

**AN EFFICIENT APPROACH FOR
GENERALIZATION OF SENSITIVE
ATTRIBUTES USING K-ANONYMITY MODEL**



**Submitted By:
Muhammad Fida
525/FBAS/MSCS/F08**

**Supervised By:
Dr. Ayyaz Hussain
Assistant Professor**

**Department of Computer Science & Software Engineering
Faculty of Basic and Applied Sciences**

**INTERNATIONAL ISLAMIC UNIVERSITY,
ISLAMABD, PAKISTAN
August 2012**



Accession No TH-9503

MCS
004.6
FIE

DATA ENTERED

July 18/62

*K-Anonymity Model
Computer Communication*

AN EFFICIENT APPROACH FOR GENERALIZATION OF SENSITIVE ATTRIBUTES USING K-ANONYMITY MODEL



Submitted By:

Muhammad Fida

A dissertation submitted as partial fulfillment of requirements for the degree of MS in
Computer Science at the Faculty of Basic and Applied Sciences
International Islamic University Islamabad, Pakistan

Supervised By:

Dr. Ayyaz Hussain

Assistant Professor, Department of Computer Science & Software Engineering
International Islamic University Islamabad, Pakistan

August 2012

**Department of Computer Science & Software Engineering
International Islamic University Islamabad, Pakistan**

Date: 12 November 2012


Final Approval

This is to certify that we have read and evaluated the thesis entitled **An efficient approach for generalization of sensitive attributes using K-anonymity model** submitted by **Muhammad Fida** under **Reg No. 525-FBAS/MSCS/F08** and that in our opinion it is fully sufficient in scope and quality as a thesis for the degree of Master of Science in Computer Science.

Committee

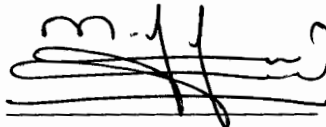
External Examiner

Dr. Waseem Shahzad
Assistant Professor,
Department of Computer Science
National University of Computer Engineering Sciences,
(FAST), H-11/4, Islamabad


Dr. Waseem Shahzad

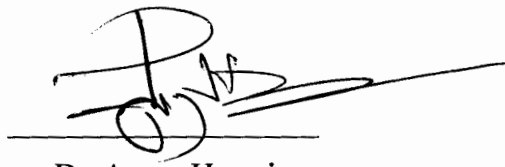
Internal Examiner

Mr. Imran Saeed
Assistant Professor,
Department of Computer Science & Software Engineering
International Islamic University, Islamabad


Mr. Imran Saeed

Supervisor

Dr. Ayyaz Hussain
Assistant Professor,
Department of Computer Science & Software Engineering
International Islamic University, Islamabad


Dr. Ayyaz Hussain

In the Name of

Allah,

The most merciful and compassionate, the most gracious and beneficent


Whose help and guidance we always solicit at every step and every moment.

*Dedicated to my Parents, Teachers
and
Ali Bhai.*

A Dissertation Submitted to
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University, Islamabad
As a partial Fulfillment of the Requirement for the Award of the
Degree of MS in Computer Science

Declaration

I hereby declare that the thesis “**An efficient approach for generalization of sensitive attributes using K-anonymity model**” neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research with the accompanied research report entirely on the basis of my personal efforts, under the proficient guidance of my teachers especially my supervisor Dr. Ayyaz Hussain. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any training institute or educational institutions, I shall stand by the consequences.


Muhammad Fida
525-FBAS/MSCS/F08

12-12-2012

Acknowledgement

In the name of Allah, The Most Gracious, The Most Merciful

Thanks Almighty ALLAH for giving me the courage and patience to carry out this work. I am very thankful to International Islamic University for providing such a good research environment.

I wish to thank my supervisor Dr. Ayyaz Hussain for his continuous advice, support and encouragement throughout this work. He has instilled in me a state of confidence, with which I now feel that I can do research of any new topic following his research guidelines.

I am grateful to department of Computer Science and Software Engineering IIU Islamabad and faculty members for providing healthy environment for research.

I would be failing in my duties if I would not remember to thank my elder brother Eng. Ali Rehman and my family members for their continuous motivational support. I truly believe that their endless support encouragement and stimulation have been a true source of strength and inspiration for me. I would also like to thank my dear friends specifically Mr. Tariq Sadad for his valuable cooperation at every step of my work, Mr. Assad Mehmood Khan and Mr. Rahat Ali Shah who has been a continuous motivation behind my success.



Muhammad Fida

525-FBAS/MSCS/F08

ABSTRACT

K-anonymity is an efficient approach that ensures the protection of sensitive data to be published. In this technique data is generalized in such a way that a tuple in quasi identifier group is indistinguishable from others in published table because there are at least $k-1$ same other tuples. K-anonymity significantly protects identity disclosure while to some extent compromising on attribute disclosure because of homogeneity and background knowledge attacks. Different algorithms have been proposed so far to minimize attribute disclosure but due to their limitations sensitive information regarding individuals could not be protected in efficient manner. Our proposed model (p, β) sensitive k-anonymity reduces the deficiencies of (p, α) -sensitive k-anonymity model by maintaining two separate tables for sensitive and non sensitive tuples. First all sensitive and non sensitive tuples are separated from each other and stored in corresponding tables. We introduced Discriminated union β of sensitive values of entire QI-group which ensures all sensitive values in QI-group are distinct. Union-fined algorithm is used to calculate Discriminated union β which first determine disjoint values from categories and then join them in single subset discriminated union or quasi group. We implemented our technique on Adult Dataset using Java and compared our results with existing techniques. Experimental results show that our proposed technique solve similarity attack problem and is much efficient from the existing technique in term of distortion ratio and accuracy of published data.

Table of Contents

CHAPTER 1: INTRODUCTION	1
1.1 Motivation.....	3
Table1.1(a): Table containing micro data without modification	3
Table1.1 (b): Modified data table without Name attribute	3
Table1.1(c): data publically available in Motor vehicle driving and voter’s database	4
1.2 Research Objective	4
1.3 Scope of study	5
1.4 Thesis Organization	5
CHAPTER 2: BACKGROUND	6
2.1 K-Anonymity	7
Table2.2: Raw data	7
Table2.3: 4-anonymous view of table 2.1	7
2.2 Quasi-Identifiers	7
2.3 Equivalence Class	7
2.4 Disclosure	8
2.4.1 Generalization	8
2.4.2 Suppression	8
2.4.3 Suppression limit	8
2.4.4 Domain Generalization	8
2.4.5 Domain Generalization Hierarchy	8
2.4.6 Generalization Lattice	10
CHAPTER 3: RELATED WORK.....	11
3 Related work	12
3.1 K-anonymity	12
Table3.1 (a): Medical data	13
3.1.1 Homogeneity problem	13
Table3.1 (b): 4-anonymous view of table 3.1(a).....	14
3.1.2 Background Knowledge based attack	14
Table3.1 (c): Background knowledge attack	15
3.2 L-diversity principle.....	15
3.2.1 Probabilistic l-diversity	16
3.2.2 Recursive (c, l)-diversity.....	16

3.2.3 Drawbacks of l-diversity	16
3.2.4 Skewness Attack	16
3.2.5 Probabilistic Inference threat	17
Table3.2: 10-anonymous data set with 3-diversity	17
3.2.6 Similarity Attack	17
Table3.3: Micro data	18
Table3.4: 3-anonymous data set with 3-diversity	18
3.4 t-closeness	19
3.4.2 Limitation of t-closeness	19
3.5 P-sensitive k-anonymity model.....	19
Table3.5: Categories	19
3.5.1 Similarity Attack	20
Table3.6: Micro data	20
Table3.7: 2-sensitive 4-anonymous data.....	21
3.6 (p, α)-sensitive k-anonymity	21
Table3.8: Raw data in hospital.....	22
Table3.9: (3, 1)-sensitive 4-anonymous table.....	22
3.7 Problem Statement	23
CHAPTER 4: PROPOSED SOLUTION	24
4.1 Proposed technique (p, β) sensitive k-anonymity	25
Table4.1 (a): Sensitive attribute partition into categories	26
Table 4.1 (b): Raw data.....	26
4.1.1 Discriminated Union of Sensitive values.....	27
4.1.2 Discriminated or Disjoint Union of sensitive values	27
Table 4.2: data set with Discriminated union of sensitive values $p=k=4$	28
4.1.3 Solved Example	29
4.2 Anonymization Algorithms	30
4.2.1 Global Recoding Incognito Algorithm	30
Table 4.4 (a): data set D.....	35
Table 4.4 (b): Table T with only sensitive tuples	36
Table 4.4 (c): data set containing only non sensitive tuples	36
Table 4.4 (d): after adding $2/3^{\text{rd}}$ distortion.....	37

Table 4.5(a): Raw data.....	38
4.4 Successful specialization	39
CHAPTER 5: RESULTS & ANALYSIS.....	40
5 Dataset.....	41
Table5.1: Attributes description of Adult Data base [10].....	41
Table5.2: Attributes with corresponding distinct values	42
5.1 Performance Measures.....	45
Table5.2: Results comparison.....	45
CHAPTER 6: CONCLUSION & FUTURE DIRECTIONS	48
6.1 Conclusion and Future directions	49

List of Tables

Table1.1(a): Table containing micro data without modification	3
Table1.1 (b): Modified data table without Name attribute	3
Table1.1(c): data publically available in Motor vehicle driving and voter's database.....	4
Table2.2: Raw data	7
Table2.3: 4-anonymous view of table 2.1	7
Table3.1(a): Medical data	13
Table3.1 (b): 4-anonymous view of table 3.1(a).....	14
Table3.1 (c): Background knowledge attack	15
Table3.2: 10-anonymous data set with 3-diversity	17
Table3.3: Micro data	18
Table3.4: 3-anonymous data set with 3-diversity	18
Table3.5: Categories	19
Table3.6: Micro data	20
Table3.7: 2-sensitive 4-anonymous data.....	21
Table3.8: Raw data in hospital.....	22
Table3.9: (3, 1)-sensitive 4-anonymous table.....	22
Table4.1 (a): Sensitive attribute partition into categories.....	26
Table 4.1 (b): Raw data.....	26
Table 4.2: data set with Discriminated union of sensitive values $p=k=4$	28
Table 4.4 (a): data set D.....	35
Table 4.4 (b): Table T with only sensitive tuples	36
Table 4.4 (c): data set containing only non sensitive tuples	36
Table 4.4 (d): after adding $2/3^{\text{rd}}$ distortion.....	37
Table 4.5(a): Raw data.....	38
Table5.1: Attributes description of Adult Data base [10].....	41
Table5.2: Attributes with corresponding distinct values	42

List of Figures

Figure2.1(a): DGH and corresponding VGH for individual's Marital Status	9
Figure2.1 (b): DGH and VGH for Race of the individual	9
Figure2.1(c): DGH for Age.....	9
Figure2.1 (d): DGH and corresponding VGH for Gender of individuals.....	9
Figure2.2: Generalization Lattice for Zip Code and Sex attributes	10
Figure 4.1(a): Discriminated set	27
Figure 4.1(b): Discriminated union.....	27
Figure4.1(c).....	31
Figure 4.2(a): First Search	32
Figure 4.2 (b): Second Search.....	33
Figure4.3: Anonymization process for Zip code [9].....	39
Figure5.1: Anonymized data.....	44
Figure5.2: proposed methodology and previous model.....	46

List of Acronyms

Acronym	Definition
K	Number of tuples to be Anonymized in QI group
QI	Quasi Identifier
UI	Unique Items
SSN	Social Security Number
EC	Equivalence Class
HIPAA	Health Insurance Portability and Accountability Act
Raw data	Data to be processed
MaxSup	Maximum limit of suppression
UCI	University of California Irvine
DGH	Domain Generalization Hierarchy
VGH	Value Generalization Hierarchy
PHIPA	Personal Health Information Protection Act

CHAPTER 1: INTRODUCTION

Privacy preservation is the most wanted right of the individuals in data sharing glob. With the advent of new technologies communication became very fast, likewise the research grew faster than ever. Some organizations share their micro data for research purpose which may contain confidential information regarding entities which need to be protected before it is released. Organizations who publish non aggregated data usually come across challenges regarding identity and attribute disclosure [1,2]. When a respondent is connected to a record in published table is known as identity disclosure while if an individual is accessed by combining published data to the information collected from other sources is called attribute disclosure. K-anonymity is performing impressively well to reduce identity revelation but fails to cope with attribute revelation. Intruders can recognize the individual if homogeneous records are found against particular group, while background knowledge can also lead towards privacy breach.

A US news agency concluded through telephonic poll that majority of the people demands for privacy protection. They were very anxious about their medical information containing in hospital database. People demanded that organizations must have prior approval of individuals before disclosing their personal information and asked to legalize this in order to force organizations protect sensitive data [3].

Medical statistics of individuals have been regulated in some countries; HIPAA is the example in United States while PHIPA enforce privacy protection in Canada [4]. Organizations holding sensitive data mostly eliminate identifying attributes i.e. name, CNIC number, and cell phone numbers to enhance privacy of the entities but still they fail to prevent privacy breach. Intruders may synchronize zip code etc with other fields to access the individual again [5].

