On Minimality Attack in Privacy Preserving Data Publishing

by

K. Hemantha
201511024

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

April, 2017

Declaration

I hereby declare that

i) the thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) due acknowledgment has been made in the text to all the reference material used.

K. Hemantha

Certificate

This is to certify that the thesis work entitled On Minimality Attack in Privacy Preserving Data Publishing has been carried out by K. Hemantha for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

Dr. Maniklal Das
Thesis Supervisor

Acknowledgments

I would like to very gratefully acknowledge the direction, support, and the encouragement provided to me by my thesis advisor, Dr. Maniklal Das, throughout this work. He has always given me plenty of freedom while making sure that the work progresses towards the right direction. It has been a very enriching experience for me to work with him. I specially thank him for his sacrifice of time and scholarly advice.

I would like to thank my committee members, Prof. Anish Mathuria and Prof. Sasidhar kalyan for their constructive criticism, which made me, work harder.

I would like to express deep thanks to Nidhi Desai (Ph.d student, DA-IICT) for spending her valuable time in discussion at times when I got stuck in various definitions.

Last but not the least, I would like to thank my family members. Whatever I have achieved in my life, I owe to them. They never let me see through the numerous sacrifices they made and the hardships they went through for me.

Contents

Abstract

Data Publishing has become much concern in recent yerars for protecting the individual privacy. Information about the individuals is collected from various domains and is being published public. We can extract lot of wealth of information by collecting and sharing personal information. For example traditional organization like hospitals and census collect information from individuals and publish them. Census data provides us information which is used for demographic an economic research. Hospitals provide information to let us know how diseases spread and the diseases according to age, gender and so on. And nobody want to leak from which disease they are suffering from.For this reason, these organizations strive to publish the data such that it discloses as much statistical information as possible while preserving the privacy of individuals who contribute to the data. Data collection agencies publish information to facilitate research. The protection of individual privacy is much important while publishing the data. From the recent studies it shows that the adversary may possess a lot of extra knowledge called background knowledge about the individuals. The knowledge of the adversary and the algorithm used for protecting the privacy may lead to loss of much more information from the published table. In order to preserve privacy at the same time balancing the utility is a difficult task. Therefore, all the mechanisms try to minimise the level of anonymization thus becoming a reason to launch attacks and such kind of attack is called minimality attack.

In this thesis work, we devise an algorithm to provide a feasible solution against Minimality Attack. The algorithm is built on k-anonymity principle and l-diversity principle. The algorithm mainly concentrates on removing the attack despite the attack being present in many existing algorithms. We experiment our algorithm on medical data set which available on the public repository.

List of Tables

List of Figures

CHAPTER 1
Introduction

Privacy-preserving data publishing has become much of concern in the recent years. Micro data contains the records which provide information of a specific entity or an individual.For example hospital data contains information like name, age, gender,zip code and the disease of the particular individual. Publishing the data for any purposes such as research should be done in such a way an adversary must not gain any information about an individual whose data is present in the published data set. The micro data generally consists of number of attributes which are classified into 3 parts namely identifying attributes, quasi identifiers and sensitive attributes. Identifying attributes are those which give unique information about the individual. Quasi identifiers give partial information, such as zip code, birth date , gender. Sensitive attributes are those which leak the private data of an individual. The following figure shows an example of hospital database mechanism.

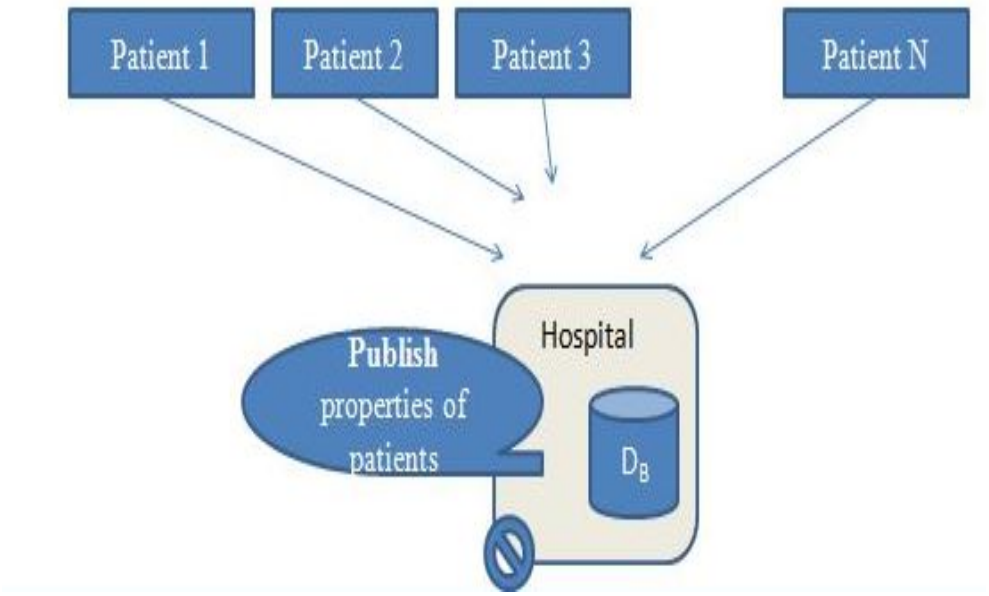


Figure 1.1: Example of a Hospital Database

For example disease can be considered as a sensitive attribute in case of medical data. So, publishing this micro data provides the researchers and the policy-makers analyze the data and learn some information which benefits the society, such as factors causing the diseases, effective of the medicine or treatment etc., Thus publishing the micro data gives some utility for the society.

At the same time, the published micro data should not damage the privacy of the individual. Thus, the micro data should be made public with the loss of privacy level which is acceptable. This can be done with the help of anonymization. The main goal of anonymization algorithm is to apply some anonymizing operations to protect the privacy of the micro data set. For example generalization and suppression are such techniques. With in the last decade a large number of anonymization algorithms have been described [1],[2],[3],[4]. But, there has been lot of limitations for these proposed algorithms. There are many class of attacks built on the minimality principle which is primarily known as Minimality Attack[5].

This attack assumes that the anonymization mechanism and its parameters are priorly known which is a sensible assumption, in accordance with many similar assumptions made in the world of security. Various anonymization methods publish a data set which encodes in a way not directly expressed in to number of possible worlds, among which each one being a candidate for the original input. It is assumed that each candidate is feasible, and by enforcing properties of these sets of possible worlds, the algorithm concludes that the attacker's ability to infer any facts about the original data is limited. Most of anonymization mechanisms try to reduce or minimize the information loss; more precisely, the algorithms should not generalize,distort or suppress the microdata more than the necessary requirement to achieve the privacy model. However, the minimization allows a "minimality attack" to argue that some of these candidates are infeasible, had the algorithm been executed on these inputs, the output would have been different. Hence, by ruling out candidates, the adversary's beliefs about the input can violate the claimed privacy requirements. For example, the simple l-diversity[5] requirement insists that the adversary cannot link any sensitive value to any particular individual with certainity greater than $1/l$, but under minimality attack, this confidence may become greater than $1/l$.

This attack was proposed by Wong et al.[6], where huge number of algorithms which require particular privacy guarantees were shown to be defenseless to this attack. Indeed, in some situations, an adversary can withdraw some credibly un-

seen facts with confidence!

Minimality Principle:

Any privacy preserving algorithm should not generalize,distort or suppress the microdata more than the necessary requirement to achieve the privacy model.

