

A Study of Inference Control Techniques

A thesis submitted in partial fulfillment of the requirements for the degree

of

Bachelor of Technology

in

Computer Science and Engineering

by

Santosh Kumar Chauhan

Roll No. 10606012



Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela-769008, Orissa, India

May 2010

A Study of Inference Control Techniques

A thesis submitted in partial fulfillment of the requirements for the degree

of

Bachelor of Technology

in

Computer Science and Engineering

by

Santosh Kumar Chauhan

Roll No. 10606012

Under the guidance of

Prof. Korra Sathya Babu



Department of Computer Science and Engineering

National Institute of Technology Rourkela

Rourkela-769008, Orissa, India

May 2010



National Institute of Technology Rourkela

CERTIFICATE

This is to certify that the thesis entitled, “**A Study of Inference Control Techniques**” submitted by **Santosh Kumar Chauhan** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** during session **2009-10** in **Computer Science and Engineering** at the **National Institute of Technology, Rourkela** is an authentic work carried out by him under my supervision and guidance.

And to the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any Degree or Diploma.

PLACE: NIT Rourkela

DATE: 12/05/2010

Prof. Korra Sathya Babu

Assistant Professor

Department of Computer Science and Engineering

National Institute of Technology

Rourkela – 769008

ACKNOWLEDGEMENT

I wish to express my deep sense of gratitude and indebtedness to **Prof. Korra Sathya Babu**, Department of Computer Science and Engineering, **N.I.T. Rourkela** for introducing the present topic and for his inspiring guidance, valuable suggestions and support throughout this project work.

I am thankful to all my Professors and Lecturers and members of the department for their generous help in various ways for the completion of the thesis work.

I would like to thank my family members for encouraging me at every stage of this project work. Last but not least, my sincere thanks to all my friends who have patiently extended all sorts of help for accomplishing this undertaking.

Santosh Kumar Chauhan

Table of Contents

Abstract	7
Chapter 1 Introduction	8
1.1 Introduction.....	9
1.2 Motivation.....	10
Chapter 2 Overview	11
2.1 Inference Control.....	12
2.2 Inference Channel.....	12
2.3 Application of Inference Control Techniques.....	12
2.4 Challenges to Inference Control.....	13
2.5 Data Types.....	13
2.6 Inference Types.....	15
2.7 Protection Type.....	15
2.8 Example of Disclosure Risk.....	16
Chapter 3 Implementation and Case Study	18
3 Inference Control Techniques.....	19
3.1 Perturbative (Noise Addition).....	19
3.1.1 Rounding.....	19
3.1.1.1 Systematic Rounding.....	22
3.1.1.2 Random Rounding.....	24
3.1.1.3 Controlled Rounding.....	26
3.1.2 Microaggregation.....	27
3.2.3 Data Swapping.....	36
3.1.4 Data Perturbation.....	38
3.1.5 Output Perturbation.....	39
3.2 Non-Perturbative.....	40
3.2.1 Table Restriction.....	40
3.2.1.1 Relative Table Size Control.....	40
3.2.1.2 Order Control.....	43
3.2.2 Query Restriction.....	43
3.2.2.1 Query set-size control.....	43
3.2.3 Cell Suppression.....	45
Chapter 5 Results and Discussion	48
Chapter 4 Conclusion and Future Work	51
References	53

List of Figures

Fig. 1: Working principle of Inference control techniques.....	10
Fig. 2: Disclosure risk plot for Rounding.....	21
Fig. 3: Information Loss plot for different base values.....	23
Fig. 4: Information loss plot for random rounding.....	24
Fig. 5: Comparison of results of Systematic and Random Rounding.....	25
Fig. 6: Microaggregation Model.....	27
Fig. 7: Comparison of results of fixed-size and variable-size grouping.....	35
Fig. 8: Data Perturbation.....	38
Fig. 9: Output Perturbation.....	39
Fig. 10: Query set size Control.....	44
Fig. 11: Variation of no. of Suppressed cells w.r.t Table-size.....	47

List of Tables

Table- 1: Records in microdata.....	13
Table- 2: Masked microdata.....	14
Table- 3: External information.....	14
Table- 4: Tabular Data.....	15
Table- 5: Original cell counts before rounding.....	22
Table- 6: Tabular data after systematic rounding.....	22
Table- 7: Results of random rounding.....	24
Table- 8: Microdata before micro-aggregation.....	28
Table- 9: Microdata after micro-aggregation.....	28
Table- 10: Microdata before grouping.....	30
Table- 11: Microdata after fixed-size grouping.....	31
Table- 12: Microdata after microaggregation with fixed-size groups.....	32
Table- 13: Microdata after variable-size grouping.....	33
Table- 14: Microdata after microaggregation with variable-size groups.....	34
Table- 15: Microdata before swapping.....	36
Table- 16: Microdata before swapping.....	37
Table- 17: Different table sizes and computation of disclosure risk.....	42
Table- 18: Tabular data after cell suppression.....	45
Table- 19: Secondary Cell suppression (Hypercube method).....	47

Abstract

Security is a major issue in every field and it impacts more when intruders get the information of an individual from the database either directly or indirectly. There are two approaches to break the confidentiality, either directly or indirectly. Here we study about different types of techniques which protect the confidentiality from indirect disclosure. These techniques are called Inference Control Techniques and also known as Statistical Disclosure Control methods. Indirect disclosure differs from the other security problems because in this presumptive intruders or external users deduce the information by set of available queries having low security risk i.e., they do computations on set of available information which is non sensitive and from that information they get the sensitive information. Inference control techniques protect the publicly released statistics of companies and institutions, such that presumptive user could not get any private information about any individual entity. Inference control in statistical databases is a part of information security which tries to prevent published statistical information (tables, individual records) from disclosing the contribution of specific respondents. Here we shall analyze about information loss, disclosure risk measures and performance of the various techniques. The major challenge to Statistical Disclosure Control is that modified data should be such that it provides more useful information with less disclosure risk i.e., protection should be maximized and information loss should be minimized. Since there is a trade-off between information loss and disclosure risk i.e., generally it happens that if disclosure risk is less then information loss will be more and vice versa. Here we propose some ideas which may provide optimal results.

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Introduction

Before going to the introduction of different Inference Control Techniques first we shall put light on existing security problem for databases and for released statistics of companies and firms. There are two ways to break the confidentiality, first way is, presumptive intruders get the sensitive information from unauthorized access and it is direct way of getting the confidential information but when users get the sensitive information from the set of non sensitive queries then it is called the indirect disclosure.

It is not a security problem like the first problem because in case of direct disclosure the intruders and hackers are addicted to access the private information in unauthorized way. In case of indirect disclosure even users do not have any type of unauthorized access but they just collect some information from a set of non sensitive queries and from that they deduce some confidential information. There are many techniques which provide protections from indirect disclosure and their classification depends upon the type of data on which they are applied. Some techniques are applied on microdata and some on tabular data. As stated in [13] that statistical database protection is a part of information security which tries to prevent published statistical information (tables, individual records) from disclosing the contribution of specific respondents. Also data should be available to researchers and statistical agencies so that the necessary research and planning activities can be conducted. If data is not available to the external world then it is useless. However, on the other hand, the right of respondents to privacy must be protected. As we know technology is growing day by day and the massive use of computers for organizing and distributing electronic data, the amount of stored information has had a spectacular increase in a very diverse set of fields, namely finance, health care and engineering. This enormous increase of stored data in sensitive areas such as health care has led to a new set of needs related to information management and privacy protection. Thus, the way to guarantee the rights of the respondents whose information is being stored has undergone a substantial change. It has become necessary to develop new techniques able to keep the balance between

sharing information and protecting individual privacy [11]. Thus the best technique will protect the privacy efficiently as well as minimize the information loss.

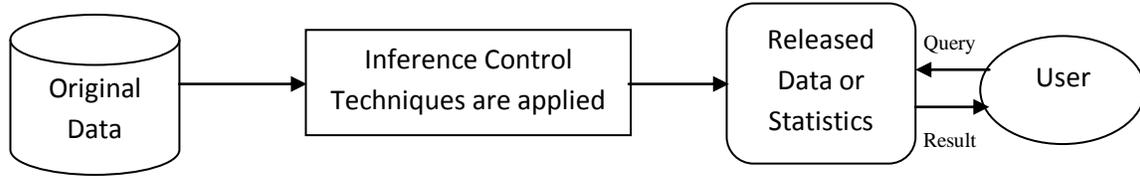


Fig. 1: Working principle of inference control techniques

In the above **Fig. 1**, it is shown that before publication of statistics of any organization the original data passes through the operation of inference control techniques and after that the resultant of those operations is released in public domain.

1.2 Motivation

The motivation behind this work is to study about different inference control techniques for performance measure and propose new ideas to produce efficient results with the help of these techniques i.e., minimizing both the information loss and disclosure risk to an optimal. It is a comparative study of different inference control techniques and here we analyze the performance of different techniques that which technique gives the better protection with less information loss. Because a data set with major information loss will be useless, i.e. it cannot be used for a research purpose unless otherwise it is more accurate and also if original information will be published then it will cause to disclosure problem. Sometimes it might be better if we combine more than one technique to operate on the same data set which can provide better results than applying individual techniques.

Chapter 2

Overview

Chapter 2

Overview

2.1 Inference Control

Inference control (disclosure control) aim at protecting data from indirect detection. It ensures queries of non-sensitive data when put together do not reveal sensitive information. The goal of inference control is to prevent the user from completing any inference channel.

2.2 Inference Channel

A sequence of queries whose responses allow a sensitive inference to be made is known as an inference channel. An inference channel is a channel where users can find an item X and then use X to derive Y as $Y = f(X)$.

2.3 Application of Inference Control Techniques

There are wide ranges of application areas where Inference Control Techniques are used to protect the private information of an individual entity from indirect disclosure. Here we refer some of the areas of application which are stated in [10].

2.3.1 Official Statistics

Most of the organizations, institutions take the services from many statistical agencies to assure the confidentiality of released statistics from many organizations and people of the countries [10]. Because maintain the confidentiality is the primary issue of any organization so that no can get the sensitive information of an individual.