K-anonymity addresses the issue and provides solution to minimize privacy breach by generalization of the data set in such a way that a tuple in QI-group is not left recognizable amongst other tuples containing in the table.

Different methods were introduced to hide the identity of person like swapping, sampling and adding noise to micro data so that confidentiality may be achieved. These methods were inefficient and precision and accuracy of published data was also compromised [6, 7, 8].

1.1 Motivation

Many organizations conduct research on different problems of the community and publish their micro data for this purpose. Micro data published by these organizations consists of confidential information regarding individuals and become a privacy leakage setback.

An American scientist L.Sweeney claimed that more than 87% of Americans can be recognized through gender and zip code etc by linking it to publically available data repositories like voter registration list motor vehicle driving database. She identified many persons including Governor of Massachusetts by linking data available at voter registration list with the one obtained from hospital [9].

If we examine data publish by a hospital, it can easily be seen that by eliminating directly identifying attribute e.g. name, SSN number or passport number does not ensure privacy exposure protection. Confidential information of the respondent can be obtained through other sources

Table1.1(a): Table containing micro data without modification

S.No	Name	Age	Zip code	Gender	Nationality	Health Status
1	Rashid	23	34235	Male	Pakistani	Cardiac disease
2	Chris	43	43067	Male	US	Cardiac disease
3	George	33	43053	Male	Canadian	Stomach ulcer
4	Huejo	34	85667	Female	Japanese	Blood Cancer

The above table shows original data stored in hospital database while Table 1.2 is modified by removing identifying attribute name for privacy reason.

Table1.1 (b): Modified data table without Name attribute

S.No	Age	Zip code	Gender	Nationality	Health Status
1	23	34235	Male	Pakistani	Cardiac disease
2	43	43067	Male	US	Cardiac disease
3	33	43053	Male	Canadian	Stomach ulcer
4	34	85667	Female	Japanese	Blood Cancer

But these parameters are insufficient for privacy preservation because the attribute removed by the data holder can be obtained through other sources e.g. motor vehicle driving database or voter's database. If an attacker gets some information from hospital data and the rest from voter's database then the individuals can easily be identified which is not tolerable. Some parameters of voter's table can be seen in the following table available publicly.

Table 1.1(c): data publicly available in Motor vehicle driving and voter's database

S.No	Name	Age	Zip code	Gender	Nationality
1	Rashid	23	34235	Male	Pakistani
2	Chris	43	43067	Male	US
3	George	33	43053	Male	Canadian
4	Huejo	34	85667	Female	Japanese

The information available in publicly available data repository can be linked to medical data published by hospital. The data in table 1.1(b) encrypted by data holders in hospital can be retrieved from table 1.1(c) clearly identifying the individuals. For example it can be found that Chris is 43 years old American whose territory zip code is 43067 having some cardiac disease.

1.2 Research Objective

The main objective of the study is to introduce an efficient approach for generalization of sensitive attributes by introducing new privacy protection model (p, β) sensitive k -anonymity. Our proposed technique eliminates the deficiencies of the (p, α) -sensitive k -anonymity. The key features are as follows.

Discriminated Union of sensitive values is introduced which satisfy privacy measures in much sophisticated manner.

Similarity attacks problem is addressed and the tuples having couples of identical records are replaced by distinct values by the help of discriminated union. Thus greater privacy is achieved. Those tuples in equivalence class having homogeneous values revealing confidential information are organized in such a way that no two identical sensitive values can occur in QI-group. In this way homogeneity attack problem is fixed up.

Quality of the data being published is enhanced as only sensitive tuples are generalized and non sensitive tuples are allowed to be published without any modification. Having accurate data without modification may help researchers to address community problem more efficiently.

1.3 Scope of study

A new approach for generalization of sensitive tuples (p, β) sensitive k-anonymity has been presented where discriminated union β is introduced which ensures distinct sensitive values in QI-group. . Simulation software is developed using Java in Net beans 2007 environment. Experiments were performed on real data set called as Adult Dataset [10] and results comparison with previous techniques is also provided. Distortion ratio of the algorithm is also calculated and presented with the help of graph.

1.4 Thesis Organization

Rest of the dissertation is organized as follows.

Chapter 2: in this chapter basic concepts of the key topics and preliminary fundamentals are discussed which will be used in rest of the dissertation.

Chapter 3: In this chapter related work previously done by different scientists is discussed in detail. Different models of k-anonymity and privacy preservation methods are evaluated and their strengths and limitations are described. After critical analysis of the literature, problem statement is formulated.

Chapter 4: This chapter describes proposed methodology and the way it is implemented. Proposed method is implemented on real data set and explained through trivial example.

Chapter 5: In this chapter, critical analysis of the results obtained from experiment is presented. Type of data used in experimentation, performance measure to validate the technique and comparison with previous techniques is described.

Chapter 6: Finally, conclusion of the research and future directions are presented in this chapter. Scope of our study and contribution is also mentioned.

CHAPTER 2: BACKGROUND

In this chapter basic definitions of the key terms are described which will be used in rest of the study.

2.1 K-Anonymity

A table is said k-anonymous if every tuple it contains is indistinguishable from at least k-1 tuples in same QI-group [9].consider the following table containing raw data.

Table2.2: Raw data

S#	Quasi-identifiers			Sensitive information
	Zip code	Age	Nationality	Medical Condition
1	34235	23	US	Cardiac disease
2	34067	43	US	Cardiac disease
3	34053	33	Canadian	Stomach ulcer
4	34667	34	Japanese	Blood Cancer

The above table contains original information regarding individuals; if we apply k-anonymity then it will be looking like this.

Table2.3: 4-anonymous view of table 2.1

S#	Quasi-identifiers			Medical Condition
	Zip code	Age	Nationality	Health Status
1	34***	<50	*	Cardiac disease
2	34***	<50	*	Cardiac disease
3	34***	<50	*	Stomach ulcer
4	34***	<50	*	Blood Cancer

Here all tuple in same quasi identifier groups i.e. zip code, age, nationality are same and cannot be distinguished.

2.2 Quasi-Identifiers

Those attributes which can be coupled with set of information externally available to re-identify an entity is known as quasi identifiers [11]. For example it can be gender, region, date of birth, CNIC number and zip code etc. [12, 13].

2.3 Equivalence Class

Quasi Identifier group having identical values is called equivalence class.

2.4 Disclosure

It is the inferable piece of information about an individual which was published without intention.

2.4.1 Generalization

In generalization a value is encrypted in such a way that it becomes less specific but consistent and difficult to recognized amongst the others [3]. In table 2.2 zip code and age are generalized.

2.4.2 Suppression

Suppression is also anonymization technique in which a value is completely encrypted. Suppression can be implemented on different levels e.g. cell level or tuple level. Generally a (*) is replaced by the original value. In table 2.2 nationality attribute is suppressed.

2.4.3 Suppression limit

This defines the level a tuple can be suppressed to; during anonymization process is called suppression limit.

2.4.4 Domain Generalization

During anonymization processes we deal with attribute belong to different domains. Values are transformed to more general one by eliminating less significant figure. E.g. 43287, 43268 can be generalized to 432**, 432**.

2.4.5 Domain Generalization Hierarchy

Domain generalization is performed on categorical attributes, which consists of all available groups for a particular attribute. DGH consist of only prefix of the values based on an ordered relation among various domains which may connect to an attribute. Various domains values can be represented in generalization hierarchy tree.

Edges and paths of the generalization tree denote direct generalization and indirect generalization respectively.

Below are few examples of Domain Generalization Hierarchy (DGH) and Value Generalization hierarchy (VGH).

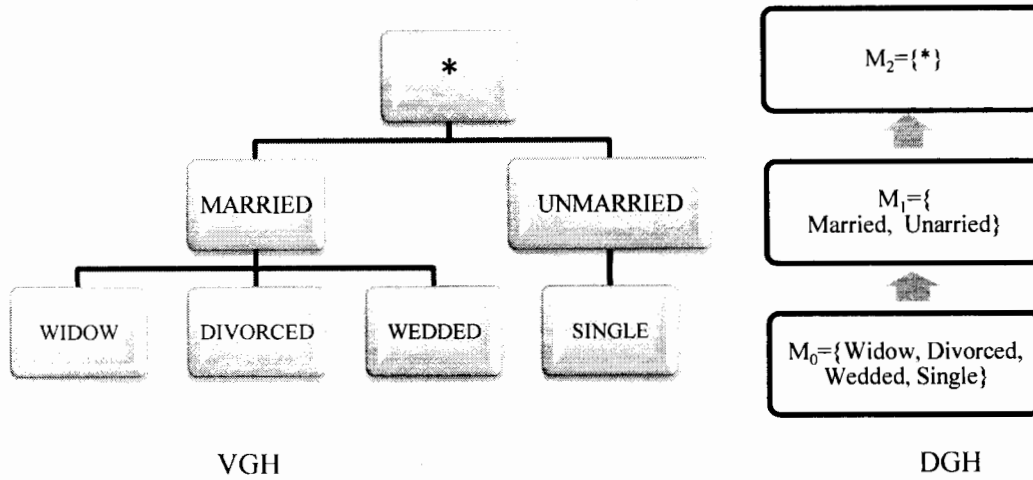


Figure2.1(a): DGH and corresponding VGH for individual's Marital Status

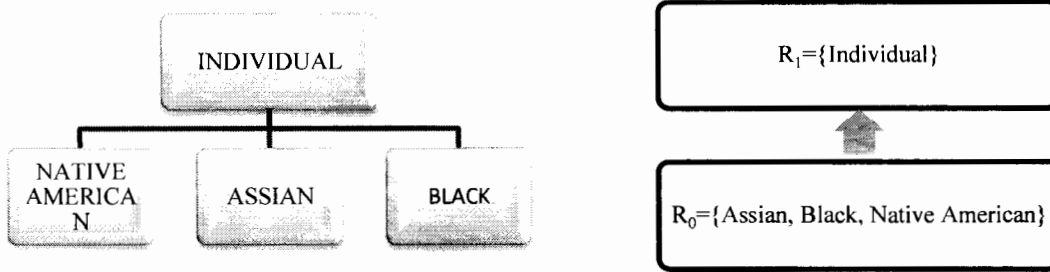


Figure2.1 (b): DGH and VGH for Race of the individual

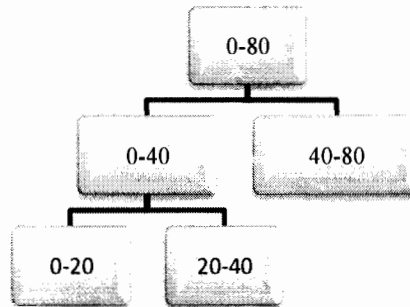


Figure2.1(c): DGH for Age



Figure2.1 (d): DGH and corresponding VGH for Gender of individuals

2.4.6 Generalization Lattice

When a data owner is supposed to generalize more than one attribute, a generalization lattice may be created to visualize maximum possible combinations for generalized domains. It consists of distance vector and the way they are interconnected to each other showing up to which level generalization can be performed and proves useful. For example consider table 2.1 and figure 2.1(d) a generalization lattice can be created to visualize attribute Sex and Zip code [14]

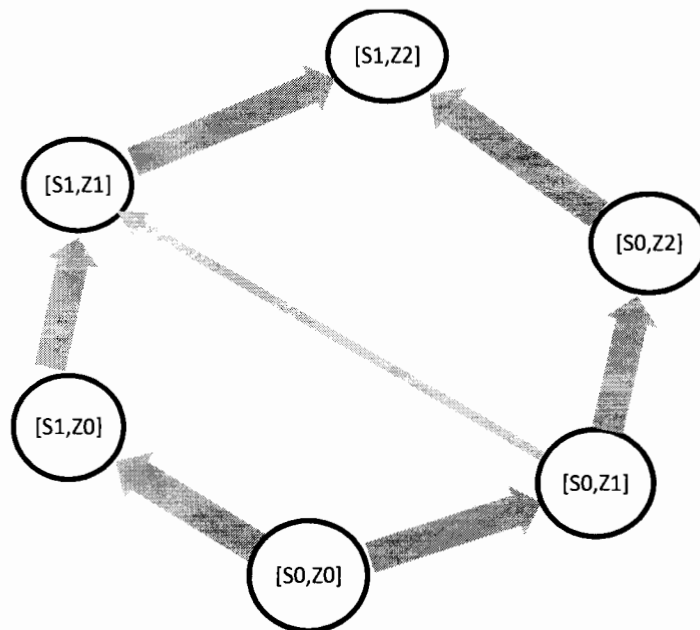


Figure2.2: Generalization Lattice for Zip Code and Sex attributes

2.4.7 Generalization Lattice Level

Vectors set with equal length in generalization lattice is called generalization lattice. Generalization lattice level can grow from low level to high starting from level 0 to some higher value upward in the lattice. In figure 2.2 generalization lattice has four levels. At level 0, there is vector $[S0, Z0]$, at level 1 there is vector $[S1, Z0]$ and so on.

