# PREVENTING SENSITIVE ATTRIBUTE DISCLOSURE USING IMPROVED K-ANONYMITY MODEL

**Submitted By:**

**Tariq Sadad**

504/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
**Jun 2012**

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

**Dated: 28 Jun 2012**

## Final Approval

This is to certify that we have read and evaluated the thesis entitled **Preventing Sensitive Attribute disclosure using improved K-anonymity model** submitted by **Tariq Sadad** under **Reg No. 504-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. Abdul Basit Siddiqui
Assistant Professor,
Foundation University Institute of Engineering
and Management Sciences (FUIEMS)
New Lalazar, Rawalpindi

Dr. Abdul Basit Siddiqui

**Internal Examiner**
Dr. Tariq Abdullah
Lecturer,
Department of Computer Science & Software Engineering
International Islamic University, Islamabad

Dr. Tariq Abdullah

**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.

# Dedication

Dedicated To My Family and Teachers

Tariq Sadad
504-FBAS/MSCS/F08

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 "**Preventing Sensitive Attribute disclosure using improved 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.

Tariq Sadad
504-FBAS/MSCS/F08

## ACKNOWLEDGMENT

**In the name of Allah, Most Gracious, Most Sympathetic**

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

For the continuous support, advice and encouragement during this work, I desire to thank my supervisor Dr. Ayyaz Hussain. He has inspired me a state of confidence, with which I now experience that following his research guidelines, I can do research of any new topic.

I am thankful to department of computer science & software engineering, IIU, Islamabad and faculty members for providing strong environment for research.

I would also thank to my best teacher Dr. Azhar Rauf, department of Computer Science, University of Peshawar, Mr. Asim Ali Manager MIS and Mr. Imran Riaz, Finance Officer, Fauji Foundation for their continuous motivational support. I am looking forward to a continuous collaboration with them in the future.

I would also like to thank my dear friends Mr. Zaheer U Din Baber, Mr. Muhammad Fida, Mr. Gulzar Mehmood, Mr. Gohar Ali and Dr. Khalid Khan who have been a continuous motivation behind my success.

Finally I am eternally grateful to my parents, life partner and whole family. Their endless support encouragement and stimulation have been a true source of strength and inspiration for me.

Tariq Sadad
504-FBAS/MSCS/F08

## ABSTRACT

*K*-anonymity is a model which protects the individual's privacy. In k-anonymity the data is shown in such a manner that there are at least *k* identical kinds of tuples in the microdata for every single tuple, but it is not sufficient to protect revelation of attribute due to two types of attacks occur in *k*-anonymity; one is called homogeneity attack and other is called background knowledge attack. To solve this problem, several models were proposed [2, 6, 8]. But these enhanced properties have some restrictions which still allow the information to be disclosed. To enhance privacy and reduce similarity attack another technique has been proposed called ($p$, $\alpha$)-sensitive *k*-anonymity model [9]. We have identified that, to reduce similarity attack, ($p$, $\alpha$)-sensitive *k*-anonymity model is not enough. To overcome the shortcoming of ($p$, $\alpha$)-sensitive *k*-anonymity, a new technique has been proposed called enhanced ($p$, $\alpha$)-sensitive *k*-anonymity model. This enhanced property states that in every quasi-identifier group there is at least $p$ distinct sensitive attribute categories with its total weight at least $\alpha$. The proposed technique uses a top-down local recoding algorithm [9]. The concept of top-down local recoding algorithm is that in initial step all tuples are generalized into one quasi-identifier group completely. Then, in every iteration tuples are specialized and enhanced ($p$, $\alpha$)-sensitive *k*-anonymity has been maintained during specialization. The proposed algorithm has been implemented on well known data set called Adult Dataset [10].
The proposed algorithm is compared with existing techniques based on well known performance measures which include similarity attack, distortion ratio and running time. Simulation result shows that proposed algorithm gives superior results in term of similarity attack and distortion ratio; where as its running time is slightly higher than the existing approaches.

# Table of Contents

# List of Tables

# List of Figures

# List of Acronyms

| **Acronym** | **Definition** |
|---|---|
| DGH | Domain Generalization Hierarchy |
| VGH | Value  Generalization Hierarchy |
| PHIPA | Personal Health Information Protection Act |
| SSN | Social Security Number |
| NIC# | National Identity Card Number |
| EC | Equivalence Classes |
| K | Level of Anonymization |
| QI | Quasi Identifier |
| UI | Unique Items |
| HIPAA | Health Insurance Portability and Accountability Act |
| Raw data | Input data to processing |
| MaxSup | Maximum Suppression allowed (or Suppression limit) |
| UCI | University of California Irvine |

# CHAPTER 1: INTRODUCTION

When releasing microdata (after applying anonymization methods on the data to be released is called microdata), the sensitive information of the entities is also compulsory to prevent it from being released. Information revelations have been happening of two types [4, 5]: identity revelation and attribute revelation. Identity revelation arises when a particular record about entity is linked in the microdata. When some new information about individuals is exposed, the attribute revelation happens. *K*-anonymity covers the problem of identity revelation, but it is insufficient to prevent attribute revelation due to two types of attacks that occur in *k*-anonymity; one is homogeneity attack and other is background knowledge attack.

## 1.1 Motivation

Currently different organizations such as a hospital issue its raw non-aggregated data (also called micro data), for a variety of different reasons. However, such data may contain private information as in the case of medical record, where the identities of the entities should be kept secret.

In the United States a telephonic poll was conducted by TIME/CNN in 1996, in which 88% of the respondents replied that without their permission medical information about them should not be released. In a second question, 87% said that organizations should be restricted from giving out medical information without patient's permission. The public prefers that directly involved people and employees can only have access to their personal records and it should be bounded to restrict further disclosure of their data by ethical and legal standards [13].