2.3.2 E-Commerce

Today the internet has provided a lot of facilities regarding the business activities. Electronic commerce results in the automated collection of large amounts of consumer data. This wealth of information is very useful to companies, which are often interested in sharing it with their subsidiaries or partners. Such consumer information transfer should not result in public profiling

of individuals and is subject to strict regulation. Here protecting the confidentiality and integrity of e-commerce data is very essential.

2.3.3 Health Information

Health statistics is very important for medical research purpose and for other issues regarding the privacy of an individual. There requires the strict regulation of protected health information for use in medical research.

2.4 Challenges to Inference Control Techniques

The major challenge to inference control techniques is to provide better protection from disclosure with very less information loss. The protection provided by ICTQs normally involves some degree of data modification, which is an intermediate option between no modification (maximum utility, but no disclosure protection) and data encryption (maximum protection but no utility for the user without clearance). The challenge for ICTQ is to modify data such that sufficient protection is provided while keeping at a minimum the information loss. That is, efficient statistical disclosure methods provide less information loss with least disclosure risk.

2.5 Data Types

2.5.1 Microdata

It represents a series of records, each record containing information on an individual entity such as a person, a firm, an institution, etc. Here we can see some attributes are sensitive (e.g. SSN, Diagnosis type and Income) and no company or organization would like to publish this type of sensitive information so before publication of statistics the data should be masked so that external users could not get any private information of individuals [14].

Name	SSN	Age	Zip	Diagnosis	Income
Ajay	11543	23	223621	AIDS	17300
Rakhi	11536	19	223622	AIDS	12456
Vivek	11524	26	223621	SARS	45213
Raj	11511	31	223625	SF	15482

Table-1: Records in microdata

2.5.2 Masked Microdata

Microdata released for use by third parties, after the data has been masked to limit the possibility of disclosure. In this the identifying information such as names, SSN and other identifying information are removed from microdata [14]. We use different inference control techniques to modify the data in such a way that from the available information no one can deduce the sensitive information. Thus these statistical disclosure control methods give grantee of disclosure control and also for optimal result the disclosure risk and information loss should be minimized.

Age	Zip	Diagnosis	Income
23	223621	AIDS	17300
19	223622	AIDS	12456
26	223621	SARS	45213
31	223625	SF	15482

Table -2: Masked microdata

2.5.3 External Information

Any known information by a presumptive intruder related to some individuals from initial microdata. This is the information by the help of this the preemptive intruders can deduce other more sensitive information [14].

Name	SSN	Age	Zip
Alice	12458	45	458621
Bob	12478	39	458622
Rosan	12574	45	458621
Raj	12635	39	458625
Alok	14852	45	458621

Table -3: External information

2.5.4 Tabular Data

In tabular data representation, each cell in the table represents the number of records for a set of attributes associated with each other. Here we will refer to an example which illustrates the tabular data. Here we see attributes are of categorical type and we see some cell counts are very less (e.g. 0, 3...) and by this small count values one can deduce information about an individual in merely 2 to 3 steps. So tabular-data protection is a big issue. In the **Table-4** we show how the

tabular data can be presented and the given cell values are the number of records for a particular category.

Sex_MS/Age	0-15	16-25	26-35	36-50	>50
Male_Married	12	9	0	4	21
Male_Single	0	5	11	8	3
Female_Married	9	13	47	25	21
Female_Single	0	14	25	7	17

Table-4: Tabular data

2.6 Inference Types

Inferences may be of two types [14], identity disclosure or attribute disclosure, are explained below.

Identity Disclosure: Identification of an entity (person, institution). **For example:** *Vivek is the third record*, is an identity disclosure in the **Table-1**.

Attribute Disclosure: The intruder finds something new about the target person. **For example:** *Ajay has AIDS*, is an attribute disclosure in the **Table-1**.

2.7 Protection Types

Here we refer, what types of protection is needed for different types of statistics released by an organization [10].

2.7.1 Tabular Data Protection

Previously most of the companies, statistical agencies and other organizations used this method to release the statistics. It was the simple method to represent and understand the statistics. Tabular data protection is a good method to protect the data because tabular data have been the traditional output of national statistical offices. The goal here is to publish static aggregate

information, i.e. tables, in such a way that no confidential information on specific individuals among those to which the table refers can be inferred. And also in this category there are many methods to protect the tabular data.

2.7.2 Microdata Protection

As we have discussed about the microdata in the previous section. Here the aim of microdata protection is to minimize the disclosure risk for an individual. There are many techniques which provide protection for micro data. This method is about protecting static individual data, also called microdata. It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. We will discuss protection techniques in the next section under this category.

2.7.3 Dynamic Databases

User can submit statistical queries (sums, averages, count etc.) to the database. The aggregate information obtained by a user as a result of successive queries should not allow him to infer information on specific individuals. In the next section we will discuss about two methods; in case of one method one can get approximate results for each query but in the second one can get the exact information but not for all queries. Some queries will be restricted which will not satisfy the criteria.

2.8 Example of Disclosure Risk

Here, we refer some examples [1] regarding, how inferences can be made?

2.8.1 Example 1

Suppose in a department there are many Profs in different- different age groups and the statistics is published about the average salary of all Profs and average salary of all Profs with **age** > 20. Think about the case, there is only one Prof whose **age** < 20. So from the released statistics one can easily deduce the salary of Prof whose **age** < 20.

Q1=> **select** avg (salary) **from** staff;

Q2=> **select** avg (salary) **from** staff **where** age > 20;

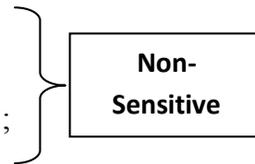
These both queries make an inference channel and deduce sensitive information. Here Prof whose age is less than 20 can be easily deduced by subtraction. Thus this is an indirect disclosure in which a user can get the salary of an individual (whose age is less than 20) by simple computation on these both queries which are all most non-sensitive in nature.

2.8.2 Example 2

Consider a Ship-Port management system, here it is shown that with the help of two nonsensitive queries, how one can deduce the confidential information.

Q1 => **select** name, port_name **from** ship;

Q2 => **select** port_name **from** port **where** portWeapon = 'yes';



These two queries together constitute an inference channel, because if both are made then it's possible to infer exactly which ships are carrying weapons, and this may be sensitive information.

Chapter 3

Implementation and Case study

Chapter 3

Implementation and Case study

3 Inference Control Techniques

Here we will discuss about different inference control techniques which help the statistical agencies, firms and institutions in statistical disclosure control. By the help of these techniques the released statistics is protected from inference problem. Most of the inference control techniques are classified in two broad categories first is Perturbative (Noise Addition) and second is Non-Perturbative [15].

3.1 Perturbative (Noise Addition)

The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset [15]. Add noise to the data or to the released statistics. Noise addition provides answers to all queries but the answers are approximate. This is further classified as:

- Rounding Techniques
- Microaggregation
- Data Swapping
- Data Perturbation
- Output Perturbation

3.1.1 Rounding

These are the techniques which are frequently used for protecting the tabular data. Rounding methods replace original values of attributes with rounded values. When the table is protected by rounding, the cell entropy conditional to the rounded table depends on the rounding base \mathbf{b} . Here

we will see, how information loss and disclosure risk vary for different base values. We will show the process of getting a rounded table from the original table.

Disclosure Risk

The protection provided by rounding can be measured by the uncertainty about the true values determined by it. Therefore, we take the width of the interval containing the possible true value for a given rounded value, called existence interval, as measure of protection. We assume that the rounding base and the number of jumps allowed are released together with the rounded table [13]. In a rounded table without marginals, if the value of a cell x'_i is $n_i b$ (i.e. n times the rounding base), then we know that the original cell x_i must lie in the interval:

$$I_i = [(n_i - 1/2) b, (n_i + 1/2) b]$$

Seeing the released statistics one can easily deduce the base values because for a table cell values will be the multiple of base so from that one can get the base value (*For example*, suppose the cell values are 0, 10, 15, 5..etc. thus here one can directly deduce that the base value will be 5) and after getting base values one can deduce the interval by the help of above formula. And after getting the base one can also compute the interval I_i and finally the disclosure risk is equal to the number of cells in I_i and will be given by the relation highlighted in [13]:

$$DR(x_i) = 1/ \log_2 m(I_i)$$

Where $m(I_i)$ is the number of possible cell values in I_i and $m(I_i)$ depends upon the value of base (b). Now from the above formula we calculate the disclosure risk for different $m(I_i)$ values. If number of cell values is more then probability of disclosure risk will be less and if interval length is less then disclosure risk is more. For a large base value the disclosure risk will be less but the information loss will be more. For example suppose original count of a cell is 3 and for base values 5 and 10 it will be rounded to 0, 5 and 10 depending upon the rounding technique used. Now we can see for base 5 the interval length $m(I_i)$ will be less as compared to the case of base 10. Here we see a plot between $m(I_i)$ and disclosure risk. In the figure we have taken possible cell counts and for different $m(I_i)$ and we see that the disclosure risk is decreasing for the increasing interval length.