CHAPTER 3: RELATED WORK

3 Related work

We have studied many research papers and articles to learn about privacy preservation and potential threats to confidential information regarding individuals. Different privacy protection models presented by researchers, their strengths and weaknesses were explored and evaluated with respect to well known privacy measures i.e. similarity attack, background knowledge attack, distortion ratio and running time etc. Amongst many, we have selected few techniques of privacy disclosure prevention to study in detail and evaluate the enhancements done so far.

We discussed t-closeness, ρ -sensitive k-anonymity, L-diversity and finally (ρ, a) sensitive k-anonymity in remaining section of this chapter.

It is noticed that k-anonymity is working well against identity disclosure while attribute disclosure is still a challenging problem need to be addressed. Previous work [6,7,8,9,12] made various attempts to solve attribute disclosure problem but still left some limitation. We proposed an optimum solution to afore mention problem presented at the end of this chapter.

3.1 K-anonymity

Whenever there is a debate on privacy protection in today's information sharing society, k-anonymity is definitely the only model that manifests assurance of confidentiality of sensitive information regarding individuals. It has been the most popular tool widely used to prevent privacy disclosure [5, 9, 14, 18, 20]. Our literature study shows that k-anonymity is working efficiently against identity disclosure but it could not provide any satisfactory solution to prevent attribute disclosure [1, 2]. Identity disclosure is the phenomenon when a respondent is directly linked to certain record in published table while attribute disclosure is the dilemma rise when some information from external source are disclosed and by synchronizing it with the published data, confidential information of the entity can be obtained. This is the main drawback of k-anonymity. The root cause of the failure of the k-anonymity is homogeneity and background knowledge attack. It is quite helpless meeting these challenges. The following examples clearly illustrate the problem. Consider table given below.

Table3.1 (a): Medical data

S.No	ZIP Code	Age	Nationality	Health position
1	23021	30	Russian	Cardiac problem
2	23023	31	US	Cardiac problem
3	23065	23	Japanese	Stomach Ulcer
4	23059	25	US	Stomach Ulcer
5	54824	52	Indian	Blood Cancerr
6	54827	57	Russian	Cardiac problem
7	54828	49	US	Stomach Ulcer
8	54829	45	US	Stomach Ulcer
9	33076	41	US	Blood Cancer
10	33053	43	Indian	Blood Cancer
11	33068	40	Japanese	Blood Cancer
12	33068	44	US	Blood Cancer

Table 3.1(a) shows a fictitious table belong to some hospital, published after removal of identifying attributes to ensure privacy measures. This table comprises of three non sensitive attributes namely zip code, nationality and age while one sensitive attribute medical position. These attributes are called quasi identifiers. The data holder manipulated the table before release hence it is slightly de-identified by eliminating the name attribute. If k-anonymity is applied on the above table, it can be transformed as following.

Consider table 3.1(b) which is a 4-anonymous view of table above table.

In above table data is generalized by applying 4-anonymity. '*' represents a suppressed digit, hence zip code 5484* depicts that it lies between 54820 to 54830 and 4* denotes the age range between 40 to 50 years, while attribute nationality is completely suppressed. Table is partitioned in group of four tuples indistinguishable from each other within the same QI group.

3.1.1 Homogeneity problem

Suppose an intruder acquires SSN number of a person and enter it in voter registration to search data against it. By doing this all non sensitive data about an individual can be obtained including name, age, zip code and nationality. This data may be synchronized with micro data published by hospitals to identify a targeted individual. If an intruder comes to know that Chris is 43 years old US national living in zip code 33068. The intruder can easily know that

the confidential record of Mr. Chris lies in Serial No. 9 to 12. As all patients have similar health position as Blood Cancer so it can firmly be assumed that Mr. Chris is suffering from blood cancer.

Likewise other confidential information like monthly income, savings etc can also be retrieved. This is a huge privacy breach and must be taken into account in order to preserve confidentiality of individual.

From this example it is quite clear that k-anonymity is working efficiently against identity revelation and does not deal with sensitive attributes disclosure.

Table3.1 (b): 4-anonymous view of table 3.1(a)

S.No	ZIP Code	Age	Nationality	Medical Condition
1	230**	<=35	*	Cardiac problem
2	230**	<=35	*	Cardiac problem
3	230**	<=35	*	Stomach Ulcer
4	230**	<=35	*	Stomach Ulcer
5	5482*	>=45	*	Blood Cancer
6	5482*	>=45	*	Cardiac problem
7	5482*	>=45	*	Stomach Ulcer
8	5482*	>=45	*	Stomach Ulcer
9	33076	4*	*	Blood Cancer
10	33053	4*	*	Blood Cancer
11	33068	4*	*	Blood Cancer
12	33068	4*	*	Blood Cancer

3.1.2 Background Knowledge based attack

Sometimes background knowledge of different people helps extra ordinary information retrieval. Culture, ethnicity, traditions and mores reveal worthwhile knowledge which may be used to infer particular entity. Suppose an entity named Umiko had been to hospital some day because of some disease. If an intruder wants to know what kind of disease Umiko is suffering from is not a difficult task to do. If he obtains the zip code, age and name of the Umiko, then he can get rest of the records from micro data published by hospital. From the above table it easily be identified that Umiko is 23 years old Japanese lady living in zip code

23065. Umiko record lies in first four of the table. If intruder does not know about background information, he may not be able to distinguish whether Umiko has stomach disease or cardiac issue. The one having background knowledge can easily conclude that Umiko may have stomach ulcer because it is obvious that Japanese has very low tendency towards heart disease.

The above example proves that k-anonymity cannot resolve background knowledge attack issue. So some stronger methodologies should be brought forward to cope with these problems.

Table3.1 (c): Background knowledge attack

S.No	ZIP Code	Age	Nationality	Medical Condition
1	1485*	> = 40	*	Cancer
2	1485*	> = 40	*	Heart Disease
3	1485*	> = 40	*	Viral Infection
4	1485*	> = 40	*	Viral Infection
5	130**	3*	*	Cancer
6	130**	3*	*	Cancer
7	130**	3*	*	Cancer
8	130**	3*	*	Cancer
9	130**	< 30	*	Heart Disease
10	130**	< 30	*	Heart Disease
11	130**	< 30	*	Viral Infection
12	130**	< 30	*	Viral Infection

To solve above mentioned problems of k-anonymity, Machanavajjhala et al. [6] proposed a model called as L- diversity.

3.2 L-diversity principle

A quasi identifier group satisfies *l*-diversity if it consist of at least *l* well represented values against sensitive attribute [6]. Here well represented means the arrangement of sensitive tuples in such a way that it is guaranteed that there are at least *l* distinct values of sensitive attributes in equivalence class or quasi identifier group. The main objective of this principle was to develop a balance amongst sensitive value in a quasi identifier group such that

different sensitive values appear against particular records. Parameter l represents distinct sensitive values in equivalence class.

3.2.1 Probabilistic l -diversity

Probabilistic l -diversity describe that the density of sensitive values in an equivalence class/QI group must be $1/l$, such table is said to have probabilistic l -diversity. This definition ensures that an attacker cannot identify particular tuple regarding individual with probability greater than $1/l$.

3.2.2 Recursive (c, l) -diversity

The concept of Recursive (c, l) -diversity enforces even distribution of sensitive attribute values in equivalence class/QI group i.e. the values appear most commonly should be restricted to be shown less frequently while the values appear less frequently should not appear too rarely in order to maintain equivalence. In Recursive (c, l) -diversity, c is a float and l is an integer number.

Despite of above mentioned measures, it is noted that l -diversity principle has some limitations due to which privacy preservation of respondents cannot be achieved in an efficient manner.

In following sections a few drawbacks of l -diversity are discussed with examples.

3.2.3 Drawbacks of l -diversity

While preventing sensitive attributes from unauthorized access, l -diversity enhanced k -anonymity model in an efficient manner. Many core issues are resolved to some extent but still rest of the problems regarding privacy need to be fixed.

3.2.4 Skewness Attack

The equivalence class having uneven distribution of sensitive attributes may lead towards skewness problem. If a data set consists of equal number of sensitive attribute values for example, students of a certain college having their result as either pass or fail. If pass/fail tuples are equal, then it may satisfy 2-diversity but still it reveal sensitive information with 50% accuracy as the probability to have an unwanted result i.e failed outcome is 50% while fail students are almost 1% of the entire students data set.

Suppose, we have an equivalence class having 400 students, there results are published with 392 students passed and 8 students stood failed. To anonymize this result by applying 2-

diversity, 98% student will consider themselves as failed. So it can be concluded that both groups proves different level of privacy leakage threats having similar level of diversity.

So the above examples show that *l*-diversity is not sufficiently capable to deal with skewed data.

3.2.5 Probabilistic Inference threat

When sensitive attributes distribution in quasi identifier group is unbalanced, probabilistic inference attack will took place, consequently the secret information of individuals will be compromised. L-diversity is not capable enough to cope with this problem.

The following 3-diversed table is manifestation of the phenomenon.

Table3.2: 10-anonymous data set with 3-diversity

Zip code	Age	Country	Medical Condition
220**	< 35	*	HIV
220**	< 35	*	HIV
220**	< 35	*	HIV
220**	< 35	*	HIV
220**	< 35	*	HIV
220**	< 35	*	Cancer
220**	< 35	*	Hepatitis
220**	< 35	*	HIV
220**	< 35	*	HIV
220**	< 35	*	HIV

Look at the table, equivalence class possess 10 records/tuples. The sensitive attribute values are 3-diversed i.e. there are at least 3 different sensitive values in equivalence class. So out of 10 sensitive values there are 8 similar values, which means an intruder can identify a targeted person disease as "HIV" with 80% accuracy. So this is a major drawback of the technique.

3.2.6 Similarity Attack

An equivalence class having different sensitive values but there meanings may be interpreted semantically identical can also reveal confidential information regarding individual. The phenomenon is illustrated in the following example by the help of given table

Table3.3: Micro data

S.No	Zip code	Age	Salary	Health status
1	66377	42	4K	Gastric ulcer
2	66302	43	4K	Gastritis
3	66378	45	6K	Stomach cancer
4	67205	27	3K	Gastritis
5	67209	23	14K	Flu
6	67206	29	11K	Bronchitis
7	87605	54	10K	Bronchitis
8	87673	51	5K	Pneumonia
9	87607	57	17K	Stomach cancer

The above table is anonymized with 3-diversity (distinct and entropy) and transformed to the following.

Table3.4: 3-anonymous data set with 3-diversity

S.No	Zip code	Age	Salary	Health status
1	663**	4*	4K	Gastric ulcer
2	663**	4*	4K	Gastritis
3	663**	4*	6K	Stomach cancer
4	672**	>20	3K	Gastritis
5	672**	>20	14K	Flu
6	672**	>20	11K	Bronchitis
7	876**	5*	10K	Bronchitis
8	876**	5*	5K	Pneumonia
9	876**	5*	17K	Stomach cancer

The table given above possesses two sensitive attributes namely Health status and Salary. Both attributes need to be kept confidential in order to preserve the secrets of individuals. But still it gives way to the intruders to access private information. For example if a person comes to know that Mr. Chris is 43 years old American and his house is located in zip code 66302, then it can easily be concluded that Mr. Chris is collecting salary between range of {4000 to

6000}. Furthermore it can also be identified that Mr. Chris is suffering from stomach problem.

3.4 t-closeness

The concept of t -closeness [7] defines a distance amongst sensitive attributes to protect against sensitive attributes revelation. In other words it represents the distribution of sensitive attribute values in any equivalence class/quasi-identifier group is close to the distribution of attributes in the whole table, i.e. the distance must be a threshold value t amongst the distribution of the attributes in the QI-group and between the rest of the tables.

To calculate the distance between two distinct distributions, t -closeness technique uses EMD (Earth Mover Distance) [31] as distance metric.

3.4.2 Limitation of t-closeness

To implement t -closeness, there must be some computational procedure which is not offered by this property, even the correlation would be destabilized between sensitive attributes and other quasi identifiers by the implementation of t -closeness property. To generalize each attribute individually will reduce their dependence on each other. Furthermore, small value of t may destabilize the utility and result an increase in computational cost.

3.5 P-sensitive k-anonymity model

This property describes a table as p -sensitive k -anonymous, if equivalence class contain at least p distinct sensitive attribute values while satisfying k -anonymity [8].

Sensitive attributes of data set is partitioned into four categories according to its level of confidentiality. We can see sensitive attribute health status is partitioned in to four categories shown in following table.

Table3.5: Categories

Category No.	Health status	Level of Confidentiality
1	HIV ,Cancer	Most Confidential
2	Phthisis ,Hepatitis	Confidential
3	Obesity ,Asthma	Less Confidential
4	Flu ,Indigestion	Non Confidential

P-sensitive k-anonymity model is a remarkable enhancement in k-anonymity but still there are few drawbacks of this model which are illustrated in the following section.

3.5.1 Similarity Attack

As we noticed in *l*-diversity, sensitive attributes were maintained distinct in an equivalence class but still has some semantically identical values. P-sensitive k-anonymity property also fails to provide acceptable solution to address similarity attack problem.