Example:

"Suppose A is an anonymization algorithm for a privacy requirement R which follows the minimality principle. Let table *T** be a table generated by *A* and *T** satisfies *R*. Then, for any *QID-EC X* in *T**, there is no specialization (reverse of generalization) of the *QID*'s in X which results in another table *T′* which also satisfies *R*[6]."

By **privacy** we mean the right of the individual to choose which kind of information involving himself he wants to share and with whom and when he wants to share it. By **security** we generally mean the level of protection we provide to control access to certain information. In some cases, that might mean complete isolation of information (no-one can access it), while in others there might be specific criteria allowing certain entities access, at particular times, etc. Formally privacy can be stated as follows:

The claim of an individual,groups or institutions to determine themselves when, how and to what extent information about them is communicated to others.

Data privacy, also called information privacy, is the aspect of information technology (IT) that deals with the ability an organization or an individual has to determine what data can be shared with third parties.

Unless we know the, quantifiable notion of what extent information is disclosed, we can not clearly check if some method for disseminating information breaches privacy or not.They are:

1. What are the privacy implications of sharing data?
2. What are the conditions for ensuring privacy?
3. Do algorithms that satisfy these privacy conditions retain useful information about the personal data?

The main goal of privacy preservation measures is to secure access to confidential information while at the same time releasing aggregate information to the

public. This can be measured as the probability of a data item being accessed, the change in knowledge of an adversary upon seeing the data, and so on. Though there are lot of privacy ensuring methods proposed, there are many kinds of attacks possible on them(shown in the following section).

1.1 Problem Definition

To devise an anonymization algorithm which provides mitigation against minimality attack.

Assumptions:

1. This algorithm targets mainly on the privacy breach called minimality attack.
2. Here we do not pre-specify any coarsening model.
3. The parameter value l is also not pre-specified. It depends on the Publisher of the dataset to set the value of l based on the privacy requirement.
4. Adverserial Knowledge:
   - Adversary knows the goal of l-diversity.
   - Adversary has the external table T^e, for example the voter registration list that maps QID's to individuals.
5. No two tuples in the table map to the same individual.

1.2 Contributions and Organisation

Background Knowledge on k-anonymity, l-diversity and the privacy breaches identified on them, the nature of minimality attack are presented in chapter 2. In chapter 3, the already existing privacy models and how minimality attack can take place on them, though they are further anonymized is shown.In chapter 4, the main reasons for the attack are identified. In chapter 5,the privacy and the utility metrics are explained. Chapter 6, provides a solution for the addressed problem with the help of an example

CHAPTER 2
Literature Review

What are the privacy implications of sharing data?

What are the **conditions for enforcing privacy?**

Do algorithms that satisfy privacy conditions **retain useful information**?

These two points are mainly important because the whole point of anonymization is to extract useful information from these privacy preserving algorithms without breaching the privacy. The two main requirements of data publishing are:

1. publish information that discloses as much statistical information as possible
2. preserves the privacy of individuals contributing to the data

Since we need to disclose as much statistical information as possible we need to publish the micro data that is the unaggregated data.

Name	**SSN**	**Zipcode**	**Age**	**Nationality**	**Disease**
Samantha	1504	34367	24	Russian	Heart Disease
Ruth	1821	14567	23	American	cancer
Prabhu	1503	24098	31	Russian	CAD
Anushka	6712	24098	27	Japanese	Gastric ulcer
Sharma	1980	34567	34	Indian	Flu
Rakul	3457	34567	20	Indian	Flu
Preet	1222	10972	18	Russian	Heart Disease
Singh	1545	10972	25	Japanese	Heart Disease
Sneha	1343	34567	32	Russian	Breast cancer

Table 2.1: Microdata: the unaggregated data

This is an example where in each individual is represented by an unique tuple. Some of these attributes like Name and SSN are identifying attributes because they can easily identify the individuals from the population. Disease is the sensitive information which an individual wants to keep secret. In all the cases the only anonymization done is removing these identifying attributes. Because in order to guarantee privacy we should not disclose the link between the identifying attributes and the sensitive attribute and the one way to do is removing them.

Zipcode	**Age**	**Nationality**	**Disease**
34367	24	Russian	Heart Disease
14567	23	American	cancer
24098	31	Russian	CAD
24098	27	Japanese	Gastric ulcer
34567	34	Indian	Flu
34567	20	Indian	Flu
10972	18	Russian	Heart Disease
10972	25	Japanese	Heart Disease
34567	32	Russian	Breast cancer

Table 2.2: Microdata: with removed identifying attributes

But eliminating the identifying attributes is not the solution to overcome the existing privacy breaches. Even that can cause serious breaches of privacy. The Massachusetts governor William Weld was exactly identified inspite the identifying attributes being removed from the published data by re-identifying the individuals by professor sweeney[7] in 2002. Governor Weld lived in Cambridge Massachusetts. She took supposedly anonymous medical data which contains records of massachusetts employees in which name and SSN were removed. She joined it with the voter registration list of massachusetts. She found out that the governor of massachusetts was uniquely determined with the combination of Zipcode, Birthdate and gender. Henceforth, there is only record corresponding to the governor in both the medical data and the voter registration list and she was easily able to link the name to the diagnosis which constituted the privacy breach.This can be named as the re-identification attack or a linking attack because the governor was uniquely found by linking the identifiers from both the lists available. In fact she found out that 87% of the United States population can be uniquely determined with these three attributes, such a combination of attributes which uniquely identify the individual are called as quasi-identifiers QI's. So inorder to guarantee privacy against such linking attacks she proposed a scheme called *k*-anonymity[7].

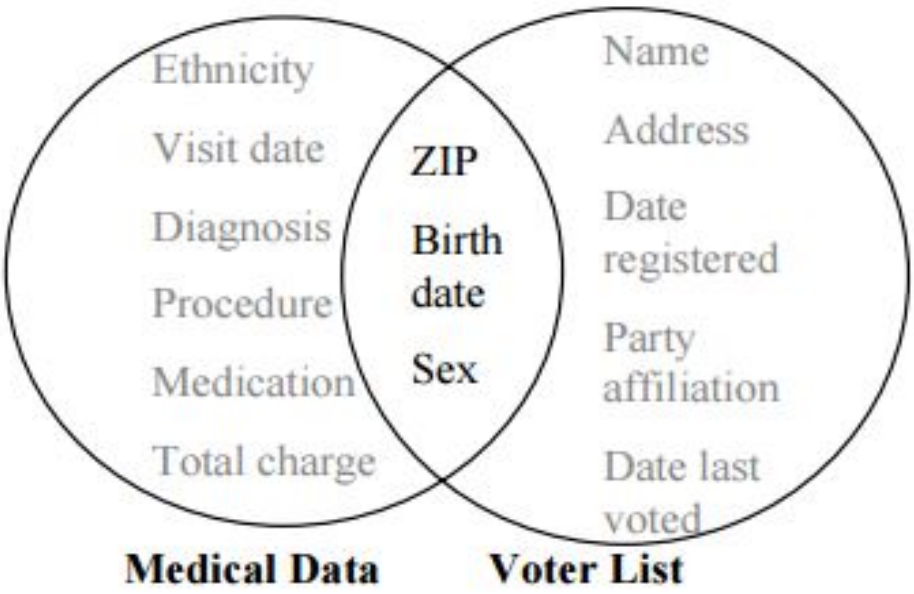


Figure 2.1: Linking to re-identify data

2.1 *k*-anonymity

Definition: A table T is said to have *k*-anonymity property if a tuple in the table shares its quasi identifiers with atleast *k*-1 tuples in the table[**?**].(The quasi identifiers are coarsened to have the mentioned property). In other words

A table T^* is said to be *k*-anonymous table if each SELECT COUNT(*) FROM T^* GROUP BY QID is $\geq k$.

