1/4 when $r_i = 1/2$. For a detailed explanation of Brier score and its components, refer [34].

## 5.2 Evaluating the probabilities

To evaluate the probability of sensitive value assignments, first we need to formulate what the probabilities are. In this context, we talk about two kinds of probabilities, namely the probability of group assignment and the probability of tuple assignment. The probability of tuple assignment in a group is the marginal of the probability of group assignment with respect to that tuple assignment. First we take a look at what is the probability of a group assignment. A group assignment is defined by a permutation of the sensitive values present in the group. Let $\pi$ denote a permutation of sensitive values in a group, and $g$ denote the number of tuples in a group. Let $\pi_m$ denote the sensitive value assigned to tuple $t_m$ by the permutation $\pi$. Note there are $g!$ number of possible permutations of sensitive values in a group. The set consisting of all possible group assignments will form the sample space. Hence the probability of a group assignment is given by

$$Pr(\pi) = \frac{\prod_{m=1}^{g} \gamma_{m\pi_m}}{\sum_{\pi} \prod_{m=1}^{g} \gamma_{m\pi_m}}$$

The probability of a tuple assignment is then given by the sum of probabilities of all the group assignments that contain that particular tuple assignment.

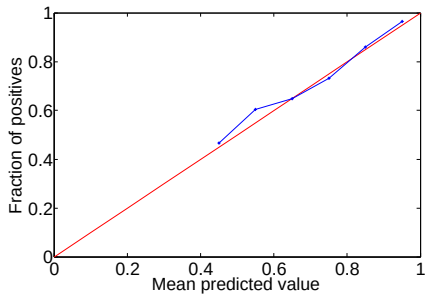



**Figure 5.2.** Reliability Diagram of probabilities obtained using EM algorithm for group size 2

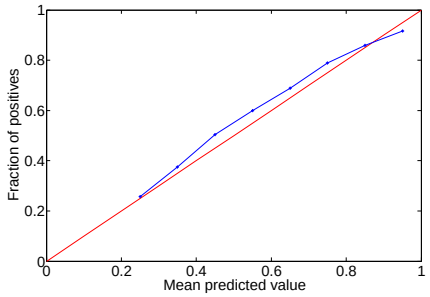



**Figure 5.3.** Reliability Diagram of probabilities obtained using EM algorithm for group size 3

$$Pr(t_m.S = s_i) = \sum_{\{\pi:\pi_m = s_i\}} Pr(\pi)$$

In this section, we evaluate the probabilities of assignments obtained in the Naive-Bayes experiments.

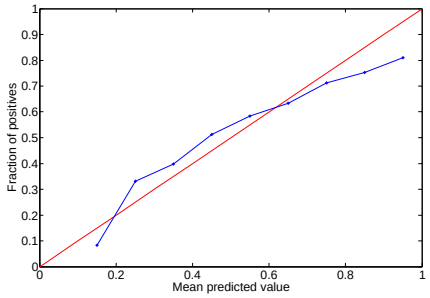



**Figure 5.4.** Reliability Diagram of probabilities obtained using EM algorithm for group size 4

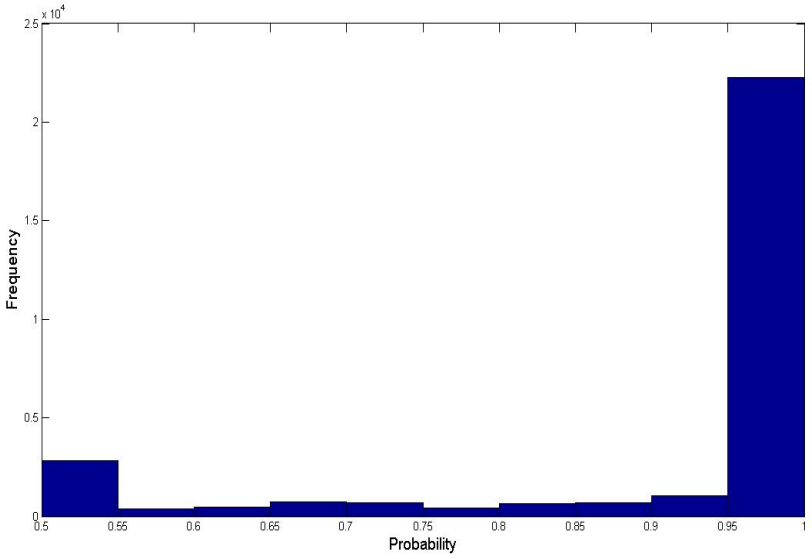



**Figure 5.5.** Histogram of assignment probabilities for group size 2

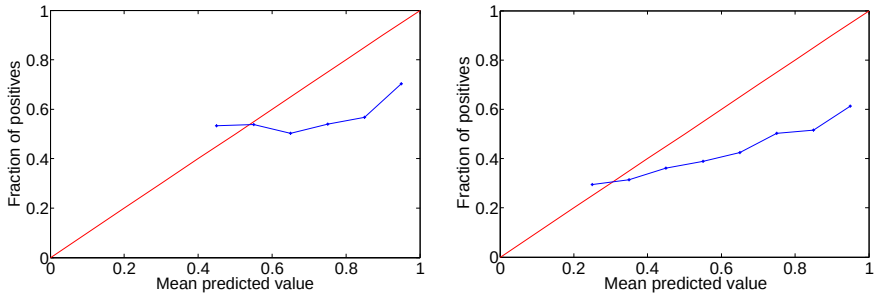
### 5.2.1 Naive Bayes Classifier using EM algorithm

Figures 5.2, 5.3 and 5.4 show the reliability diagrams of the probabilities of sensitive value assignments obtained for group sizes 2, 3 and 4 respectively. We observe that the plots are very close to the perfect calibration line which is indicated in red. This shows that the probabilities are very well calibrated and there is not really a need or calibration in this case. As we will see in the next section, the probabilities of assignments for the algorithm which does not take into consideration the permutations are not well calibrated. This shows that the effect of the Naive Bayes model, which is to push the probabilities of predictions towards 0 or 1 is negated by considering the permutations of sensitive values in each group.

### 5.2.2 Naive Bayes Classifier using Approximate EM algorithm

Figure 5.5 shows the histograms of the assignment probabilities for group sizes 2. We observe that there is a high percentage of probabilities closer to 1. The same is true for the assignment probabilities in group size 3 and 4 too. This is expected due to the Naive-Bayes structure of the classifier.

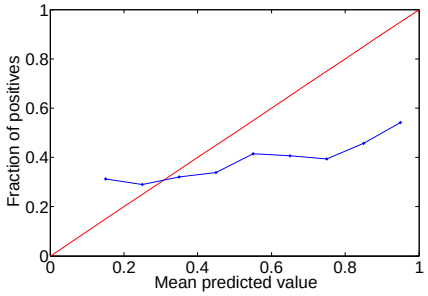
Figure 5.2.2 show the reliability diagrams of the assignment probabilities for group sizes 2, 3 and 4 respectively. The ideal case of the reliability diagram is shown in red in each of the figures. The Brier scores can be evaluated only in comparison. The comparison is shown in the next section.




(a) Group Size 2




(b) Group Size 3




(c) Group Size 4

**Figure 5.6.** Reliability Diagrams of Uncalibrated Probabilities

## 5.3 Calibration

The reliability diagrams of the predicted probabilities showed the need for calibration. The Naive-Bayes formulation of classification will always give overly optimistic prediction probabilities, i.e. the prediction probabilities will be pushed towards 1. From the reliability diagrams, we observe that a calibration method is needed which reduces the Naive-Bayes probabilities. The calibration should also take care that the prediction probabilities should still be a minimum of $\frac{1}{g!}$, since that is the baseline. One such calibration method is the average of the normalized individual prediction probabilities. Following the same notation used in the previous sections the adjusted probability of the group assignment is given by

$$\boxed{Pr(\pi) = \frac{1}{g!} \sum_{m=1}^{g} \left( \frac{\gamma_{m\pi_m}}{\sum_{q=1}^{g} \gamma_{q\pi_q}} \right)}$$

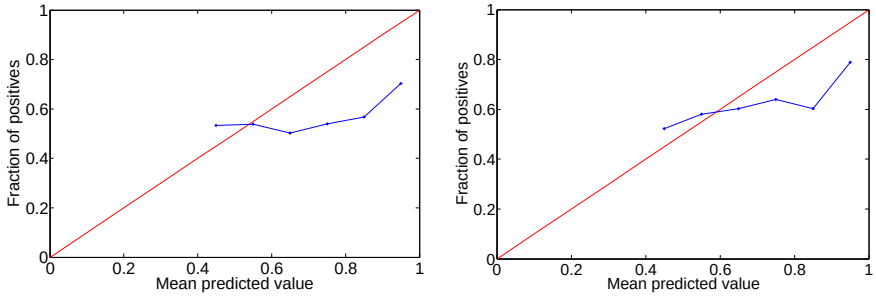
Using this formula for the probability of an assignment, the probability of a tuple assignment is again calculated by adding the probabilities of all the assignments that contain that tuple assignment.

$$Pr(t_m.S = s_i) = \sum_{\{\pi : \pi_m = s_i\}} Pr(\pi)$$

| Group Size | Measurement | BS | CAL | REF |
|---|---|---|---|---|
| 2 | $PP$ | 0.2878 | 0.0536 | 0.2177 |
| | $PP^{new}$ | 0.2269 | 0.0150 | 0.2119 |
| 3 | $PP$ | 0.2973 | 0.0572 | 0.2375 |
| | $PP^{new}$ | 0.2460 | 0.0032 | 0.2428 |
| 4 | $PP$ | 0.2683 | 0.0401 | 0.2282 |
| | $PP^{new}$ | 0.2382 | 0.0076 | 0.2306 |

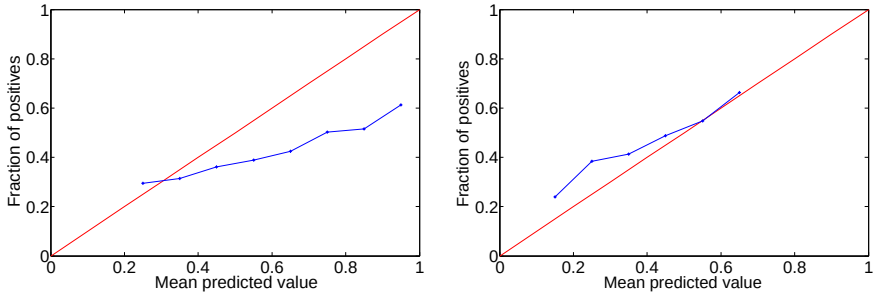
**Table 5.1.** Brier Score Evaluation

Figures 5.3, 5.3 and 5.3 show the reliability diagrams for group sizes 2, 3 and 4 of original and adjusted probabilities. Note that the curves of the adjusted probabilities are closer to the diagonal indicating that the probabilities are better calibrated.