Table3.6: Micro data

Zip code	Age	Country	Health Status
23021	30	Denmark	Flu
23023	31	France	Asthma
23065	23	Germany	Flu
23059	25	France	Indigestion
54824	52	Japan	Hepatitis
54827	57	China	Obesity
54828	49	Pakistan	Flu
54829	45	Pakistan	Phthisis
33076	41	Canada	HIV
33053	43	USA	Cancer
33068	40	Canada	Cancer
33068	44	Canada	HIV

The above table consists of micro data a hospital gathered from individuals. The sensitive attribute of the data set can be viewed as partitioned into four categories. The purpose the this effort is to enhance the privacy of respondent but you can see in table 3.7 secret information can be disclosed.

Table3.7: 2-sensitive 4-anonymous data

Zip code	Age	Country	Health status
230**	<=35	Europe	Flu
230**	<=35	Europe	Indigestion
230**	<=35	Europe	Flu
230**	<=35	Europe	Indigestion
5482*	>=45	Asia	Hepatitis
5482*	>=45	Asia	Obesity
5482*	>=45	Asia	Flu
5482*	>=45	Asia	Flu
33076	4*	America	HIV
33053	4*	America	Cancer
33068	4*	America	Cancer
33068	4*	America	HIV

From the above table it can be seen that the last four tuple of the table having sensitive attribute values belong to same most confidential category and reveal most sensitive secrets regarding individuals. This is a huge privacy breach and need to be resolved. It is proved from the above example that p -sensitive k -anonymity fails to address disclosure prevention of confidential attribute values.

To solve these problems an enhanced technique (p, α) sensitive k -anonymity [9] model was introduced.

3.6 (p, α) -sensitive k -anonymity

This property describes, a table is said to be (p, α) sensitive k -anonymous if it satisfies k -anonymity and every equivalence class must have minimum p sensitive values having total weight of the equivalence at least α [9].

This is an efficient and most enhanced version of k -anonymity but still having some limitations.

Table3.8: Raw data in hospital

Zip code	Age	Country	Health condition
75359	26	Canada	HIV
75308	25	USA	Hepatitis
75305	27	USA	Obesity
75308	24	Canada	Cancer
67264	42	USA	Asthma
67285	45	China	Phthisis
67275	48	Pakistan	HIV
67273	41	Pakistan	Flu
25306	32	Canada	Asthma
25305	35	Canada	Phthisis
25306	36	Canada	Flu

(p, α) sensitive k-anonymity is applied to the following table

Table3.9: (3, 1)-sensitive 4-anonymous table

Zipcode	Age	Country	Health condition	Weight	Total
7****	<30	*	HIV	0	1
7****	<30	*	Cancer	0	
7****	<30	*	HIV	0	
7****	<30	*	Flu	1	
672**	>40	US	Hepatitis	1/3	2
672**	>40	US	Phthisis	1/3	
672**	>40	US	Asthma	2/3	
672**	>40	US	Obesity	2/3	
--	--	--	--	--	3

From the above table it can be seen that then necessary condition of (p, α) sensitive k-anonymity property are satisfied i.e. $p=3$, weight of equivalence class is at least α , even then the first equivalence class/quasi identifier group contains sensitive attribute values distributed unevenly. Three out of four sensitive values are part of same category which helps an attacker to reach a targeted person confidential information easily with accuracy of 75%. Hence privacy of the respondents compromised.

3.7 Problem Statement

We have discussed improved versions of k-anonymity such as L-diversity, p-sensitive k-anonymity, t-closeness and (p, α) -sensitive k-anonymity. It can be noticed from the literature that these enhanced models still have many limitations which may cause privacy breach. Amongst all, (p, α) -sensitive k-anonymity offers relatively better solution as compared to the previous models. Our study identified few weaknesses of (p, α) -sensitive k-anonymity mentioned below.

a). (p, α) -sensitive k-anonymity model generalizes entire data set without any distinction which results in loss of the valuable information and needless overhead of computational cost. For small portion of sensitive attributes, all data set is passed on through anonymization process where each tuple of the data set is generalized. Thus data set is destabilized as noise is incorporated to every record of micro data while computational complexity is increased on the other hand.

b). (p, α) -sensitive k-anonymity mainly focus on sensitive values and assigns weight to each sensitive element to enhance privacy but the distribution of sensitive attribute values in quasi identifier group is uneven. Some values appear more frequently while other less. This phenomenon leads towards probabilistic attack and allow the intruder recognize particular record easily. For example:

Consider table 3.8 where first four tuples may suffer from potential threat of probabilistic attack. It can be seen that three tuples out of four i.e. (HIV, Cancer, HIV) belong to top secret category. So an individual can be identified with 75% accuracy.

CHAPTER 4: PROPOSED SOLUTION

To improve efficiency, and maximize the volume of published data and minimize the similarity attack to an optimum level, an efficient algorithm (p, β) sensitive k-anonymity is proposed. In proposed technique Discriminated union β of sensitive attribute values is introduced which ensures distinct sensitive values in each equivalence class.

A table satisfy (p, β) sensitive k-anonymity, if it satisfy k-anonymity and for each QI group there are at least $p=k$ distinct sensitive attribute values with Discriminated union β over p .

4.1 Proposed technique (p, β) sensitive k-anonymity

A table satisfy (p, β) sensitive k-anonymity, if it satisfy k-anonymity and for each QI group there are at least $p=k$ distinct sensitive attribute values with discriminated union β over p .

Before the implementation of (p, β) sensitive k-anonymity property we perform data reduction process on data set containing micro data.

Pre-processing step: First the whole data set is scanned for sensitive tuples. Then sensitive tuples are separated from non sensitive tuples and moved to another table called ST. Now there are two tables containing homogeneous non aggregated data i.e. table NST consists of non sensitive tuples while table ST contains only sensitive tuples. To ensure privacy some distortion is added to table ST and a portion of non sensitive tuples from table NST is imported to sensitive tuples table ST.

Table NST will be published directly without any modification because it contains non sensitive tuples and does not reveal any confidential information regarding individuals, while table containing sensitive tuples ST is processed for further privacy measures as below.

Generalization step: To ensure privacy protection and prevent confidential information disclosure, (p, β) sensitive k-anonymity is used, where sensitive attributes are partitioned into different categories or sets so that similarity and background knowledge probabilistic attack can be prevented.

Suppose C is an attribute set containing eight distinct sensitive values. Attribute C is further partitioned into four categories or sets (C_1, C_2, \dots, C_4) . The categories manifests the degree of sensitivity of value belong to it. $C = \bigcup_{i=1}^x C_i, C_i \cap C_j = \emptyset$ for $(i \neq j)$.

As mentioned earlier medical condition is a sensitive attribute consist of eight diseases i.e.

$D = \{\text{HIV, Hepatitis, Asthma, Indigestion, Cancer, Phthisis, Obesity, Flu}\}$ so it is partitioned into four different categories with respect to their level of confidentiality. From the table given below it is quite clear that category A is most confidential while category D is least. As we move downhill in category table from category A through D, the level of sensitivity decreases.

Table4.1 (a): Sensitive attribute partition into categories

Category	Category values	C_i	C_j
A	HIV , Cancer	HIV	Cancer
B	Hepatitis, Phthisis	Hepatitis	Phthisis
C	Asthma, Obesity	Asthma	Obesity
D	Indigestion, Flu	Indigestion	Flu

Table 4.1 (b): Raw data

S.No	ZIP Code	Age	Nationality	Medical Condition	Categories
1	23021	30	Russian	HIV	A
2	23023	31	US	HIV	A
3	23065	23	Japanese	Flue	D
4	23059	25	US	Indigestion	D
5	54824	52	Indian	Hepatitis C	B
6	54827	57	Russian	Phthisis	B
7	54828	49	US	Asthma	C
8	54829	45	US	Obesity	C
9	33076	41	US	HIV	A
10	33053	43	Indian	Cancer	A
11	33068	40	Japanese	Flu	D
12	33068	44	US	Indigestion	D

The above table contains raw data consists of original information about different patients. Some values in the table are confidential and need to be protected. After applying proposed technique, the above table can be transformed into more general but secure one as shown in table 4.2.

To minimized similarity attack a Discriminated union β of sensitive attribute values is proposed. We illustrate here how to calculate Discriminated union in section 4.2(a) and 4.2(b) while data reduction and generalization are presented in section 4.5.

Discriminated union can be defined and calculated as below. Note that the word “category” and “set” are used interchangeably having similar meaning as a set of sensitive attribute values.

4.1.1 Discriminated Union of Sensitive values

Two categories A and B are said to be discriminated if they possess no elements in common, i.e., $A \cap B = \emptyset$

For example:

$A = \{\text{HIV, Cancer}\}$ $B = \{\text{Hepatitis, Phthisis}\}$

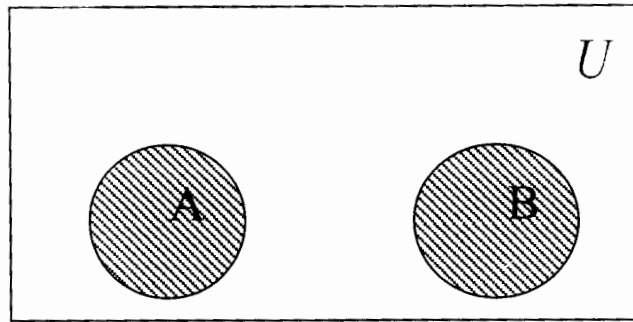


Figure 4.1(a): Discriminated set

4.1.2 Discriminated or Disjoint Union of sensitive values

Discriminated union of sensitive values can be defined as set of values belong to different categories such that each value in the discriminated union is distinct and none overlapping within the set.

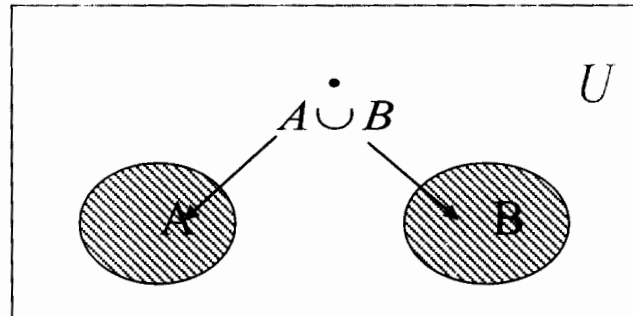


Figure 4.1(b): Discriminated union

To calculate Discriminated union of sensitive values Union-fined algorithm is use which performs the following two useful operations.

1. Find
2. Union

Find function: this function search the categories and determine which particular value belongs to it. It can also identify if two values belong to same category so it can prevent repetition.

Union function: this function joins two elements from different categories into single subset.

The methodology is illustrated in the following example where $k=p=4$ and Discriminated union over p is calculated using Union-fined algorithm.

Table 4.2: data set with Discriminated union of sensitive values $p=k=4$

S.No	ZIP Code	Age	Nationality	Medical Condition	Categories
1	230**	≤ 35	*	HIV	A
2	230**	≤ 35	*	Hepatitis	B
3	230**	≤ 35	*	Asthma	C
4	230**	≤ 35	*	Indigestion	D
5	5482*	≥ 45	*	Cancer	A
6	5482*	≥ 45	*	Phthisis	B
7	5482*	≥ 45	*	Obesity	C
8	5482*	≥ 45	*	Flue	D
9	330**	4*	*	HIV	A
10	330**	4*	*	Hepatitis	B
11	330**	4*	*	Asthma	C
12	330**	4*	*	Indigestion	D

From the above table it can easily be seen that how successfully similarity attack problem is addressed. The table shows that all the sensitive elements are disjoint and no one is repeated with in particular equivalence class.

Cardinality of the Discriminated union dependent on p and k . if $p=4$, then k is also equal to 4, consequently cardinality of Discriminated union is 4.

For $p=4$, initially disjoint/discriminated union is empty i.e. $\beta = \{\emptyset\}$.

D is the universal set contains all the disease i.e. $D = \{\text{Cancer, HIV, Hepatitis, Phthisis, Asthma, Obesity, Indigestion, Flu}\}$ which is further partitioned into four categories listed in table 4.1(a).

Name of the category is called root or representative of the category i.e. A, B, C and D are the roots/ representatives of the categories.

β is the set contains discriminated union of distinct values of different categories. Initially β is empty i.e.

$\beta = \{\}$. Our algorithm takes the first sensitive value from universal set D containing diseases, and compares it with the values already stored in discriminated union set β . First roots of the

values are compared with each other, if match is found then its Childs are compared if no match is found then the value is added to discriminated union set β , otherwise the value is ignored and next value from universal set D is processed and the process is repeated until cardinality of Discriminated union set become equal to p.

4.1.3 Solved Example

For example if $p = k = 4$ then Discriminated union for table 4.2 is determined as follows.

Initially the first element of universal data set D is taken which is *ith* element of category A i.e. ($A[i] = \text{HIV}$) and stored in discriminated union set β . Now $\beta = \{\text{HIV}\}$

On the next iteration Second element of universal set D is processed which is the first *ith* element of category B (*ith* element $B[i] = \text{Hepatitis}$). Root of $B[i]$ is then compared with the root of the element previously stored in β i.e. HIV. The fined () function match the elements with each other, if match is found, the element is discarded otherwise stored in the discriminated union set β . Now Discriminated union contains two elements i.e. $\beta = \{\text{HIV}, \text{Hepatitis}\}$.