Here the parameter k implies the degree of anonymity. The following two tables show the actual and the *k*-anonymous variant of the actual table. Here *k*=4 and QI=zipcode, Age, Nationality

Zipcode	Age	Nationality	Disease
15075	28	Russian	Heart Disease
15068	29	American	Heart Disease
15043	21	Japanese	Viral Infection
15056	23	American	Viral Infection
17643	50	Indian	Cancer
17643	55	Russian	Heart Disease
17640	47	American	Viral Infection
17640	47	American	Viral Infection
15075	31	American	Cancer
15075	37	Indian	Cancer
15068	36	Japanese	Cancer
15068	35	American	Cancer

Table 2.3: Inpatient Microdata

Zipcode	Age	Nationality	Disease
150**	≤ 30	*	Heart Disease
150**	≤ 30	*	Heart Disease
150**	≤ 30	*	Viral Infection
150**	≤ 30	*	Viral Infection
176**	≥ 40	*	Cancer
176**	≥ 40	*	Heart Disease
176**	≥ 40	*	Viral Infection
176**	≥ 40	*	Viral Infection
150**	3*	*	Cancer
150**	3*	*	Cancer
150**	3*	*	Cancer
150**	3*	*	Cancer

Table 2.4: 4-anonymous Inpatient Microdata

The formation of *k*-anonymous table is a very simple technique. It takes these Quasi identifier attributes and coarsens them such that every tuple in the table shares it's Quasi identifier values with atleast *k-1* other tuples in the table. For example Table 2.4 is a 4-anonymous table, clearly from this table you can not identify any individual with the help of the Quasi identifier attributes. Thus, this is a good quantifiable notion of privacy. Single dimension recoding, Incognito, Mondrain are some of the algorithms for achieving *k*-anonymity. The adversary might have a lot of background knowledge which leads to loss of sensitive information of the individual which is explained in the following section.

2.2 Attacks on *k*-anonymity

OBSERVATION:

Even though *k*-anonymity anonymizes data, still adversary can infer important information from the data which leads to attacks like Homogeneity attack and the other one being Background knowledge attack.

HOMOGENEITY ATTACK[8]: Suppose A is the enemy of B and B's medical status is present in Table 2.4, which A want to infer from the table.And now the zipcode and age of B are known to A i.e., zipcode is 15075 and the age is 31.

Zipcode	**Age**	**Nationality**	**Name**
15075	**32**	**Indian**	**Anushka**

Table 2.5: Adversary Knowledge

So with the help of this knowledge A can eliminate the records from 1 to 8 and can easily identify that the B's record belong to one of the tuples of record no. 9,10,11,12. And the all the records from 9 to 12 have the same sensitive attribute cancer. So A can easily conclude that B has Cancer with out any dilemma. Thus, this implies that the groups which are formed by *k*-anonymity procedure can be responsible for the information leakage. The main reason for the above is the lack of the diversity among the sensitive tuples of the sensitive attributes in the group. That is all the tuples in the group formed are having the same sensitive attribute. Thus this attack can be named as the Homogeneity attack[8] besides the reason being the homogenous nature of the sensitive attributes. From this attack one can conclude that the diverse nature of the sensitive attributes in the groups formed by the *k*-anonymity procedure is necessary. This is one of the attack which makes a drawback in the *k*-anonymity proposed by professor sweeney. The following table shows the group in which the attack happened. Thus the adversary has chance to guess the sensitive attribute with full probability inspite the table being anonymized.

Zipcode	Age	Nationality	Name
150**	3*	*	Cancer
150**	3*	*	Cancer
150**	3*	*	Cancer
150**	3*	*	Cancer

Table 2.6: Homogeneity Attack

BACKGROUND KNOWLEDGE ATTACK:

Now let us take an other scenario where C and D are neighbors. C wants to know from which disease D is suffering from that is the private information of D which is the medical status. Table 2.4 shows the 4-anonymous inpatient microdata which obeys *k*-anonymity. So, now for identifying the record of D from the published table, C finds three more records with the same quasi identifiers, thus ending up with four choices in order to infer D's medical status. Now let us think the aggressive neighbor of D has the information of zipcode, age and nationality.

Zipcode	Age	Nationality	Name
15083	24	Japanese	Umeko

Table 2.7: Adversary Knowledge

Thus with the help of these quasi identifiers as the grouping knowledge he finds out or concludes that the D's record is present in records 1,2,3,4. But here D has two options for identifying. Here C uses his background knowledge or the medical claim that japanese are less prone heart disease depending on the appetite they take. Thus C can conclude that his neighbor D is suffering from viral infection.

Zipcode	Age	Nationality	Disease
150**	≤ 30	*	Heart Disease
150**	≤ 30	*	Heart Disease
150**	≤ 30	*	Viral Infection
150**	≤ 30	*	Viral Infection

Table 2.8: Background Knowledge Attack1

Similarly, there is an other background knowledge attack[8] in the above table. If the adversary has the following knowledge of zipcode, age, nationality he can find that the individual is present in the records 5,6,7,8. Now he has totally three options of diseases because there are three distinct diseases in the group which the individual falls.

Zipcode	Age	Nationality	Name
17026	50	Russian	Rakul

Table 2.9: Adversary Knowledge

Now with the help of the background knowledge that the individual has low blood pressure and she won't take fatty meals, the adversary can conclude that the individual is suffering with heart disease by eliminating the other two options which he/she has.

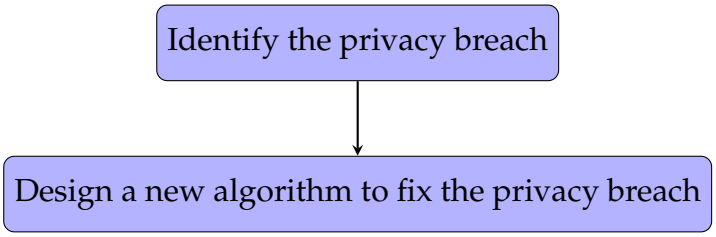
Zipcode	Age	Nationality	Disease
176**	≥ 40	*	Cancer
176**	≥ 40	*	Heart Disease
176**	≥ 40	*	Viral Infection
176**	≥ 40	*	Viral Infection

Table 2.10: Background Knowledge Attack2

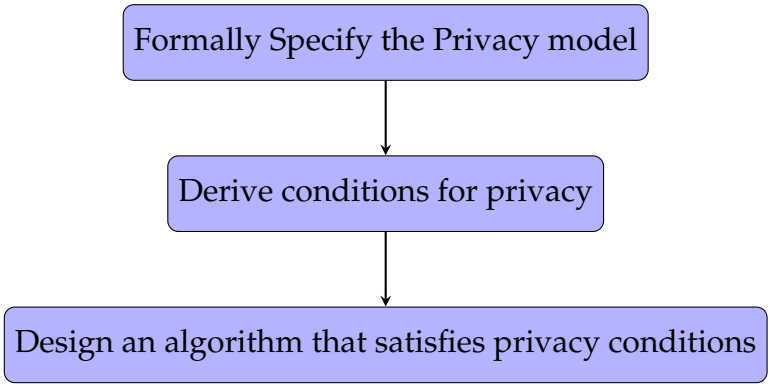
Thus even though *k*-anonymity anonymizes the data in order to provide privacy, still it is not able to eliminate the privacy breaches namely Homogeneity attack and the other one being Background Knowledge attack.

