# Comparative Study of Distance Metrics for $t$-closeness


Dilip Kumar Vallabhadas


**Department of Computer Science and Engineering**

**National Institute of Technology Rourkela**

**Rourkela-769 008, Odisha, India**

# Comparative Study of Distance Metrics for $t$-closeness

*Thesis submitted to the department of*

***Computer Science and Engineering***

*of*

***National Institute of Technology Rourkela***

*in partial fulfillment of the requirements*

*for the degree of*

## Master of Technology

*by*


**Dilip Kumar Vallabhadas**

[ Roll No. 211CS2063 ]


*under the guidance of*

**Prof. Korra Sathya Babu**



**Department of Computer Science and Engineering**

**National Institute of Technology Rourkela**

**Rourkela-769 008, Odisha, India**

Department of Computer Science and Engineering
**National Institute of Technology Rourkela**
Rourkela-769 008, India. www.nitrkl.ac.in

**Mr. Korra Sathya Babu**
Assistant Professor

# Certificate

This is to certify that the work in the thesis entitled *Comparative Study of Distance Metrics for t-closeness* by *Dilip Kumar Vallabhadas* is a record of an original work carried out by him under my supervision and guidance in partial fulfillment of the requirements for the award of the degree of *Master of Technology* in *Computer Science and Engineering*. Neither this thesis nor any part of it has been submitted for any degree or academic award elsewhere.

***Korra Sathya Babu***

# Acknowledgment

First of all, I would like to express my deep sense of respect and gratitude towards my supervisor Prof. **Korra Sathya Babu**, who has been the guiding force behind this work. I want to thank him for introducing me to the field of **Data Privacy** and giving me the opportunity to work under him. His undivided faith in this topic and ability to bring out the best of analytical and practical skills in people has been invaluable in tough periods. Without his invaluable advice and assistance it would not have been possible for me to complete this thesis. I am greatly indebted to him for his constant encouragement and invaluable advice in every aspect of my academic life. I consider it my good fortune to have got an opportunity to work with such a wonderful person.

I wish to thank all faculty members and secretarial staff of the CSE Department for their sympathetic cooperation.

I am really thankful to my all friends. My sincere thanks to everyone who has provided me with kind words, a welcome ear, new ideas, useful criticism, and their invaluable time, I am truly indebted.

My thanks and apologies to those whom I have inadvertently missed out.

Finally, Above all ***I praise God*** for helping me in completing my work.

***Dilip Kumar Vallabhadas***

# Abstract

In our present technical world we all have to submit our personal information to various organisations driven by mutual benefits. The collected data has to be published. There was a need for exchange of the published data between different parties. Given data in its original form contains sensitive information of the person. If the given data is published directly sensitive information is revealed to others which violates the privacy of individual directly. In order to publish the data without violating one's personal privacy we use a technique called $t$-closeness. In this method the metric used was Earth Mover's Distance(EMD). But this metric does not satisfies probability scaling property which makes it not to reflect the difference between the probabilities. This make EMD to produce inaccurate results which may increase the anonymization. In order to have a metrics that satisfies all the distance metric properties including probability scaling property we make a study on different metrics like Squared Root Jensen-Shannon,Pearson and Divergence. We compare these metric by taking different parameters like Discenibility Metrics,propensity score and precision.



***Keywords*:** $k$-Anonymity, $l$-diversity, $t$-closeness, Privacy, Utility

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Introduction

In our present technical world we all have to expose our personal information at various government and private organization like hospitals, voter enrolment and so on. For example, licensed hospitals in California are required to submit specific data on every patient discharged from their facility [1]. In June 2004, the Information Technology Advisory Committee released a report entitled Revolutionizing Health Care Through Information Technology. One of its key points was to establish a nationwide system of electronic medical records that encourages sharing of medical knowledge through computer assisted clinical decision support. Data publishing is equally distributed in other domains.

For example, Netflix, a popular online movie rental service, recently published a data set containing movie ratings of 500,000 subscribers, in a drive to improve the accuracy of movie recommendations based on personal preferences [2]; AOL

published a release of query logs but quickly removed it due to the re-identification of a searcher [3].

Given data in its original form contains sensitive information of individual,publishing this directly violates the one's individual privacy [4]. The data should be published in a way that one cannot get any useful information from the released data. In order to provide particular privacy a task has been introduced which is known as Privacy Preserving and Data Publishing(PPDP).

The main objective of this method is to preserve one's privacy while publishing the data [5]. Lot of methods have been proposed on PPDP but having some of the attacks on them. Here we are focusing on $t$ -closeness which is helpful in protecting against the attribut disclosure attack. It uses Earth Mover Distance(EMD) for calculating the distance. But EMD doesn't satisfy probability scaling property of a distance metrics.