Currently, the leakage of health information is thoroughly regulated in many organizations/authorities. To protect health data earlier to their revelation to researcher's organizations are needed to apply privacy protection. For example, the *HIPAA* in the United States [14], and the *PHIPA* [15] in Canada, are some of the well known privacy regulation authorities that protect the confidentiality of healthcare information.

Before releasing the data, organizations often encrypt or remove explicit identifiers such as NIC#, SSN and names, in order to protect the confidentiality of respondents, [1]. However, de-identifying these attributes give no guarantee of secrecy, because released table contains some other fields, such as age, zip code and gender which can be linked with external information to re-identify the individuals [1]. To avoid the expose of the data, some researchers tried to anonymize the data by using different methods, for example, swapping sampling and adding noise to the data in order to overcome the possibility of a privacy breach. However, this compromises the integrity or truthfulness of the released data, while maintaining an overall statistical property of the result [21, 22, 23].

L. Sweeney predicted that approximately 87% population of the United State can be exclusively recognized by the combination of gender, zip code and age, because all these records are linked with openly accessible database such as voter list records and driving records [1]. To prove this point, anonymous medical records have been re-identified by Sweeney including one of them in William Weld record, who was Governor of Massachusetts at that era [2].

Consider the medical data given in the table to be published by a hospital

Table1.1: Hospital data

| | Non-Sensitive Data | | | | Sensitive Data | |
|---|---|---|---|---|---|---|
| **S#** | **Zipcode** | **Age** | **Gender** | **Nationality** | **Name** | **Condition** |
| 1 | 24064 | 25 | M | Pakistani | Ali | HIV |
| 2 | 24078 | 26 | M | Indian | Rajesh | HIV |
| 3 | 24064 | 39 | M | Canadian | Jan | Viral Infection |
| 4 | 24078 | 32 | F | Japanese | Tina | Cancer |

Table 1.1 shows the original data while table 1.2 shows the de identified data by suppressing names in order to protect identities of respondents

Table1.2:De-identified data seems to be protected (medical record)

| | Non-Sensitive Data | | | | Sensitive Data |
|---|---|---|---|---|---|
| **S#** | **Zip code** | **Age** | **Gender** | **Nationality** | **Condition** |
| 1 | 24064 | 25 | M | Pakistani | HIV |
| 2 | 24078 | 26 | M | Indian | HIV |
| 3 | 24064 | 39 | M | Canadian | Viral Infection |
| 4 | 24078 | 32 | F | Japanese | Cancer |

But when released, the values of these attributes such as Name, Zip code, Age and Nationality were also available in various external databases, for example, in driving record, which is used to be linked for the identification of an individual's record.

Table1.3: Non de-identified publicly available table/Driving record

| S# | Name | Zip code | Age | Nationality |
|---|---|---|---|---|
| 1 | David | 13053 | 28 | Indian |
| 2 | Rajesh | 24078 | 26 | Indian |
| 3 | Katrina | 13053 | 23 | Indian |

For example, Zip code, Age, Gender and Nationality can be linked to the driving record in the above table 1.3 to re-identify person's name. Thus this identifies that the corresponding tuple belonging to "Rajesh, who is 26 years of Age living in Zip code 24078 of India, is a patient of HIV".

Various researches have been directed towards the anonymization of the data, in a different way. Although guaranteeing complete anonymity is clearly an impossible task, but the concept of *k*-anonymity has been introduced by L. Sweeney to protect the respondent identities and release truthful information

*K*-anonymity is defined as, "Change the data in such a way that for every tuple in the microdata, there are at least ($k$ -1) other tuples for the value of quasi-identifiers" [2].

## 1.2 Research Objective

Different techniques were proposed to prevent attribute revelation such as *l*-diversity [6], *p*-sensitive *k*-anonymity [8], *t*-closeness [7] and ($p$, $\alpha$)-sensitive *k*-anonymity [9]. But still, these enhancements of k-anonymity allow the information to be exposed or have various other restrictions. Following are some limitations of the existing techniques of *k*-anonymity.

The *l*-diversity model [6] says that in every quasi-identifiers group there are at least *l* "well-represented" values, but achieving this technique is not easy and may produce a large amount of data loss. Further, for prevention of similarity attack, *l*-diversity is insufficient.

The idea of *p*-sensitive *k*-anonymity [8] is that there should be at least p different sensitive attribute values for every quasi-identifier group. The limitation of this technique is that, may be the sensitive attributes are similar for any quasi-identifier group. Also, it may cause a large amount of data loss to achieve the required level of privacy.

The concept of *t*-closeness model [7] is that between sensitive attributes it defines a semantic distance to protect against sensitive attributes revelation. The semantic distance is no more than a threshold *t* between the distributions of the attributes in the group and between the whole tables. But enforcing t-closeness would damage the value of data and destroy the links between quasi-identifier group and sensitive attributes.

The ($p$, $\alpha$)-sensitive k-anonymity model [9] protects sensitive attribute revelation by defining at least *p* distinct sensitive attributes with its total weight $\alpha$, for every group of quasi-identifier. As compared to above mentioned properties, ($p$, $\alpha$) sensitive *k*-anonymity model protects sensitive information well, but it mainly focuses on specific value. So ($p$, $\alpha$) sensitive *k*-anonymity property is insufficient for privacy preservation and we proposes a solution for this problem.

So a new technique called enhanced (*p*, *α*) sensitive *k*-anonymity has been proposed to enhance the current privacy principles to protect data quality, data privacy and reduce similarity attack.