How can one ensure the privacy of a published data?

A data publisher may not be able to enumerate all the possible privacy breaches. Even if he enumerate a list of them, he might not be able to know what other privacy breaches are might be possible.



We need to have a more formal way to reason about privacy which is explained with the help of following flowchart which says that first formalizing and then deriving the conditions and then forming the algorithm which satisfies the required privacy conditions is a suitable way for the design.



2.3 *l*-diversity

The *l*-diversity Principle says that

"An equivalence class is said to have *l*-diversity if there are at least *l* "well-represented" values for the sensitive attribute[5]."

A table is said to have *l*-diversity if every equivalence class of the table satisfies *l*-diversity property. In other words

Every group of tuples with the same QID- value has $\geq$ *L* distinct sensitive values of equal proportion.

If we recall attacks on *k*-anonymity are mainly due to the homogeneity nature of the sensitive attributes of the same class. So if we eliminate that nature that if we have diverse nature of sensitive attributes in the class then there is no chance of having such kind of attacks. For example if we consider the table 2.4 it doesn't have this diverse nature. Let us take a different table which has the above property.

Zipcode	**Age**	**Nationality**	**Disease**
150**	$\leq$ 30	*	Heart Disease
150**	$\leq$ 30	*	Flu
150**	$\leq$ 30	*	Viral Infection
150**	$\leq$ 30	*	Viral Infection
176**	$\geq$ 40	*	Cancer
176**	$\geq$ 40	*	Heart Disease
176**	$\geq$ 40	*	Viral Infection
176**	$\geq$ 40	*	Viral Infection
150**	3*	*	Hear Disease
150**	3*	*	Flu
150**	3*	*	Cancer
150**	3*	*	Cancer

Table 2.11: 4-anonymous & 3-diverse Inpatient Microdata

Thus even if the adversary has the two below two pieces of information from table 2.12 which he searches for in table 2.11, he can not identify or guess the sensitive attribute. Because with the firstpiece information even if he gets to know in which class the individual is present he can find because they aren't homogenous sensitive attributes in the records 9,10,11,12. And from the second piece of information which he goes for the first four records, even if he knows that japanese are less prone to Heart Disease, but still he has two more diseases to guess. Thus from a *l*-diverse we can eliminate the attacks from which *k*-anonymity suffers.

Zipcode	**Age**	**Nationality**	**Name**
15083	24	Japanese	Umeko
17026	50	Russian	Rakul

Table 2.12: Adversary Knowledge

2.4 Attacks on *l*-diversity

MINIMALITY ATTACK :

If we recall the minimality principle, it states that

Any privacy preserving algorithm should not generalize,distort or suppress the microdata more than the necessary requirement to achieve the privacy model. Now let us assume that the quasi identifiers q_1 and q_2 can be anonymized or generalised to Q and here we take only disease from the table as the sensitive attribute, amongst which flu is the only sensitive value. Let us consider an example, Jan 1987, Z1234,F be the values of q_1 and Feb 1987, Z1234, F be the values of q_2. And these both can be generalised to Q as Jan/Feb 1987, Z1234, F. Now any tuple which is bearing flu as its sensitive attribute is said to be the sensitive tuple. Now let us the goal of 2-diversity. According to *l*-diversity, for every QID-EC at most only half of the tuples must be sensitive(in the scenario of sensitive and non-sensitive). All the approaches which are existing follow the MINIMALITY principle i.e., "For any anonymization algorithm, it is a prerequisite to define some notion minimality notion." Thus intuitively *l*-diversity should not anonymize, distort or suppress the data exceeding the necessity to achieve *l*-diversity. According to the Minimality principle table 2.13 doesn't need to be generalised more because it satisfies the criteria.

Now consider a variation of that table, that is consider table 2.14. Here the QID-EC q_1 violates the 2-diversity principle the reason being the proportion of flu in the group exceeding the limit. Thus, this requires to be generalised more. By

doing so, we can get an of the one's shown i the table 2.15, which are the global and local recoding algorithm[4] results. Global recoding recodes all the values of an attribute value to the same value. According to local recoding occurences of the same attribute value can be recoded to different attribute values. Thus these locally and globally anonymised or recoded tables satisfies the 2-diversity principle. But the question here is does this anonymization protect the privacy of the individual. Suprisingly the answer is NO. Since the adversarial knowledge includes the external table such as voter registration list that maps the individuals to the QID's and adversary knows the goal of 2-diversity. The published table is also available now to the adversary. Now the adversary draws the attack as follows:

- Since the anonymization follows minimality principle, from the table 2.15 he comes to know that there are only two tuples which are sensitive in the total table.
- From the table available externally he can see that there are only 2 records in the EC q_1 and 5 records in the EC q_2.
- So, if both the sensitive tuples are present in the EC q_2 then the proportion of sensitive tuples will be only $2/5$.
- So, there won't be an necessity to anonymize this EC.
- But, still since the anonymization is done, then definitely it might be because of q_1. Thus both the sensitive tuples are present in q_1 thus launching the attack.

QID	**Disease**
Q1	Flu
Q1	Non-sensitive
Q2	Flu
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive

Table 2.13: Good Table

QID	Disease
Q1	Flu
Q1	Flu
Q2	Flu
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive

Table 2.14: Bad Table

Local	Global
Q	Q
Q	Q
Q	Q
Q	Q
Q2	Q
Q2	Q
Q2	Q

Table 2.15: Local and Global Recoding

The damage here occurs mainly due to the taken parameter value of l= 2 is so small. In many real time practical examples, smaller values of l are not used. Only larger values of l are used.

CHAPTER 3
Existing Schemes

Minimality attacks has been shown successful on a variety of anonymization models.

3.1 Recursive (*c*,*l*)-diversity

A table satisfies recursive (*c*,*l*)-diversity if every QID-EC in the table satisfies the property. Any QID-EC meet the requirement to satisfy recursive (*c*,*l*)-diversity if it satisfies the following[5]:

let the number of occurrences of the most sensitive value be v, if we eliminate the succeeding *l*-2 top frequent sensitive records, then c times the total count of the remaining records in the group must be greater than the value of c.

Now take the bad table and the global recoding for the bad table. Let us consider the goal of recursive (3,3)-diversity. The adversary having the knowledge that the anonymization follows minimization principle, from the table 3.2 adversary steps out that the q_2 EC satisfies the required diversity and that of the q_1 EC violates it and thus it must contain two diabetes. Thus the intended requirement is not satisfied showing the adversary can launch attack on the algorithm of recursive (*c*,*l*)-diversity.

QID	**Disease**
Q1	Heart Disease
Q1	Diabetes
Q2	HIV
Q2	Diabetes
Q2	CAD
Q2	Gallstones
Q2	Breast cancer

Table 3.1: Good Table

QID	Disease
Q1	Heart Disease
Q1	Diabetes
Q1	Diabetes
Q2	HIV
Q2	CAD
Q2	Gallstones
Q2	Breast Cancer

Table 3.2: Bad Table

Local	Global
Q	Q
Q	Q
Q	Q
Q	Q
Q2	Q
Q2	Q
Q2	Q

Table 3.3: Local and Global Recoding

3.2 *(k,e)*-anonymity

The scheme (k,e)-anonymity [10] concentrates mainly on the tables which have sensitive attributes that are numeric, for anonymiation mechanism.