(a) Original (b) Adjusted

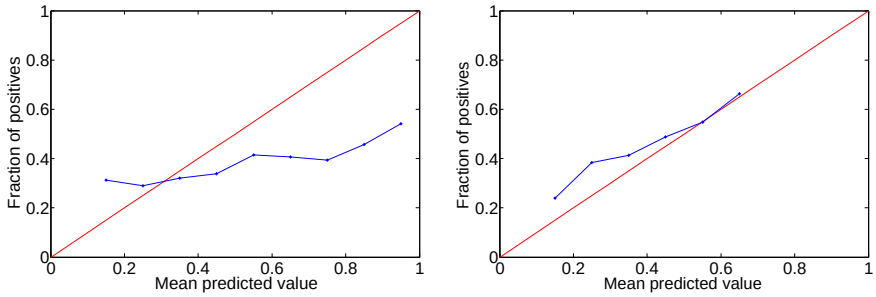
**Figure 5.7.** Reliability Diagrams for Group Size 2




(a) Original (b) Adjusted

**Figure 5.8.** Reliability Diagrams for Group Size 3

Table 5.1 shows the comparison of the brier score and its components before and after calibration for different group sizes. We observe that the Brier score of the probabilities after calibration is consistently lesser than that of the probabilities before calibration. The component values CAL(calibration) and REF(reference) show that the component REF contributes heavily to the Brier score. We also observe that the values of REF in all three cases is close 0.25.




(a) Original (b) Adjusted

**Figure 5.9.** Reliability Diagrams for Group Size 4

However, it has to be remembered that the baseline for REF is 0.1675 since we are binning the probabilities into 10 equal sized groups.

## 5.4 Conclusion

A Naive Bayes classifier was used to exploit the correlations between the sensitive and non-sensitive attributes in anatomized tables and attack the privacy of individuals in the data set. The parameters of the Naive Bayes classifier were trained using two different kinds of EM algorithms, the actual EM algorithm in which the permutations of sensitive values in each group were considered and the approximate EM algorithm in which the permutations of sensitive values in each group were not considered. The results showed that the Naive Bayes classifier using the actual EM algorithm performed better for all group sizes compared to the classifier using approximate EM algorithm and better for group size 4 compared to the deFinetti attack presented in the previous work.

The evaluation of the probabilities of assignment of sensitive values showed that the classifier using the actual EM algorithm gave very well calibrated probabilities. The classifier using the approximate EM algorithm did not give good probabilities of assignments and needed a calibration method. A suitable calibration method was proposed and it was shown that the probabilities obtained after calibration were better.

This concludes the first part of this thesis.

Chapter 6

# Differential Privacy and James-Stein Estimation

Consider a trusted party that holds a dataset of sensitive information (e.g. medical records, voter registration, email usage) with the goal of providing global statistical information about the data publicly available, while preserving the privacy of the users whose information the data set contains. Such a system is called a statistical database. A statistical database provides information in the form of queries. Differential privacy[10] aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. One way of answering queries under differential privacy, is by using the Laplace mechanism [37] which achieves differential privacy by adding independent noise to the query answers where the noise is sampled from a Laplace distribution.

In this thesis, we restrict the type of queries to simple count queries, which are counts of individuals present in the database that satisfy certain predicates. Hence the maintainer of the database gives a vector of count query answers. Since the noise is added independently to each query answer, a person who needs the query answers would just consider the noisy query answers to be the true answers. In such a case, the maximum likelihood estimate of the query answer vector is being considered. In 1961, James and Stein[38] showed that there exist estimates of a multivariate Gaussian mean, which have lower squared error compared to the maximum likelihood estimate. In the following chapters, we show that using an approach similar to James and Stein's approach, we can develop an estimate for the query answers which is better than the maximum likelihood estimate. We then use such an estimate to build a differentially private classifier.

In this chapter, we introduce the concepts of differential privacy and the James-Stein estimation.

## 6.1 $\epsilon$-Differential Privacy

Differential privacy aims to provide means to maximize the accuracy of queries from statistical databases while minimizing the chances of identifying its records. Dwork et al.[10] introduced the concept of $\epsilon$ differential privacy, which ensures that the removal or addition of a single record in a statistical database does not have a substantial impact on the output produced by a differentially private mechanism.For any input database $I$, let $NB(I)$ denote the set of neighbors of $I$, each differing from $I$ in at most one record, i.e. if $I' \in NB(I)$, then $|(I - I') \cup (I' - I)| = 1$. Then,

**Definition** Algorithm $A$ is $\epsilon$-differentially private if for all instances I, any $I' \in NB(I)$, and any subset of outputs $S \subseteq Range(A)$, the following holds:

$$Pr[A(I) \in S] \leq exp(\epsilon)Pr[A(I') \in S]$$

where the probability is taken over the randomness of the A. To answer queries under differential privacy, the Laplace mechanism [37] is used, which achieves differential privacy by adding Laplace noise to the query answers. The magnitude of the noise depends on the query's *sensitivity*. Sensitivity is defined as follows.

**Definition** The sensitivity $S$ of a query $\mathbf{Q} : I \to \mathbb{R}^p$ is given by

$$S_{\mathbf{Q}} = \max_{I, I' \in NB(I)} \|\mathbf{Q}(I) - \mathbf{Q}(I')\|_1$$

where $\|\mathbf{X} - \mathbf{Y}\|_1$ denotes the $L_1$ distance between the vectors $\mathbf{X}$ and $\mathbf{Y}$.

Given a query $\mathbf{Q}$, the Laplace mechanism first computes the query answer $\mathbf{Q}(I)$ and then adds random noise independently to each component of the answer. The noise is drawn from a zero-mean Laplace distribution with a scale defined by the Sensitivity of $\mathbf{Q}$ and the privacy parameter $\epsilon$.

**Proposition 6.1.1.** *Refer [37]. Let* $\mathbf{Q}$ *be a query sequence of length p and let* $\langle Lap(\sigma)\rangle^p$ *denote a vector of p i.i.d. samples from a Laplace distribution with zero mean and scale* $\sigma$*. The randomized algorithm* $\tilde{\mathbf{Q}}$ *that takes as input database I and outputs the following vector is* $\epsilon$*-differentially private:*

$$\tilde{\mathbf{Q}}(I) = \mathbf{Q}(\mathbf{I}) + \langle Lap(\frac{S_{\mathbf{Q}}}{\epsilon})\rangle^p$$

Note that when $\epsilon$ is high, the noise added is less and vice versa. Hence when $\epsilon$ needs to be high, the *utility* or usefulness of the query has to be sacrificed for better privacy by adding more noise to it.

## 6.2 Improving Utility of Count Queries

We have seen that by using a differentially private algorithm, the utility or usefulness of a query is reduced. To improve the utility of a query obtained from a differentially private algorithm, we

have to post-process the obtained query vector in such a way as to make it closer to the original query vector or, in other words, reduce the noise added. This is an estimation problem. To be more specific, it is the problem of estimating the means of independent Laplace distributions, given a sample from each of the distributions. The use of a maximum-likelihood estimator would result in no change in the query vector. In 1961, James and Stein[38] introduced an estimator of a multivariate normal mean for dimension 3 and higher which achieves a lower mean squared error than the maximum-likelihood estimator. We use a similar approach to derive such an estimator for the mean of a multivariate Laplace distribution.

## 6.3 The James-Stein Estimator

Consider the problem of estimating the mean of a multivariate normal distribution from a vector of uncorrelated, equal variance measurements. Specifically, let $\mathbf{x} \in \mathbb{R}^p$ be a measurement drawn from a normal distribution with mean $\theta$ and covariance $\sigma^2\mathbf{I}$. In other words, $\mathbf{x} \sim \mathbf{N}(\theta, \sigma^2\mathbf{I}$. Application of the maximum likelihood criterion gives the maximum likelihood estimate(MLE),

$$\hat{\theta}^{\mathbf{ML}} = \mathbf{x}$$

A common measure of estimator performance is the mean-squared error(MSE), defined as

$$J(\theta; \hat{\theta}) = E\left(\|\hat{\theta} - \theta\|^2\right)$$

For simplicity, let $\mathbf{J}_{ML}(\theta)$ represent the MSE of an MLE.