## 1.3 Contribution of Thesis

A new technique called enhanced (*p*, *α*) sensitive *k*-anonymity has been proposed.
The proposed technique uses top-down local recoding algorithm and reduces the similarity attack. The proposed technique also measures distortion ratio and running time of the algorithm

## 1.4Thesis Layout

In this thesis, we have critically discussed and analyzed the basic concepts, and preliminary developed theories in chapter 2, followed by previous studies related to the subject research in chapter 3. A comprehensive study and analysis led to the proposed methodology to reduce similarity attack and algorithm that are necessary for anonymization, are discussed in chapter 4. Experimentation results to enhance privacy and reduce similarity attack based upon proposed methodology and comparative analysis are described in chapter 5. Finally an overview of future potential work and conclusions are put in chapter 6.

# CHAPTER 2: BASIC DEFINITIONS

In this chapter, basic concepts relevant to the study are defined. The basic definition includes the *k*-anonymity concepts and all the terms that are related to *k*-anonymity are also defined. If a de-identified private table PT is anonymized, the rows in PT are called as tuples, and the columns in the table are called attributes. Moreover, the quasi-identifier attributes is set of PT's attributes and the table is supposed to have at least k tuples. For anonymization a concept called generalization and suppression is used. Generalization is being performed through Domain Generalization Hierarchy and Value Generalization Hierarchy. Maxsup is used to define how many cells or rows are to be suppressed in a table

## 2.1*K*-anonymity

L. Sweeney proposed a model called *k*-anonymity to protect the respondent identities and release truthful information. This states that "Change the data in such a way that for every tuple in the microdata, there are at least (*k* -1) other tuples for the value of quasi-identifiers" [2]. For example, consider a table 2.1

Table2.1: Raw data

| | **Quasi-identifiers** | | | **Sensitive data** |
|---|---|---|---|---|
| **S#** | **Zipcode** | **Age** | **Nationality** | **Condition** |
| 1 | 24064 | 25 | Pakistani | HIV |
| 2 | 24078 | 26 | Indian | Fever |
| 3 | 24064 | 39 | Canadian | Indigestion |
| 4 | 24078 | 32 | Japanese | Hepatitis |

So if we apply *k*- anonymity in above table 2.1

Table2.2: 4-anonymous data

| | **Quasi-identifiers** | | | **Sensitive data** |
|---|---|---|---|---|
| **S#** | **Zipcode** | **Age** | **Nationality** | **Condition** |
| 1 | 240** | <40 | * | HIV |
| 2 | 240** | <40 | * | Fever |
| 3 | 240** | <40 | * | Indigestion |
| 4 | 240** | <40 | * | Hepatitis |

Here 4-anonymity has been applied in table 2.2. In 4-anonymous table there are (4-1) = 3 other tuples with the same value for quasi-identifiers.

## 2.2 Generalization and Suppression

Generalization is the replacement of the original value by a semantically consistent but less specific value [3]. For example, in the above table zip codes (24078, 24064) can be generalized into 240**.

Suppression deals with the removing data from the table, such that in the microdata it is not released. Suppression can perform at cell or tuple level [3]. For example, in the above table 2.2, Nationality (Pakistani, Canadian, Japanese, and Indian) can be suppressed into *.

## 2.3 Quasi-Identifiers

A group of attributes that can be linked with other database to re-identify the individual's records is called quasi-identifiers [16, 17]. Examples of common *QI* are dates (such as birth, death, visit, admission, discharge etc), location (such as zip code, region etc), and gender [18, 19, 20].

## 2.4 Domain Generalization

Building a general domain from existing domains is called Domain Generalization [12]. For example consider a domain of zip code 23145 which is generalized into 2314* by disregarding the least significant number.

## 2.5 Domain Generalization Relationship

A $D_i \leq DD_j$ is defined as domain generalization relationship [12]. The relationship denotes that domain $D_J$ is either a domain generalization or matching of domain $D_I$. This relationship shows a *many-to-one relationship* between original domain values and resulting domain.

The function $\gamma : D_i \rightarrow D_j$ is called 'value generalization function' which shows the many-to-one relationship. $D_J$ is called the direct generalization of $D_I$, if there is an edge from $D_i$ to $D_J$. Domain generalization relationship is transitive, that is, If $D_i \leq DD_j$ and $D_j \leq DD_K$ then $D_i \leq DD_K$.

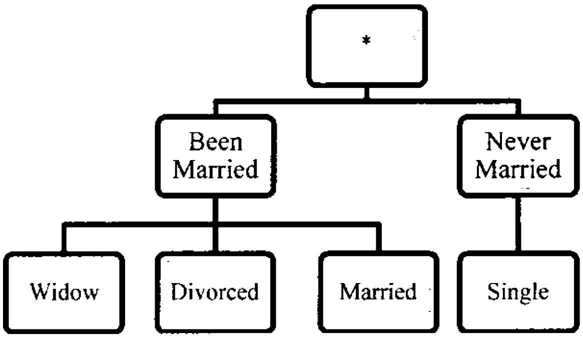
Transitivity property trends to a new definition, which is called Domain Generalization Hierarchy.

### 2.5.1 Domain Generalization Hierarchy

A series of direct generalizations in the nodes can be supposed as DGH [12]. DGH consists of

- Edges: which is direct generalizations
- Paths: which is indirect generalizations

Examples of generalization hierarchies, that is, domain generalization hierarchies and value generalization hierarchies are given in below figures from figure 2.1 to figure2.7.