## 1.2 Privacy Preserving Data Publishing

A task is to be produced in-order to publish the data without violating the privacy of individual. This technique of publishing data is known as Privacy Preserving Data Publishing(PPDP) [6]. In the past few years , several researchers have proposed different methods for PPDP. These proposed methods have their advantages and disadvantages , we need a method that balance both utility and privacy

To publish data [4, 7–9] introduced a method caleed $k$-anonymity.$k$-anonymity

says that an equivalence class on quasi-identifier should contain at least $k$ records.$k$ anonymity can protect over identity disclosure but it fails to protect against attribute disclosure.Two attacks have been identified on $k$-anonymity they are homogeneity attack and background knowledge attack. In order to overcome this another method has been proposed known as $l$-diversity.It states that an equivalence class should have atleast $l$ well represented values. A table satisfies $l$-diversity if and only if each equivalence class should satisfy $l$-diversity [10]. $l$-diversity was also not able to protect against attribute disclosure. Two attacks have been identified on $l$-diversity they are similarity attack and skewness attack [11],$k$-anonymity and $l$-diversity are not sufficient to protect against attribute disclosure [12] so a new method to protect against attribute disclosure has been introduced which is known as $t$-closeness

$t$-closeness states that the distance between distribution of attributes in each equivalence class and whole table should not be more than t [11].$t$-closeness uses Earth Mover distance(EMD). In EMD we have a drawback that it does not satisfy probability scaling. However $t$-closeness protect against attribute disclosure but it does not satisfy identity disclosure. The distance metric used in $t$-closeness does not works properly so we need a distance metric that gives accurate results. Here we are discussing different metrics that can be used on $t$-closeness for calculating the distance detween two probabilities and also the properties required for a distance metrics. We compare these metrics with the given Metric i.e EMD

## 1.3 Problem Statement

The main problem is to calculate the distance between two probability distributions. For this measure of distance we use Earth Mover Distance(EMD). When we are using EMD for calculating the distance in $t$ -closeness. Due to the lack of probability scaling in EMD the reflecting difference is not considered for some random values. This makes the metrics to give inaccurate results. So, there is a need to have a measure which satisfy all the properties of a distance metrics. So that we can have a good metrics for calculation of distance in $t$-closeness which considers all the value and gives more accurate results.

## 1.4 Motivation

The data that we are giving to different organizations contain sensitive information. In order to protect the privacy of one's personal data we need a method that does not reveals any sensitive information but can share the information between different organizations. So we use different methods that are used to publish the data without revealing sensitive information of a person. We use k-anonymity and l-diversity but they were able to handle only identity disclosure. To protect against attribute disclosure we use t-closeness

## 1.5 Organisation of Thesis

The following chapters gives the outline and organization of the thesis with an emphasis on the contributions.

**Chapter 1: Introduction** In this chapter we are going to discuss brief introduction and the fundamental concept of Privacy Preserving and Data Publishing(PPDP).It also gives the information why we are using different distance metrics for calculating $t$-closeness and also the problem of the existing metrics. It also gives the basic idea of our metrics

**Chapter 2: Literature Review** In this chapter, we are going to discuss the basic definitions that are required for PPDP. Attacks that are performed on the released data. We are going to have a description of different methods and the attacks that are existing on these methods.

**Chapter 3: Distance Metrics** In this chapter, we are going to discuss about different properties required for a distance metrics. Eath Mover Distance in detail and the proof for the failure of EMD

**Chapter 4: Distance Metrics used for study** In this chapter we are going to discuss different metrics which can also be used for $t$-closeness.Comparison of those metrics with EMD.

**Chapter 5: Results** Here we are going to give graph for comparing of these metrics on different costs like DMCost,Propensity Score and Precision. And also

we are going to have tabular form of metrics satisfying different properties to be a good distance metrics

**Chapter 6: Conclusion** This chapter gives the use of the method and limitations of the proposed schemes.

# Chapter 2

# Literature Review

In this chapter we are going to discuss about how the data is to be stored and various methods of publishing the data without violating the personal privacy of a person. The attacks that are possible on each of the method and also the advantages and disadvantages of the existing methods.

## 2.1 Basic Definitions

Generally the data has to be published are stored in the form of tables.Each table contains several records which are divided into attributes. Based on the functionality of the attributes these tables are divided into four type of attributes. They are explained using the following table.

| SSN | NAME | AGE | JOB | SEX | SALARY | DISEASE |
|---|---|---|---|---|---|---|
| 321 | A | 28 | Engineer | Male | 4K | Cancer |
| 432 | B | 32 | Engineer | Female | 3K | Heart disease |
| 543 | C | 26 | Lawyer | Male | 3K | Flu |
| 234 | D | 45 | Dancer | Male | 4K | Cancer |
| 456 | E | 32 | Lawyer | Female | 3K | Flu |
| 426 | F | 42 | Engineer | Female | 3K | Gastritis |
| 453 | G | 32 | Lawyer | Female | 3K | Stomach disease |

Table 2.1: Example

The attributes are described as follows [6]

- *Explicit Identifier* : These are the attributes that are used to identify the record owner directly. In the above table SSN and Name are Explicit Identifiers.