In the next iteration of the loop third element of universal set D which is Asthma is processed which is the first element (*ith* element $C[i] = \text{Asthma}$) of category C. root of $C[i]$ is then compared with the root of the elements previously stored in β i.e. HIV and Hepatitis. If no match is found, the element is added to Discriminated union set which now contains three elements i.e. $\beta = \{\text{HIV}, \text{Hepatitis}, \text{Asthma}\}$.

Similarly the loop will execute its last iteration and processed fourth element in data set D which is the first element (*ith* element $D[i] = \text{Indigestion}$) of category D. Root of $D[i]$ is then compared with the roots of previously stored three distinct values in Discriminated union set. If no match is found; the element is added to the Discriminated union set otherwise discarded. Loop for Discriminated union terminates after completing fourth iteration. Discriminated union of QI group in table 4.2 now contain four distinct sensitive values i.e. $\beta = \{\text{HIV}, \text{Hepatitis}, \text{Asthma}, \text{Indigestion}\}$.

Here now first equivalence class is completed for $p = k = 4$. For the rest of QI groups *jth* elements of the categories are taken and processed in the above mentioned manner.

The process continues until all data set is processed.

For $k = 4$, initially every first element of each category is taken and placed in equivalence class while for the second equivalence class every second element of each category is taken and placed in equivalence class.

The first disjoint union of sensitive values with respect to corresponding equivalence class consists of A_i, B_i, C_i and D_i which are HIV, Hepatitis, Asthma and Indigestion respectively.

TH 9503

The second disjoint union consists of values A_j , B_j , C_j and D_j where the corresponding values are Cancer, Phthisis, Obesity and Flue. The algorithm uses values C_i and C_j on alternative iterations as can be observed in table 4.4.

4.2 Anonymization Algorithms

To anonymize a given data set, many algorithms have been proposed so far with their own efficiency and limitations. However fewer got much popularity because of efficiency, running time, precision and accuracy which are given below.

4.2.1 Global Recoding Incognito Algorithm

Global recoding is one of the consistent methods that generalize a table at domain level. Most anonymization methods are based on global recording incognito model. Incognito generates k -anonymous full-domain generalized set of attributes. It provides another option to put a threshold on suppression to ensure specified level of tuple suppression. Using subset property, it evaluates single attribute subsets contained in QI group and in rest of the iterations it checks k -anonymity for progressively large subsets. During iterations, two important tasks are performed as illustrated below.

First iteration: Graph of candidate nodes (C_i) generated from subset of QI having size i is considered and processed accordingly and using rollup property and generalization, bottom up-breath first search is performed represented by S_i .

Second iteration: In second iteration on obtaining S the algorithm develops set of candidate nodes C_{i+1} for QI having size $i+1$ and use subset property to avoid the nodes which cannot be helpful while size of the set of the attribute is large.

When bottom up breath first algorithm runs, in every iteration it is searching for QI attribute and determine whether it is k -anonymous or not according to candidate generalization(C_i).

For example, in iteration 1 the algorithm check k -anonymity for every single attribute contained candidate generalization; if a node does not prove k -anonymity, it is eliminated in second iteration. If a node satisfies k -anonymity than all of its direct generalizations are marked true and not checked in subsequent iterations using generalization property. All generalizations that fulfill k -anonymity property are paired.

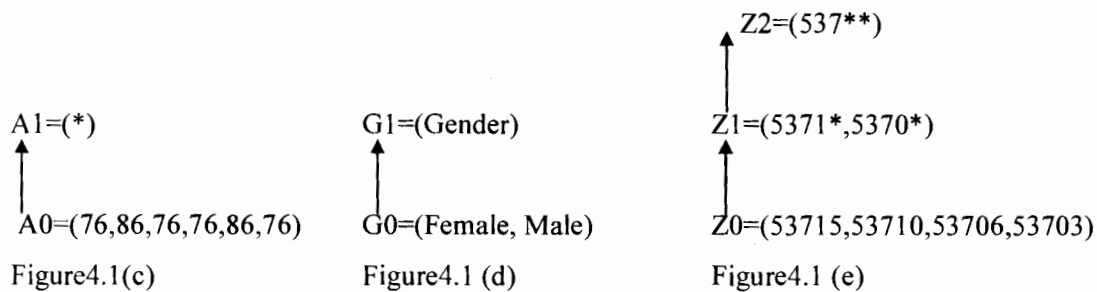
Consider the following table with quasi identifiers Zip code, Gender and Age where $k=2$ and the maximum suppression threshold is 2.

Table 4.3: Micro data for generalization hierarchy

Zip code	Gender	Age	Medical Condition
53715	Male	76	Hypertension
53715	Female	86	Hypertension
53703	Male	76	Obesity
53703	Male	76	HIV
53706	Female	86	Obesity
53706	Female	76	Flu

In the below figures 4.1(c)(d) and (e) value generalization hierarchies of given quasi-identifiers of the all of the subsets are presented on the left side and the corresponding sub-hierarchies computed by incognito algorithm during execution of every iteration are shown towards the right side with respect to above table in the following figure.

In the figures, Age, Gender and Zip code are represented by A, G and Z respectively



Here are the domain generalization hierarchies shown in above figures (c) (d) (e).

Attribute Age can be generalized to two levels i.e. level $A_0 = (76, 86, 76, 76, 86, 76)$ and level $A_1 = (*)$, and it cannot be generalized further. Likewise attribute Gender can be generalized to level $G_0 = (\text{Female}, \text{Male})$ and level $G_1 = (\text{Gender})$ and attribute Zip code can be generalized up to three levels i.e. level $Z_0 = (53715, 53710, 53706, 53703)$, level $Z_1 = (5371*, 5370*)$ and level $Z_2 = (537**)$. Maximum suppression threshold is 3.

First iteration:

In the first iteration Incognito determines that table T is k-anonymous with respect to un-generalized quasi identifiers Age, Gender, Zip code (A_0), (G_0) and (Z_0) respectively at initial stage.

Second iteration:

In second iteration three bottom up breath first searches are performed to find table T is 2-anonymous with respect to Quasi identifiers (Age, Gender), (Age, Zip code) and (Gender, Zip

code). In first search the algorithm analyzes the nodes and generates the frequency set of table T with respect to (G_0, Z_0) and determines that it does not satisfy 2-anonymity. This frequency set is rolled up and generate another frequency according to (G_1, Z_0) and (G_0, Z_1) . The table is 2-anonymous with respect to (G_1, Z_0) , so all of its direct generalizations (G_1, Z_1) and (G_1, Z_2) are marked and considered 2-anonymous according to generalization property. Table T does not fulfill 2-anonymity conditions with respect to (G_0, Z_1) so this edge is removed from graph while frequency set (G_0, Z_2) satisfy 2-anonymity and the breath first search is completed for multi-attribute generalization with respect to $(\text{Gender}, \text{Zip code})$.

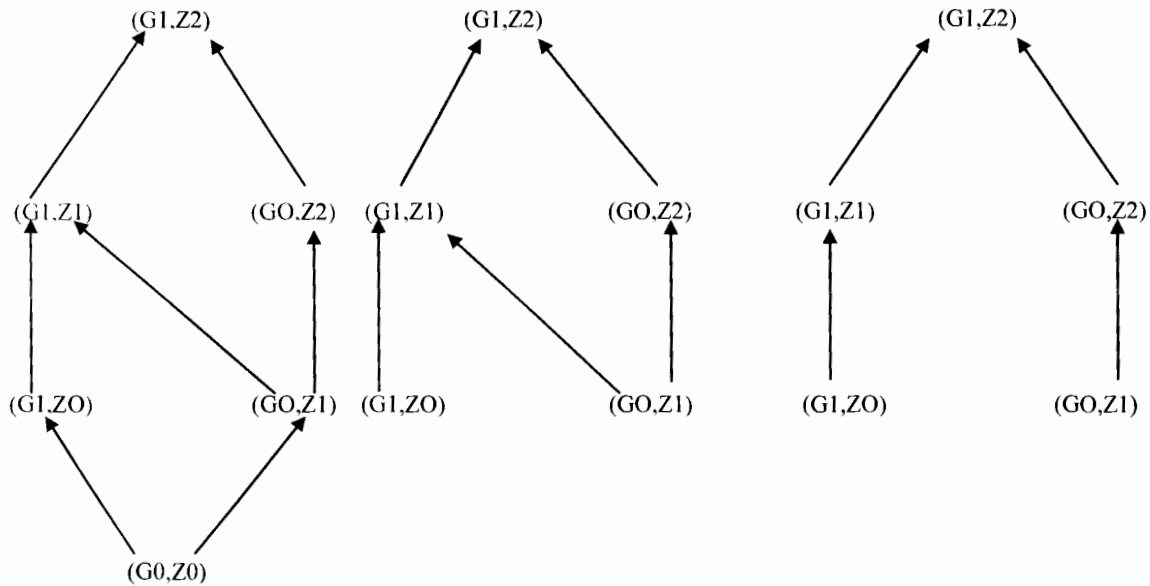
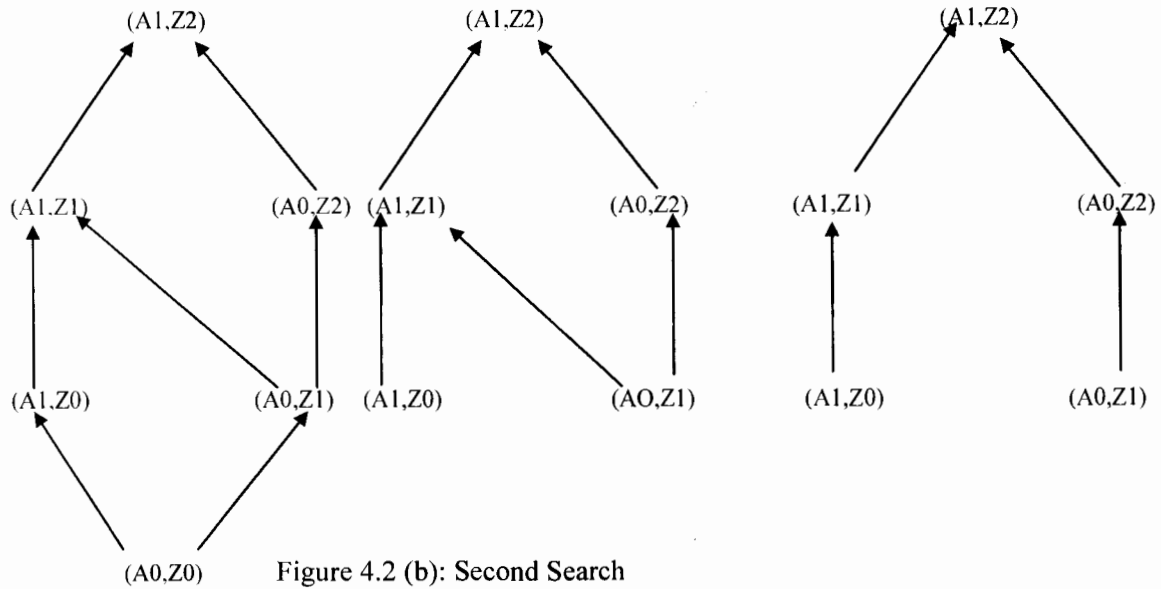


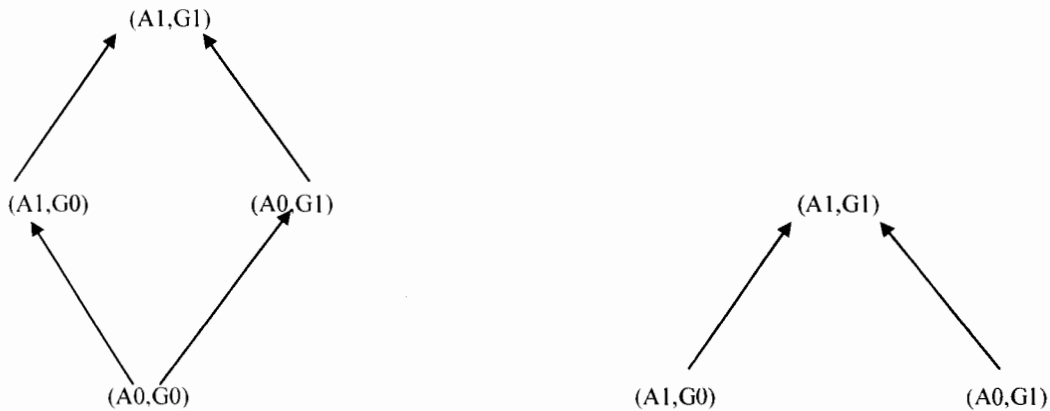
Figure 4.2(a): First Search

In second search the algorithm evaluates the nodes and generates the frequency set of table T with respect to (A_0, Z_0) and determines that it does not satisfy 2-anonymity. This frequency set is rolled up and generate another frequency according to (A_1, Z_0) and (A_0, Z_1) . The table is 2-anonymous with respect to (A_1, Z_0) , so all of its direct generalizations (A_1, Z_1) and (A_1, Z_2) are marked and considered 2-anonymous according to generalization property. table T does not fulfill 2-anonymity conditions with respect to (A_0, Z_1) so this edge is removed from graph while frequency set (A_0, Z_2) satisfy 2-anonymity and the breath first search is completed for multi-attribute generalization with respect to $(\text{Gender}, \text{Zip code})$.