In 1961, James and Stein [38] introduced an estimator that achieves uniformly lower MSE than the MLE for all parameter values $\theta$. That is, they constructed an estimator $\hat{\theta}$ for which

$$\mathbf{J}_{JS}(\theta) \leq \mathbf{J}_{ML}(\theta) \quad \forall\theta$$

when $p > 2$. The JSE takes the form

$$\hat{\theta}^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\mathbf{x}\|^2}\right)\mathbf{x}$$

A brief overview of the James-Stein estimator and its applications can be found in [39]. More about the James-Stein estimator can be found in [40] [38] [41] and [42].

## 6.4 Heuristic Justifications of the James-Stein Estimator

In this section we present the arguments suggesting the superiority of the JSE that stop short of being a mathematical derivation for the JSE.

### 6.4.1 Stein's original argument

Stein's original argument[38] was based on the comparison of $\|\theta\|^2$ to $\|\mathbf{x}\|^2$ when $p$ is large. Intuitively, a good estimate $\hat{\theta}$ of $\theta$ should satisfy $\hat{\theta}_i \approx \theta_i$. This implies that $\hat{\theta}$ should also satisfy $\hat{\theta_i}^2 \approx \theta_i^2$ for $i = 1, ..., p$, and thus,

$$\|\hat{\theta}\|^2 \approx \|\theta\|^2$$

We would hope that the chosen estimator satisfies this condition. Suppose now, we calculate the expected value of the quantity $\|\mathbf{x}\|^2$ noting that $\hat{\theta}^{\mathbf{ML}} = \mathbf{x}$ and $x_i \sim \mathbf{N}(\theta, \sigma^2\mathbf{I}$. We get,

$$E(\mathbf{x}'\mathbf{x}) = p\sigma^2 + \theta'\theta \tag{6.1}$$

Application of the Chebyshev's inequality to $\frac{\mathbf{x}'\mathbf{x}}{p}$ gives

$$\frac{\mathbf{x}'\mathbf{x}}{p} \to \sigma^2 + \frac{\theta'\theta}{p} \tag{6.2}$$

in probability as $p \to \infty$. In other words, for very large $p$, it is very likely that $\mathbf{x}'\mathbf{x}$ is larger than $\theta'\theta$. This suggests that, to form a good estimate of $\theta$, we would need to shrink $\hat{\theta}^{\mathbf{ML}} = \mathbf{x}$ towards $\mathbf{0}$, which is exactly what the JSE does.

A similar argument can be applied to the case where $x_i \sim L(\theta, b)$ to show that the we would need to shrink $\hat{\theta}^{\mathbf{ML}} = \mathbf{x}$ towards $\mathbf{0}$ for the Laplace case too.

### 6.4.2 An Empirical Bayesian Argument

Another argument for the JSE is based on the Bayesian formulation. This argument is presented in [43] and [39]. In the estimation framework up to this point, we assumed that $\theta$ is a nonrandom parameter. Now, Let us now model $\theta$ as a random quantity. Specifically suppose that we model

$$\theta \sim \mathbf{N}(\mathbf{0}, \tau^2\mathbf{I})$$

where $\tau^2$ is assumed to be known. With the introduction of a prior density, $\theta$ could be estimated in a Bayesian setting. The Bayes least squares estimate BLSE of $\theta$ is given by

$$\hat{\theta}^{BLS} = \frac{\tau^2}{\sigma^2 + \tau^2}\mathbf{x} = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)\mathbf{x} \tag{6.3}$$

The right hand side of the second equality is intended to evoke the form of the JSE. The BLSE is used here since it achieves the smallest expected MSE of any estimator of $\theta$ by construction.

The quantity $\tau^2$ is still unknown. From the given data, we can get an empirical estimate of the quantity $\tau^2$ first and then use it to produce an empirical BLSE of $\theta$. Such an approach in which a prior density does not exist or is unknown but is estimated from the data in order to apply the Bayesian framework is known as empirical Bayesian estimation. The form of 6.3 suggests that we could forgo estimation of $\tau^2$ and instead estimate $\frac{\sigma^2}{\sigma^2+\tau^2}$ directly. It can be

shown[43] that $\frac{\sigma^2(p-2)}{\|\mathbf{x}\|^2}$ is an unbiased estimator of $\frac{\sigma^2}{\sigma^2+\tau^2}$. Substituting this in equation 6.3, we have,

$$\hat{\theta}^{BLS} = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)\mathbf{x} = \left(1 - \frac{\sigma^2(p-2)}{\|\mathbf{x}\|^2}\right)\mathbf{x} \tag{6.4}$$

Thus one interpretation of the JSE is that it is an attempt to perform Bayes estimation when no prior is present but one of the form of is assumed. A similar approach is used to derive an empirical Bayes' estimator for the mean of a multivariate Laplacian distribution and is presented in Chapter-7.

Chapter 7

# Estimation of Mean of a Multivariate Laplacian

There has been previous work on developing a James-Stein estimator for distributions other than the Gaussian[44, 45, 46, 47, 48]. However, most of the previous work was aimed at distributions that belong to the exponential family. Unfortunately, none of the previous estimators can be used for estimating the mean of a multivariate Laplacian distribution since the Laplacian distribution does not belong to the exponential family. In this chapter, we focus on deriving and evaluating an estimator for estimating the mean of a multivariate Laplacian.

It has been shown in Chapter-6 that the James-Stein estimator can be derived in an empirical Bayes' formulation of estimation, where the mean of the multivariate normal is assumed to have a zero mean Gaussian distribution. In this chapter, we try to derive an estimator for the mean of a multivariate Laplacian distribution using a similar approach.

## 7.1 An Empirical Bayesian Framework

The main idea behind an empirical Bayesian framework is that we treat the mean of the multivariate Laplacian as an unknown quantity with a prior distribution.

Let $\bar{z} \in \mathbb{R}^p$ be a random variable having a multivariate laplace distribution with mean vector $\theta$, i.e. $x_i \sim \mathbf{L}(\theta_i, \frac{1}{\lambda})$. The probability density function of the vector $\mathbf{x}$ given $\theta$ is given by

$$P(\mathbf{x}|\theta) = \frac{\lambda^p}{2^p} e^{-\lambda \|\mathbf{x}-\theta\|_1} \tag{7.1}$$

Let the $\theta$ have a multivariate laplace distribution with zero mean and scale $\frac{1}{\alpha}$.

$$P(\theta) = \frac{\alpha^p}{2^p} e^{-\alpha \|\theta\|_1} \tag{7.2}$$

Using 7.1 and 7.2, the joint distribution of $\mathbf{x}$ and $\theta$ will be as follows.

$$\begin{aligned} P(\mathbf{x},\theta) &= P(\mathbf{x}|\theta)\ P(\theta) \\ &= \frac{\lambda^p}{2^p} e^{-\lambda\|\mathbf{x}-\theta\|_1}\ \frac{\alpha^p}{2^p} e^{-\alpha\|\theta\|_1} \\ &= \frac{\lambda^p\alpha^p}{4^p} e^{-\lambda\|\mathbf{x}-\theta\|_1-\alpha\|\theta\|_1} \end{aligned}$$

$$\boxed{P(\mathbf{x},\theta) = \frac{\lambda^p\alpha^p}{2^{2p}} e^{-\lambda\|\mathbf{x}-\theta\|_1-\alpha\|\theta\|_1}} \tag{7.3}$$

The marginal distribution of $\mathbf{x}$ is then given by,

$$\begin{aligned} P(\mathbf{x}) &= \int_{-\infty}^{\infty}\cdots\int p(\mathbf{x},\theta)\,\mathrm{d}\theta_1...\mathrm{d}\theta_p \\ &= \frac{\lambda^p\alpha^p}{2^{2p}}\prod_i\left(\int_{-\infty}^{0} e^{-\lambda x_i+(\lambda+\alpha)\theta_i}\mathrm{d}\theta_i + \int_0^{x_i} e^{-\lambda x_i+(\lambda-\alpha)\theta_i}\mathrm{d}\theta_i + \int_{x_i}^{\infty} e^{\lambda x_i-(\lambda+\alpha)\theta_i}\mathrm{d}\theta_i\right) \\ &= \frac{\lambda^p\alpha^p}{2^{2p}}\prod_i\left(\frac{e^{-\lambda x_i}}{\lambda+\alpha} + \frac{e^{-\alpha x_i}-e^{-\lambda x_i}}{\lambda-\alpha} + \frac{e^{-\alpha x_i}}{\lambda+\alpha}\right) \\ &= \frac{\lambda^p\alpha^p}{2^p(\lambda^2-\alpha^2)^p}\prod_i\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right) \end{aligned}$$

$$\boxed{P(\mathbf{x}) = \frac{\lambda^p\alpha^p}{2^p(\lambda^2-\alpha^2)^p}\prod_i\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)} \tag{7.4}$$

From (7.3) and (7.4), the posterior distribution of $\theta$ is given by

$$\begin{aligned} P(\theta|\mathbf{x}) &= \frac{p(\mathbf{x},\theta)}{p(\mathbf{x} \\ &= \frac{(\lambda^2-\alpha^2)^p}{2^p\prod_i\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)} e^{-\lambda\|\mathbf{x}-\theta\|_1-\alpha\|\theta\|_1} \qquad (7.5) \\ & \qquad (7.6) \end{aligned}$$