- *Quasi Identifier* : These are the attributes that are used to potentially identify records owner. Age,Job and Sex are quasi Identifiers in the above table.

- *Sensitive Attributes* : These are the attributes that contains the person-specific information.Salary and Disease are the sensitive attributes in the above table.

- *Non-Sensitive Attributes* : All the attributes that does not fall into these three categories are considered as Non-Sensitive attributes.

When releasing the micro data different disclosures [13] have been identified they are

- *Identity Disclosure* : Disclosure occurs when a data subject is identified from a released file
- *Attribute Disclosure* : Attribute disclosure occur when confidential information is revealed exactly or when it can be closely estimated.
- *Inferential Disclosure* : The released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible

## 2.2 *k*-Anonymity

Definition:

Let RT$(A_1, A_2, ..., A_n)$ be a table and $QI_{RT}$ be the quasi-identifier associated with it. RT is said to satisfy $k$-anonymity if and only if each sequence of values in RT$[QI_{RT}]$ appears with at least $k$ occurrences in RT$[QI_{RT}]$. [7]

The basic idea of $k$-anonymity is easy and simple to understand. If a person knows quasi-identifier value of an individual he cannot finds an individual record from a table satisfying $k$-anonymity with confidence greater than 1-k corresponding to that individual. $k$-anonymity was able to protect against identity disclosure but it cannot protect against attribute disclosure. [6, 7, 10, 11] has recognize attacks on $k$-anonymity

Example of $k$-anonymous table with

- k=2,QID= Race, Birth, Gender, Zip
- Sensitive attribute=Problem

| Race | Birth | Gender | Zip | Problem |
|---|---|---|---|---|
| Black | 1965 | m | 02141 | short breath |
| Black | 1965 | m | 02141 | chest pain |
| Black | 1964 | f | 02138 | obesity |
| Black | 1964 | f | 02138 | chest pain |
| White | 1964 | m | 02138 | chest pain |
| White | 1964 | m | 02138 | obesity |
| White | 1964 | m | 02138 | short breath |

Table 2.2: 2-Anonymous [7]

### 2.2.1 Attacks on *k*-Anonymity

Several authors [4, 7–9, 11] has discussed attacks on $k$-anonymity. There are two types of attacks $k$-anonymity. They are

- Homogeneity Attack
- Background Knowledge Attack

**Homogeneity Attack:**

Lets have Alice and Bob are two neighbours [10].One day that bob feels sick and admitted in a hospital. Seeing that she fells that Bob has some disease. She get's 4-anonymous table that was released by the hospital.

Alice being his neighbour she knows the age of bob and also the zip code. By having these data she gets an idea that bob record is in 9,10,11 or 12 record. Since each record have disease condition as cancer she was able to know that bob is suffering from cancer

| S No | Zip Code | Age | Nationality | Condition |
|---|---|---|---|---|
| 1 | 130** | $< 30$ | * | Heart Disease |
| 2 | 130** | $< 30$ | * | Heart Disease |
| 3 | 130** | $< 30$ | * | Viral Infection |
| 4 | 130** | $< 30$ | * | Viral Infection |
| 5 | 1485* | $\geq 40$ | * | Cancer |
| 6 | 1485* | $\geq 40$ | * | Heart Disease |
| 7 | 1485* | $\geq 40$ | * | Viral Infection |
| 8 | 1485* | $\geq 40$ | * | Viral Infection |
| 9 | 130** | 3* | * | Cancer |
| 10 | 130** | 3* | * | Cancer |
| 11 | 130** | 3* | * | Cancer |
| 12 | 130** | 3* | * | Cancer |

Table 2.3: 3-Anonymous [7]

**Background Knowledge Attack:**

Alice contain a friend Umeko who is been admitted in the same hospital as bob admitted [10]. As Umeko is her friend she know her age and also the address where she lives. From this information Alice know that Umeko record should in be in 1,2,3 or 4 record in Table 2.3.