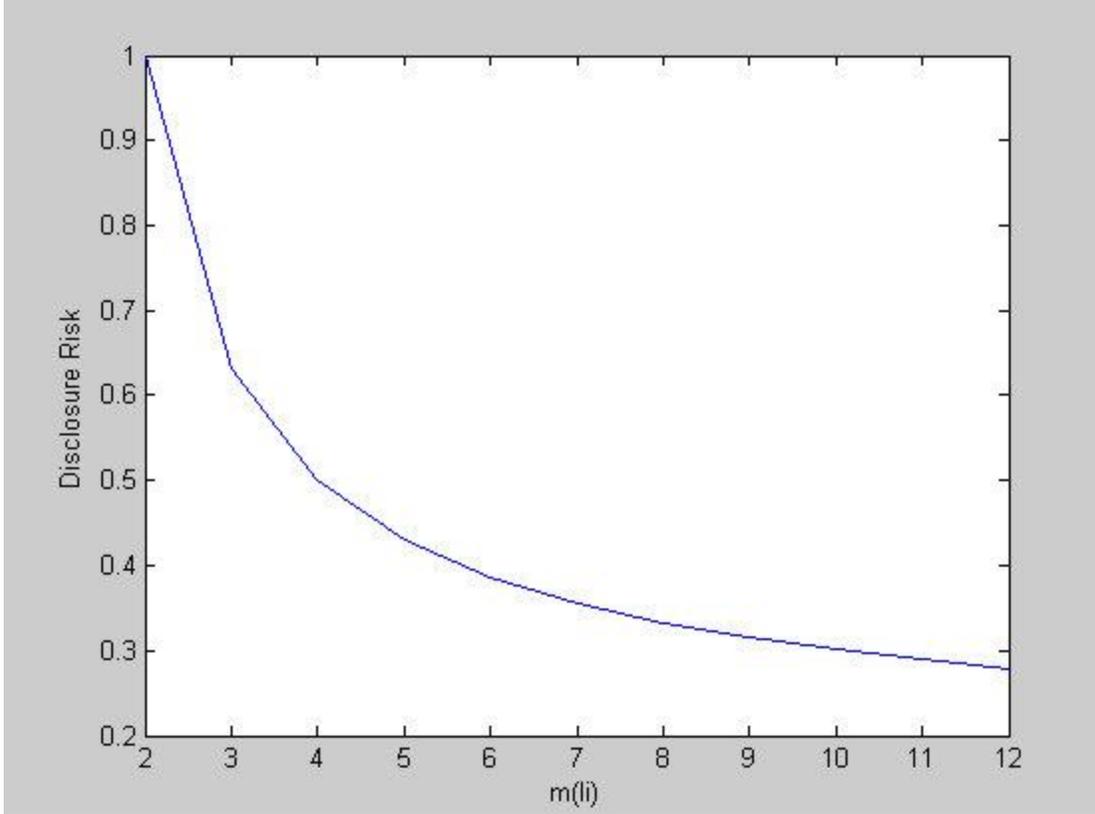


Fig. 2: Rounding disclosure risk plot

Information Loss

The information loss of a cell is defined as the distance between the original and the modified values of this cell. It can be shown this by the formula given below:

$$IL(x_i) = x'_i - x_i$$

Where $x'_i = n*b$ (cell value after rounding) and x_i = original cell count. Thus from this linear relation we can imagine that if base value will be large then information losses will be more.

Note: In further explanations, we will indicate the original cell values by z and the rounded values by a , b be the rounding base.

3.1.1.1 Systematic Rounding

In systematic rounding we round a statistic to the closest integer multiple of a fixed rounding base b . In this method all values are rounded but this has a demerit that this is not additive. For example explained in [1] we see for base $(b) = 5$, two tables are shown below.

Original Counts:

a/b	b1	b2	b3	b4	b5	Total
a1	10	0	11	3	9	33
a2	5	8	7	12	27	59
a3	3	16	21	7	4	51
a4	9	10	11	3	9	42
Total	27	34	50	25	49	185

Table-5: Original cell counts before rounding

After Systematic Rounding:

a/b	b1	b2	b3	b4	b5	Total
a1	10	0	10	5	10	35
a2	5	10	5	10	25	60
a3	5	15	20	5	5	50
a4	10	10	10	5	10	40
Total	25	35	50	25	50	185

Table-6: Tabular data after systematic rounding

Disclosure Problem

In conventional (systematic) rounding users know that the true value has been rounded to the nearest multiple of the base [12]. Hence, the existence intervals that can be computed are:

$$z \in [0, \lfloor b/2 \rfloor] \quad \text{if } a = 0$$

$$z \in [a - \lfloor b/2 \rfloor, a + \lfloor b/2 \rfloor] \quad \text{if } a > 0,$$

Where symbol denotes the floor value. For example [12], if the base is $b=5$ and $a=0$, then a user can determine that if $a=5$, then a user can determine that $z \in [3, 7]$.

$$\begin{aligned} \lfloor b/2 \rfloor & \quad \text{if } a = 0 \\ b - 1 & \quad \text{if } a > 0, \end{aligned}$$

From the expressions above it is clear that conventional rounding gives less protection than controlled rounding. Broadly speaking, we could say that it is half of that of controlled rounding. Information loss plot for different base (b) values:

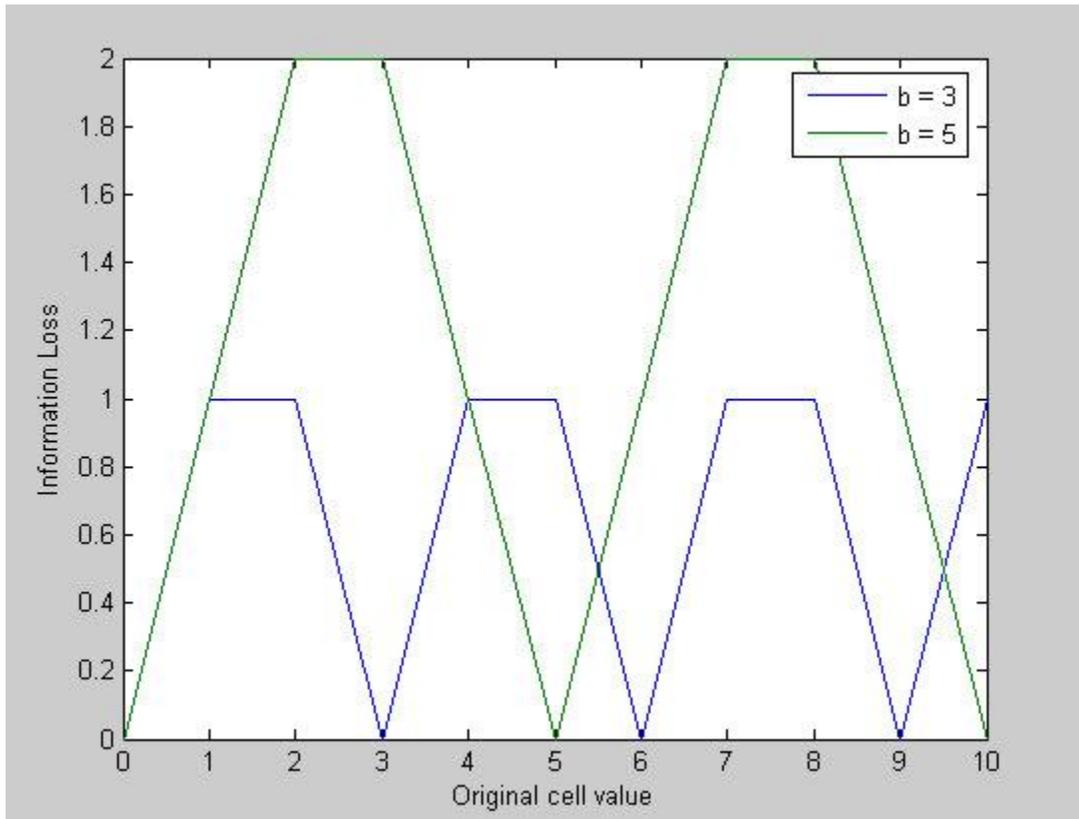


Fig. 3: Information loss plot for $b = 3$ and 5 .

From the **Fig. 3** we perceive that for the large base information loss is more and for the small base information loss is less.

3.1.1.2 Random Rounding

This technique randomly rounds a statistics either to the next higher or the next lower multiple of the rounding base. Considering the previous example in systematic rounding, the table after random rounding becomes like given below.

a/b	b1	b2	b3	b4	b5	Total
a1	10	0	15	0	10	30
a2	5	5	10	10	25	55
a3	0	15	25	10	5	50
a4	10	10	15	5	5	45
Total	30	30	50	25	50	185

Table-7: Results after random rounding

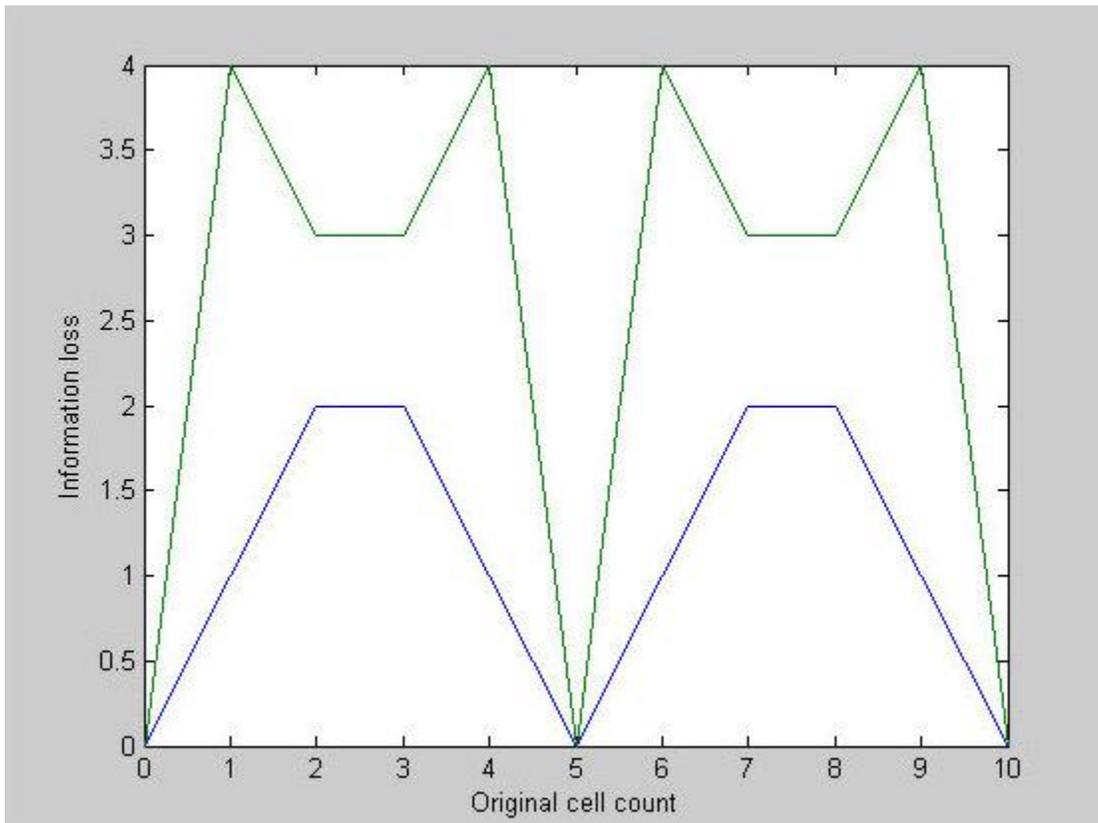


Fig. 4: Information loss plot for random rounding

As we know in case of random rounding for a cell values there can be two possible rounded values; one upper rounded value and second is lower rounded value. For example, suppose cell value is 2 and base is 5 then after random rounding the result will be either 0 or 5, both possibilities are there. Thus in the above plot we have shown both sided rounding i.e., upper as well as lower.