In the third and final search our algorithm generates frequency set of the above table T with respect to (A_0, G_0) and found that it does not satisfy 2-anonymity so this frequency is rolled up and another frequency set (A_1, G_0) and (A_0, G_1) of the is processed. Table T is 2-anonymous with respect to (A_1, G_0) so using generalization property its direct generalization (A_1, G_1) is also 2-anonymous and marked as generalized.

On the other hand (A_0, G_1) does not satisfy 2-anonymity so it is discarded from Sub hierarchy graph and the search are finished.



4.3 Algorithm

- Step 1: Generalize all tuples such that all tuples become equal.
- Step 2: G is the set consist of generalized tuples and
 $U \leftarrow \{G\}; 0 \leftarrow \emptyset$
- Step 3: **Repeat**
- Step 4: $U \leftarrow \emptyset$
- Step 5: **For all** $G \in U$ **do**
- Step 6: specialize all tuples contained in G and push them one level down in generalization hierarchy constructing a child node while satisfying conditions of (p, β) sensitive k -anonymity
- Step 7: if a node does not satisfy (p, β) sensitive k -anonymity, move it back to the parent node G and unspecialize it
- Step 8: if parent node G does not satisfy the condition, then unspecialized some child node and move upward in hierarchy to parent so that parent node fulfill (p, β) sensitive k -anonymity
- Step 9: End of **if** structure
- Step 10: For all non-empty branches R of G , **do** $U' \leftarrow U' \cup \{R\}$
- Step 11: $U \leftarrow U'$
- Step 12: If G is non-empty then $0 \leftarrow 0 \cup \{G\}$
- Step 13: End of **for** loop
- Step 14: Until $U = \emptyset$
- Step 15: **Return** 0 .

Illustration:

The above algorithm runs in two phases i.e. data reduction phase and generalization phase.

In the data reduction phase the algorithm determines tuples which are mandatory for generalization and for this purpose entire data set D as in table 4.4 (a) is traversed to find sensitive tuples. When all the data set is scanned and sensitive tuples are found then these sensitive tuples are separated from the data set and moved to another table called T as in Table 4.4 (b). This process continues until all tuples are moved to table T and data set D has no more sensitive tuples. To ensure privacy of the individuals a portion of sensitive tuples is imported from data set D and added to table T containing only sensitive tuples in such a way that in table T density of non sensitive tuples is 2/3 of the whole Table 4.4 (c), i.e. In table T if there are 10 sensitive tuples then there must be 20 non sensitive tuples as distortion.

In generalization phase table T is processed by Top down local recording algorithm and generalized in such a way that all tuples are indistinguishable from each other.

From step 6 specialization processes begins and all tuples are specialized one by one and moved one level down in generalization hierarchy. On specialization of every node (p, β) sensitive k-anonymity is necessarily satisfied. If a child node does not satisfy (p, β) sensitive k-anonymity, it is moved upward to parent node. If parent node does not satisfy (p, β) sensitive k-anonymity, some other child nodes are moved to the parent node to satisfy the condition. This process continues until all nodes are specialized. Table 4.4(d) consists of the data set containing micro data before anonymization. It can be seen that the data set contain contains sensitive and non sensitive tuples and the density of sensitive tuples in the entire data set is about 10%. So the sensitive tuples are separated and stored in table T as in table4.4 (b).

Table 4.4 (a): data set D

S.No	ZIP Code	Age	Nationality	Health position
1	23021	30	Russian	HIV
2	23023	31	US	Indigestion
3	23065	23	Japanese	Flu
4	23059	25	US	Indigestion
5	54824	52	Indian	Flue
6	54827	57	Russian	Phthisis
7	54828	49	US	Asthma
8	54829	45	US	Flue

9	33076	41	US	Cancer
10	33053	43	Indian	Indigestion
11	33068	40	Japanese	Flue
12	54829	45	US	Flue
13	33076	41	US	Indigestion
14	33053	43	Indian	Cancer
15	33068	40	Japanese	Flue
16	33068	44	US	Indigestion

Sensitive tuples are separated from above data set and stored in following table.

Table 4.4 (b): Table T with only sensitive tuples

S.No	ZIP Code	Age	Nationality	Health position
1	23021	30	Russian	HIV
9	33076	41	US	Cancer
6	54827	57	Russian	Phthisis
14	33053	43	Indian	Cancer

In the following table only non sensitive tuples are left after separating sensitive tuples from it. This table is supposed to be published without any major modification because it does not contain any confidential information regarding individuals.

Table 4.4 (c): data set containing only non sensitive tuples

S.No	ZIP Code	Age	Nationality	Health position
2	23023	31	US	Indigestion
3	23065	23	Japanese	Flu
4	23059	25	US	Indigestion
5	54824	52	Indian	Flue
7	54828	49	US	Flu
8	54829	45	US	Flue
10	33053	43	Indian	Indigestion
11	33068	40	Japanese	Flue
12	54829	45	US	Flue
13	33076	41	US	Indigestion
15	33068	40	Japanese	Flue
16	33068	44	US	Indigestion

Table 4.4(d) is modified by adding non sensitive tuples to it. And the ratio of the non sensitive tuples imported to it is $2/3^{\text{rd}}$ of the whole data set G as we can see in following table.

Table 4.4 (d): after adding $2/3^{\text{rd}}$ distortion

S.No	ZIP Code	Age	Nationality	Health position
1	23021	30	Russian	HIV
9	33076	41	US	Cancer
6	54827	57	Russian	Phthisis
14	33053	43	Indian	Cancer
5	54824	52	Indian	Flue
6	54827	57	Russian	Indigestion
7	54828	49	US	Flu
8	54829	45	US	Flue
10	33053	43	Indian	Indigestion
11	33068	40	Japanese	Flue
12	54829	45	US	Flue

The remaining part of the algorithm can be illustrated by the help of the following example.

Step 4 and onward, working mechanism of the algorithm is illustrated by the following tables and diagram.

Let us process only one quasi identifier i.e. Zip code in the table 4.5(a) containing micro data with four tuple where two of them are sensitive i.e. HIV and Cancer. For the data set threshold values are set to $k = 2, p = 2$. First all tuples are generalized as shown in figure 4.3(a) where we take only quasi identifier Zip code and converted to more general value *****.

After generalization, specialization process begins; this is reversal of the generalization. A tuples is specialized if it satisfy (p, β) sensitive k-anonymity and moved down in generalization hierarchy forming a child node i.e.5***** as shown in figure 4.3(b).

Table 4.5(a): Raw data

Zip code	Gender	Health condition
54824	Male	HIV
54824	Male	Hepatitis
54824	Female	Asthma
54823	Female	Indigestion

Table 4.5(b): Generalized Table

S#	Zip code	Medical condition
1	54824	HIV
2	54824	Hepatitis
3	5482*	Asthma
4	5482*	Indigestion

Further more in the next iteration, this process is continue obtaining branches with Zip codes 54***,548**,5482* shown in figures 4.3(c), 4.3(d) and 4.3(e) respectively. Tuples are further specialized into two branches as we see in figure 4.3(f). Specialization process will be successful if all tuples in generalization hierarchy are specialized, however there are exists some problematic leaves which does not satisfy (p, β) sensitive k-anonymity and need to be pushed upward in the hierarchy. In figure 4.3(g) it is quite clear that Zip code 54824 is alone and do not satisfy (p, β) sensitive k-anonymity , so it is pushed back to the parent node 5482*.

If parent node does not satisfy (p, β) sensitive k-anonymity, then push some of the child nodes upward in generalization hierarchy so that parent node satisfy the condition and specialized. In figure 4.3(g) , it is cleared that parent node54824* we moved tuple 54824 upward which satisfy (p, β) sensitive k-anonymity. In figure h we determined the data set with Zip code 3and 4 are generalized to 5482* and 5482* respectively while Zip code 1,2 are generalized to 54824, 54824 at the end table 4.5(b) is obtained which satisfy (p, β) sensitive k-anonymity.

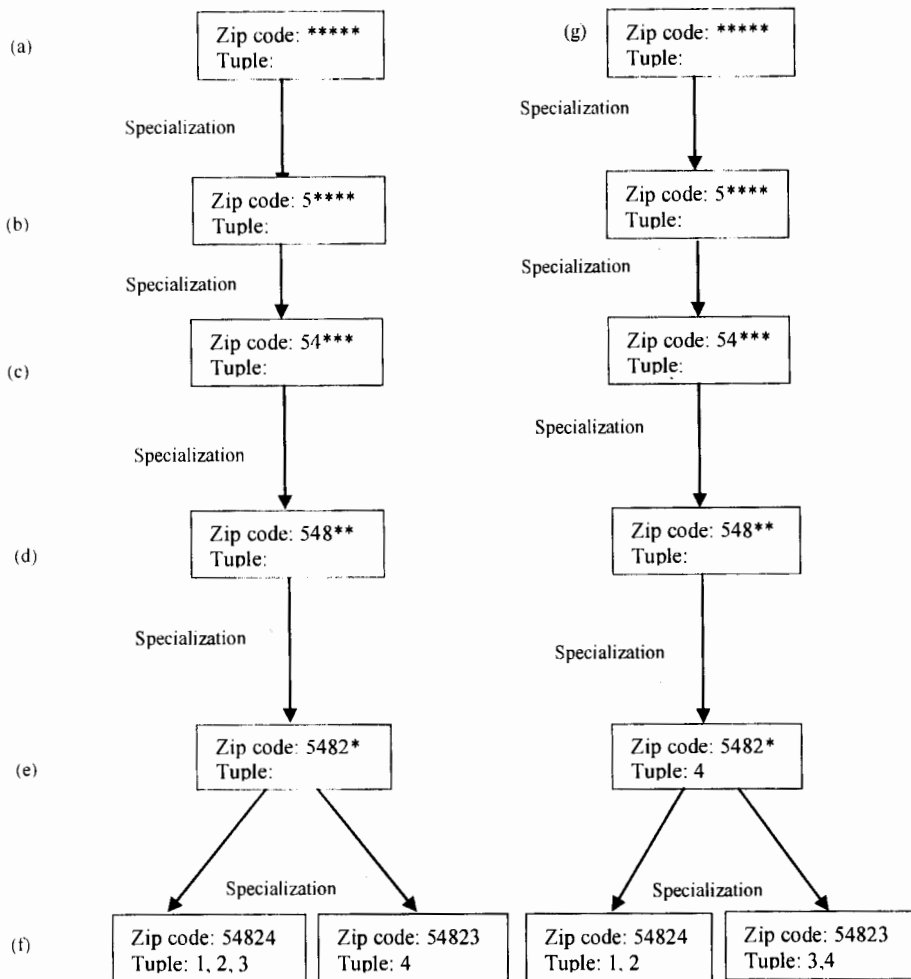


Figure4.3: Anonymization process for Zip code [9]

In this way all tuples in a data set are generalized. If there are some tuples which fulfill (p, β) sensitive k -anonymity and still un-specialized, then the specialization process is repeated again and again until all tuples satisfying the necessary condition are specialized. In this process information loss is minimized and maximum accurate data is published.

4.4 Successful specialization

During specialization, some tuples usually left un-specialized which is not acceptable and cause lose of useful information. Thus the specialization in which maximum data is specialized is the best specialization. For this purpose attributes of data set are analyzed and optimal quasi identifier attributes are selected so that information loss is reduced and the published data is precise.

CHAPTER 5: RESULTS & ANALYSIS

In this chapter results are calculated and compared with previous models. Data set which is known as Adult data set was obtained from UCI machine Learning Repository which is publically available. We determined running time, similarity attack and distortion ration of our technique and the results were compared with previous models.

Apart from the fact that the proposed model (p, β) sensitive k-anonymity provided optimum solution for similarity attack by using Discriminated union of sensitive attribute values but also working efficiently with respect to distortion ratio. Accuracy of the published data is more than 50% greater than other models. More than 60% non sensitive data is published without any modification so distortion ratio of the mention portion of the data set is reduced.

5 Dataset

UCI Machine Learning Warehouse [10] contains standard data repositories like adult data set mostly used in majority of experiments [6, 12]. We have also used the said real world data set in our implementation to prove its efficiency and usefulness. The data set is available publically, initially there were more than 45000 records in the data set; we have removed missing values and the tuples with unknown values. The data set after eliminating needless values left with 30000 records. Data set is versatile in nature, having numerical and categorical attributes which provide greater support in anonymization process. We summarize adult data base attributes in the following table. Table 5.1 shows all attributes of the data set including name of the attributes, types of the attributes, corresponding distinct values and heights.

Table5.1: Attributes description of Adult Data base [10]

Name of attribute	Type of attribute	Distinct values	Height
Gender	Categorical	2	2
Work class	Categorical	14	3
Education	Categorical	16	4
Marital Status	Categorical	7	3
Race	Categorical	5	3
Health_condition	Sensitive	8	1
Age	Numeric	78	4
Country	Categorical	41	3

Attribute values of the above table are shown in the following table. All distinct values are taken from real world data set.