From the records she was not sure that Umeko has heart disease or cancer. However she knows that Japanese were extremely low incidence of having heart disease.So that she concludes with more probability that Umeko is suffering from viral infection

## 2.3 *l*-Diversity

Definition:
An equivalence class is said to have $l$-diversity if there are at least $l$ "well represented" values for the sensitive attribute. A table is said to have $l$-diversity if every equivalence class of the table has $l$-diversity [10]

$l$-diversity works one step ahead of $k$-anonymity in preventing attribute disclosure.But l-diversity is more difficult to achieve and also it is not able to provide sufficient protection for privacy. $l$-diversity also does not provide protection against attribute disclosure. In [11] attacks have been discussed on $l$-diversity. In Table 2.4 even though Alice knows bob's age and zip code she was not able to infer from the 3-diverse table

| S No | Zip Code | Age | Nationality | Condition |
|---|---|---|---|---|
| 1 | 1305* | ≤40 | * | Heart Disease |
| 4 | 1305* | ≤40 | * | Viral Infection |
| 9 | 1305* | ≤40 | * | Cancer |
| 10 | 1305* | ≤40 | * | Cancer |
| 5 | 1485* | > 40 | * | Cancer |
| 6 | 1485* | > 40 | * | Heart Disease |
| 7 | 1485* | > 40 | * | Viral Infection |
| 8 | 1485* | > 40 | * | Viral Infection |
| 2 | 1306* | ≤40 | * | Heart Disease |
| 3 | 1306* | ≤40 | * | Viral Infection |
| 11 | 1306* | ≤40 | * | Cancer |
| 12 | 1306* | ≤40 | * | Cancer |

Table 2.4: 3-Diverse [10]

When it comes to Umeko Alice thinks that Japanese are most low in heart disease. She was not sure that Umeko has cancer or viral infection.

### 2.3.1 Attacks on *l*-Diversity

In [10, 11] authors have discussed the attacks on $l$-diversity. They are

- Skewness Attack
- Similarity Attack

### Skewness Attack

In a given $l$-diverse table when the overall distribution is been skewed even though the table satisfies $l$-diversity it was not able to prevent against attribute disclosure.

For example consider a table with equal number of positive and negative record [11]. Even though the table was able to satisfy $l$-diversity this contains a serious privacy threat. Everyone in the table has a probability of 50% of being positive or negative which is as compared with 1%.

### Similarity Attack

In a given $l$-diverse table, even though sensitive attributes in a given equivalence class are distinct but they are semantically similar so the adversary can get the required information. Let us consider a table with sensitive attribute as disease.

| S No | Zip Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 47677 | 29 | 3K | gastric ulcer |
| 2 | 47602 | 22 | 4K | gastritis |
| 3 | 47678 | 27 | 5K | stomach Cancer |
| 4 | 47905 | 43 | 6K | gastritis |
| 5 | 47909 | 52 | 11K | flu |
| 6 | 47906 | 47 | 8K | bronchitis |
| 7 | 47605 | 30 | 7K | bronchitis |
| 8 | 47673 | 36 | 9K | pneumonia |
| 9 | 47607 | 32 | 10K | stomach cancer |

Table 2.5: Original Table [10]

In the given 3-diverse Table 2.6 one know that the salary of bob is in the range of [3K,4K] which is in the first 3 records. By having these information one conclude that bob's record is in first equivalence class. By this he conclude that bob has stomach related disease.

## 2.4 *t*-closeness

Definition:

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

| S No | Zip Code | Age | Salary | Disease |
|---|---|---|---|---|
| 1 | 476* | 2* | 3K | gastric ulcer |
| 2 | 476** | 2* | 4K | gastritis |
| 3 | 476** | 2* | 5K | stomach Cancer |
| 4 | 4790* | ≥40 | 6K | gastritis |
| 5 | 4790* | ≥40 | 11K | flu |
| 6 | 4790* | ≥40 | 8K | bronchitis |
| 7 | 476** | 3* | 7K | bronchitis |
| 8 | 476** | 3* | 9K | pneumonia |
| 9 | 476** | 3* | 10K | stomach cancer |

Table 2.6: 3-Diverse [10]

have $t$-closeness if all equivalence classes have $t$-closeness [11]

At first an observer has a belief B0 which is prior belief about sensitive information of an individual. After he has given generalized table then his belief is influenced by Q which changes to B1. After release of the anonymous table by using the quasi-identifier he can find the equivalence class P then his belief changes to B2. We try to limit the gain of B1 to B2 which there by limits the distance between P and Q.

Here we are going to consider two probability distributions p(x) and q(x). Here we try to limit the distance between p(x) and q(x).It is given as

$$D[p(x), q(x)] < t \tag{2.1}$$

In this given work we use Earth Mover Distance(EMD) to calculate the distance between two probability distributions. But EMD does not work well due to the lack of probability scaling property.