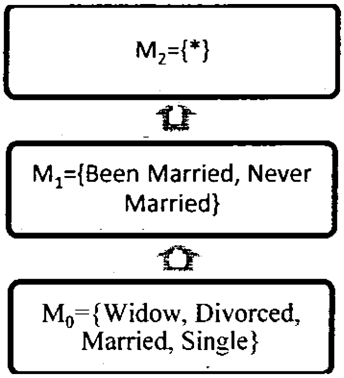
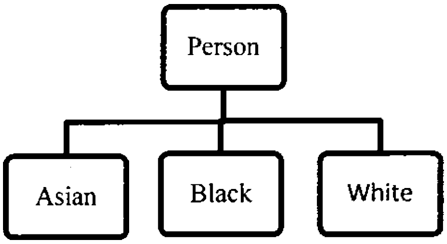

M2={*}
M1={Been Married, Never Married}
M0={Widow, Divorced, Married, Single}

Figure2.1: DGH for Marital Status

Figure2.2: VGH for Marital Status



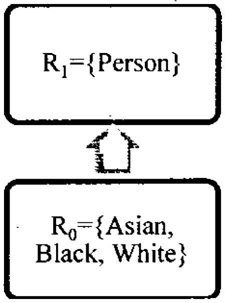


Figure2.3: DGH for Race

Figure2.4: VGH for Race

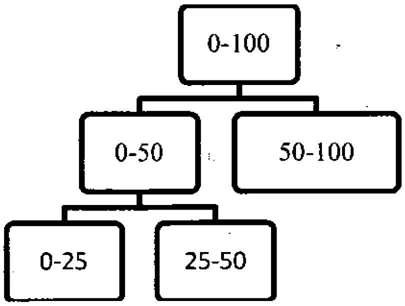
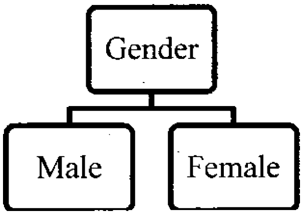


Figure2.5: DGH for Age



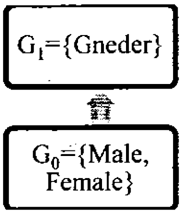


Figure2.6: DGH for Gender

Figure 2.7: VGH for Gender

## 2.6 Unique Items (UI)

The distinct data items belonging to an attribute is called UI [16]. For example, in table 2.3, the UI of the variable Race are: Asian, Black and White. These UI are leaves of the corresponding hierarchy (here in figure 2.3)

## 2.7 Equivalence Classes (EC)

The tuples of quasi identifier that are uniquely distinguishable from other tuples is called EC [16]. For example, in below table, we have 3 EC: {Asian, Been married, [45-50]}, {Black, Never married, [20-25]} and {White, Been married, [45, 50]}.

Table2.3: 3-Anonymized data

| Race | Marital Status | Age |
|---|---|---|
| Asian | Been married | [45-50] |
| Black | Never married | [20-25] |
| Asian | Been married | [45-50] |
| White | Been married | [45-50] |
| White | Been married | [45-50] |
| White | Been married | [45-50] |
| Black | Never married | [20-25] |
| Asian | Been married | [45-50] |

## 2.8 Suppression limit

Suppression limit is the maximum number of tuples that we are allowed to suppress in order to achieve k-anonymity [24].

## 2.9Frequency Set

Let *T* be a relation and *Q* be set of quasi-identifier size with *n* attribute. The frequency set of *T* with respect to *Q* is a mapping from every unique combination of values of '$< q_0, q_1, \ldots\ldots, q_n >$ of Q in T (the value groups), to the total number of tuples in *T* with certain values of *Q* (the counts) [12].

The frequency set from *T* is obtained with respect to a set of attributes *Q* by assigning a COUNT (*) query, with *Q* as the attribute list in the GROUP BY clause, in SQL. For example, in order to check whether the above table is 3-anonymous with respect to Race, Marital Status, Age. A query is given "SELECT COUNT (*) FROM TABLE GROUP BY Race, Marital Status, Age". Since the output contains groups with count equal to 3. So with respect to Race, Marital Status, Age the above table 2.3 is 3-anonymous.

## 2.10 Generalization Property

Let *T* be a relation, and *P* and *Q* be sets of attributes in *T* such that $D_P < D_Q$ . If *T* is *k*-anonymous with respect to *P*, then *T* is also anonymous with respect to *Q* [12].

## 2.11 Subset Property

Let $T$ be a relation, and $Q$ be a set of attributes in $T$. If $T$ is $k$-anonymous with respect to $Q$, then $T$ is $k$-anonymous with respect to any set of attributes $P$ such that $P \leq Q$ [12].

## 2.12 Rollup Property

Let $T$ be a relation, and let $P$ and $Q$ be sets of attributes such that $D_P \leq D_Q$. If we have f1, the frequency set of $T$ with respect to $P$, then we can generate each count in f2, the frequency set of $T$ with respect to $Q$, by summing the set of counts in f1 associated by generalization function $\gamma$ with each value set of f2 [12].

Consider $P$ is <M, R, $G_0$> and Q is <M, R, $G_1$>. The Frequency set of $P$ is calculated by a COUNT (*) query with Marital Status, Race and Gender attributes in the GROUP BY clause While the Frequency set of Q is calculated by summing the counts of groups formed by a GROUP BY clause with Marital Status, Race and $G_1$

## 2.13 Distance vector

This is the measure of the level of generalizations of each attribute [16].