Table5.2: Attributes with corresponding distinct values

Attribute name	Total	Distinct values
Nationality	36	United-States, Peru, Scotland, South, Taiwan, Philippines, Poland, Portugal, Puerto-Rico, China, Columbia, Cuba, Cambodia, Canada, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Netherlands, Honduras, Guatemala, Haiti, Holland, Hong, Italy, Jamaica, Japan Hungary, India, Iran, Ireland, , Laos, Vietnam, Yugoslavia, Mexico, Nicaragua, Thailand ,
Marital status	7	Married-AF-spouse, Never-married, Divorced, Married-civ-spouse, Separated, Widowed, Married-spouse-absent
Medical condition	9	HIV, Obesity, Flu, Cancer, Phthisis, Indigestion, Hepatitis, Asthma ,Chest infection
Age	78	97,11,12,13,14,15,16,17 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 96
Work class	15	Priv-house-serv, Craft-repair, Farming-fishing ,Adm-clerical, Exec-managerial,Tech-support, Transport-moving, Prof-specialty, Handlers-cleaners ,Protective-serv, Sales, Machine-op-inspct, Other-service, Armed-Forces, Doctor
Race	5	White, Asian-Pac-Islander ,Black ,American, Other
Education	16	Doctorate , Masters,Preschool, Prof-school,1 st -4 th , 5 th -6 th , 7 th -8 th , 9 th , 10 th , 11 th , 12 th , Assoc-acdm, Assoc-voc, Some-college, HS-grad, Bachelors,
Gender	2	Female ,Male

In our experiment six attributes were obtained from adult database while health condition with distinct values was added to the data set implicitly.

In following figure a snap shot of adult data set is shown.

Age	Education	Marital_status	Workclass	Race	Gender	Country	Health_conc
39	Bachelors	Never-married	Adm-clerical	White	Male	United-States	HIV
50	Bachelors	Married-civ-spouse	Exec-managerial	White	Male	United-States	Phthisis
38	HS-grad	Divorced	Handlers-cleaners	White	Male	United-States	Obesity
53	11th	Married-civ-spouse	Handlers-cleaners	Black	Male	United-States	HIV
28	Bachelors	Married-civ-spouse	Prof-specialty	Black	Female	Cuba	Phthisis
37	Masters	Married-civ-spouse	Exec-managerial	White	Female	United-States	Phthisis
43	9th	Married-spouse-abs	Other-service	Black	Female	Jamaica	Phthisis
52	HS-grad	Married-civ-spouse	Exec-managerial	White	Male	United-States	HIV
31	Masters	Never-married	Prof-specialty	White	Female	United-States	Asthma
42	Bachelors	Married-civ-spouse	Exec-managerial	White	Male	United-States	Flu
37	Some-college	Married-civ-spouse	Exec-managerial	Black	Male	United-States	Obesity
30	Bachelors	Married-civ-spouse	Prof-specialty	Asian-P	Male	India	Cancer
23	Bachelors	Never-married	Adm-clerical	White	Female	United-States	Phthisis
32	Assoc-acdm	Never-married	Sales	Black	Male	United-States	Indigestion
34	7th-8th	Married-civ-spouse	Transport-moving	Amer-ir	Male	Mexico	Obesity
25	HS-grad	Never-married	Farming-fishing	White	Male	United-States	Obesity
32	HS-grad	Never-married	Machine-op-inspct	White	Male	United-States	HIV
38	11th	Married-civ-spouse	Sales	White	Male	United-States	HIV
43	Masters	Divorced	Exec-managerial	White	Female	United-States	Cancer
40	Doctorate	Married-civ-spouse	Prof-specialty	White	Male	United-States	Indigestion
54	HS-grad	Separated	Other-service	Black	Female	United-States	Phthisis
35	9th	Married-civ-spouse	Farming-fishing	Black	Male	United-States	Cancer
43	11th	Married-civ-spouse	Transport-moving	White	Male	United-States	Hepatitis
59	HS-grad	Divorced	Tech-support	White	Female	United-States	Phthisis
56	Bachelors	Married-civ-spouse	Tech-support	White	Male	United-States	Flu
19	HS-grad	Never-married	Craft-repair	White	Male	United-States	Phthisis

Record: 1 of 30169 | Filter | Search

Figure 5.1: Snap shot of the adult data set

Anonymization algorithm was applied on data set shown in above snap shot and the resultant anonymized data obtained is shown following figure. Where K and P=4.

```

k30. 66321243523316:48. 6275067787054, Adm-clerical, white, 0.0:1.0, United-States, HIV
<30. 73770491803279:27. 996775060467634, Exec-managerial, white, 0.0:1.0, United-States, Hepatitis
<30. 66321243523316:48. 6275067787054, Handlers-cleaners, white, 0.0:1.0, United-States, Asthma
>40. 73770491803279:27. 996775060467634, Handlers-cleaners, black, 0.0:1.0, United-States, Indigestion
>40. 47747747747748:47. 06030354679008, Prof-specialty, black, 1.0:0.0, Cuba, Cancer
>40. 47747747747748:47. 06030354679008, Exec-managerial, white, 1.0:0.0, United-States, Phthisis
>40. 47747747747748:47. 06030354679008, Other-service, black, 1.0:0.0, Jamaica, Obesity
2*. 456. 73770491803279:27. 996775060467634, Exec-managerial, white, 0.0:1.0, United-States, Flu
2*. 636. 47747747747748:47. 06030354679008, Prof-specialty, white, 1.0:0.0, United-States, HIV
2*. 436. 66321243523316:48. 6275067787054, Exec-managerial, white, 0.0:1.0, United-States, Hepatitis
2*. 036. 66321243523316:48. 6275067787054, Exec-managerial, black, 0.0:1.0, United-States, Asthma
>30. 66321243523316:48. 6275067787054, Prof-specialty, Asian-Pac-Islander, 0.0:1.0, India, Indigestion
>30. 0:4. 769230769230769, Adm-clerical, white, 1.0:0.0, United-States, Cancer
>30. 66321243523316:48. 6275067787054, Sales, black, 0.0:1.0, United-States, Phthisis
>30. 66321243523316:48. 6275067787054, Transport-moving, Amer-Indian-skimo, 0.0:1.0, Mexico, Obesity
<30. 66321243523316:48. 6275067787054, Farming-fishing, white, 0.0:1.0, United-States, Flu
<30. 66321243523316:48. 6275067787054, Machine-op-inspct, white, 0.0:1.0, United-States, HIV
<30. 66321243523316:48. 6275067787054, Sales, white, 0.0:1.0, United-States, Hepatitis
>40. 47747747747748:47. 06030354679008, Exec-managerial, white, 1.0:0.0, United-States, Asthma
>40. 66321243523316:48. 6275067787054, Prof-specialty, white, 0.0:1.0, United-States, Indigestion
>40. 285714285714285:26. 48979591836735, Other-service, black, 1.0:0.0, United-States, Cancer
>40. 66321243523316:48. 6275067787054, Farming-fishing, black, 0.0:1.0, United-States, Phthisis
2*. 436. 66321243523316:48. 6275067787054, Transport-moving, white, 0.0:1.0, United-States, Obesity
2*. 658. 285714285714285:26. 48979591836735, Tech-support, white, 1.0:0.0, United-States, Flu
2*. 456. 73770491803279:27. 996775060467634, Tech-support, white, 0.0:1.0, United-States, HIV
2*. 021. 0:5. 3076923076923075, Craft-repair, white, 0.0:1.0, United-States, Hepatitis
>30. 66321243523316:48. 6275067787054, Exec-managerial, white, 0.0:1.0, United-States, Asthma
>30. 66321243523316:48. 6275067787054, Craft-repair, white, 0.0:1.0, United-States, Indigestion
>30. 0:5. 3076923076923075, Protective-serv, white, 0.0:1.0, United-States, Cancer
>30. 0:5. 3076923076923075, Sales, black, 0.0:1.0, United-States, Phthisis
<30. 66321243523316:48. 6275067787054, Exec-managerial, white, 0.0:1.0, United-States, Obesity
<30. 66321243523316:48. 6275067787054, Adm-clerical, white, 0.0:1.0, United-States, Flu
<30. 0:5. 3076923076923075, Other-service, black, 0.0:1.0, United-States, HIV
>40. 66321243523316:48. 6275067787054, Machine-op-inspct, white, 0.0:1.0, Puerto-Rico, Hepatitis
>40. 0:5. 3076923076923075, Machine-op-inspct, white, 0.0:1.0, United-States, Asthma
>40. 0:4. 769230769230769, Adm-clerical, white, 1.0:0.0, United-States, Indigestion
>40. 66321243523316:48. 6275067787054, Prof-specialty, white, 0.0:1.0, United-States, Cancer
2*. 436. 66321243523316:48. 6275067787054, Machine-op-inspct, white, 0.0:1.0, United-States, Phthisis
2*. 656. 73770491803279:27. 996775060467634, Prof-specialty, white, 0.0:1.0, United-States, Obesity
2*. 421. 0:5. 3076923076923075, Tech-support, white, 0.0:1.0, United-States, Flu
2*. 036. 47747747747748:47. 06030354679008, Adm-clerical, white, 1.0:0.0, United-States, HIV
>30. 66321243523316:48. 6275067787054, Handlers-cleaners, white, 0.0:1.0, United-States, Hepatitis
>30. 73770491803279:27. 996775060467634, Prof-specialty, black, 0.0:1.0, United-States, Asthma
>30. 73770491803279:27. 996775060467634, Machine-op-inspct, white, 0.0:1.0, United-States, Indigestion
>30. 47747747747748:47. 06030354679008, Exec-managerial, white, 1.0:0.0, United-States, Cancer
<30. 66321243523316:48. 6275067787054, Craft-repair, white, 0.0:1.0, United-States, Phthisis
<30. 66321243523316:48. 6275067787054, Prof-specialty, white, 0.0:1.0, United-States, Obesity
<30. 47747747747748:47. 06030354679008, Exec-managerial, other, 1.0:0.0, United-States, Flu
>40. 47747747747748:47. 06030354679008, Prof-specialty, white, 1.0:0.0, Honduras, HIV
>40. 73770491803279:27. 996775060467634, Exec-managerial, white, 0.0:1.0, United-States, Hepatitis
>40. 66321243523316:48. 6275067787054, Exec-managerial, white, 0.0:1.0, United-States, Asthma
>40. 66321243523316:48. 6275067787054, Tech-support, white, 0.0:1.0, United-States, Indigestion
2*. 436. 66321243523316:48. 6275067787054, Machine-op-inspct, white, 0.0:1.0, Mexico, Cancer
2*. 636. 66321243523316:48. 6275067787054, Other-service, white, 0.0:1.0, Puerto-Rico, Phthisis

```

Figure 5.1: Anonymized data

Proposed model (p, β) sensitive k-anonymity implemented on data set acquired from UC Irvine Machine Learning Repository with more than 30000 tuples. Initially it was containing nine quasi identifier attributes but we selected only seven attributes which were considered most significant to analyze the performance of the algorithm. The result obtained are compared with the previous anonymization technique namely (p, α)-sensitive k-anonymity [12] to evaluate our technique with respect to efficiency, distortion ratio and effectiveness. We introduced new technique called Discriminated union to calculate distinct sensitive value in each quasi identifier group. From experimental results it is proved that the proposed methodology is much efficient than the previous technique and the quality of the published data is much better. Probabilistic attack is reduced to an acceptable level while unnecessary computational cost is reduced and needless generalization and suppression of information is avoided. Quasi identifier attribute Medical condition with sensitive tuples were assigned 8 sensitive numbered values e.g. (HIV=1, Cancer=2, Obesity=3, and so on).

The experiment was performed on hardware similar to incognito [12] with Dual Core CPU, 2 GB RAM and 1.7 processor. Java SE was used in Net beans 2007 environment.

5.1 Performance Measures

Our technique evaluates performance of the algorithm in term of similarity attack and distortion ratio. We described performance measures along with result obtained with particular scenario in the following section.

Scenario 1: Similarity attack

When more sensitive values of quasi identifier group fall in one category then similarity attack occurs and an individual can be easily identified. Sensitive information can be retrieved by inferring quasi identifier attributes to the attribute having similar values against particular record. (p, α) sensitive k -anonymity generate 30 minimal $(2, 2)$ sensitive 4-anonymous tables where $p = 2, k = 4, \alpha = 2$. 23 tables were satisfactory while 7 tables were suffering from similarity attack and supposed to be vulnerable with 23% of the total. Proposed technique (p, β) sensitive k -anonymity generates 26 tables and all were satisfactory no table came across similarity attack. So the similarity attack is reduced to 0% by our proposed technique. For second and third set of experiment, parameters and its corresponding output with percentage is given in the following table.