# Chapter 3

# Distance Metrics

The problem here is to measure the distance between p(x) and q(x) where p(x) and q(x) are two probability distribution . p(x) is the distribution of the first equivalence class with x value and q(x) is distribution of whole table having x values. Here we have to limit the distance between p(x) and q(x) which should be less than the threshold t. In order to have a good privacy we have to maintain low t value. If t value is low the more privacy we can get. A good metric is that it does not reveal any information. So that a 0-closed table does not reveal any kind of information. In this chapter we first discuss the properties that are required to be a good distance metric then we go are going to discuss about the EMD distance metric which is used to calculate t value then we are going to show how EMD failed to provide good privacy

## 3.1 Properties of Distance Metrics

Let us consider a function F:XxR is said to be a distance function if and only if it satisfies the following properties [14]

- Non Negativity:

  The adversary have non negative information gain after the release of the table. It can be shown mathematically as $D[P, Q] \geq 0$.

- Identity of discernibles:

  The adversary has no information gain until his belief change. It is shown mathematically as D[P,Q]=0 if and only if P=Q.

- Triangular Inequality:

  For any given three distributions P,Q and R it is defined as $D[P, Q] \leq D[Q, R] + D[R, P]$

- Probability Scaling:

  The belief change from probability P is more significant than that of Q when t is small. D[P,Q] should consider reflecting the difference

- Zero Probability Definability:

  When there are zero probable values in both P and Q. The distance between P and Q should be well defined i.e D[P,Q]

## 3.2 EMD Distance Metric

We have a metric for ground distance is defined for any pair of values. In $t$-closeness we have two distributions for which we have to find the distance. We need that distance should be dependent on the ground distance which makes us to go for Earth Mover's Distance(EMD) [11, 15] which is Monge-Kantovich transportation distance [16]. In $t$-closeness we are going to find the distance between P and Q where P is the distribution of values in equivalence class and Q is the distribution of values in the whole table. The main disadvantage of EMD is that it does not satisfy probability scaling property of a distance metric and also the computation required for EMD is more. In general we are going to use data set which is not normalized but EMD requires normalized data set. In the next section we are going to discuss about how EMD fails probability scaling property.

## 3.3 Failure of EMD

Let us consider two distributions (0.01,0.99) and (0.11,0.89) the EMD measure between these two distributions is 0.1. Consider another two distributions (0.4,0.6) and (0.5,0.5) the EMD measure between these two distributions is also 0.1. One may argue that belief change in first two distributions is much more significant than in second. In first the probability of taking the first value is from 0.1 to 0.11 which is a 1000 percent [11]. But in second case it is only 25 percent even though the EMD measure gives 0.1 for both the cases. This happens because of the lack of probability scaling, EMD does not reflects each value some values are been

neglected. Which produces inaccurate distance In result the tuples are been placed in improper equivalence class. Which results in increasing the anonymization. So there is a need for a good metric which satisfies the above all properties and should produce a close t value.

# Chapter 4

# Distance Metrics used for Study

In privacy preserving and data publishing th main challenge is to have a good trade of between utility and privacy. From the existing model we are able to handle attribute disclosure which means attribute disclosure can be handled by satisfying t-closeness. However privacy and utility cannot get compromised they have to be so that while publishing the data we use to have a utility along with privacy. In [11,17] t-closeness is been satisfied which provides a protection against attribute disclosure. In [11] we know that authors have taken EMD distance measure for calculating the distance between the two distributions P and Q. P is the distribution of the equivalence class and Q is the distribution of whole table. In order to satisfy $t$-closeness the distance between P and Q does not exceed the threshold value t. EMD does not satisfies probability scaling property. So there is a need for metric which satisfies all the properties of a distance metric [18]. Here we are going to discuss various metrics [19, 20] that $t$-closeness. How they are giving the distance between p and q.

Let us consider two multinomial populations p(x) and q(x), each consisting of N classes with respective probabilities p(x=1),p(x=2),.....,p(x=N) and q(x=1),q(x=2),.....,q(x=N). Since both p(x) and q(x) are probability distribution we have

$$\sum_{x=1}^{N} p(x) = \sum_{x=1}^{N} q(x) = 1$$

## 4.1 Squared Root of JSD

The measure for squared Root of Jensen-Shannon Divergence [21] is given as

$$D_{sjsd} = \sqrt{JSD} \tag{4.1}$$

Where JSD is Jensen-Shannon divergence. This equation gives the metric which is been derived from jensen-shannon divergence [22] which satisfies all the properties of the distance metrics.

The coefficient JSD is given as

$$JSD(p,q) = \frac{1}{2}[\sum_{x=1}^{N} p(x) log \frac{2p(x)}{p(x)+q(x)} + \sum_{x=1}^{N} q(x) log \frac{2q(x)}{p(x)+q(x)}] \tag{4.2}$$

Where JSD is the symmetric version of Kullback-Leibler Divergence.

$$KL(p,q) = \sum_{x=1}^{N} \frac{p(x)}{q(x)} q(x) \tag{4.3}$$

Squared Root Jensen-Shannon Divergence(SJSD) satisfies all the properites of a distance metrics [21]. The SJSD distance between two probability distributions

(0.01,0.99) and (0.11,0.89) is 0.16032. The distance between another two distributions (0.4,0.6) and (0.5,0.5) is 0.0707. By this we can say that the change of first two probability distributions is much more significant when compared to the second two distributions. So SJSD satisfies the probability scaling property.