Consider below table 2.4

Table2.4: Raw data

| Race | Marital_status | Age |
|---|---|---|
| Asian | Married | 47 |
| Black | Single | 21 |
| Asian | Married | 49 |

The vector [0,1,1] generalize the second attribute (that is, Marital_status) once regarding to its corresponding hierarchy according to figure 2.1, the third attribute (that is, Age) once regarding the hierarchy according to figure 2.5, while the first attribute named Race not generalize shown in below table 2.5

Table2.5: Hierarchical generalization with regard to the vector [0,1,1]

| Race | Marital_status | Age |
|---|---|---|
| Asian | Been married | [45-50] |
| Black | Never married | [20-25] |
| Asian | Been married | [45-50] |

## 2.14 Lattice

A collection of distance vectors and their interconnections is called Lattice; it is arrangement of hierarchy going from null vector to the maximum allowed generalizations. For example, consider above table 2.5, but without the age column. The corresponding hierarchies of the first two attribute are according to figure 2.1 and figure 2.3, vector [1,2] is the maximum permitted generalization and the corresponding lattice is shown in figure 2.8 [16.]

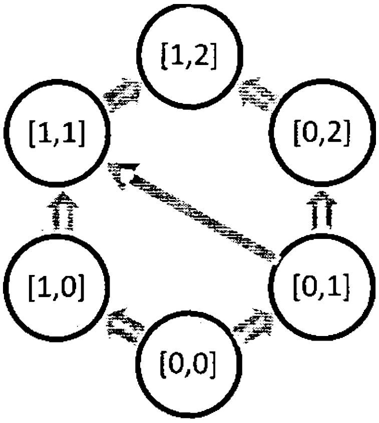


Figure2.8: A Lattices

## 2.15 Generalization Strategy

Every path in the lattice going from bottom vector to the topmost vector with respect to the corresponding arrows is called generalization strategy. In figure 2.8, one strategy could {[0,0] → [0,1]→[1,1] →[1,2]}

## 2.16 Lattice Level

The set of vectors with equal length in the lattice is its lattice level. Consider a figure 2.8, there are four levels. At level 0, we have vector [0,0], at level 1 we have vector [1,0]&[0,1] and so.

## 2.17 Summary

In this chapter, the basic definitions and concepts that are used in this research are reviewed. The idea of *k*-anonymity and the terms that are related to *k*-anonymity are defined. Anonymization techniques called generalization and suppression has been defined. The mechanism through which generalization is performing is also discussed. Also different techniques that are required for generalization and terms that are used in the Incognito algorithm are defined.

# CHAPTER 3: LITERATURE REVIEW

Literature has identified different information disclosure constraints regarding publishing of micro data [4, 5]. Among many, one information disclosure constraint is attribute disclosure. It arises, when an individual is assigned a sensitive attribute value. Others are membership and identity disclosure [4, 5]. Membership disclosure is the learning of whether an entity (individual) is incorporated in the relevant Database. Also, identity is to link an entity to a particular record in the Database. The previous studies [6, 7, 8, 9, 12] are related to limiting of attribute disclosures in data publishing. Following this, this study attempts to prevent the sensitive attributes against an entity.

## 3.1 K-anonymity

A lot of work has been done in k-anonymity to achieve privacy. In data publishing, k-anonymity has been extensively highlighted as a possible definition of privacy. K-anonymity has developed a reputation, because of algorithmic advances in generating k-anonymous forms of a dataset [1, 2, 3, 12, 22, 25, 26]. Furthermore, this study will try to prove that k-anonymity is insufficient regarding privacy. Information revelations are of two types [4, 5]: identity revelation and attribute revelation. Identity revelation arises when a particular record about entity is linked in the microdata. When some new information about individuals is exposed, then attribute revelation happens. *K*-anonymity cover the problem of identity revelation, that is, a tuple cannot be linked back in the k-anonymized dataset to the equivalent record in the original dataset [30]. But it is insufficient to prevent attribute revelation because of two type of attack that take place in *k*-anonymity; one is homogeneity attack and other is background knowledge attack.

Consider a hospital record in below table 3.1

Table3.1: Hospital record

| **Non-sensitive data** | | | **Sensitive Data** |
|---|---|---|---|
| **Zip code** | **Age** | **Nationality** | **Health condition** |
| 24064 | 39 | Chinese | Viral Infection |
| 24064 | 32 | Indian | Viral Infection |
| 24079 | 33 | Japanese | Viral Infection |
| 24079 | 38 | Japanese | Viral Infection |
| 24064 | 26 | Indian | Hepatitis |
| 24079 | 22 | Canadian | Hepatitis |
| 24079 | 27 | American | Headache |
| 24064 | 29 | Japanese | Headache |
| 25964 | 53 | Russian | Heart Disease |
| 25967 | 51 | Russian | Heart Disease |
| 25963 | 42 | Japanese | Cancer |
| 25961 | 41 | Japanese | Cancer |

Above table 3.1 shows patients' records of a hospital. Uniquely recognizing attributes like NIC#, Name, SSN etc are not present in this table. This table is separated into two groups: one is sensitive attributes which contains only health condition and other is non-sensitive attributes such as Age, Zip code and Nationality. If attacker is restricted to find out the value of any entity, then such type of attribute is called sensitive attribute, while attributes other than sensitive are called non-sensitive attribute. Additionally, consider the set of quasi-identifiers for this table is non-sensitive attributes such as Nationality, Zipcode, and Age. Below table 3.2 shows 4-anonymous table resulting from the above table 3.1.