Here we put a plot showing the deviation from the original values in systematic rounding and random rounding.

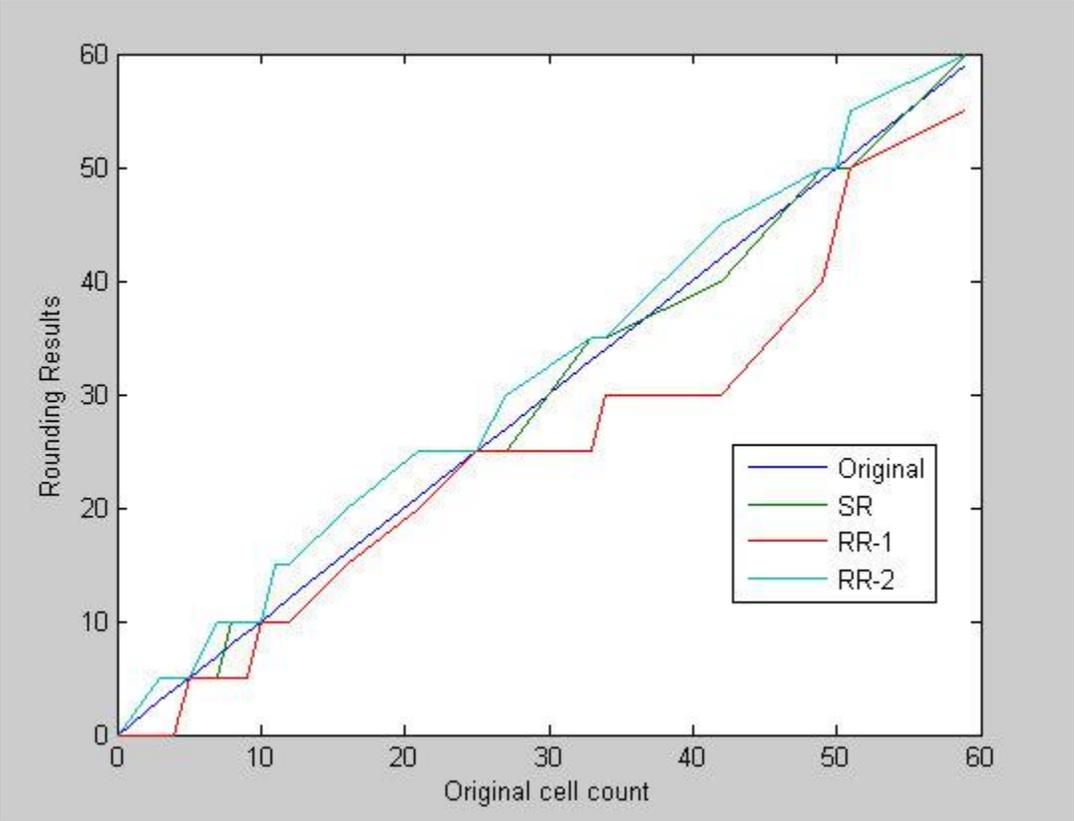


Fig. 5: Comparison of results of Systematic and Random Rounding

3.1.1.3 Controlled Rounding

This method of rounding provides an even smaller variance and greater consistency than random ranges by requiring the marginal sums of rounded statistics to equal their rounded sum. *For example*, the table in systematic rounding could be modified to meet the requirements for controlled rounding by replacing the *column-1* sum of 25 with 30 and in *column-3* sum of 50 with 45. Thus controlled rounding is additive i.e. the values in the marginal cells coincide with those calculated by adding the relevant interior cells [12].

Note: In the above example a_1, a_2, a_3, a_4 are the attributes values of Attribute **A** and b_1, b_2, b_3, b_4, b_5 are the attributes values of Attribute **B** and each cell is represented by the number of records corresponding to each combination of a_i and b_i ($i=1, 2, \dots$).

For example:

If **A** represents **Sex_Martial-Status** then a_1, a_2, a_3, a_4 may be represented by **Male_Married, Male_Unmarried, Female_Married** and **Female_Unmarried**. And if **B** represents **AGE** then b_1, b_2, b_3, b_4, b_5 may be represented by the age groups **0-20, 21-35, 36-55, 56-70, >71**. And the value of each cell represents the number of records associated with the corresponding attributes values.

3.1.2 Microaggregation

Microaggregation comes under perturbative inference control techniques for continuous microdata. Here continuous microdata means microdata which attributes are such that one can perform numerical and arithmetic operations on that. As stated in [10], the rationale behind microaggregation is that confidentiality rules in use allow publication of microdata sets if records correspond to groups of k or more individuals, where no individual dominates i.e. contributes too much to the group and k is a threshold value. This is the basic principle of microaggregation. Raw (individual) data is grouped into small aggregates before publication. The average value of the group replaces each value of the individual. Groups are formed using a criterion of maximal similarity. Once the procedure has been completed, the resulting (modified) records can be published. This helps to prevent disclosure of individual data. A model for microaggregation as described in [17] is given below.

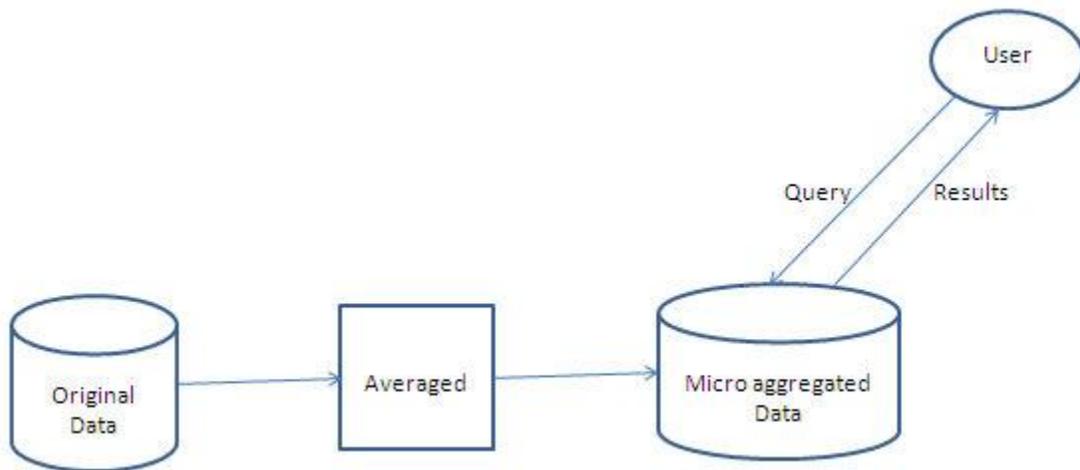


Fig. 6: Micro-aggregation

In this method user can get the results for every query made to database but the results are approximate. First original database is converted into averaged database; from that microaggregated database the users make query and get the results which are approximate. Here information loss depends upon the number of data elements in group formation and how much are there values close to each other. Data with the most similarities are grouped together to maintain data accuracy.

Here we refer an example in which grouping is done of 3-3 data sets and we take average of each group (which has sensitive numerical data). Like in the given example we have assumed income as sensitive information.

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Alice	12458	45	458621	AIDS	17300
2	Bob	12478	39	458622	AIDS	12456
3	Rosan	12574	45	458621	SF	19213
4	Raj	12635	39	458625	SF	15482
5	Alok	12852	45	458621	SF	12635
6	Vijay	12845	44	458628	AIDS	12500
7	John	12863	45	458652	SARS	15600
8	James	12477	33	458629	AIDS	17500
9	Alin	12356	33	458622	SARS	17900
10	Jassica	12475	39	458621	SF	15600
11	Hagon	12412	50	458623	SF	19200

Table -8: Microdata before micro-aggregation

Microaggregation for attribute Income and minimum size 3. The total sum for all Income values remains the same. Table after Microaggregation:

ID	Age	Zip	Diagnosis	Income
1	45	458621	AIDS	17566
8	33	458629	AIDS	17566
9	33	458622	AIDS	17566
2	39	458622	AIDS	12530
5	45	458621	SF	12530
6	44	458628	AIDS	12530
4	39	458625	SF	15560
7	45	458652	AIDS	15560
10	39	458621	SF	15560
3	45	458621	SF	19206
11	50	458623	SF	19206

Table-9: Microdata after micro-aggregation

If we think from the information loss point of view then an optimal k -partition is defined to be the one that maximizes within-group homogeneity; the higher the within group homogeneity, the lower the information loss, since microaggregation replaces values in a group by the group centroid. Mean to say that if the values in a group will closer to each other then the averaged value also will be very close to each value and hence the information loss will be less.

Suppose for $k = 3$, grouping is done for set of values (2.65, 2.78, 2.70), in this case the averaged value will be 2.71 and each three values will be replaced by 2.71 and correspondingly the information loss will be 0.06, 0.07 and 0.01 which is almost a low value. Now consider an example in which values are 2.67, 5.89 and 7.98 for $k = 3$. Here averaged value is 5.51 and information loss will be 2.84, 0.38 and 2.47 respectively which are quite large so information loss is more due to non homogeneity within the group.