## 4.2 Pearson $\chi^2$

Let us consider two probability distribution p(x) and q(x) where p(x) is probability of equivalence class and q(x) is the distribution of whole class. The pearson [19] measure for this distributions is given as

$$D_p = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{q_i} \tag{4.4}$$

Pearson also satisfies the metric properties we are required.The Pearson [23] distance between two probability distributions (0.01,0.99) and (0.11,0.89) is 0.102. The distance between another two distributions (0.4,0.6) and (0.5,0.5) is 0.0404. Pearson also shows the reflection of the values so it also satisfies the probability scaling property.

## 4.3 Divergence

Let us consider two probability distributions p and q over a variable x. Then the equation for the divergence [19, 20] measure is given as

$$D_d = \sum_{i=1}^{m} \frac{(p_i - q_i)^2}{(p_i + q_i)^2} \tag{4.5}$$

The above equation also satisfies all the properties of a distance metrics.The Divergence distance [24] between two probability distributions (0.01,0.99) and (0.11,0.89) is 0.6727. The distance between another two distributions (0.4,0.6) and (0.5,0.5) is 0.0206. This measure also satisfies probability scaling but the measure produce a large value. In order to have a good privacy we need a value which is much less i.e around 0.1.

## 4.4 Comparision of Metrics for distributions

In this section we are going to discuss how these metrics are going to reflect for a given distributions. Here we are giving how other metrics satisfies probability scaling property.

Let us consider two cases of probability distribution. In case1 we are going to take p(x)=(0.01,0.99) and q(x)=(0.11,0.89). The comparison table is given as follows

EMD - Earth Mover's Distance

SJSD - Squared Root of Jensen-Shannon

| Methods | Case 1 | Case 2 |
|---|---|---|
| EMD | 0.1 | 0.1 |
| SJSD | 0.16032 | 0.0707 |
| Pearson | 0.102 | 0.0404 |
| Divergence | 0.6727 | 0.0206 |

Table 4.1: Comparison of Metrics

Case 1:p=(0.1,0.99) q=(0.11,0.89)

Case 2:p=(0.4, 0.6) q=(0.5, 0.5)

**F** rom the Table 4.1 we can clearly say that EMD does not satisfies probability scaling property and the other metrics which we used were able to satisfy all the properties So, we can use this metrics for calculating the distance between two distributions in $t$-closeness.

## 4.5 Analysis of Distances

Our experiment using R environment on the adult set gives the clear view of all the metrics. The figure 4.1 gives the difference in distance produced by various metrics.When we compute the distance between two distribution p(x) and q(x) on the adult set we got the distance for EMD as 0.266 for SJSD the distance is 0.199255 for divergence the distance is 0.1822 and for pearson it is 0.4096. The difference of these metrics were plotted.

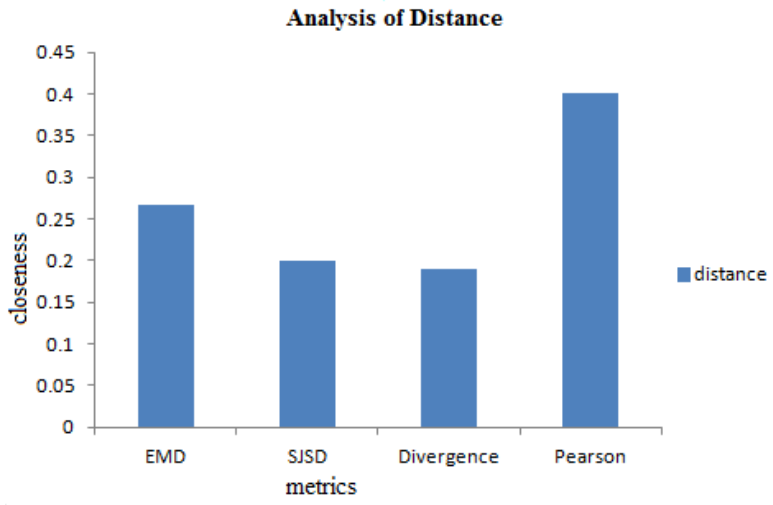


Figure 4.1: Analysis of distance with closeness

# Chapter 5

# Results

The main goal of this study is to have a good privacy metric that balance both privacy and utility. Here we are looking for a metric which works good than that of EMD.