Table3.2: 4-anonymous table

| **Quasi-Identifier** | | | **Sensitive Data** |
|---|---|---|---|
| **Zip code** | **Age** | **Nationality** | **Health condition** |
| 240** | 3* | * | Viral Infection |
| 240** | 3* | * | Viral Infection |
| 240** | 3* | * | Viral Infection |
| 240** | 3* | * | Viral Infection |
| 240** | < 30 | * | Hepatitis |
| 240** | < 30 | * | Hepatitis |
| 240** | < 30 | * | Headache |
| 240** | < 30 | * | Headache |
| 2596* | > 40 | * | Heart Disease |
| 2596* | > 40 | * | Heart Disease |
| 2596* | > 40 | * | Cancer |
| 2596* | > 40 | * | Cancer |

Suppressed value is represented by "*", thus if "age=3*" shows that the range of age is between 30 and 39 "and zip code = 2596*" shows that the range of zip code is between 25960 and 25969.

### 3.1.1 Homogeneity Attack

Consider two neighbors Jan and David. One day David falls ill, Jan wants to determine what kind of disease David is suffering from. About Patients, Jan determines the 4-anonymous released data of the hospital as shown in table 3.2, and he come to know that in this table one of the records contains David's data. Since Jan knows that David is a 39-year-old Chinese male living in the zip code 24064. Therefore, Jan recognizes that David's record occurred in tuple number 1, 2, 3 or 4. Since all these four tuples have same health condition, that is, Viral Infection, thus Jan draws a result that David is a patient of Viral Infection. Such type of attack is called homogeneity attack.

**Observation 1**: K-anonymity generates quasi-identifier groups that reveal information because of less variety in the sensitive attribute.

Assume a dataset having 6,000 rows where three different values are taken by the sensitive attribute. If we apply 5-anonymization, this table will have around 1,200 groups and out of every 8 groups, 1 group will have no diversity. So we can state that about 148 groups there is no diversity. Thus, information is comprомised by a homogeneity attack of about 740 people

### 3.1.2 Background Knowledge Attack

Consider two friends, Jan and Liza. Let Liza is admitted in the same hospital where David is admitted. So in the above table 3.2, medical record of Liza is also appearing. Jan recognizes that Liza is a 41 year-old Japanese female and she is presently living in the zip code 25961. According to above information about Liza, Jan learns that Liza's record is occurred in row number 9, 10, 11, or 12. Without extra information, Jan is not clear whether Liza is a patient of heart disease or cancer. But it is very famous that Japanese have a less occurrence of heart disease. Therefore Jan draws a conclusion with near certifies that Liza has a cancer. This attack is also considered probabilistic attack.

**Observation2**: From above point of view it has been observed that using the background knowledge and homogeneity attacks, K-anonymity does not protect against background knowledge attack.

To overcome the above mentioned limitations of k-anonymity, Machanavajjhala et al. [6] initiated a stronger idea of privacy called L-diversity.

## 3.2L-diversity principle

A quasi-identifier group is supposed to contain l-diversity if there are at least 'well-represented' values for the sensitive attribute. A table is supposed to contain l-diversity if all quasi-identifiers groups of the table have l-diversity [6].

In this principle a number of definitions of the word "well represented" have been introduced by Machanavajjhala et al [6].

The word "well represented" would make certain that in every quasi-identifier group there are at least l distinct values for the sensitive attribute. But this definition does not stop probabilistic inference attacks. For example an equivalence class, appearing one value more frequently than other values, enabling an attacker to find out that an equivalence class/quasi-identifier group individual is probable to contain that value. This aggravated the improvement of the better ideas of '-'diversity

### 3.2.1 Probabilistic l-diversity

If the occurrence of a sensitive value in every quasi-identifier group is at most 1/l, then an anonymized table is said to have probabilistic l-diversity. This definition warrants that a viewer cannot conclude an individual's sensitive value with possibility greater than 1/l.

### 3.2.2 Entropy l-diversity

Any quasi-identifier/equivalence class group, the entropy of equivalence class E is said to be

$$Entropy\,(E) = -\sum_{s \epsilon S} P_{(q^*,s)} \log(P_{(q^*,s')}) \geq \log(l)$$

Where $P_{(q^*,s)}$ shows the division of records in the quasi-identifier group with sensitive attribute identical to s. In order to achieve entropy l-diversity for every equivalence class in a table, the entropy of the whole table must be at least $\log(l)$. That is,

$Entropy\,(E) \geq \log(l)$. But this may be excessively restrictive, because if a few values are very frequent, then entropy of the whole table may be low. This directs to another concept of l-diversity.

### 3.2.3 Recursive (c, l)-diversity

This concept says that the most common value does not show too regularly, and do not show too rarely the less common values. In Recursive (c, l)-diversity, c is a float and l is an integer number.

## 3.3 Limitations of l - diversity

In protecting against attribute exposure, the l-diversity principle leads to a significant step beyond k-anonymity, but consists of many weaknesses.

**Observation:** To get l-diversity it may be tricky and may not give satisfactory privacy safety.

Consider a dataset containing just one sensitive attribute, let this particular attribute consist of pass and fail values only. Suppose that there are 2,000 students with their corresponding records, say 99% of them have passed, and only 1% of students have failed values. So, these two values contain very diverse degrees of sensitivity.

A student may not mind to know others if he is passed, but he may not like to know others if he is failed. In such situation, 2-diversity for a quasi-identifier group does not provide privacy that has only records that have passed value.

**Thus l-diversity is not sufficient to prevent attribute exposure.**

### 3.3.1 L-diversity Attacks

There are three types of attacks on l-diversity.

#### 3.3.1.1 Skewness Attack

L-diversity does not stop attribute revelation, when the entire distribution of sensitive attribute is skewed. For example, suppose that there is an equal number of pass and fail records for a quasi-identifier group. It satisfies any ‘-‘diversity constraint that can be applied, that is, entropy 2-diversity, distinct 2-diversity and any recursive (c, 2)-diversity. However, this depicts a privacy threat, because anybody in the class might be supposed to contain 50 percent probability of being failed, as compared with 1 percent of the entire students.