It forms a table where every QID-EC size is at least k and where the numeric sensitive values range is atleast e.

In the tables in Table , we show the bucketization in terms of *QID* values, the individuals with the same *QID* value are in the same bucket. Consider the tables (where Income is a sensitive numeric attribute). From table(3.8), the adversary concludes that the records of EC q_1 are violating (*k,e*)-anonymity and both of them must be bearing 90k as their sensitive values, thus opening the door for the minimality attack. We can draw a similar argument even for local recoding.

QID	Salary
Q1	90k
Q1	80k
Q2	90k
Q2	70k
Q2	100k

Table 3.4: Good Table

QID	Salary
Q1	90k
Q1	90k
Q2	80k
Q2	70k
Q2	100k

Table 3.5: Bad Table

Local	Global
Q	Q
Q	Q
Q	Q
Q2	Q
Q2	Q

Table 3.6: Local and Global Recoding

3.3 *t*-closeness

A table T satisfies t-closeness[9] if , "the distribution P of every QID-EC in the table T is roughly equal to the distribution Q of the complete table T with reference to the sensitive attribute. More specifically, the difference between the each equivalence class distribution in T and the whole table distribution T, denoted by *D[P,Q]*, is a value not exceeding t".

From the explanation in [9], $D[P,Q] = 1/2\sum |pi - qi|\ 1 \leq i \leq m$. Consider global recoding table 3.6. For each possible sensitive value distribution P for *QID-EC q2*, the adversary computes *D[P,Q]*.From Table(3.5), the adversary concludes that the distribution *D[P,Q]* is everytime lesser than 0.2. Hence we need anonymization on the EC *q1*. Thus the adversary concludes that the both records present in the EC q_1 suffers from the disease flu, thus opening the door for the minimality attack. Similar explanation can be given to local recoding also.

QID	Disease
Q1	Flu
Q1	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Flu
Q2	Flu

Table 3.7: Good Table

QID	Disease
Q1	Flu
Q1	Flu
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Flu

Table 3.8: Bad Table

Local	Global
Q	Q
Q	Q
Q	Q
Q2	Q
Q2	Q
Q2	Q

Table 3.9: Local and Global Recoding

3.4 The Flow of Privacy Braches

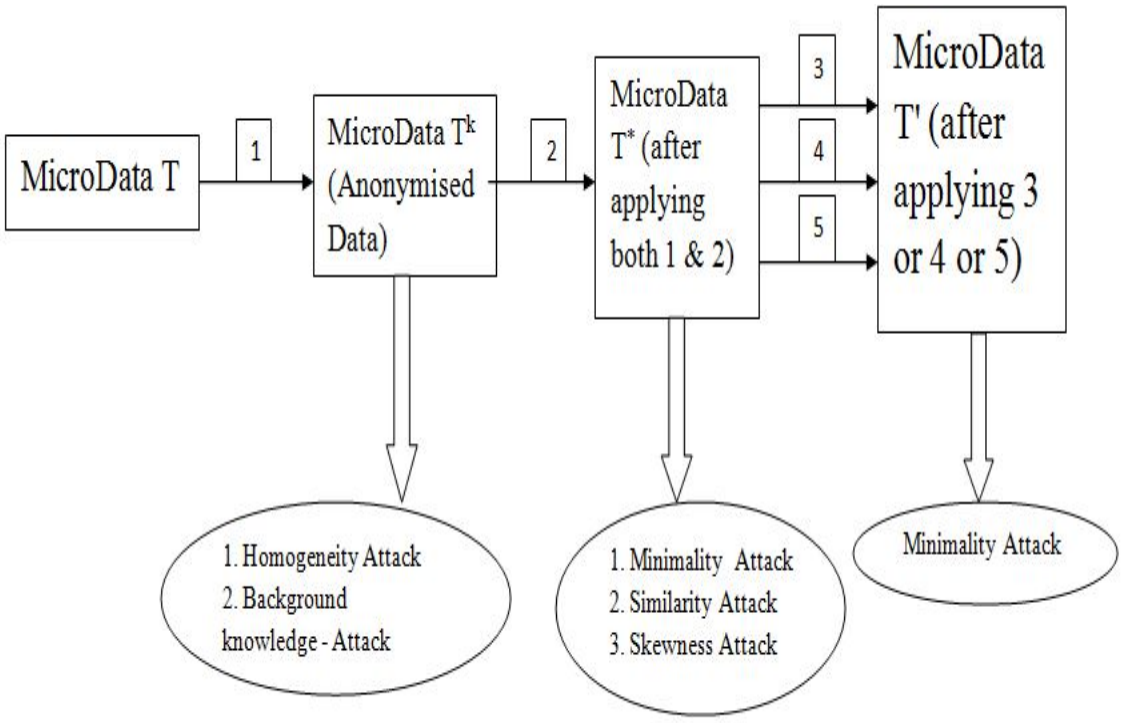


Figure 3.1: Figure representing the privacy breaches on the existing algorithms 1.*k*-anonymity, 2. *l*-diversity, 3. *t*-closeness 4. *(k,e)*-anonymity, 5. recursive*(c,l)*-diversity

The above figure can be drawn as a conclusion of the total literature survey done. The figure shows us that MINIMALITY ATTACK is possible in all the available anonymization algorithms which shows as a major attack.

CHAPTER 4
Defense against Minimality Attack

4.1 Reasons behind the Attack

1) Deterministic Behavior:

Deterministic behavior of the algorithm provides a plan for the adversary to work back from the table which is published, to know which possible inputs might suffice[11]. For example if the adversary has the knowledge of 2-diversity goal, then he can easily come to know which EC's of the table are anonumized more to achieve 2-diversity and might infer the EC with the possible attack.

2) Asymmetric Group Choices:

This is an important or a key observation made by an adversary by looking at the published table and the external table available with the adversary. If there are smaller groups in the table(external table) and in the published table if there are no such kind of groups(due to global or local recoding), the reason behind such decision to merge groups might be because, smaller groups are much more likely prone to the attack when compared with the larger groups[11]. Because violation of diversity constraint is more in smaller groups due to less number of records. From the below example, since there are six tuples in Q_2 and there are total of three sensitive attributes and even if all of them are present in second group indicates that there is no necessity for generalising. Since the table is globally recoded, the adversary can conclude that Q_1 has the attack possibility.

3) Consideration of QIs and SAs together:

If QI's and their respected SA's are together, then there is high chance for the adversary to restore the exact mapping with a lot confidence[11]. Algorithms which consider the above scenario leak information to a lot of extent. But removing the link between both of them leads to lack of attribute co-relation as in the case of bucketization[18]. Thus grouping or removing the link must be selectively chosen.

Explanation:

QID	Disease
Q1	HIV
Q1	HIV
Q2	HIV
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive

Table 4.1: Bad Table

Local	Global
Q	Q
Q	Q
Q	Q
Q	Q
Q2	Q
Q2	Q
Q2	Q
Q2	Q

Table 4.2: Explained reason for attack

From the Global Recoding table the adversary can analyse as follows:

- Global recoding is a deterministic algorithm and since the adversary has the knowledge of the goal of *l*-diversity he/she can easily construct the table back.
- The adversary can get to know that Q1 is a smaller group and is more prone to attack(by the reasoning of the adversary explained in the minimality attack) and thus global recoding is done.

4.2 Proposed Algorithm

Our algorithm is based on *k*-anonymity principle and *l*-diversity principle even though there are questions for the utility of *k*-anonymity. We use them because they have been successful in some practical applications. And these algorithms doesn't eliminate the attribute co-relation when compared to bucketization. And

here privacy of the individual is the main goal of the algorithm.