The expected value of $\theta$ with respect to the posterior distribution is given by,

$$\begin{aligned} E_{\theta_i|\mathbf{x}}[\theta_i] &= \frac{(\lambda^2-\alpha^2)}{2\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\int_{-\infty}^{\infty}\theta e^{-\lambda|x_i-\theta_i|-\alpha|\theta_i|}d\theta_i \\ &= \frac{(\lambda^2-\alpha^2)}{2\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\left(\int_{-\infty}^{0}\theta e^{(\lambda+\alpha)\theta_i-\lambda x_i}d\theta_i + \int_0^{x_i}\theta e^{(\lambda-\alpha)\theta_i-\lambda x_i}d\theta_i\right. \\ &\quad \left. + \int_{x_i}^{\infty}\theta e^{-(\lambda+\alpha)\theta_i+\lambda x_i}d\theta_i\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{(\lambda^2-\alpha^2)}{2\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\left[\left(-\frac{e^{-\lambda x_i}}{(\lambda+\alpha)^2}\right)+\left(\frac{x_i e^{-\alpha x_i}}{\lambda-\alpha}-\frac{e^{-\alpha x_i}}{(\lambda-\alpha)^2}+\frac{e^{-\lambda x_i}}{(\lambda-\alpha)^2}\right)\right.\\
&\quad\left.+\left(\frac{x_i e^{-\alpha x_i}}{\lambda+\alpha}+\frac{e^{-\alpha x_i}}{(\lambda+\alpha)^2}\right)\right]\\
&= \frac{(\lambda^2-\alpha^2)}{2\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\left[\frac{2\lambda x_i}{\lambda^2-\alpha^2}e^{-\alpha x_i}+\frac{4\alpha\lambda}{(\lambda^2-\alpha^2)^2}\left(e^{-\lambda x_i}-e^{-\alpha x_i}\right)\right]\\
&= \frac{\lambda}{\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\left[x_i e^{-\alpha x_i}+\frac{2\alpha}{(\lambda^2-\alpha^2)}\left(e^{-\lambda x_i}-e^{-\alpha x_i}\right)\right]
\end{aligned}
$$

$$
\boxed{E_{\theta_i|\mathbf{x}}[\theta_i]=\frac{\lambda}{\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}\left[x_i e^{-\alpha x_i}+\frac{2\alpha}{(\lambda^2-\alpha^2)}\left(e^{-\lambda x_i}-e^{-\alpha x_i}\right)\right]} \tag{7.7}
$$

Equation 7.7 can also be written as follows.

$$
\boxed{E_{\theta_i|\mathbf{x}}[\theta_i]=\frac{\lambda}{\left(\lambda-\alpha e^{(\alpha-\lambda)x_i}\right)}\left[x_i-\frac{2\alpha}{(\alpha^2-\lambda^2)}\left(e^{(\alpha-\lambda)x_i}-1\right)\right]} \tag{7.8}
$$

The unknown quantity in the above estimation formula is $\alpha$. $\alpha$ is to be estimated empirically from the given data $\mathbf{x}$. This is done by finding the maximum likelihood or maximum log-likelihood estimate of $\alpha$ with respect to the distribution of $\mathbf{x}$. From (7.4), we have,

$$
P(\mathbf{x})=\frac{\lambda^p\alpha^p}{2^p(\lambda^2-\alpha^2)^p}\prod_i\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)
$$

$$
\Rightarrow log(P(\mathbf{x}))=\log\frac{\lambda^p}{2^p}+p\log\alpha-p\log(\lambda^2-\alpha^2)+\sum_i\log\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right) \tag{7.9}
$$

Differentiating 7.9 with respect to $\alpha$ and equating to zero, we get,

$$
\frac{p}{\alpha}+\frac{2p\alpha}{\lambda^2-\alpha^2}-\sum_i\frac{\lambda x_i e^{-\alpha x_i}-e^{-\lambda x_i}}{\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}=0
$$

$$
\boxed{\Rightarrow\frac{p(\lambda^2+\alpha^2)}{\alpha(\lambda^2-\alpha^2)}=\sum_i\frac{\lambda x_i e^{-\alpha x_i}-e^{-\lambda x_i}}{\left(\lambda e^{-\alpha x_i}-\alpha e^{-\lambda x_i}\right)}} \tag{7.10}
$$

Assume a diffuse prior on the mean $\theta_i$ or, in other words, let $\alpha << \lambda$. Using this assumption, (7.10) becomes

$$
\begin{aligned}
\Rightarrow\frac{\cancel{p(\lambda^2+\alpha^2)}}{\cancel{\alpha(\lambda^2-\alpha^2)}} &= \sum_i\frac{\cancel{\lambda}x_i\cancel{e^{-\alpha x_i}}}{\cancel{\lambda e^{-\alpha x_i}}}\\
\Rightarrow\alpha &= \frac{p}{\sum_i x_i}=\frac{p}{|\mathbf{x}|}
\end{aligned} \tag{7.11}
$$

Using the assumption $\alpha << \lambda$, (7.8) becomes

$$\begin{aligned}\hat{\theta}_i &= \frac{\lambda}{\cancel{\left(\lambda + \alpha e^{(\alpha-\lambda)x_i}\right)}^{1}}\left[x_i - \frac{2\alpha}{\lambda^2}\left(1 - \cancel{e^{(\alpha-\lambda)x_i}}^{0}\right)\right] \\ &= x_i - \frac{2\alpha}{\lambda^2}\end{aligned}$$

Using the maximum likelihood estimate of $\alpha$ from equation 7.11 in the above equation, we get,

$$\hat{\theta}_i = x_i - \frac{2p}{\lambda^2|\mathbf{x}|} \tag{7.12}$$

Note that the estimation formula given by equation 7.12 does not impose any intrinsic restriction on $p$ like in the case of the Normal distribution. For simplicity, $\frac{1}{\lambda}$ which is the scale of the Laplacian distributions, is replaced by $\sigma$, thus giving the following estimator.

$$\boxed{\hat{\theta}_i = x_i - \frac{2p\sigma^2}{|\mathbf{x}|}} \tag{7.13}$$

For simplicity, we will refer to this estimator as the Empirical Bayes Estimator(EBE) from now on.

## 7.2 Evaluation of Estimators

In this section, we evaluate the performance of two candidate estimators in comparison with the MLE experimentally. The estimators considered are listed as follows.

1. The original James-Stein estimator(JS)

$$\hat{\theta}^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\mathbf{x}\|^2}\right)\mathbf{x}$$

2. The Empirical Bayes'(EB) estimator derived in the previous section(EB).

$$\hat{\theta}^{EB} = \mathbf{x} - \frac{2p\sigma^2}{|\mathbf{x}|}\mathbf{I}$$

The evaluation is done by trying to estimate the mean of a multivariate Laplacian given a p-dimensional sample point. The sample point was generated randomly. Since the EBE was derived with the assumption $\alpha << \lambda$, the random generation of the p-dimensional sample point had to be done in such a way that the inequality given by the assumption is satisfied for all dimensions. One idea was to generate the p-dimensional sample point from a uniform distribution over a wide range. But this does not ensure high variance for smaller dimensions. To overcome this problem, a random number 'r'$(0 < r < p)$ of points were generated using a uniform distribution

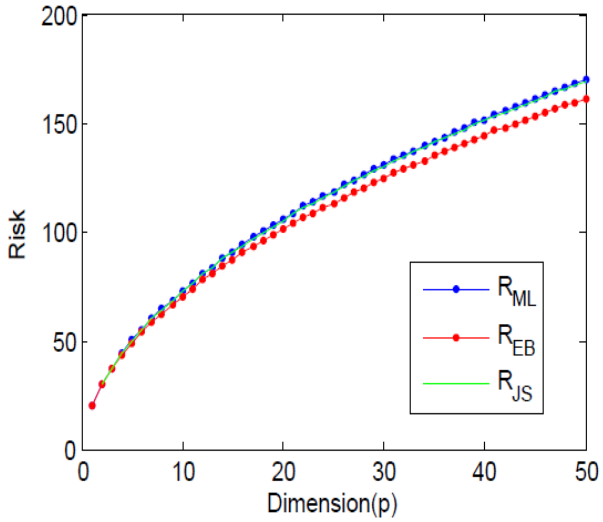



**Figure 7.1.** Risk vs Dimension plot for different estimators

$U(0, N_1)$ and the rest of the points generated from the distribution $U(0, N_2)$, where $N_1$ is very small compared to $N_2$. This ensures that the variance of the generated multidimensional point is high for any dimension.

### 7.2.1 Dependence on the dimension of the data

To evaluate the dependence of dimension of the data on the performance of the estimators, the following experimental setup is used. First, a p-dimensional random vector is generated according to the method described in the previous section. Laplace noise is then added to each component of the generated vector independently. Then the original vector is estimated using the two estimators. The squared norm of the difference of the estimated vectors and the original vector was then used to evaluate the performance. The average risk for 50,000 initializations of the p-dimensional vector is noted. Figure 7.1 shows the plot of the average risk of different estimators versus the dimension of the randomly generated vector. It is observed that the EBE outperforms the MLE for dimensions 3 and higher when the assumption $\alpha << \lambda$ holds i.e. the variance of the p-dimensional vector is large compared to the variance of the Laplace noise added to each of its components. The figures show that of the three estimators MLE, EBE and JSE, the EBE has the best performance. The risk of JSE looks to almost coincide with the risk of MLE. To have a better idea, we plot the difference of risks of the estimators against the dimension of the data.