Now, suppose a quasi-identifier group that contains only 1 pass and 49 fail records. This fulfills ‘-‘diversity that may apply, anybody would be considered 98% chance of having failed in the quasi-identifier group, rather than 1%. Although the two groups show dissimilar levels of privacy threats, however, this quasi-identifier group contains accurately the identical diversity as a group that contain 49 passed and 1 failed records.

#### 3.3.1.2 Probabilistic Inference Attack

For such type of attack, l-diversity is not sufficient. For example, consider below table 3.3, which satisfies 3-diversity.

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

| Zip code | Age | Nationality | Disease |
|---|---|---|---|
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | Cancer |
| 240** | < 30 | * | Hepatitis |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |
| 240** | < 30 | * | HIV |

In above table 3.3, if each group consists of ten tuples, then in the “Disease” area, one of them is “Cancer”, one is “Hepatitis” and the remaining eight are “HIV”. This satisfies 3-diversity, but the attacker can still confirm that the target person’s disease is “HIV” with the accuracy of 80%.

### 3.3.1.3 Similarity Attack

In a quasi-identifier group, when the sensitive attribute are different but semantically identical, an attacker can get significant information. Consider below table 3.4

Table3.4: Raw data

| Zip code | Age | Monthly salary | Health condition |
|---|---|---|---|
| 78677 | 26 | 60,000 | Gastric malignancy |
| 78602 | 28 | 80,000 | Stomach inflammation |
| 78678 | 29 | 70,000 | Gastric ulcer |
| 78905 | 51 | 40,000 | HIV |
| 78909 | 55 | 1,00,000 | Indigestion |
| 78906 | 53 | 90,000 | Fever |
| 78605 | 33 | 80,000 | Headache |
| 78673 | 39 | 70,000 | Flu |
| 78607 | 30 | 1,10,000 | Stomach cancer |

Below table 3.5 shows an anonymized version of above table 3.4, satisfying distinct and entropy 3-diversity.

Table3.5: 3-anonymous with 3-diversity

| Quasi-identifier | | Sensitive data | |
|---|---|---|---|
| Zip code | Age | Monthly salary | Health condition |
| 786** | 2* | 60,000 | Gastric malignancy |
| 786** | 2* | 80,000 | Stomach inflammation |
| 786** | 2* | 70,000 | Gastric ulcer |
| 789** | >50 | 40,000 | HIV |
| 789** | >50 | 1,00,000 | Indigestion |
| 789** | >50 | 90;000 | Fever |
| 786** | 3* | 80,000 | Headache |
| 786** | 3* | 70,000 | Flu |
| 786** | 3* | 1,10,000 | Stomach cancer |

Above anonymous table 3.5 consists of two sensitive attributes. One is Monthly salary and other is Health condition. If attacker gets information that Jan is 29 years of age living in zip code 78678, then attacker easily concludes that Jan's salary is in the range [60,000-80,000] and he has some stomach-related disease.

To overcome the drawback of l-diversity, t-closeness [7] was proposed

## 3.4 t-closeness

The concept of *t*-closeness [7] is that it defines a distance between sensitive attributes to protect against sensitive attributes revelation. In other words it defines that the distribution of sensitive attribute in any quasi-identifier group is close to the distribution of attribute in the entire table, that is, the distance is no more than a threshold $t$ between the distributions of the attribute in the group and between the whole tables.

Now, if a distance metric between sensitive attributes is required, the Earth Mover Distance (EMD) [31] has been used in t-closeness to calculate the distance among the two distributions.

### 3.4.1 Earth Movers Distance

The EMD transform one distribution to another via distribution mass among each other, such that the minimum amount of work is required. EMD could be defined using the transportation problem.

Let $X = (x_1, x_2, \ldots, x_m), Y = (y_1, y_2, \ldots, y_m)$ are the rows of the dataset and $d_{ij}$ is the ground distance between $i^{th}$ and $j^{th}$ element of $X$ and $Y$ respectively. In order to find the flow $F = [f_{ij}]$ where $f_{ij}$ represents the flow of mass from element $i$ of row $X$ to element $j$ of row $Y$. such that entire work is significantly minimized.

$$WORK\ (X, Y, F) = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} f_{ij}$$

Subject to below constraints

$$f_{ij} \geq 0\,, 1 \leq i \leq n, 1 \leq j \leq n \qquad \rightarrow (A)$$

$$x_i - \sum_{j=1}^{n} f_{ij} + \sum_{j=1}^{n} f_{ji} = y_i\,, 1 \leq i \leq n \quad \rightarrow (B)$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = 1 \qquad \rightarrow (C)$$

There are several advantages for the use of this measure. This measure could be easily integrated with the Incognito algorithm because of its generalization and subset properties. It implies that monotonicity with respect to both the generalization level and number of attributes is chosen.

#### 3.4.1.1 Limitation of t-closeness

To enforce t-closeness, there is no computational procedure. Also co-relation between different attributes is lost, because every attribute is generalized individually and so we lose their dependence on each other. Another limitation is that, using small value of t damaged data utility and will result increase in computational time.

So another technique called p-sensitive k-anonymity [8] was suggested.

## 3.5 P-sensitive k-anonymity

A released table satisfies p-sensitive k-anonymity, if every group of quasi-identifier consists of at least p different sensitive values and it also satisfies k-anonymity [8].

For protecting against attribute revelation, p-sensitive k-anonymity leads to a significant step beyond k-anonymity, but it has still several shortcomings. Below, we will illustrate that p-sensitive k-anonymity is not satisfactory for preventing similarity attack.