ADVERSARIAL MODEL:
Formally, in the attack by minimality principle, the attacker is believed to have the knowledge of

- The *QID* values of all records in the table i.e., the external table T^e which can include voter registration list,
- The anonymization algorithm used for forming the table (including the parameters),
- The published table.
- The attacker has at most *l*-2 negotiation statements(assumed in the case of *l*-diversity)
- The adversary can guess which QID-EC the particular individual belongs to.

GOAL OF THE ADVERSARY:
The main goal of the adversary is to infer the value of Sensitive attribute for a particular individual or for a particular *QI* value, and the attack effectiveness can be calculated based on the adversary's ability.

Algorithm 1 Defending Minimality Attack

INPUT Micro-data Table with n rows. $T = QI_1, \ldots, QI_j, SA_1, \ldots, SA_k$, $j > 0$, $k > 0$, $QI_1 \cap \ldots \cap QI_j = \Phi$ and $SA_1 \cap \ldots \cap SA_k = \Phi$. T
OUTPUT Micro data table T'

1: Create a table which follows the *k*- anonymity principe, T^k from the given table T.
2: **for** each QID-EC in T^k **do**
3: if $count_distinct$ of SA is less than a predefined value *l*,then
4: Add spurious records()
5: Else
6: *z*= the number of tuples in the EC
7: **for** each SA in EC **do**
8: if $\frac{count(SA)}{z}$ is less than $\frac{1}{l}$
9: exit from the loop
10: end if
11: Else
12: Add spurious records() till count(SA_i) is less than $\frac{1}{l}$
13: **end for**
14: end if
15: **end for**

Algorithm 2 Add spurious records

1: Create set $S = SA_1, \ldots, SA_j$
2: random(S) $\rightarrow$ z
3: if z is not in the QID list then
4: Add z to QID-EC
5: end if
6: if $countdistinct(SA) \geq l$ then
7: exit
8: else
9: Repeat the procedure from 2
10: end if

ILLUSTRATION:

The following table is formed by applying the steps in the above algorithm. From the table, the adversary can not get any knowledge, though he know the EC of the particular individual(or any other background knowledge).

Zipcode	**Age**	**Gender**	**Disease**
14034	24	Female	HIV
14089	23	Female	HIV
14056	29	Male	BloodCancer
13200	31	Male	HIV
13212	38	Female	HIV
15012	21	Female	Asthma
15078	21	Male	Heart Disease
15011	25	Female	CAD
15070	25	Female	Bronchitis
18120	43	Male	HIV
18171	49	Female	Heart Disease
18129	41	Male	HIV
18191	50	Male	Blood cancer

Table 4.3: Microdata of the patients

Zipcode	Age	Gender	Disease
140**	[21,30]	*	HIV
140**	[21,30]	*	HIV
140**	[21,30]	*	BloodCancer
132**	[31,40]	*	HIV
132**	[31,40]	*	HIV
150**	[21,30]	*	Asthma
150**	[21,30]	*	Heart Disease
150**	[21,30]	*	CAD
150**	[21,30]	*	Bronchitis
181**	[41,50]	*	HIV
181**	[41,50]	*	Heart Disease
181**	[41,50]	*	HIV
181**	[41,50]	*	Blood cancer

Table 4.4: Microdata: with applied *k*-anonymity

Zipcode	Age	Gender	Disease
140**	[21,30]	*	HIV
140**	[21,30]	*	HIV
140**	[21,30]	*	BloodCancer
140**	[21,30]	*	Heart Disease
132**	[31,40]	*	HIV
132**	[31,40]	*	HIV
132**	[31,40]	*	Heart Disease
132**	[31,40]	*	CAD
150**	[21,30]	*	Asthma
150**	[21,30]	*	Heart Disease
150**	[21,30]	*	CAD
150**	[21,30]	*	Bronchitis
181**	[41,50]	*	HIV
181**	[41,50]	*	Heart Disease
181**	[41,50]	*	HIV
181**	[41,50]	*	Blood cancer

Table 4.5: Microdata: with the applied defense

Thus from the above table the following conclusions can be drawn which eliminates the attack.

1. The table satisfies *l*-diversity.

2. The table didn't undergo any kind of recoding, which doesn't give the adversary a chance to attack on smaller groups.

3. Addition of spurious records disguises the adversary thus making the table strong.

Below is the example where the table is having both sensitive and non-sensitive attributes given below with the help of following tables. Since there is only one disease HIV in the table the only option for the parameter l is 2.

QID	**Disease**
Q1	HIV
Q1	HIV
Q2	HIV
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive

Table 4.6: Table with both sensitive and nonsensitive attributes

QID	**Disease**
Q1	HIV
Q1	HIV
Q1	Non-sensitive
Q1	Non-sensitive
Q2	HIV
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive
Q2	Non-sensitive

Table 4.7: Newly formed Table with the removed attack

4.3 Privacy Proof and the Utility Metric

PRIVACY PROOF:

Here the privacy metric of the table is explained with the help of the following privacy proof. Now let us take table 4.4 as an example

QID	Zipcode	Age	Gender	Disease
Q1	140**	[21,30]	*	HIV
Q1	140**	[21,30]	*	HIV
Q1	140**	[21,30]	*	BloodCancer
Q2	132**	[31,40]	*	HIV
Q2	132**	[31,40]	*	HIV
Q3	150**	[21,30]	*	Asthma
Q3	150**	[21,30]	*	Heart Disease
Q3	150**	[21,30]	*	CAD
Q3	150**	[21,30]	*	Bronchitis
Q4	181**	[41,50]	*	HIV
Q4	181**	[41,50]	*	Heart Disease
Q4	181**	[41,50]	*	HIV
Q4	181**	[41,50]	*	Blood cancer

Table 4.8: Microdata: Having Minimality Attack

From the table clearly Q2 suffers from the attack. Since both the sensitive attributes are same. The adversary since have the knowledge about which group the particular individual belongs to and also the external table T^e (from the adversarial Model), he easily launches the attack. Now the following table shows the result when local recoding(the normally followed mechanism in all the schemes explained in chapter 3) is applied and when the proposed scheme is applied.

QID	Disease
Q1	HIV
Q1	HIV
Q1	BloodCancer
Q	HIV
Q	HIV
Q	Asthma
Q	Heart Disease
Q	CAD
Q	Bronchitis
Q4	HIV
Q4	Heart Disease
Q4	HIV
Q4	Blood cancer

Table 4.9: Microdata: Local Recoding

If any scheme is taken in to consideration, if the table is not satisfying the criterion prior, then the scheme follows the recoding process which might be

one among the local or global technique. Now the adversary looks at the newly formed table and finds that the EC's Q2 and Q3 are merged to Q. The adversary reasons as, the repeated sensitive attribute is HIV and since Q4 has 4 tuples,even if both the HIV's are present in Q3, it still satisfies the goal of 3-diversity. Thus the problematic EC is Q2, and he conforms that both the sensitive values are from HIV and thus launching the attack. Now appling the proposed defense on the table, the output will be the following:

QID	**Disease**
Q1	HIV
Q1	HIV
Q1	BloodCancer
Q1	Heart Disease
Q2	HIV
Q2	HIV
Q2	Heart Disease
Q2	CAD
Q3	Asthma
Q3	Heart Disease
Q3	CAD
Q3	Bronchitis
Q4	HIV
Q4	Heart Disease
Q4	HIV
Q4	Blood cancer

Table 4.10: Microdata: with the applied defense

Now the adversary looks at the newly published table and compares it with the external table available to him. The main aim here is to disguise the adversary, so that he can exactly map both the tables. since spurious records have been added to more than one group, he can not identify the noise and he doesn't even know the number of spurious records added because generally an external table is always a super of the published table(Sensible assumption). Thus the adversary can't launch the attack by looking at the table and the background knowledge he/she has. Thus we can prove that launching the attak on the table which is formed from our scheme is difficult. More specifically, we can say that the formed table is free from the reasons which we have mentioned above.