Table5.2: Results comparison

Technique	Parameters	Total tables generated	Similarity attack	Percentage
(p, α) -sensitive k -anonymity	$k=4, p= 2, \alpha = 2$	30 tables	7 tables	$7/30 = 23\%$
	$k=8, p= 4, \alpha = 4$	27 tables	8 tables	$8/27 = 29\%$
	$k=16, p=8, \alpha = 4$	21 tables	7 tables	$7/21 = 33\%$
(p, β) sensitive k -anonymity	$k=4, p= 4$	26 tables	0 tables	$0/26= 0\%$
	$k=4, p= 4$	22 tables	0 tables	$0/22= 0\%$
	$k=4, p= 4$	18 tables	3 tables	$3/18= 16\%$

Scenario 2: Distortion Ratio

Distortion ratio measures the amount of information lost during generalization process. It determines the difference between the original data and the data obtained after generalization. It is the most important utility measure of the techniques used for anonymization. An algorithm with less distortion

ratio is more efficient as it preserve more information, maximum data is published and information loss is reduced. During anonymization an attribute value is replaced by more general value to hide confidential information regarding entity. It is defined in term of height i.e. the height of the value is 0 if it is not generalized and there will be no distortion at all. If height of the generalized value is turned to one, it is supposed to move one level up in generalization hierarchy. It moves upward in the hierarchy as the height of the generalized value is increased level by level.

In the proposed technique distortion ratio is minimized significantly up to 60% of the whole data set as compared to the previous methodology(p, α)-sensitive k-anonymity. First all sensitive tuples are separated from data set and then a portion of non sensitive tuples is incorporated to the table containing only sensitive tuples separated from data set. The tuples left behind this process are non sensitive and up to 70% of the whole micro data so it is published directly without adding any distortion to it.

To determine height of the generalized value of an attribute, suppose attribute A_i , tuple t_{ij} and their corresponding height $h_{i,j}$. The distortion ratio of the entire data set the sum of all individual distortions of the tuples in generalized data set $D = \sum_{i,j} h_{i,j}$

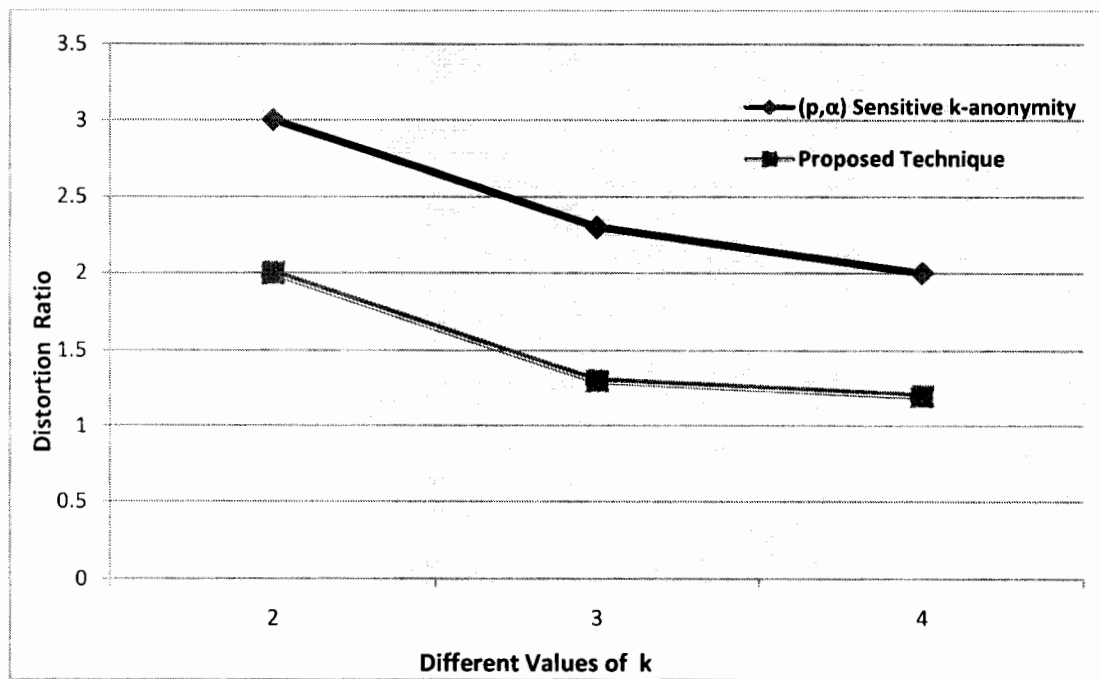


Figure5.2: proposed methodology and previous model

Distortion ratio can be calculated by

Distortion ratio of the data set = (Distortion of the generalized dataset) / (Distortion of the fully generalized dataset)

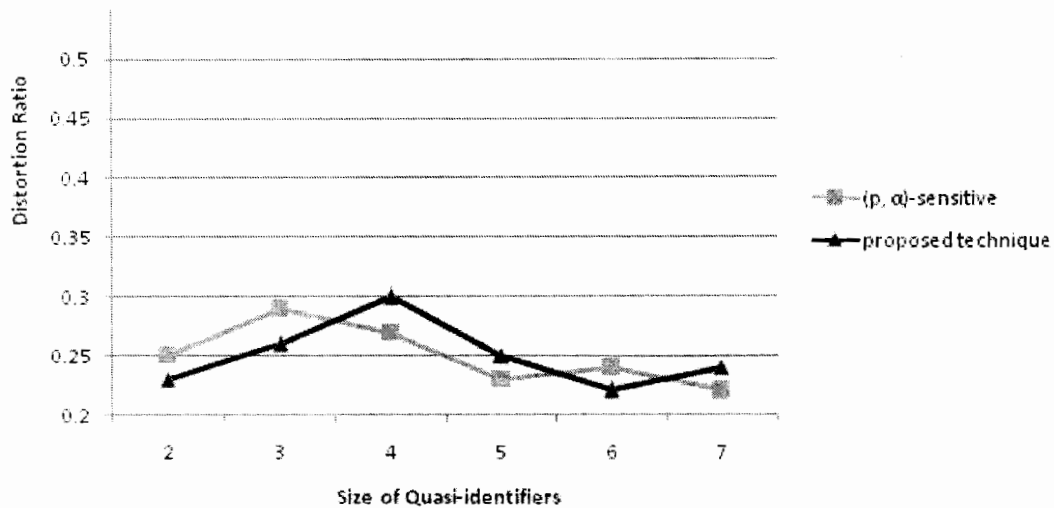


Figure5.3: comparison based on different values of Quasi Identifiers

In taxonomy tree if all values are generalized to root is called fully generalized data set while generalized data set is the one which is generalized to some lower level in the hierarchy with minimum height. Distortion ratio is dependent on the size of quasi identifier. If quasi identifier has more attributes then the distortion ratio will be high and vice versa. More attributes means more distinct values so more chances to generalize the tuples.

As we apply generalization on some specific portion of the data set, which is less than 40% of the entire data set so our technique performs less generalization as compared to the previous models.

**CHAPTER 6: CONCLUSION
&
FUTURE DIRECTIONS**

6.1 Conclusion and Future directions

(p, α) -sensitive k -anonymity is a mandatory property for non aggregated data to satisfy before it is published. However our research shows that this technique has some limitations which results in loss of the valuable information and needless overhead of computational cost by generalization entire data set without any distinction. For small portion of sensitive attributes, all data set is passed on through anonymization process where each tuple of the data set is generalized. Thus data set is destabilized as noise is incorporated to every record of micro data and computational complexity is increased on the other hand. Secondly, this technique mainly focused on sensitive values but with uneven distribution resulting probabilistic attack which lead to privacy breach.

Proposed methodology (p, β) sensitive k -anonymity, introduces new technique “Discriminated union” of the sensitive attribute values which ensure that every tuple in equivalence class have distinct sensitive values. This property guarantees the privacy of the respondent to be protected and intruders can never be able to disclose confidential information. The technique works in two phases. In first phase, size of data set is reduced by separating sensitive tuples from non sensitive tuples. As we know from the literature sensitive tuples are always less than 15% of the whole data set so it is isolated and dealt separately. To minimize similarity attack and improve privacy of the individual, a portion of non sensitive tuples is added to sensitive tuples. In second phase of the algorithm, (p, β) sensitive k -anonymity is applied on reduced dataset and results are obtained.

Experimental results show that our technique is efficient in term of distortion ration and similarity attack. Privacy of the individual is ensured as well as quality of the published data is improved. Computational cost is reduced significantly up to 60% by publishing non sensitive data directly without any modification; on the other hand data loss is also minimized.

Our proposed technique is performing well for size of equivalence class equal to total number to diseases. Its performance degrades if size of equivalence class i.e. 'k' is significantly greater than total number of diseases.

In future the proposed technique can further be extended for large size of equivalence class.

References

References

1. L. Sweeney, "Uniqueness of Simple Demographics in the U.S. Population," Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.
2. L. Sweeney, "K-anonymity: A model for protecting Privacy," in *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no.5, pp. 557-570, 2002.
3. L. Sweeney, "Achieving K-anonymity Privacy Protection using generalization and suppression," in *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 571-588, 2002
4. G.T. Duncan and D. Lambert, "Disclosure-limited data dissemination," in *Journal of the American Statistical Association*, vol. 81, pp. 10-27, 1986
5. D. Lambert, "Measure of disclosure risk and harm," in *Journal of official statistics*, vol. 9, no. 2, pp. 313-331, 1993
6. A. Machanvajhala et al, "L-Diversity: Privacy beyond k-anonymity," in *International Conference on Data Engineering*, 2006
7. N. Li et al, "t-closeness: Privacy Beyond k-anonymity and l-diversity," in *International Conference on Data Engineering*, pp.106-115, 2007
8. T.M. Traian and V. Bindu, "Privacy Protection: p-sensitive k-anonymity," in *International Conference on Data Engineering*, Atlanta, 2006
9. Xiaoxun Sun et al, "Extended k-anonymity models against sensitive attribute disclosure," in *Computer Communications*, vol. 34, pp.526-535, 2011
10. D.J. Newman et al, "UCI Repository of Machine Learning Databases," Available at www.ics.uci.edu/~mllearn/MLRepository.html, University of California, Irvine ,1998
11. B. Fung et al, "Top Down Specialization for Information and Privacy Preservation," in *International Conference on Data Engineering (ICDE)*, Tokyo, Japan, 2005
12. K. LeFevre et al, "Incognito: efficient full-domain k-anonymity," in *SIGMOD Conference*, pp. 49-60, June 2005.
13. L. Sweeney, "Guaranteeing Anonymity When Sharing Medical Data, The Data fly System," in *Proc AMIA Annual Fall Symposium*, 51-5, 1997
14. Health Insurance Portability and Accountability Act, Available online at <http://www.hhs.gov/ocr/hipaa>
15. Personal Health Information Protection Act, available online at www.e-laws.gov.on.ca/html/statutes/english/elaws_status_04p03_e.htm

16. P. Samarati, "Protecting Respondent's identity in Microdata Release," in *IEEE Transactions on Knowledge and Data Engineering*, vol.13, no. 6, pp. 1010-1027, 2001
17. T. Dalenius, "Finding a needle in a haystack - or indentifying anonymous census record," in *Journal of official Statistics*, vol. 2, no. 3, pp. 329-336, 1986
18. CIHR (Canadian Institutes of Health Research) Best Practices For Protecting Privacy in Health Research, available at <http://www.cihr-irsc.gc.ca>, Sep 2005
19. K.EL Eman et al, "Evaluating Predictors of Geographic Area Population Size Cutoffs to Manage Re-identification Risk," in *Journal of the American Medical Informatics Association*, 2008
20. K.EL Eman et al, "Evaluating common de-identification heuristics for personal health information," in *Journal of Medical Internet Research*, vol. 8, no. 4, 2006
21. N.R. Adam, and J.C. Wortman, "Secuirty-Control Methods for Statistical Databases: A Comparative Study", *ACM Computing Surveys*, vol. 21, no.4, 1989
22. P. Samarati, "Protecting Respondent's Identity in Microdata Release," in *IEEE Transactions on Knowledge and Data Engineering*, vol.13, no. 6, pp. 1010-1027, 2001
23. L. Willenborg and T. DeWaal, "Statistical Disclosure Control in Practice," Lecture Notes in Statistics, in *Springer*, Vol. 111, 1996
24. J. Kim and J. Curry, "The treatment of missing data in multivariate analysis," *Social Methods & Research*, vol. 6, pp. 215-240, 1977
25. G. Aggarwal, et al, "Anonymizing tables," in *International Conference on Database Theory (ICDT)*, Edinburgh, Scotland, pp. 246-258, 2005
26. R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymity," in *International Conference on Data Engineering (ICDE)*, 2005
27. A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems*, Paris, France, pp. 223-228, 2004
28. M.R. Garey and D.S. Johnson, "Computers and Intractability: A Guide to the Theory of NP-Completeness," Freeman, San Francisco, 1979
29. K. Fung and P. Wang, "Top-down specialization for information and privacy preservation," in *International Conference on Data Engineering*, Tokyo, Japan, 2005
30. J. Domingo-Ferrer et al, "A Critique of k-Anonymity and Some of Its Enhancements," in *The Third IEEE International Conference on Availability, Reliability and Security*, pp. 990-993, 2008