Figure 7.2 shows a plot of the difference in risks of the estimators versus the dimension of the data. From the figure, it is clear that the risk of the EBE is lesser than that of MLE for dimensions 3 and higher. As the dimension increases, it is observed that the difference between the risk of the EBE and that of the MLE increases, implying that the improvement is more for higher dimensions. It is also observed that the difference between the risks of the JSE and the

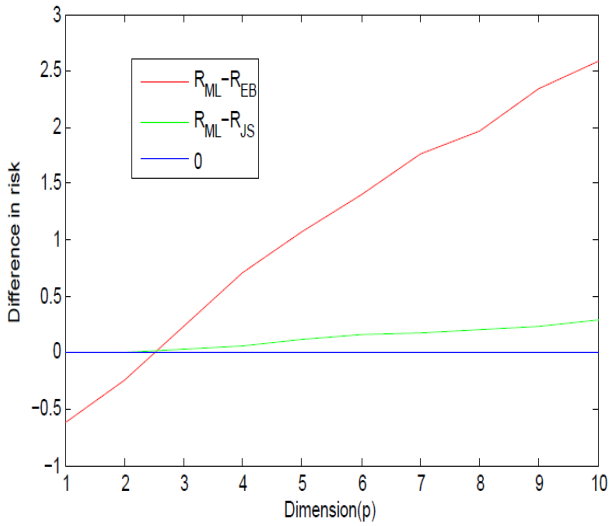



**Figure 7.2.** Difference in risk vs Dimension

MLE also increases with dimension.

Chapter 8

# Applications of the Laplace James-Stein Estimator

In the previous chapter, we derived and evaluated estimators for the mean of a multivariate Laplacian distribution. In this chapter, we will use the estimators to improve the performance of a differentially private Naive-Bayes classifier.

Consider an $\epsilon$ differential private database. The data is made publicly available through a statistical database on which only aggregate queries are permitted. The task is to use this query information to construct a Naive-Bayes classifier. This would involve extracting the required Naive-Bayes parameters. Since, under the setting of $\epsilon$ differential privacy, Laplace noise is added to each individual query answer, the utility of the classifier constructed using these query answers is also affected. In the previous chapter, we derived estimators that can reduce such noise. In this chapter, we try to see the practical use of such estimators by using them on the query answers provided and then using the result to construct the classifier.

## 8.1 Description

In this section, we first formulate a simple Naive-Bayes classifier. Let $T$ represent the tuples present in a dataset. Let it consist of the features $f_j$ and the class label $C$. Let $f_{jk}$ represent the $k^{th}$ value that the feature $f_j$ can take. Let $c_i$, $i = 1, 2, ..|C|$ represent the class labels. The Naive Bayes parameters are

$$\pi_i = P(c_i)$$

$$w_{ijk} = P(f_j = f_{jk}|C = c_i)$$

where $c_i$ represents the class $i$ and $f_{jk}$ is the $k^{th}$ distinct value of the $j^{th}$ attribute.

The likelihood of the data given the parameters is given by,

$$\begin{aligned} P(T|W) &= \prod_{n=1}^{N}\prod_{i=1}^{|C|}\left(P(c_i)\prod_{j=1}^{|f_j|}\prod_{k=1}^{|f_{jk}|}P\left(f_j=f_{jk}|C=c_i\right)^{\gamma_{njk}}\right)^{z_{ni}} \\ &= \prod_{n=1}^{N}\prod_{i=1}^{|C|}\left(\pi_i\prod_{j=1}^{|f_j|}\prod_{k=1}^{|f_{jk}|}{w_{ijk}}^{\gamma_{njk}}\right)^{z_{ni}} \end{aligned} \tag{8.1}$$

where $z_{ni}$ is an indicator for the $n^{th}$ tuple belonging to class $c_i$ and $\gamma_{njk}$ is an indicator to the $j^{th}$ feature of tuple $n$ being $f_{jk}$. That is,

$$z_{ni} = \begin{cases} 0 & \text{if } t_n.C \neq c_i; \\ 1 & \text{if } t_n.C = c_i; \end{cases}$$

$$\gamma_{njk} = \begin{cases} 0 & \text{if } t_n.f_j \neq f_{jk}; \\ 1 & \text{if } t_n.f_j = f_{jk}; \end{cases}$$

The log-likelihood is then given by

$$\log(P(T,Z|W)) = \sum_{n=1}^{N} z_{ni}\left(\sum_{i=1}^{|C|}\log\pi_i + \sum_{j=1}^{|f_j|}\sum_{k=1}^{|f_{jk}|}\gamma_{njk}\log w_{ijk}\right)$$

The constraints on the parameters are as follows.

1. Each tuple belongs to at least one of the classes.

$$\sum_{i=1}^{|C|}\pi_i - 1 = 0$$

2. Given a tuple belongs to class $c_i$, its $j^{th}$ feature should take one of the possible values.

$$\sum_{k=1}^{|f_{jk}|} w_{ijk} - 1 = 0$$

Therefore, the Lagrange is given by,

$$\begin{aligned} L &= \log(P(T,Z|W)) + \alpha\left(\sum_{i=1}^{|C|}\pi_i - 1\right) + \beta\left(\sum_{k=1}^{|f_{jk}|} w_{ijk} - 1\right) \\ &= \sum_{n=1}^{N} z_{ni}\left(\sum_{i=1}^{|C|}\log\pi_i + \sum_{j=1}^{|f_j|}\sum_{k=1}^{|f_{jk}|}\gamma_{njk}\log w_{ijk}\right) + \alpha\left(\sum_{i=1}^{|C|}\pi_i - 1\right) + \beta\left(\sum_{k=1}^{|f_{jk}|} w_{ijk} - 1\right) \end{aligned}$$

The maximum likelihood estimates of the parameters are obtained by maximizing the lagrange with respect to the parameters. The resulting parameter expressions are as follows.

$$\pi_i = \frac{\sum_{n=1}^{N} z_{ni}}{\sum_{i=1}^{|C|}\sum_{n=1}^{N} z_{ni}} = \frac{n_i}{N} \tag{8.2}$$

$$w_{ijk} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{|C|} z_{ni}\gamma_{njk}}{\sum_{k=1}|f_j|\sum_{n=1}^{N}\sum_{i=1}^{|C|} z_{ni}\gamma_{njk}} = \frac{n_{jk}}{n_j} \tag{8.3}$$

## 8.2 Count Queries

The queries needed to form the above classifier are the histogram or count queries of the data set. Specifically, for each feature $f_j$, the counts of the number of tuples that take the value $f_{jk}$ and have the class value $c_i$. Let us represent this as $c_{ijk}$. Thus the histogram query vector is formed as follows

$$Q(I) = \left[\left(c_{111}, ...c_{11|f_1|}\right), ..., \left(c_{1|f|1}, ...c_{1|f||f_{|f|}}\right), ..., \left(c_{|C|1}, ...c_{|C|1|f_1|}\right), ..., \left(c_{|C||f|1}, ...c_{|C||f||f_{|f|}|}\right)\right]$$

By using the Laplace mechanism, we add Laplacian noise to each component of the above query vector. A differentially private query vector $\tilde{Q}(I)$ is thus obtained.

$$\tilde{Q}(I) = Q(I) + \langle Lap(\frac{S_{\mathbf{Q}}}{\epsilon})\rangle^p \tag{8.4}$$

## 8.3 Sensitivity Calculation

The sensitivity $S_Q$ of a query $Q$, is defined by

$$S_{\mathbf{Q}} = \max_I, I' \in NB(I)\|\mathbf{Q}(I) - \mathbf{Q}(I')\|_1$$

If a single record is removed(/added) from(/to) the database, one component of every feature's histogram is reduced(/increased) by 1. Thus, the sensitivity, in the present case is equal to the number of features.

$$S_{\mathbf{Q}} = |f| \tag{8.5}$$

## 8.4 Stein Correction

Stein correction is applied to the histogram query vector obtained after Laplace noise is added(Laplace mechanism). The estimators described in the previous section are used to obtain the estimates of $\hat{Q}_{JS}(I)$ and $\hat{Q}_{EB}(I)$ of $Q(I)$. The estimates are then evaluated by the result of the classification accuracy that each of them give.

## 8.5 Algorithm

Table 8.5 shows the pseudo code of the algorithm.

| | |
|---|---|
| 1: | **for** $i = 1$ to num_outer_iterations **do** |
| 2: | Form the training and the test sets from the data set. |
| 3: | Form the query vector Q(=$Q_{ML}$) from the training set (concatenate the histograms of each feature given each class). |
| 4: | **for** $j = 1$ to num_inner_iterations **do** |
| 5: | $\tilde{Q} \leftarrow$ **LaplaceMechanism**$(Q, S_Q, \epsilon)$ |
| 6: | $\hat{Q}_{EB} \leftarrow EBE(\tilde{Q})$ <br> $\hat{Q}_{JS} \leftarrow JSE(\tilde{Q})$ |
| 7: | If any of the components of $\hat{Q}_{EB}$ or $\hat{Q}_JS$ are negative, make them 0. |
| 8: | Extract $\pi_i$ and $w_{ijk}$ from each query answer $\hat{Q}$, $\hat{Q}_{EB}$ and $\hat{Q}_{JS}$. <br> $\pi_i = \frac{n_i}{N} = \frac{\#\text{ of tuples in class } c_i}{\text{total }\#\text{ of tuples}}$ <br> $w_{ijk} = \frac{n_{ijk}}{n_i} = \frac{\#\text{ of tuples with } f_j{=}f_{jk} \text{ and class } c_i}{\#\text{ of tuples in class } c_i}$ |
| 9: | Compute the class probabilities for each tuple. <br> $P(t_n.C = c_i) = \pi_i \prod_j w_{ijk}^{\gamma_{njk}}$ |
| 9: | Assign class to each tuple. <br> $t_n.C = \arg\max_i P(t_n.C = c_i)$ |
| 10: | **end for** |
| 11: | **end for** |
| | $LaplaceMechanism(Q, S_Q, \epsilon)$ |
| 1: | **for** $i = 1$ to length_of_Q **do** |
| 2: | $\tilde{Q}(i) = Q(i) + Lap\left(\frac{S_Q}{\epsilon}\right)$[1] |
| 3: | if $\tilde{Q}(i) < 0$ then $\tilde{Q}(i) = 0$ |
| 4: | **end for** |

**Table 8.1.** Algorithm for Differentially Private Naive-Bayes Classifier

## 8.6 Experiments

The experiments are performed using 5,10 and 20-fold cross-validation techniques. Each of the cross validations was done 20 times(Step 1) and in each cycle, 500 iterations of the inner loop(Step 4) were performed. The algorithm was tested on two datasets from the UCI machine learning repository[32] namely the Adult dataset and the Nursery dataset.