1. The anonymization mechanism is not deterministic due to the addition of spurious in random fashion.

2. The groups formed at the end of the algorithm are of almost equal sizes.

UTILITY:

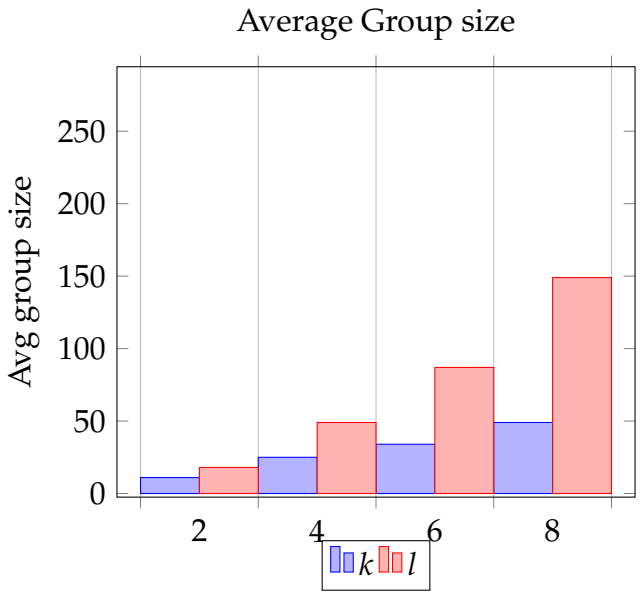
There are many metrics to measure the utility of a published table. The utility here calculated based on the aggregate query answering which has been used widely for calculating the data utility[12]. Here we incorporate or choose Average Relative Error to measure the utility of the published table because the table formed by algorithm includes spurious records which is treated as noise. The "count" operator is taken in to consideration when sensitive attribute is contained in the predicate of the query. For every query, it is queried or run on original table and also the anonymized one. The number of records obtained from the original table, that is the count constitutes the *act_count* and the number of records that are the output of the query from the anonymized table form the *rec_count* that is the reconstructed count.The average relative error is calculated with the help of the following equation:

$$|\textbf{act\_count} - \textbf{rec\_count}| / \textbf{act\_count}$$

The more the average relative error indicates less utility. Because if the error is more it indicates that there is lot of noise or the generalization has exceeded the minimality limit. Thus for any anonymization algorithm, we require less relative error.

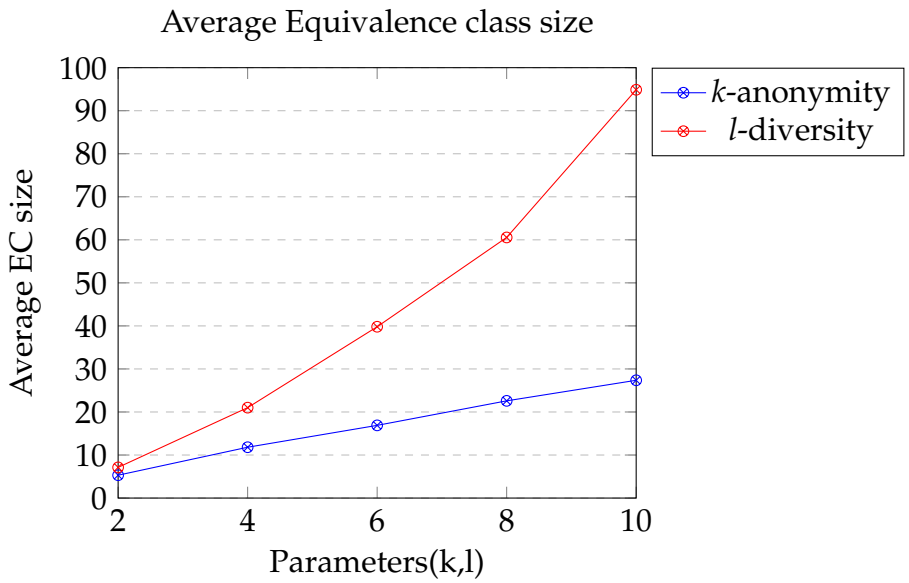
CHAPTER 5
Experimental Results

We measure the metrics like loss, average equivalence class size, non-uniform entropy, for both *k*-anonymity and *l*-diversity since our scheme is build upon these two models. We performed these two models with different values of *k*, *l* on adult data set[14] in ARX tool to get the above metrics. We have done so to identify which would be the suitable parameters to apply, for getting good amount off privacy and utility. We measure the metrics like loss, average equivalence class size, non-uniform entropy, for both *k*-anonymity and *l*-diversity since our scheme is build upon these two models. We performed these two models with different values of *k*, *l* on adult data set[14] in ARX tool to get the above metrics. The following is a basic graph which indicates that increase in the parameter value increases the size of the QID-EC.

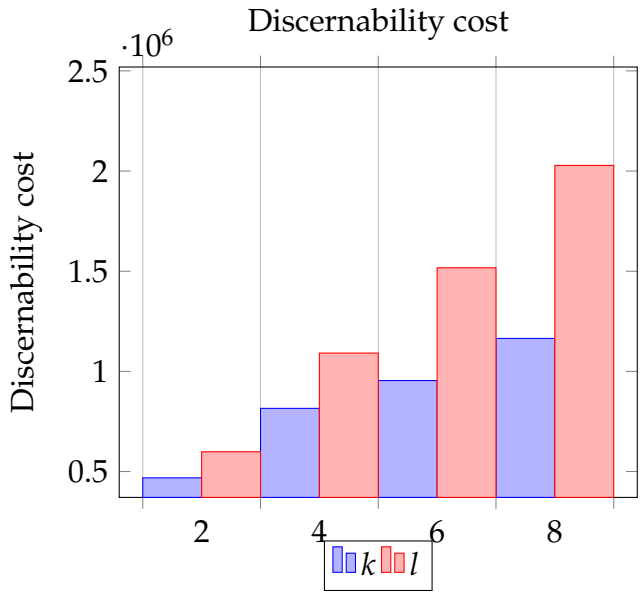


Average Equivalence class size metric is an utility metric which measures the information loss based on the depending on the equivalence classes size in the table formed after the transformation[13]. The following graph shows the comparison of Average EC metric for both the algorithms. This metric also measures

the quality of the anonymization technique based on the EC size.