## 5.1 Computational Environment and Dataset

Here we are working on Adult dataset which is a repository of US census. From the dataset we are taking 8 attributes. The attributes which we are considering are Age which is numeric, Work class which is categorical, Education which is categorical, Marital status which is categorical, race which is categorical,gender is categorical, Occupation is sensitive attributes

| S.No | Attributes | Type |
|---|---|---|
| 1 | Age | Numeric |
| 2 | Work Class | Categorical |
| 3 | Zip Code | Numeric |
| 4 | Education | Categorical |
| 5 | Marital Status | Categorical |
| 6 | Race | Categorical |
| 7 | Gender | Categorical |
| 8 | Occupation | Sensitive |

Table 5.1: Attributes taken from Dataset

## 5.2 Evaluation Parameters

In this we are going to evaluate by using three parameters namely Propensity Score(PS), Discernibility Metric Cost(DM cost) and Precision.By using these three parameters we are going know how these metrics are been working. Here we are using occupation as sensitive attribute

### 5.2.1 Discernibility Metric

The first parameter we use is one that attempts to capture in a straight forward way the desire to maintain discernibility between tuples as much as is allowed by a given setting of *t*. DM Cost assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it. If an unsuppressed tuple falls into an induced equivalence class of size j, then that tuple is assigned

a penalty of j . If a tuple is suppressed, then it is assigned a penalty of $|D|$, the size of the input dataset. This penalty reflects the fact that a suppressed tuple cannot be distinguished from any other tuple in the dataset [25]. The metric can be mathematically stated as follows:

$$C_{DM}(g,k) = \sum_{\forall E s.t. |E| \geq k} |E|^2 + \sum_{\forall E s.t. |E| < k} |D||E| \tag{5.1}$$

In this expression, the sets E refer to the equivalence classes of tuples in D induced by the anonymization g . The first sum computes penalties for each non-suppressed tuple, the second for suppressed tuples.

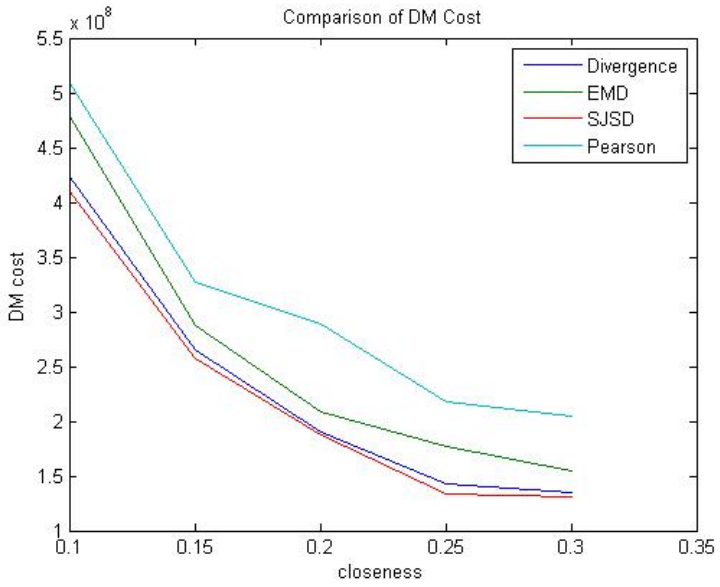


Figure 5.1: DM Cost for various Metrics

### 5.2.2 Propensity Score

The propensity score , in the literature, is defined as the probability of assigning a treatment given a covariate, x. for any propensity score,the treatment assignment and covariates are conditionally independent. Thus, its evident that two large groups with same distributions of propensity scores also have similar distributions of covariates. This helps us in measuring the utility .In calculating propensity score [26] first we have have to merge both original data and masked data to a variable by giving one value for masked and zero value for original data.Then we compute the probability of each record to be in masked data. Then we compute the probability of having in original and masked. When both the distributions are equal then utility will be high.

The similarity of the propensity scores for the masked and original observations can be assessed in numerous ways. A simple summary is to compute

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[P_i - C]^2 \tag{5.2}$$

where N is is total number of records in the merged data and p is the estimated propensity score for unit i and c equals the proportion of units with masked data in the merged data set.In many cases, the original and masked data sets would have the same size N0, in which case, N = 2N0 and c = 1/2. When the original and masked data have the same distribution, the propensity scores for all units should approximately equal c, so that Up is near zero. At the other extreme, if p is nearly one for units i from the masked data and nearly 0 for units from the original data, then the two data sets are completely distinguishable and Up 1/4.

| Method | Propensity Score |
|---|---|
| EMD | 0.21 |
| SJSD | 0.23 |
| Divergence | 0.25 |
| Pearson | 0.25 |

Table 5.2: Comparison of Propensity Score

### 5.2.3 Precision

The precision of a generalization scheme is the average height of generalization (measured over all cells). The precision is 1 if the is no generalization that is if we use same values as given and is 0 if all values are generalized. [27]