The Nursery dataset 12960 tuples and 8 attributes. The attribute *health* was required to be predicted.

The accuracies of the algorithm obtained for three different estimators was recorded. The first estimator is the maximum likelihood estimator, in which case no correction is applied to the differentially private count query. The other two estimators are the Laplace James-Stein estimator(EB) and the original James-Stein estimator(JS). The experiments were run on MATLAB.

[1] $Lap\left(\frac{S_Q}{\epsilon}\right)$ denotes a sample from a Laplace distribution with zero mean and scale $\frac{S_Q}{\epsilon}$

Since the experiments are not computationally expensive, the details of the machine configuration and time taken are omitted.

The code was implemented in MATLAB and run on LION-XC PC clusters of The Pennsylvania State University. The LION-XC cluster consists of 128 compute nodes with Dual 3.0 GHz Intel Xeon 3160 (Woodcrest) Dual-Core Processors out of which 64 have 8GB ECC RAM and 64 have 16GB ECC RAM. Only one dedicated processor on one node was used to run the code. Ultimately, the code involved minimal computational complexity.

### 8.6.1 Choice of $\epsilon$

The privacy parameter $\epsilon$ defines the amount of noise being added to the query answer. The amount of noise added increases as the value of $\epsilon$ decreases. However there exists a lower limit to the amount of noise that can be added to any query answer and thus on the value of $\epsilon$. If the noise added is too much, the result of adding noise to the query answer may distort it to an extent of not having any relation to the original query answer. This limit on the amount of noise that can be added is defined by the query answer itself which inturn depends on the data set. For example, if the number of records in the data is very large, the histogram query vector will be composed of large numbers. This will allow addition of a lot of noise while maintaining the similarity to the original query answer. The values of $\epsilon$ for a data set are chosen by examining the histogram query answers for that data set.

## 8.7 Results

### 8.7.1 Results for the Adult Dataset

The adult dataset has 16 different attributes. The continuous attributes were removed and the tuples with missing values were deleted. This resulted in 30162 tuples. The attributes work-class, education, marital-status, occupation, relationship, race, sex, native country and salary were retained. The salary attribute was required to be predicted. The eight non-sensitive attribute constitute a total of 102 values. The sensitive attribute salary has 2 different values. Hence, the query vector has the size $102 * 2 = 204$. The scale of the Query vector is approximately 2500. Hence, the scale of the Laplace noise was varied between 8-800, or, the $\epsilon$ value was varied between $0.01 - 1$. Table8.2 shows the accuracies of the three kinds of estimators for 5-fold, 10-fold and 20-fold cross validation. We observe that the accuracies of the algorithms with Stein correction ($A_{EB}$ and $A_{JS}$) are consistently better than or as good as the accuracies of the algorithm without Stein correction. We also observe that the $EBE$ performs the best among the three estimators. The improvement in the accuracy given by the EBE increases with decreasing $\epsilon$. This means that the improvement is more when there is more noise. This is expected because the more the noise added to the query vector, the more the scope to improve. This does not hold true when the value of $\epsilon$ is less than a threshold, which in this case is 0.005. This happens because for values of $\epsilon$ lesser than 0.005, the noise added to the query vector is very high and it totally destroys most

of the information in the query vector, thus making it totally unrelated to the original query vector. It has been observed that the accuracies of the EBE and the MLE do not follow any particular trend for values of $\epsilon$ below the threshold. Note that threshold value of $\epsilon$ is different for different datasets and depends on the components of the count query.

| | 5-fold | | | 10-fold | | | 20-fold | | |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ |
| 0.01 | 74.63 | **75.84** | 74.65 | 75.22 | **76.43** | 75.22 | 75.11 | **76.2** | 75.12 |
| 0.02 | 76.87 | **77.44** | 76.87 | 77.04 | **77.51** | 77.05 | 77.59 | **78** | 77.6 |
| 0.05 | 78.30 | **78.36** | 78.30 | 77.93 | **78.23** | 77.95 | 78.35 | **78.72** | 78.36 |
| 0.1 | 78.67 | **78.72** | 78.67 | 78.57 | **78.62** | 78.58 | 79.15 | **79.2** | 79.15 |
| 0.5 | 79.03 | **79.05** | 79.04 | 79.04 | **79.1** | 79.07 | 79.44 | **79.45** | 79.44 |
| 1 | 79.34 | **79.35** | 79.34 | 79.24 | **79.25** | 79.24 | 79.75 | **79.76** | 79.75 |

**Table 8.2.** Comparison of Accuracy of the Naive Bayes classifier for varying $\epsilon$ for the Adult dataset

It is observed that the magnitude of correction introduced by the original James-Stein estimator is very small and hence, the results of the MLE and the JSE do not differ by much. For simplicity, we only compare the MLE and EBE in the rest of the results.

Tables 8.3, 8.4 and 8.5 show the ratio of the number of times the accuracy of the EBE is greater than, equal to, and less than the accuracy of the MLE for 5-fold, 10-fold and 20-fold cross validations respectively. The tables show that the accuracy of the algorithm using the EBE does better than that using the MLE most of the time.

**Table 8.3.** Dominance statistics(%) in 5-fold cross validation

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 69.6 | 0.2 | 30.2 |
| 0.02 | 68.8 | 0.2 | 31.0 |
| 0.05 | 56.6 | 2.2 | 41.2 |
| 0.10 | 70.8 | 4.8 | 24.4 |
| 0.50 | 41.4 | 37.2 | 21.4 |
| 1.00 | 18.6 | 72.2 | 9.2 |

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 71.4 | 0.6 | 28 |
| 0.02 | 66.9 | 1 | 32.1 |
| 0.05 | 61.4 | 2.5 | 36.1 |
| 0.10 | 66.1 | 7.2 | 26.7 |
| 0.50 | 26.4 | 59.1 | 14.5 |
| 1.00 | 10.4 | 85.4 | 4.2 |

**Table 8.4.** Dominance statistics(%) in 10-fold cross validation

Figures 8.1,8.2, 8.3 show the accuracies of EBE and MLE vs $log(\epsilon)$ for 5-fold, 10-fold and 20-fold cross validations respectively. It can be clearly seen that the empirical Bayes' estimator

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 71.2 | 0.95 | 27.85 |
| 0.02 | 63.1 | 2.2 | 34.7 |
| 0.05 | 57.8 | 5.7 | 36.5 |
| 0.10 | 58.7 | 13.5 | 27.8 |
| 0.50 | 18 | 73.7 | 8.3 |
| 1.00 | 5.1 | 92 | 2.9 |

**Table 8.5.** Dominance statistics(%) in 20-fold cross validation

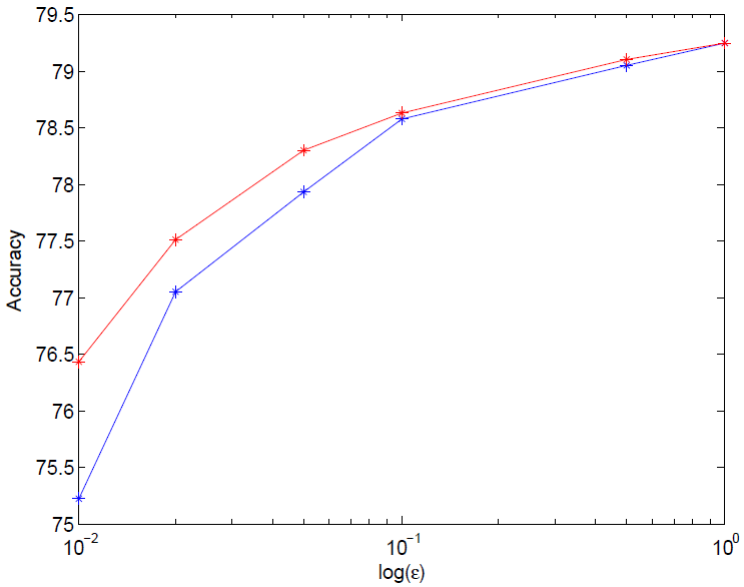



**Figure 8.1.** Accuracy graphs for 5-fold cross validation

performs better than the maximum-likelihood estimator and the difference in accuracies is more for higher $\epsilon$. In other words, more the noise added, the better the empirical Bayes' estimator performs compared to maximum-likelihood estimator.