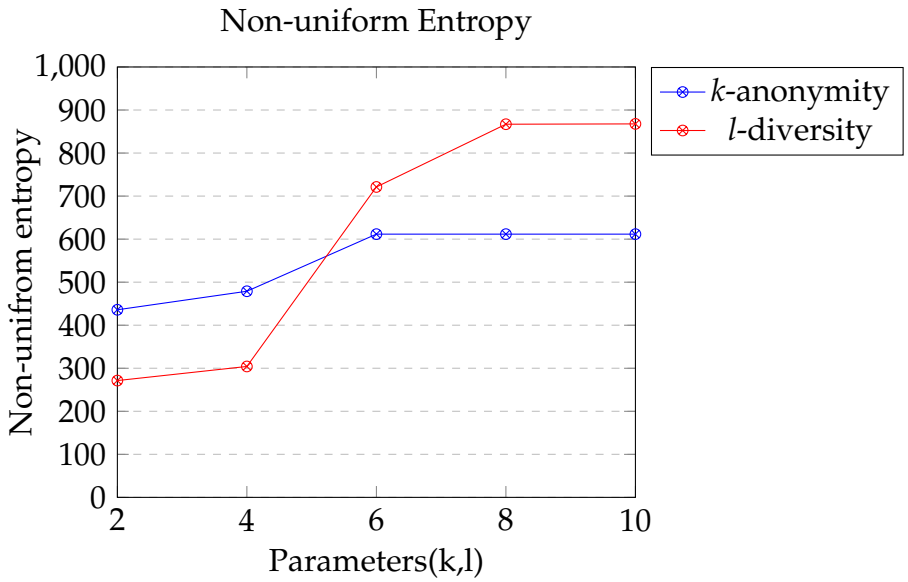


Discernibility cost metric also measures the quality of the anonymization technique based on the equivalence classes size. Discernibility cost metric assigns a penalty to every record in the table depending on the number of tuples which can not be distinguished from other records in the transformed table which is published[15].



Non-uniform Entropy metric is an utility metric which measures the loss of information based on the entropy loss, i.e., information content. It utilizes the mutual information concept to quantify the amount of information which can be

obtained about the original variables in the input dataset by observing the variables in the output dataset[16]. The comparison results of this metric on both the algorithms *k*-anonymity and *l*-diversity is shown below.



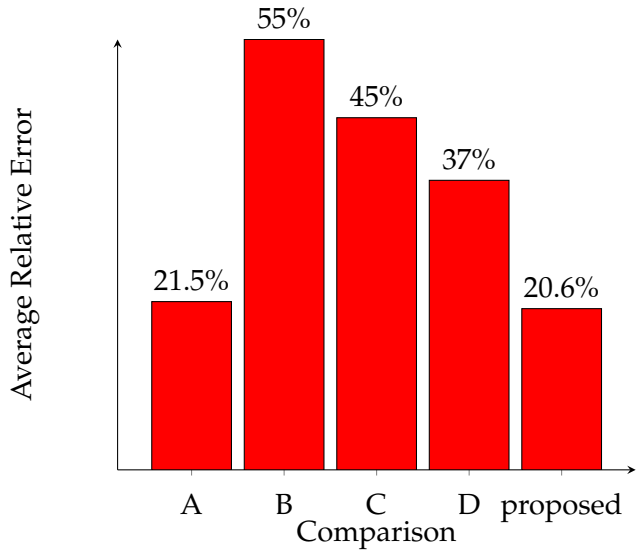
An attack on the published table can be said as a successful attack if re-identification can be done on larger portion of records in the table[17]. Loss is the measure which "summarizes the coverage of the domain of an attribute" i.e., general loss of information. The following table shows the trade off between the re-identification risks and the loss metric which explains that as the risk of identifying the individual decreases resulting in the rise of the utility loss.

Re-identification Risk (%) and Loss				
parameters(k,l)	***k*-anonymity**	***l*-diversity**	***k*-anonymity**	***l*-diversity**
2	**50**	**50**	**0.211**	**0.167**
4	**25**	**25**	**0.271**	**0.235**
6	**7.69**	**10**	**0.297**	**0.288**
8	**7.69**	**0.9009**	**0.315**	**0.318**
10	**7.69**	**0.6993**	**0.339**	**0.388**

Table 5.1: Trade off between Re-identification risks and Loss

From all the above metrics, we come to a conclusion that if the parameter value of *k* and *l* increases the corresponding values of utility is being decreased. The main reason being the increase in the level of anonymization with higher parameter values. Hence it is advisable to take to take the values between [3,6] depending on the size of data set and also the nature of the data set, reason being most of the real time data is sparse and also skewed in nature.

In this work, we have taken the value of the parameter values k and l as 3. We measure the utility of the current scheme and compare it with the other existing schemes.



From the above graph we can say that the utility of the proposed scheme is comparable with other existing schemes mainly with the k-anonymity which is a widely used mechanism, with the extra feature of providing the mitigation against the MINIMALITY attack.

CHAPTER 6
Conclusions and Future work

In the existing privacy preserving models for publishing the data, minimality in information loss is an underlying principle. This report presents an approach to solve the Minimality Attack, a privacy breach which is prevalent in most of the privacy preserving data publishing schemes. We use the help of *k*-anonymity and *l*-diversity and build our scheme upon that to eliminate this attack. We illustrate how the scheme successfully eliminates the attack and provides a mitigation against the attack while preserving the privacy the data.

For future work one can aim to determine other kinds of attacks related to the anonymization process. The proposed scheme works for single sensitive attribute, but analyzing today's scenario the micro data consists of multiple sensitive attributes. The proposed work can be extended to address the above scenario. Now-a-days privacy in social networking site has become a bigger challenge, hence the current scenario can be extended to work with the social networking data.

References

[1] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D. and Zhu, A., 2005, January. Anonymizing tables. In International Conference on Database Theory (pp. 246-258). Springer Berlin Heidelberg.

[2] Kisilevich, S., Rokach, L., Elovici, Y. and Shapira, B., 2010. Efficient multidimensional suppression for k-anonymity. IEEE Transactions on Knowledge and Data Engineering, 22(3), pp.334-347.

[3] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., 2005, June. Incognito: Efficient full-domain k-anonymity. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 49-60). ACM.

[4] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R., 2006, April. Mondrian multidimensional k-anonymity. In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 25-25). IEEE.

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

[6] Wong, R.C.W., Fu, A.W.C., Wang, K. and Pei, J., 2007, September. Minimality attack in privacy preserving data publishing. In Proceedings of the 33rd international conference on Very large data bases (pp. 543-554). VLDB Endowment.

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

[8] Hamza, N. and Hefny, H.A., 2013. Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing. Journal of Information Security, 4(02), p.101.

[9] Li, N., Li, T. and Venkatasubramanian, S., 2007, April. t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE.

[10] Zhang, Q., Koudas, N., Srivastava, D. and Yu, T., 2007, April. Aggregate query answering on anonymized tables. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 116-125). IEEE.

[11] Cormode, G., Srivastava, D., Li, N. and Li, T., 2010. Minimizing minimality and maximizing utility: analyzing method-based attacks on anonymized data. Proceedings of the VLDB Endowment, 3(1-2), pp.1045-1056.

[12] Li, T. and Li, N., 2009, June. On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 517-526). ACM.

[13] Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K., 2003, November. On the privacy preserving properties of random data perturbation techniques. In Data Mining, 2003. ICDM 2003. Third IEEE International Conference on (pp. 99-106). IEEE.

[14] https://archive.ics.uci.edu/ml/datasets/Adult

[15] Bayardo, R.J. and Agrawal, R., 2005, April. Data privacy through optimal k-anonymization. In Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on (pp. 217-228). IEEE.

[16] http://arx.deidentifier.org/overview/metrics-for-information-loss/

[17] Iyengar, V.S., 2002, July. Transforming data to satisfy privacy constraints. InÂăProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data miningÂă(pp. 279-288). ACM.

[18] Li, T., Li, N., Zhang, J. and Molloy, I., 2012. Slicing: A new approach for privacy preserving data publishing. IEEE transactions on knowledge and data engineering, 24(3), pp.561-574.