Grouping in microaggregation can be done in two ways [10], one fixed size grouping and second variable size grouping. There are advantages of variable-sized groups i.e. variable-size grouping results in more homogeneous groups as a result information loss will be less as compared to fixed-size grouping. Now here we put an example which might be able to make distinction between fixed-size groups and variable-groups i.e. advantage of variable-size groups over the fixed-size groups regarding homogeneity which is a major factor deciding information loss.

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Alice	12458	45	458621	AIDS	17450
2	Bob	12478	39	458622	AIDS	12556
3	Rosan	12574	45	458621	SF	19213
4	Raj	12635	39	458625	SF	15482
5	Alok	12852	45	458621	SF	13635
6	Vijay	12845	44	458628	AIDS	12500
7	John	12863	45	458652	SARS	15600
8	Almas	12477	33	458629	SF	13500
9	Jassica	12475	39	458621	SF	15600
10	Hagon	12412	30	458623	SF	19200
11	Roja	15628	36	421622	SF	19206
12	Bali	15674	35	421621	AIDS	19213
13	Nikil	15635	49	421625	SF	17412
14	Suraj	15652	25	421621	AIDS	12590
15	Shyam	15663	35	421652	SARS	12600
16	James	15677	30	421629	AIDS	17400
17	Srijan	15563	33	421656	SARS	17380
18	Suman	15546	38	421697	SARS	19200

Table-10: Microdata before grouping

Suppose group size (k) = 3, now after making fixed-size groups with maximum similarity of values. Here grouping is done over the income attribute and we got table as:

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Bob	12478	39	458622	AIDS	12556
2	Shyam	15663	35	421652	SARS	12600
3	Suraj	15652	25	421621	AIDS	12590
4	Vijay	12845	44	458628	AIDS	12500
5	Alok	12852	45	458621	SF	13635
6	Almas	12477	33	458629	SF	13500
7	Raj	12635	39	458625	SF	15482
8	John	12863	45	458652	SARS	15600
9	Jassica	12475	39	458621	SF	15600
10	Srijan	15563	33	421656	SARS	17380
11	James	15677	30	421629	AIDS	17400
12	Nikil	15635	49	421625	SF	17412
13	Rosan	12574	45	458621	SF	19213
14	Suman	15546	38	421697	SARS	19200
15	Hagon	12412	30	458623	SF	19200
16	Roja	15628	36	421622	SF	19206
17	Bali	15674	35	421621	AIDS	19213
18	Alice	12458	45	458621	AIDS	17450

Table-11: Microdata after fixed-size grouping

Table after microaggregation is:

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Bob	12478	39	458622	AIDS	12582
2	Shyam	15663	35	421652	SARS	12582
3	Suraj	15652	25	421621	AIDS	12582
4	Vijay	12845	44	458628	AIDS	13211
5	Alok	12852	45	458621	SF	13211
6	Almas	12477	33	458629	SF	13211
7	Raj	12635	39	458625	SF	15560
8	John	12863	45	458652	SARS	15560
9	Jassica	12475	39	458621	SF	15560
10	Srijan	15563	33	421656	SARS	17397
11	James	15677	30	421629	AIDS	17397
12	Nikil	15635	49	421625	SF	17397
13	Rosan	12574	45	458621	SF	19204
14	Suman	15546	38	421697	SARS	19204
15	Hagon	12412	30	458623	SF	19204
16	Roja	15628	36	421622	SF	18623
17	Bali	15674	35	421621	AIDS	18623
18	Alice	12458	45	458621	AIDS	18623

Table-12: Microdata after microaggregation with fixed-size groups

In general value of $k \geq 3$ but here we assume that $k \geq 2$ thus in case of variable-size grouping the table will look like as given below:

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Bob	12478	39	458622	AIDS	12556
2	Shyam	15663	35	421652	SARS	12600
3	Suraj	15652	25	421621	AIDS	12590
4	Vijay	12845	44	458628	AIDS	12500
5	Alok	12852	45	458621	SF	13635
6	Almas	12477	33	458629	SF	13500
7	Raj	12635	39	458625	SF	15482
8	John	12863	45	458652	SARS	15600
9	Jassica	12475	39	458621	SF	15600
10	Srijan	15563	33	421656	SARS	17380
11	James	15677	30	421629	AIDS	17400
12	Nikil	15635	49	421625	SF	17412
13	Alice	12458	45	458621	AIDS	17450
14	Rosan	12574	45	458621	SF	19213
15	Suman	15546	38	421697	SARS	19200
16	Hagon	12412	30	458623	SF	19200
17	Roja	15628	36	421622	SF	19206
18	Bali	15674	35	421621	AIDS	19213

Table-13: Microdata after variable-size grouping

Table after microaggregation is:

ID	Name	SSN	Age	Zip	Diagnosis	Income
1	Bob	12478	39	458622	AIDS	12561
2	Shyam	15663	35	421652	SARS	12561
3	Suraj	15652	25	421621	AIDS	12561
4	Vijay	12845	44	458628	AIDS	12561
5	Alok	12852	45	458621	SF	13567
6	Almas	12477	33	458629	SF	13567
7	Raj	12635	39	458625	SF	15560
8	John	12863	45	458652	SARS	15560
9	Jassica	12475	39	458621	SF	15560
10	Srijan	15563	33	421656	SARS	17410
11	James	15677	30	421629	AIDS	17410
12	Nikil	15635	49	421625	SF	17410
13	Alice	12458	45	458621	AIDS	17410
14	Rosan	12574	45	458621	SF	19206
15	Suman	15546	38	421697	SARS	19206
16	Hagon	12412	30	458623	SF	19206
17	Roja	15628	36	421622	SF	19206
18	Bali	15674	35	421621	AIDS	19206

Table-14: Microdata after microaggregation with variable-size groups

Here we show a plot comparing fixed-size grouping and variable-size grouping with respect to the original dataset.

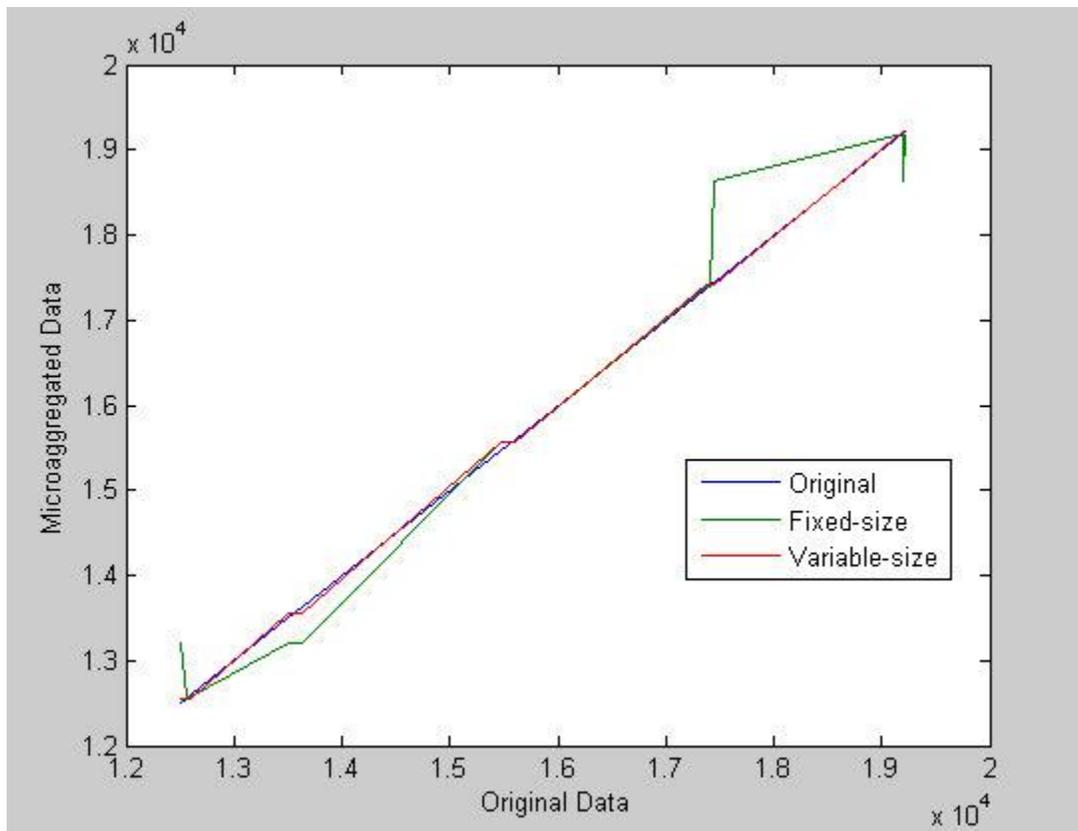


Fig. 7: Comparison of results of fixed-size and variable-size grouping

As we see in the plot that curve of microaggregation with fixed-size groups is more distorted than the variable-size groups. Hence information loss will be more in case of fixed-size grouping than the variable-size grouping due to maximum similarity of records in variable-size groups i.e. due to homogeneity of the groups.

3.1.3 Data Swapping

In the current time user's needs are increasing towards the more detailed and larger microdata files and accordingly their demand is increased now days. Previously disclosure control techniques are not enough sufficient to mask the anonymity of the respondents. When used in conjunction with the other techniques, data swapping may be a feasible procedure to adequately mask data files while providing users with more information.

As stated in [19] one of the first references to data swapping is found in Reiss (1980). In his article, the author describes interchanging values of individual records within a highly visible field. The swap protects the univariate distribution of that variable. Since its value belongs to some other respondent, each respondent's anonymity is protected.

In this disclosure method a sequence of so-called elementary swaps is applied to a microdata. An elementary swap consists of two actions: **(i)** a random selection of two records *i* and *j* from the microdata. **(ii)** A swap (interchange) of the values of the attribute being swapped for records *i* and *j*.