### 8.7.2 Results for the Nursery Dataset

The nursery dataset has 8 different attributes and 12960 records. The salary attribute was required to be predicted. The eight non-sensitive attribute constitute a total of 27 values. The sensitive attribute 'health' has 5 different values. Hence, the query vector has the size $27*5 = 135$. The scale of the Query vector is approximately 700. Hence, the scale of the Laplace noise was varied between 8-80, or, the $\epsilon$ value was varied between $0.1 - 1$.

Table8.6 compares the accuracies obtained using the three estimates for 5-fold, 10-fold and 20-fold cross validation methods. Again, as in the case of the Adult dataset, it is observed that the EBE performs better or at least as good as the MLE. Also, the different between the accuracies of the EBE and the MLE reduces as the value of $\epsilon$ increases. The threshold value of $\epsilon$ for the Nursery dataset is observed to be 0.09. For an $\epsilon$ below 0.09, the noise is so much that

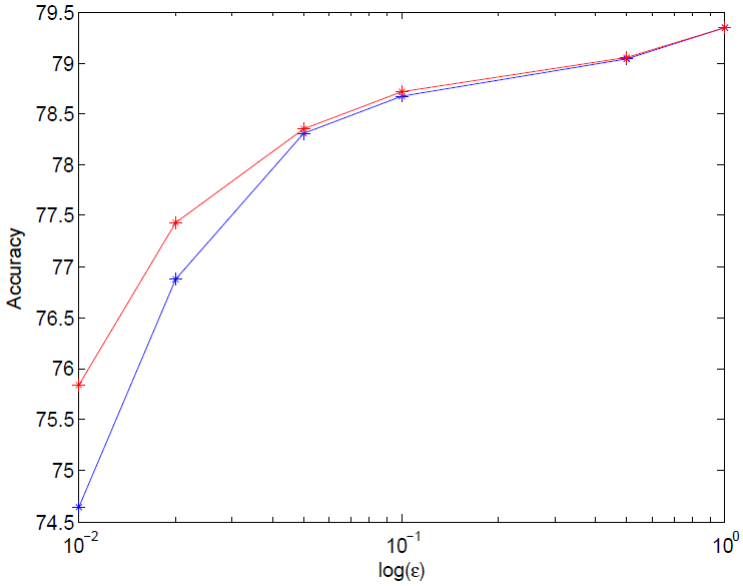



**Figure 8.2.** Accuracy graphs for 10-fold cross validation

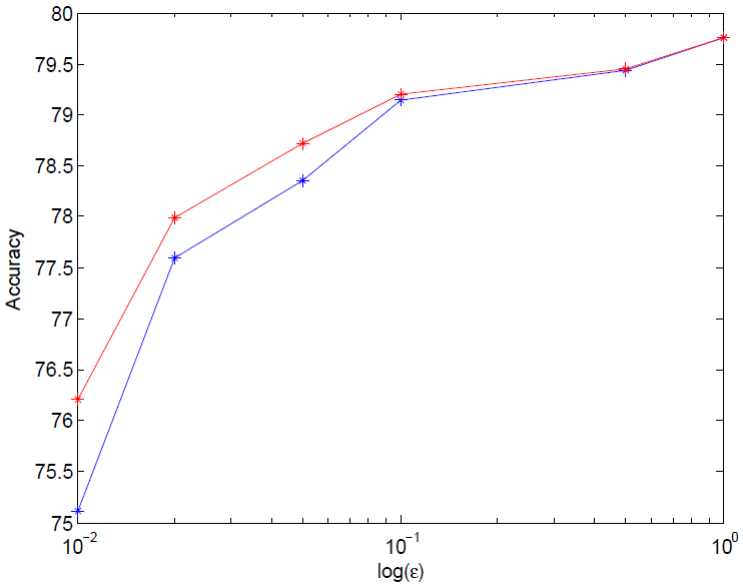



**Figure 8.3.** Accuracy graphs for 20-fold cross validation

it destroys most of the information present in the count query answer. Also, the accuracies of the EBE and the MLE no longer follow any trend for $\epsilon \leq 0.09$. Tables 8.7, 8.8 and 8.9 show the dominance statistics of the EBE over the MLE. Observe that the accuracies obtained using EBE are greater than or equal to those obtained using the MLE most of the time.

| | 5-fold | | | 10-fold | | | 20-fold | | |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ | $A_{ML}$ | $A_{EB}$ | $A_{JS}$ |
| 0.1 | 87.76 | **87.87** | 87.76 | 87.91 | **87.98** | 87.92 | 89.45 | **88.51** | 88.45 |
| 0.25 | 89.69 | **89.88** | 89.80 | 89.49 | **89.53** | 89.49 | 89.95 | **89.96** | 89.95 |
| 0.5 | 90.07 | **90.08** | 90.07 | 89.76 | **89.77** | 89.76 | 90.24 | **90.25** | 90.24 |
| 0.75 | 90.13 | **90.14** | 90.13 | 89.79 | **89.8** | 89.79 | 90.30 | **90.31** | 90.30 |
| 1 | 90.16 | **90.16** | 90.16 | 89.8 | **89.8** | 89.8 | 90.34 | **90.34** | 90.34 |

**Table 8.6.** Comparison of Accuracy of the Naive Bayes classifier for varying $\epsilon$ for the Nursery dataset

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 65.9 | 6.2 | 27.9 |
| 0.25 | 36.7 | 27.2 | 36.1 |
| 0.5 | 17.1 | 65.3 | 17.6 |
| 0.75 | 8.6 | 83.4 | 8.0 |
| 1 | 5.1 | 90.2 | 4.7 |

**Table 8.7.** Dominance statistics(%) in 5-fold cross validation for Nursery Dataset

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 54.3 | 13.1 | 32.6 |
| 0.25 | 24.7 | 48.4 | 26.9 |
| 0.5 | 8.8 | 82.8 | 8.4 |
| 0.75 | 3.8 | 92.8 | 3.4 |
| 1 | 2 | 96.1 | 1.9 |

**Table 8.8.** Dominance statistics(%) in 10-fold cross validation for Nursery Dataset

| $\epsilon$ | $A_{EB} > A_{ML}$ | $A_{EB} = A_{ML}$ | $A_{EB} < A_{ML}$ |
|---|---|---|---|
| 0.01 | 46.2 | 23.1 | 30.7 |
| 0.25 | 15.5 | 69.4 | 15.1 |
| 0.5 | 4.5 | 91.5 | 4.0 |
| 0.75 | 1.9 | 96.3 | 1.8 |
| 1 | 1.2 | 97.9 | 0.9 |

**Table 8.9.** Dominance statistics(%) in 20-fold cross validation for Nursery Dataset

## 8.8 Conclusion

An estimate for a multivariate Laplacian has been derived which is better than the maximum likelihood estimate. This estimate was used to estimate the query answer obtained from a $\epsilon$-differentially private database and was used to build a Naive Bayes classifier. It has been observed that the estimate gives better results for classification compared to the maximum likelihood estimate. Effectively, a better way of estimating the query answer has been proposed which boasts of better utility or usefulness compared to the naive maximum likelihood estimate.

Chapter 9

# Conclusions

The work presented in this Masters thesis is in two different areas. In the first part, we present an attack algorithm for Anatomy sanitization method and then calibrate the probabilities of prediction of the sensitive values. The attack algorithm is simple and is shown to be able to predict the sensitive values with probabilities more than intended by Anatomy. In the second part, we describe a post-processing step to improve the utility of differentially private histogram queries. Experiments show that classifiers constructed using the processed queries are better than the ones constructed with queries obtained directly from differential privacy mechanism.

## 9.1 Conclusions

### 9.1.1 Attacking Anatomy

1. A Naive-Bayes' model was used to build a classifier for the purpose of predicting the sensitive value given the non-sensitive attribute values of tuples. The Naive-Bayes parameters were initialized using a Dirichlet prior and updated using the Expectation-Maximization algorithm. Two variations of the EM algorithm, the EM and the approximate EM(AEM), were used. The AEM algorithm does not consider the constraint on the possible sensitive values for a tuple in the M step where are the EM algorithm does.

2. The algorithm was tested on anatomized versions of the Adult data set with group sizes 2, 3 and 4. The accuracy of the classifier was observed to be significantly higher than that intended by Anatomy sanitization. Specifically, for the adult data set, the accuracy was approximately 71% for group size of 2, 54% for group size of 3 and 43% for group size 4.

3. A method for calibration of probabilities was proposed for the algorithm using approximate EM. The calibrated probabilities were evaluated using reliability diagrams and brier scores and were shown to be better than the uncalibrated probabilities.

### 9.1.2 Constructing Improved Differentially Private Classifiers

1. A post-processing step was proposed to improve the utility of differentially private histogram queries. Since the differential privacy mechanism involves adding independent Laplace noise to each component of the query, the problem was converted to that of estimating the mean of a multi-variate Laplacian random variable.

2. An empirical Bayes' estimator was derived for a multi-variate Laplacian variable which outperforms the maximum-likelihood estimator for dimensions 3 and higher under certain constraints.

3. The estimator was used to post-process the differentially private histogram queries which were in turn used to construct a Naive-Bayes' classifier to predict the sensitive values. The accuracy of such a classifier was compared to the accuracy of a classifier which does not use the post-processing step. The results showed that the classifier which uses the processed histogram queries performs better.

4. The squared error of the empirical Bayes estimate was lesser than that of the maximum likelihood estimate. Hence, any differentially private classifier constructed using the derived empirical-Bayes' estimate would perform at least as good as a classifier which does not.