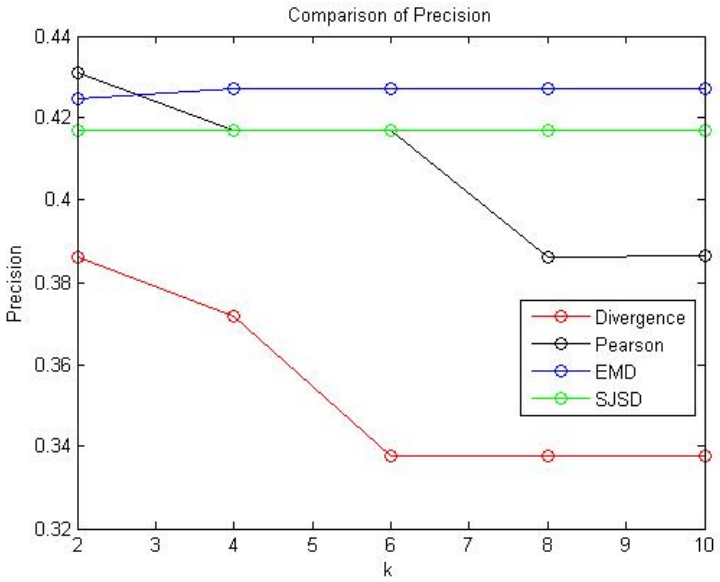


Figure 5.2: Comparision of precision

# Chapter 6

# Conclusion

In order to have a good privacy metric we need to balance both utility and privacy which are inverse terms. Existing metrics does not provide a good privacy due to the lack of probability scaling it does not reflect all the values. In order to have a good metrics that balance both privacy and utility we use different metrics. By satisfying $t$-closeness we are able to handle attribute disclosure. Here we have given different metric which can be used in calculating the distance between probability distributions in $t$-closeness. A method should be produced which can handle both identity disclosure and attribute disclosure.In order to satisfy both identity disclosure and attribute disclosure we have to combine both $k$-anonymity and $t$-closeness but that will become more complex.

# Bibliography


[1] J C Maxwell and B A Rutkowski. The prevalence of methamphetamine and amphetamine abuse in north america: a review of the indicators. *Drug and alcohol review*, 27(3):229–235, sep 2008.

[2] http://www.time.com/time/magazine/article/0,9171,336006,00.html.

[3] Barbaro and Zeller. http://www.nytimes.com/2006/08/09/technology/09aol.html, 2006.

[4] P Samarati. Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on*, 13(6):1010–1027, Dec 2001.

[5] S V Vassilios, B Elisa, N F Igor, P P Loredana, S Yucel, and T Yannis. State of the art in privacy preserving data mining, 2004.

[6] C M F Benjamin, W Ke, C Rui, and S Y Philip. Privacypreserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):589–592, June 2010.

[7] L Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, Sep 2002.

[8] J L Lin and M C Wei. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 international workshop on Privacy and anonymity in information society*, pages 46–50. ACM, 2008.

[9] R J Bayardo and R Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[10] A Machanavajjhala, D Kifer, J Gehrke, and M Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, march 2007.

[11] N Li, T Li, and S Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[12] S P Reiss. Practical data-swapping: the first steps. *ACM Transactions on Database Systems (TODS)*, 9(1):20–37, 1984.

[13] D Lambert. Measures of disclosure risk and harm. *Jounal of official statistics-stockholm*, 9:313–313, 1993.

[14] A V Arkhangel'skii and Pontryagin L S. *General topology I*, volume 1. Springer, 1990.

[15] Y Rubner, C Tomasi, and L J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, Nov 2000.

[16] C R Givens and R M Shortt. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, march 1984.

[17] N Li, T Li, and S Venkatasubramanian. Closeness: A new privacy measure for data publishing. *Knowledge and Data Engineering, IEEE Transactions on*, 22(7):943–956, July 2010.

[18] R Motwani and Y Xu. Efficient algorithms for masking and finding quasi-identifiers. In *Proceedings of the Conference on Very Large Data Bases (VLDB)*, 2007.

[19] S H Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):300–307, 2007.

[20] C D Manning and H Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[21] D M Endres and J E Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, July 2003.

[22] J Basak. Detection of neural activities in fmri using jensen-shannon divergence. In *Advances in Pattern Recognition, 2009. ICAPR '09. Seventh International Conference on*, pages 39–42, 2009.

[23] Robin L Plackett. Karl, p. *International Statistical Review/Revue Internationale de Statistique*, pages 59–72, 1983.

[24] T F Cox and M A Cox. *Multidimensional scaling*, volume 88. CRC Press, 2001.

[25] M J Woo, J P Reiter, A Oganian, and A F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1):111–124, 2009.

[26] P R Rosenbaum and D B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[27] C C Aggarwal and S Yu Philip. *A general survey of privacy-preserving data mining models and algorithms*. Springer, 2008.