Here we refer this example; a microdata file for 11 respondents is provided. In order to protect the anonymity of the respondents, income values are randomly swapped among the records.

ID	Age	Zip	Diagnosis	Income
1	45	458621	AIDS	17500
2	39	458622	AIDS	12500
3	45	458621	SF	19200
4	39	458625	SF	15482
5	45	458621	SF	12500
6	44	458628	AIDS	12456
7	45	458652	AIDS	15600
8	33	458629	AIDS	17500
9	33	458622	AIDS	17900
10	39	458621	SF	15600
11	50	458623	SF	19200

Table-15: Microdata before swapping

ID	Age	Zip	Diagnosis	Income
1	45	458621	AIDS	17500
2	39	458622	AIDS	12456
3	45	458621	SF	19200
4	39	458625	SF	15600
5	45	458621	SF	12500
6	44	458628	AIDS	12500
7	45	458652	AIDS	15482
8	33	458629	AIDS	17500
9	33	458622	AIDS	17900
10	39	458621	SF	15600
11	50	458623	SF	19200

Table-16: Microdata after swapping

We see in the **Table-16** that records 1 & 3 are swapped with records having the same income therefore for these respondents, swap has provided no masking.

As stated in [19], the probability that a swap has masked a particular record is inversely proportional to the frequency of its value appearing in the file. For large data files, this is acceptable. An income which appears frequently in a microdata file does not as easily identify the respondent as one which appears very rarely. The reason is that more number of records having same income decreases the disclosure risk of a particular individual as many respondents have the same income so a presumptive intruder cannot identify a record by income.

The basic idea behind the method is to transform a database by exchanging values of confidential attributes among individual records. Records are exchanged in such a way that low-order frequency counts or marginals are maintained. Here data swapping is introduced to protect continuous and categorical microdata, respectively. In data swapping exact disclosure is not possible but partial disclosure is possible.

In a microdata file swap can be performed on more fields which are deemed to be sensitive. Like in the **Table-15**, suppose age is also treated as sensitive then it may also be swapped randomly. Remember that random swap of age can differ from the random swap of income. Independent multiple random data swaps can be used in succession [19]. These independent multiple random data swaps further ensures that the resulting file has adequately masked accurate information about each respondent.

3.1.4 Data Perturbation

In this method first noise is added to the raw data and perturbed database is made. Replaces original data with another sample drawn from the same probability distribution. User can query to the perturbed database rather than the original database. Thus here user can get the approximate result. The pro and cons are like; **pro**: With perturbed databases, if unauthorized data is accessed, the true value is not disclosed. **Con**: Data perturbation runs the risk of presenting biased data.

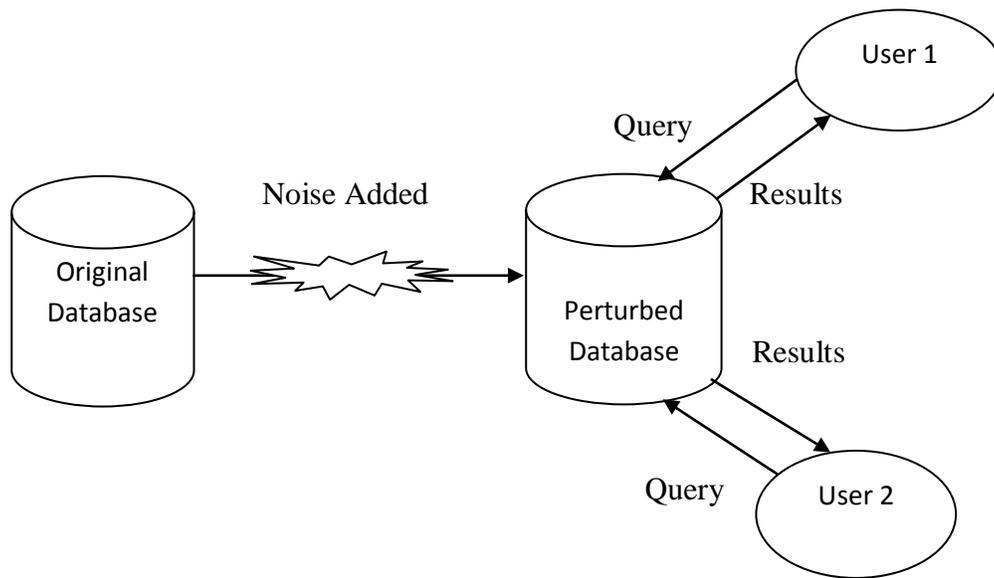


Fig. 8: Data perturbation

As described in [17] a model for data perturbation is given in **Fig. 8**. The data-perturbation approach usually requires that a dedicated transformed database is created for statistical research. If the original database is also used for other purposes, the original database and the transformed SDB are both maintained by the system. It is important to note that this approach has a large risk of introducing bias to quantities such as the conditional means and frequencies.

In data perturbation method bias problem is much severe. Bias refers to the difference between a true statistic and the expectation of its perturbed estimate. The bias should be zero or at least as small as possible [1]. As described in [1] consistency means lack of contradictions and paradoxes in the perturbed statistics. Inconsistency arises, e.g. when repetitions of the same query give the different results. Paradoxes arise when negative or fractional values are returned for counts.

These both factors bias and consistency play a vital role in measuring the performance of the perturbation techniques.

3.1.5 Output Perturbation

Instead of the raw data being transformed as in Data Perturbation, only the output or query results are perturbed. As stated in [17], the interaction of users with the database is given below.

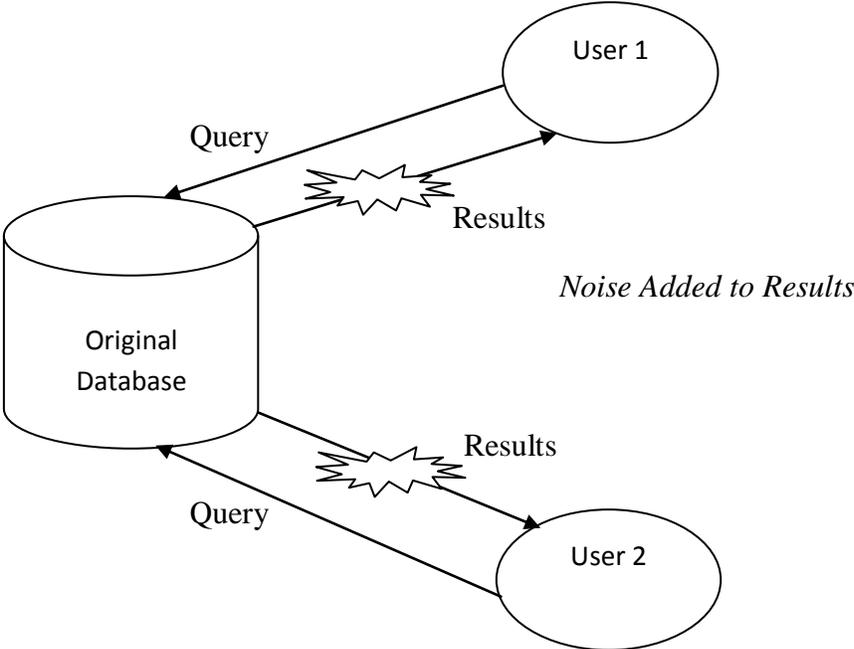


Fig. 9: Output perturbation

The bias problem is less severe than with data perturbation. This is due to the fact that the query set selected under the output-perturbation approach is based on the original values, not the perturbed values. In this method noise is added to the results of the query.

3.2 Non-Perturbative

Non-Perturbative methods do not alter data; rather, they produce partial suppressions or reductions of detail in the original dataset. In this category if a query satisfies the condition then the accurate result will be come. There are many techniques in this category which we will discuss one by one.

- Table Restriction
- Query Restriction
- Cell Suppression

3.2.1 Table Restriction

Table-level controls restrict complete tables of statistics, namely those higher dimensional tables in the lattice that have, or are likely to have, a positive disclosure or identification risk (i.e., one or more cells corresponding to a single individual). This method incurs large information loss as for protecting few elements in the table one has to restrict the whole table data.

3.2.1.1 Relative Table Size Control

Restricts an \mathbf{m} -table of counts if its relative size S_m/N ^[1] exceeds a threshold $1/k$, where k is a parameter of the database chosen to reduce, but not necessarily eliminate, the possibility of disclosure. Thus, an \mathbf{m} -table is permitted if the average number of records falling into each cell is at least k . S_m is the number of cells in a table and S_m is the product of domain sizes for the attributes named in the characteristic formula (query formula). Here a table is published only if it is k permitted. We refer the method of protecting the table statistics, i.e. what computations are made for a table and on what basis a table is permitted or restricted.

We refer this example as stated in [1]. Let $k = 10$, $N = 185$ (total no. of records in database), domain sizes of attributes are $|A| = 4$, $|B| = 5$, $|C| = 8$, $|D| = 2$ and applying above criteria only 8 of the 16 tables are permitted. We calculate the S_m value by multiplying the domain sizes of the attributes.

Results

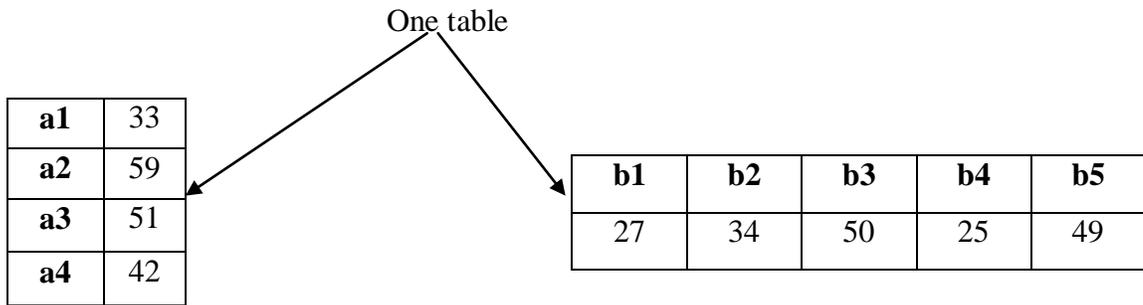
<u>Table</u>	<u>S_m</u>	<u>S_m/N</u>	<u>Status</u>
T _{ALL}	1	0.005	Permitted
T _A	4	0.022	Permitted
T _B	5	0.027	Permitted
T _C	8	0.043	Permitted
T _D	2	0.011	Permitted
T _{AB}	20	0.108	Restricted
T _{BC}	40	0.216	Restricted
T _{CD}	16	0.086	Permitted
T _{AC}	32	0.173	Restricted
T _{AD}	8	0.043	Permitted
T _{BD}	10	0.054	Permitted

As we see here that tables which have the S_m/N values less than $1/k$ are permitted. All other tables have the S_m/N value greater than $1/k$ ($=0.1$), so they all are restricted. Identification risk in an m -table will be approximately e^{-N/S_m} .