#### 9.1.2.1 Limitations

The derived estimator for the mean of a multi-variate Laplacian has a constraint that the variance of the differentially private histogram query has to be very high compared to the variance of Laplace noise added to each component. In other words, the histogram query vector should be diffuse. The experiments showed that this constraint was satisfied for the data sets over the range of $\epsilon$ values. In general, the constraint will be satisfied for data sets with large query vector size.

# Bibliography


[1] Arrington, M., "AOL Proudly Releases Massive Amounts of Private Data," .
URL http://www.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

[2] Barbaro, M. and T. Zeller (2006), "A Face Is Exposed for AOL Searcher No. 4417749," .
URL http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=2&ei=5088&en=996f61c946da4d34&ex=1312776000&partner=rssnyt&emc=rss

[3] Narayanan, A. and et al. (2008), "Robust De-anonymization of Large Sparse Datasets," .

[4] Sweeney, L. (1000), "Uniqueness of Simple Demographics in the U.S. Population," .

[5] Kifer, D. (2009) "Attacks on privacy and deFinetti's theorem," in *Proceedings of the 35th SIGMOD international conference on Management of data*, SIGMOD '09, ACM, New York, NY, USA, pp. 127–138.
URL http://doi.acm.org/10.1145/1559845.1559861

[6] Sweeney, L. (2002) "k-anonymity: a model for protecting privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, **10**(5), pp. 557–570.

[7] Samarati, P. and L. Sweeney (1998) *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression*, *Tech. rep.*

[8] Machanavajjhala, A., D. Kifer, J. Gehrke, and M. Venkitasubramaniam (2006) "l-diversity: Privacy beyond k-anonymity," in *In ICDE*.

[9] Xiao, X. and Y. Tao (2006) "Anatomy: Simple and Effective Privacy Preservation." in *VLDB'06*, pp. 139–150.

[10] Dwork, C. (2008) "Differential Privacy: A Survey of Results," *Theory and Applications of Models of Computation*, pp. 1–19.
URL http://dx.doi.org/10.1007/978-3-540-79228-4_1

[11] Gagan Aggarwal, K. K. R. M. R. P. D. T. A. Z., Tomas Feder (2004) *k-Anonymity: Algorithms and Hardness*, *Tech. rep.*, Stanford University.

[12] Bayardo, R. J. (2005) "Data privacy through optimal k-anonymization," in *In ICDE*, pp. 217–228.

[13] Lefevre, K., D. J. Dewitt, and R. Ramakrishnan, "ABSTRACT Incognito: Efficient Full-Domain K-Anonymity," .

[14] Meyerson, A. and R. Williams (2004) "On the complexity of optimal K-anonymity," in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '04, ACM, New York, NY, USA, pp. 223–228.
URL http://doi.acm.org/10.1145/1055558.1055591

[15] Samarati, P. (2001) "Protecting Respondents Identities in Microdata Release," *IEEE Transactions on Knowledge and Data Engineering*, **13**, pp. 1010–1027.

[16] Zhong, S. and Z. Y. R. N. Wright (2005) "Privacy-enhancing k-anonymization of customer data," in *In PODS*, ACM Press, pp. 139–147.

[17] Lefevre, K., D. J. Dewitt, and R. Ramakrishnan (2006) "Mondrian multidimensional k-anonymity," in *In ICDE*.

[18] Ghinita, G., P. Karras, P. Kalnis, and N. Mamoulis (2007) "Fast Data Anonymization with Low Information Loss," in *in VLDB, 2007*, pp. 758–769.

[19] Aggarwal, G., T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu (2006) "Achieving anonymity via clustering," in *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '06, ACM, New York, NY, USA, pp. 153–162.
URL http://doi.acm.org/10.1145/1142351.1142374

[20] Niculescu-mizil, A. and R. Caruana (2005) "Obtaining Calibrated Probabilities from Boosting," in *In: Proc. 21st Conference on Uncertainty in Artificial Intelligence (UAI 05), AUAI Press*, AUAI Press.

[21] Zadrozny, B. and C. Elkan (2001) "Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers," in *In Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann, pp. 609–616.

[22] Drish, J. (1998), "Obtaining calibrated probability estimates from support vector machines," .

[23] Barak, B., S. Kale, K. Chaudhuri, F. Mcsherry, C. Dwork, and K. Talwar (2007) "K.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release," in *In: Proc. of the 26th Symposium on Principles of Database Systems (PODS*, pp. 273–282.

[24] Hay, M., V. Rastogi, G. Miklau, and D. Suciu (2010) "Boosting the accuracy of differentially private histograms through consistency," *Proc. VLDB Endow.*, **3**, pp. 1021–1032.
URL http://portal.acm.org/citation.cfm?id=1920841.1920970

[25] Blum, A., K. Ligett, and A. Roth (2008) "A learning theory approach to non-interactive database privacy," in *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, ACM, New York, NY, USA, pp. 609–618.
URL http://doi.acm.org/10.1145/1374376.1374464

[26] Hay, M., C. Li, G. Miklau, and D. Jensen, "Accurate Estimation of the Degree Distribution of Private Networks," .

[27] Xiao, X., G. Wang, and J. Gehrke (2009) "Differential Privacy via Wavelet Transforms," *CoRR*, **abs/0909.5530**.

[28] Chan, T.-H. H., E. Shi, and D. Song (2010) "Private and continual release of statistics," in *Proceedings of the 37th international colloquium conference on Automata, languages and programming: Part II*, ICALP'10, Springer-Verlag, Berlin, Heidelberg, pp. 405–417.
URL http://portal.acm.org/citation.cfm?id=1880999.1881044

[29] Li, C., M. Hay, V. Rastogi, G. Miklau, and A. McGregor (2009) "Optimizing Histogram Queries under Differential Privacy," *CoRR*, **abs/0912.4742**, informal publication.
URL http://dblp.uni-trier.de/db/journals/corr/corr0912.html#abs-0912-4742

[30] ——— (2010) "Optimizing linear counting queries under differential privacy," in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems of data*, PODS '10, ACM, New York, NY, USA, pp. 123–134.
URL http://doi.acm.org/10.1145/1807085.1807104

[31] Kuhn, H. W. (1955) "The Hungarian method for the assignment problem," *Naval Research Logistic Quarterly*, **2**, pp. 83–97.

[32] Frank, A. and A. Asuncion (2010), "UCI Machine Learning Repository," .
URL http://archive.ics.uci.edu/ml

[33] Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales (2002) "Confidence Builders: Evaluating Seasonal Climate Forecasts from User Perspectives." *Bulletin of the American Meteorological Society*, **83**, pp. 683–698.

[34] Brier, G. W. (1950) "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, **78**, pp. 1–+.

[35] Sanders, F. (1963) "On Subjective Probability Forecasting." *Journal of Applied Meteorology*, **2**, pp. 191–201.

[36] Murphy, A. H. (1972) "Scalar and Vector Partitions of the Probability Score: Part I. Two-State Situation." *Journal of Applied Meteorology*, **11**, pp. 273–282.

[37] Dwork, C., F. Mcsherry, K. Nissim, and A. Smith (2006) "Calibrating noise to sensitivity in private data analysis," in *In Proceedings of the 3rd Theory of Cryptography Conference*, Springer, pp. 265–284.

[38] James, W. and J. Stein (1961) "Estimation with Quadratic Loss," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), University of California Press, pp. 361–379.
URL http://projecteuclid.org/euclid.bsmsp/1200512173

[39] Richards, J. A. (1999) *An Introduction to James-Stein Estimation*, *Tech. rep.*
URL http://ssg.mit.edu/group/johnrich/docs/jse.ps.gz

[40] Stein, J. (1956) "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), University of California Press, pp. 197–206.
URL http://projecteuclid.org/euclid.bsmsp/1200501656

[41] Lindley, D. V. (1962) "Discussion of Professor Stein's paper," *Journal of the Royal Statistical Society*, **Series B**(24), pp. 285–287.

[42] Efron, B. and C. Morris (1977) "Stein's Paradox in Statistics," *Scientific American*, **236**(5), pp. 119–127.

[43] ——— (1972) "Limiting the Risk of Bayes and Empirical Bayes Estimators–Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, **67**(337), pp. 130–139.
URL http://dx.doi.org/10.2307/2284711

[44] Berger, J. (1980) "Improving on Inadmissible Estimators in Continuous Exponential Families with Applications to Simultaneous Estimation of Gamma Scale Parameters," *The Annals of Statistics*, **8**(3), pp. pp. 545–571.
URL http://www.jstor.org/stable/2240592

[45] Hwang, J. T. (1982) "Improving Upon Standard Estimators in Discrete Exponential Families with Applications to Poisson and Negative Binomial Cases," *The Annals of Statistics*, **10**(3), pp. pp. 857–867.
URL http://www.jstor.org/stable/2240909

[46] Matsumura, E. M. and K.-W. Tsui (1982) "Stein-Type Poisson Estimators in Audit Sampling," *Journal of Accounting Research*, **20**(1), pp. pp. 162–170.
URL http://www.jstor.org/stable/2490768

[47] Hudson, H. M. (1978) "A Natural Identity for Exponential Families with Applications in Multiparameter Estimation," *The Annals of Statistics*, **6**(3), pp. pp. 473–484.
URL http://www.jstor.org/stable/2958553

[48] Olkin, I. and M. Sobel (1979) "Admissible and Minimax Estimation for the Multinomial Distribution and for K Independent Binomial Distributions," *The Annals of Statistics*, **7**(2), pp. pp. 284–290.
URL http://www.jstor.org/stable/2958810