We refer this example of [1] and calculate the disclosure risk for the given tabular data:

Zero table: \longrightarrow

185



Two-table

a/b	b1	b2	b3	b4	b5
a1	10	0	11	3	9
a2	5	8	7	12	27
a3	3	16	21	7	4
a4	9	10	11	3	9

Table-17: Tables of sizes 0, 1, 2.

<u>N/S_m</u>	<u>e^{-N/S_m}</u>
9.25	9.61E-05
37	8.53E-17
46.25	8.20E-21
185	4.52E-81

3.2.1.2 Order Control

As stated in [1] it restricts all \mathbf{m} -tables of counts for $\mathbf{m} > \mathbf{d}$ where the threshold \mathbf{d} is a parameter of the database. The parameter \mathbf{d} is chosen so that most, if not all, statistics in the permitted \mathbf{m} -tables are nonsensitive (or are expected to be nonsensitive). The control is easily implemented by counting the number of attributes appearing in the characteristic formula of a query, if there are more than \mathbf{d} attributes, the query is not answered.

Results

<u>\mathbf{d} – Value</u>	<u>Table-size</u>	<u>Status</u> (<i>table statistics</i>)
0	0	Permitted
	> 0	Restricted
1	0, 1	Permitted
	> 1	Restricted
2	0, 1, 2	Permitted
	> 2	Restricted

3.2.2 Query Restriction

Query restriction rejects a query which can lead to a compromise, but the answers in query restriction are always accurate. In this the main techniques is Query Set-Size Control which is explained in this section.

3.2.2.1 Query Set-Size Control

As stated in [2] a query is not answered if it contains too few or too many records. A query-set size control can limit the number of records that must be in the result set. The query-set-size control method permits a statistic to be released only if the size of the query set \mathbf{R} (i.e., the number of entities included in the response to the query) satisfies the condition. It allows the query results to be displayed only if the size of the query set satisfies the condition.

Setting a minimum query-set size can help protect against the disclosure of individual data.

- Let \mathbf{K} represents the minimum number or records to be present for the query set.
- Let \mathbf{R} represents the size of the query set.
- The query set can only be displayed if

$$K \leq R \leq L-K$$

Where L is the size of the database (the number of entities represented in the database) and K is a parameter set by the **DBA**. K should satisfy the condition:

$$0 \leq K \leq (L/2)$$

Notice that K cannot exceed $L/2$; otherwise no statistics would ever be released.

We refer this example which shows the overview of query set- size control approach. Here we see if the result of a query exceeds the k value then the result will not be published.

Let $k=3$

Q1 => COUNT (birth place = 'rkl' and street no = 'A123') = 5; is allowed.

Q2 => COUNT (birth place = 'rkl' and street no. not 'A123') = 2; is not allowed.

In this figure taken from [17], the process of query set restriction is explained.

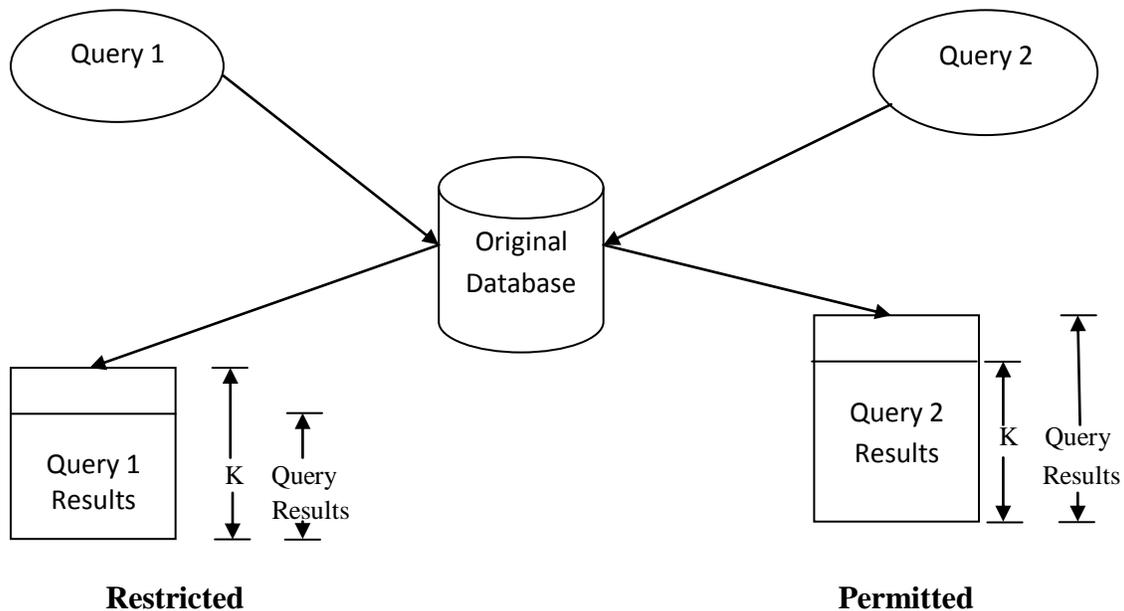


Fig. 10: Query set-size control

In the above example we see that if the query result size is less than the k value then that query (also query result) is not published.

Results of Query set size control

With query set size control the database can be easily compromised within a frame of 4-5 queries. For query set control, if the threshold value k is large, then it will restrict too many queries. And still does not guarantee protection from compromise.

3.2.3 Cell Suppression

This method is used to protect data published in tabular form. The basic idea behind this approach is to remove the sensitive cells which can cause disclosure of individual data. The linear relations among all cells of a table (including the marginal sums in the parent tables) are analyzed to determine whether sensitive cells can be deduced (exactly or approximately) from those that are released; additional cells, called complementary suppressions, are suppressed for as long as possible [1].

In tabular data, the general procedure to define a cell as sensitive (disclosure cell) is that if cell contains statistics computed over less than t (a threshold defined) individual records. When cells contain frequencies, are less than t then those cells are called disclosure cells. Once a disclosure cell X has been identified, a disclosure interval $I(X)$ is determined [16]. Here we refer an example as explained in [13] to compute the disclosure risk for cell suppression. Suppose the table after suppression with marginal sums is given as:

MS/AGE	0-15	16-25	26-35	36-50	>50	Total
Single	1234	656	415	125	698	3128
Married	457	x_1	789	896	x_2	2956
Divorced	856	x_3	587	621	x_4	3894
Total	2547	3057	1791	2436	2215	8878

Table-18: Tabular data after cell suppression

Let's all cells contain the positive integer values. From the above table it can be made linear equations like:

$$\begin{aligned}
 x_1 \mid x_1 + x_2 = 814, \quad x_1 + x_3 = 2401, \quad x_i \geq 0 \\
 x_2 \mid x_1 + x_2 = 814, \quad x_2 + x_4 = 1517, \quad x_i \geq 0 \\
 x_3 \mid x_1 + x_3 = 2401, \quad x_3 + x_4 = 1830, \quad x_i \geq 0 \\
 x_4 \mid x_3 + x_4 = 1830, \quad x_2 + x_4 = 1517, \quad x_i \geq 0
 \end{aligned}$$

As stated in [13] the disclosure risk for each suppressed cell is computed by the following formula:

$$\mathbf{DR}(X) = 1 / \log_2 \mathbf{m} (\mathbf{S}_y(X))$$

where $\mathbf{m} (\mathbf{S}_y(X))$ is the number of possible values of the cell in $\mathbf{S}_y(X)$. $\mathbf{S}_y(X)$ is computed by solving two linear programming (LP) problems (one maximization and one minimization) and then subtracting the solutions.

Cell suppression is basically of two types; one is primary suppression and second is secondary suppression. In primary suppression the cell only which is disclosure cell (sensitive) is suppressed but due to linear relations between the cell values (marginal sums) some additional non-sensitive (non-disclosure cells) are also suppressed to provide required interval protection to X, thus determining and suppressing those additional cells is called secondary suppression [16]. If we consider these both patterns then we see that in case of secondary suppression information loss is more so to get good performance it is needy that information loss should be less and for that we have to choose secondary suppression such that the number of suppressed cells is minimal.

As stated in [1] a secondary suppression methodology, in which suppressed cells fall into hypercube of size 2^m , where \mathbf{m} is the size of the table. Some attempt is made in selecting cells for suppression to minimize information loss.

Here we refer an example of [1] which shows how cell suppression can be used to protect the sensitive cell $\mathbf{T} [3, 1]$.

MS/AGE	0-15	16-25	26-35	36-50	>50	Total
Single	1234	656	415	125	698	3128
Married			789	896	856	2956
Divorced			587	621	456	3894
Total	2547	3057	1791	2436	2215	8878

Table-19: Secondary Cell suppression (Hypercube method)

In hypercube method of secondary cell suppression number of suppressed cells depends upon the size of the table i.e. 2^m where m is the size of the table [1]. We can show a plot of variation of number of suppressed cells versus table size.

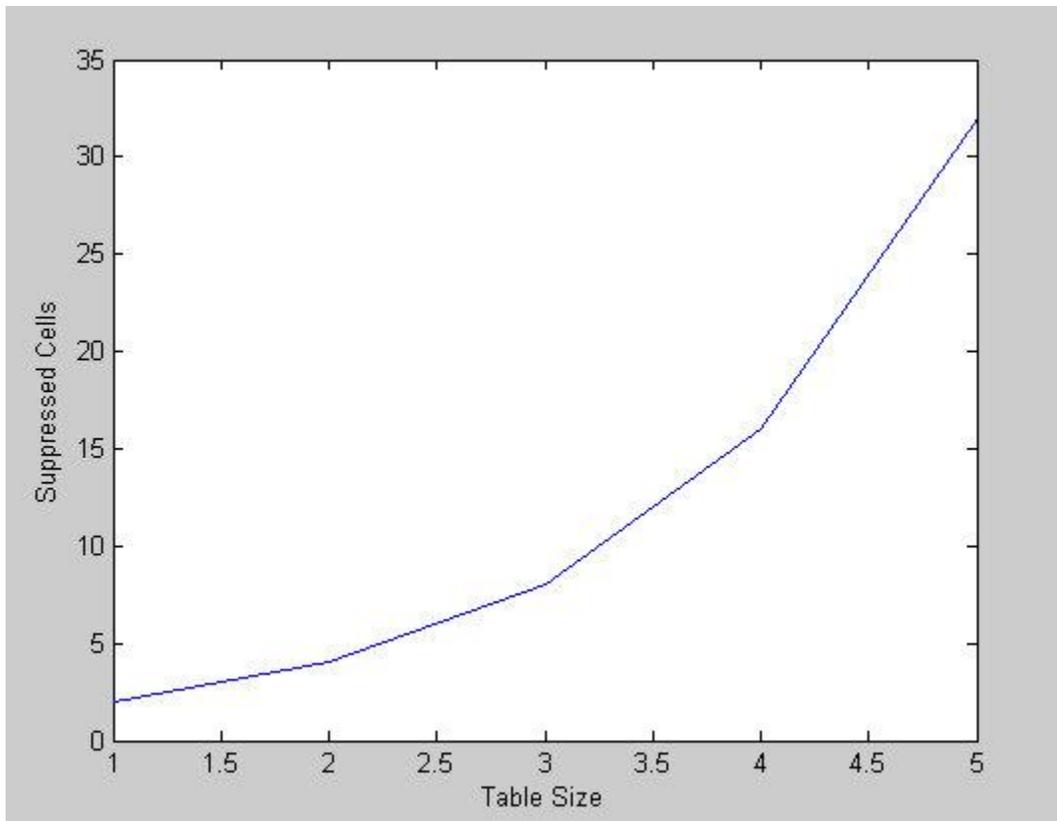


Fig. 11: Variation of no. of Suppressed cells w.r.t Table-size.

Chapter 4

Results and Discussion

Chapter 4

Results and Discussion

Study after so many techniques we came to know about their performance. Also we came to know that no single inference control method alone satisfies the conflicting aim of high protection, low information loss, and efficiency, thus to achieve these all properties at the same time one has to combine more control methods. A scheme that combines simple restriction criteria, such as query-set size and relative table size, with simple perturbation techniques can provide an acceptable level of protection for many applications without being overly restrictive or costly. For applications requiring extremely high levels of security or low levels of information loss, the techniques such as cell suppression and controlled rounding can be used although at the rate of increased cost. The best strategy for a particular application will depend on its objectives and risks.

If we consider rounding methods then we see that out of three methods explained here controlled rounding is better than the other two in the sense of minimizing information loss. Rounding techniques may be either **zero**-restricted or **k**-restricted [13] and if we consider then in case of **k**-restricted disclosure is difficult than in case of **zero**-restricted. Because in case of **k**-restricted a value is rounded **k** times of the rounding base. For example; suppose rounding base is 5 and value to be rounded is 3 then in case of **zero**-restricted it will be either 0 or 5 depending upon the 3 methods used but in case of **k**-restricted it will become any of these values like 0, 5, 10, 15,...(depending upon **k** value). Thus it is difficult to find the existence interval in **k**-restricted rounding than the **zero**-restricted.

In case of microaggregation as it is explained that variable-size grouping is more preferable than the fixed-size grouping regarding the information loss.

Considering data perturbation and output perturbation, output perturbation is better than the data perturbation because in case data perturbation a different perturbed database is created and from that database users can query only and in this method bias problem is more severe than the output perturbation method as in case of output perturbation only the results of the queries are perturbed rather than whole database is perturbed i.e. noise is added to the results of the queries.

In output perturbation the bias problem is less severe as compared to the output perturbation.

In non-perturbative techniques, it may be all queries are not answered but the results of these techniques are accurate i.e. if a query is satisfying the criteria then result will be published and if not then nothing will be come out. As it is explained in perturbative techniques the results of each and every query is answered but the only thing is that all results are approximate. So it can be said that in case of perturbative techniques information loss has no significant meaning i.e. it is meaningless because for a particular data we get the accurate results not the perturbed one, so only thing here matters that up to what extent one can get the information and what is the percentage of disclosure risk.

Finally so many techniques are there but the protection and information loss depends upon the type of method used for a particular dataset and it varies from one technique to another technique.

Chapter 5

Conclusion and Future work

Chapter 5

Conclusion and Future Work

After study of different Inference Control Techniques , we come to know that Information Loss and Disclosure Risk both varies reversely, that is; if information loss is more then disclosure risk will be less and vice versa. So far, here we come to a conclusion that, for achieving good performance, that is the Information loss as well as Disclosure risk both to be minimized, we will have to combine more techniques to get an optimal result and no single method can provide full protection as well as minimize the information loss simultaneously. Like if we consider a case in very few cells are not satisfying the k criteria in a table and due to that whole table is restricted. Suppose instead of applying relative table size control we apply cell suppression method in this we can avoid from more information loss as compared to relative size control. So a better choice of a disclosure control method gives better performance in the sense of low information loss and low disclosure risk. Also combining more techniques on the same dataset will give more valuable results.

References

- [1] Dorothy E. Denning and Jan Schlorer. Inference Control for Statistical databases. IEEE Journal, Vol. 16, No. 7: pp 69 - 82, July 1983.
- [2] Adam, Nabil R.; Wortmann, John C.; Security-Control Methods for Statistical Databases: A Comparative Study; ACM Computing Surveys, Pages: 515-556, Vol. 21, No. 4, December 1989, ISSN: 0360-0300.
- [3] John B. Kam and Jeffrey D. Ullman; A model of statistical database their security. Vol. 2, No. 1: pp 1 - 10, March 1977.
- [4] Joe Ryan, Mirka Miller, Yogesh Mehta. Inference Protection on Statistical Databases, December 2005.
- [5] Josep Domingo-Ferrer, Josep M. Mateo-Sanz and Vicenç Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk". Pre-proceedings of ETK-NTTS'2001 (vol. 2), pages 807-826. 2001. ISBN: 92-894-1176-5.
- [6] Thomas D. Garvey; The Inference Problem for Computer Security. IEEE, (1): pp 78 – 81, Computer Security Foundations Workshop V, 1992. Proceedings, ISBN: 0-8186-2850-2.
- [7] Rajagopal Chakravarthi, Glenn Shafer, S.Raman. Inference Detection Using Clustering. PHD Thesis, 13 December 2001.
- [8] Jessica Staddon. Dynamic inference control. ACM New York, NY, USA, Journal, (1): 94 – 100, 2003.
- [9] <http://gatton.uky.edu/faculty/muralidhar/maskingpapers/>
- [10] Josep Domingo-Ferrer, "Microaggregation for protecting individual data privacy". Proceso de Toma de Decisiones, Modelado y Agregación de Preferencias TIC-2002-11492-E, pages 171-178. 2005. ISBN: 84-934654-1-0.
- [11] Josep Domingo-Ferrer, Agusti Solanas and Antoni Martí nez-Ballesté, "Privacy in statistical databases: k-anonymity through microaggregation". Proceedings of IEEE Granular Computing 2006, pages 774-777. 2006. ISBN: 1-4244-0133-X.
- [12] <http://www.statistics.gov.uk/events/downloads/SessionF2.doc> (The application of controlled rounding for tabular data with particular reference to the Tau- Argus software by Philip Lowthian, Giovanni Merola, Office for National Statistics).

- [13] Josep Domingo-Ferrer, Anna Oganian and Vicenç Torra, "Information-theoretic disclosure risk measures in statistical disclosure control of tabular data". Proceedings of the 14th Intl. Conf. on Scientific and Statistical DataBase Management, pages 227-231. 2002. ISBN: 0-7695-1632-7.
- [14] http://dimacs.rutgers.edu/Workshops/HIPAA2009/Slides/Truta_P1_DIMACS2009: (Traian Marius Truta – DIMACS Tutorial, Overview of Statistical Disclosure Control and Privacy-Preserving Data Mining).
- [15] Josep Domingo-Ferrer and Vicenç Torra, "Disclosure control methods and information loss for microdata". Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pages 91-110. 2001. ISBN: 0-444-50761-2.
- [16] Josep Domingo-Ferrer and Josep M. Mateo-Sanz, "On the security of cell suppression in contingency tables with quantitative factors". Proceedings of the 3rd International Seminar on Statistical Confidentiality, pages 208-217. 1996. ISBN: 86-81141-41-4.
- [17] <https://users.cs.jmu.edu/aboutams/Web/SDB%20Presentation.ppt> (Security Methods for Statistical Databases by Karen Goodwin).
- [18] Anna Oganian and Josep Domingo-Ferrer, "A posteriori disclosure risk measure for tabular data based on conditional entropy". Vol. 27, No. 2, pages 175-190, 2003.
- [19] Controlled data-swapping techniques for masking public use microdata sets. R.A. Moore, Jr. SRD Report RR 96-04, U.S. Bureau of the Census, 1996.
- [20] Joe Ryan. Mirka Miller. Yogesh Mehta. A seminar on Inference Protection on Statistical Databases. School of Information Technology and Mathematical Sciences University of Ballarat. December 2005.
(<http://webpages.ull.es/users/cryptull/Doctorado/ProtectiononStatisticalDatabases.ppt>.)