### 3.5.1 Similarity Attack

An attacker can get significant information, when in a quasi-identifier group the sensitive attribute are different but possess identical sensitivity

According to their sensitivity, the sensitive attributes in p-sensitive k-anonymity are partitioned and placed into different categories. Consider a table 3.6, the Health condition attribute of which are separated into four classes according to table 3.7

Table3.6: Raw Data

| Zip code | Age | Country | Health condition |
|---|---|---|---|
| 25359 | 25 | Denmark | Flu |
| 25308 | 29 | 'France | Asthma |
| 25305 | 23 | Germany | Flu |
| 25308 | 26 | France | Indigestion |
| 24064 | 42 | Japan | Hepatitis |
| 24085 | 49 | China | Obesity |
| 24075 | 44 | Pakistan | Flu |
| 24073 | 41 | Pakistan | Phthisis |
| 25306 | 35 | Canada | HIV |
| 25305 | 39 | USA | Cancer |
| 25306 | 32 | Canada | Cancer |
| 25359 | 31 | Canada | HIV |

Below table 3.7 shows different disease and its category

Table3.7: List of Categories

| Category # | Health condition | Sensitivity |
|---|---|---|
| 1 | Cancer, HIV | Most secret |
| 2 | Hepatitis, Phthisis | Secret |
| 3 | Asthma, Obesity | Less secret |
| 4 | Indigestion, Flu | Non secret |

Different types of Health condition are organized into a category according to their sensitivity according to above table 3.7. For example, most secret information about individuals depicts HIV and Cancer. Organization is concerned to protect not only these top secret diseases but also the category of those top secret diseases. Let's suppose p-sensitive k-anonymity property is applied and the microdata have Health condition attribute which contain specific sensitive values, it may be possible that all the p distinct sensitive values in each quasi-identifier group belong to the one category. For example, below table 3.8 is 2-sensitive 4-anonymous (means that there is 2 sensitive values in each group) view of above table 3.6

Table3.8: 2-sensitive 4-anonymous data

| Zip code | Age | Country | Health condition |
|---|---|---|---|
| 253** | <30 | Europe | Flu |
| 253** | <30 | Europe | Indigestion |
| 253** | <30 | Europe | Flu |
| 253** | <30 | Europe | Indigestion |
| 240** | >40 | Asia | Hepatitis |
| 240** | >40 | Asia | Obesity |
| 240** | >40 | Asia | Flu |
| 240** | >40 | Asia | Flu |
| 2530* | 3* | America | HIV |
| 2530* | 3* | America | Cancer |
| 2530* | 3* | America | Cancer |
| 2530* | 3* | America | HIV |

According to above table 3.8, it satisfy p-sensitive k-anonymity property but the all sensitive value {HIV, Cancer, Cancer, HIV } in last quasi-identifier group belong to one category. The information of an individual belong to most secret category needs to be protected, no issue either it is Cancer or HIV. From this point of view p-sensitive k-anonymity does not provide sufficient protection for sensitive attribute. To protect sensitive values and avoid similarity attack, another technique called (p, α) sensitive k-anonymity has been defined.

## 3.6 (p, α)-sensitive k-anonymity

A released table satisfies (p, α) sensitive k-anonymity, if every group of quasi-identifier consists of at least p different sensitive values with its total weight at least α and also it satisfies k-anonymity [9].

(p, α) sensitive k-anonymity model can well protect sensitive information as compared to previous model, but it still focuses on specific value.

Consider below table 3.9

Table3.9: Raw data

| Zip code | Age | Country | Health condition |
|---|---|---|---|
| 25359 | 26 | Canada | HIV |
| 25308 | 25 | USA | Hepatitis |
| 25305 | 27 | USA | Obesity |
| 25308 | 24 | Canada | Cancer |
| 24064 | 42 | USA | Asthma |
| 24085 | 45 | China | Phthisis |
| 24075 | 48 | Pakistan | HIV |
| 24073 | 41 | Pakistan | Flu |
| 25306 | 32 | Canada | Asthma |
| 25305 | 35 | Canada | Phthisis |
| 25306 | 36 | Canada | Flu |

If $(p, \alpha)$ sensitive k-anonymity is apply to above table 3.9, we get

Table3.10: (3, 1)-sensitive 4-anonymous table

| Zipcode | Age | Country | Health condition | Weight | Total |
|---|---|---|---|---|---|
| 2**** | <50 | * | HIV | 0 | 1 |
| 2**** | <50 | * | Cancer | 0 | |
| 2**** | <50 | * | HIV | 0 | |
| 2**** | <50 | * | Flu | 1 | |
| 253** | <40 | America | Hepatitis | 1/3 | 2 |
| 253** | <40 | America | Phthisis | 1/3 | |
| 253** | <40 | America | Asthma | 2/3 | |
| 253** | <40 | America | Obesity | 2/3 | |
| -- | -- | -- | -- | -- | 3 |

Since in above table, each group consists of three distinct sensitive values and the total weight of each quasi-identifier group is at least 1. As shown in above first group, three out of four values belong to same category, so attacker can still confirm that the target person's disease is "most secret" that is either HIV or Cancer with the accuracy of 75%.

## 3.7 Problem Statement

For the protection of sensitive attributes, various models such as l-diversity [6], p-sensitive k-anonymity [8] and $(p, \alpha)$-sensitive k-anonymity [9] have been introduced. But these improved versions of k-anonymity still allow the sensitive values to be exposed or contain several limitations. $(p, \alpha)$ sensitive k-anonymity model provide well protection for sensitive values as compared to earlier enhanced versions of k-anonymity, but it is mainly focused on specific value due to which probabilistic attack may occur and privacy of the individual may be compromised.

## 3.8 Summary

To prevent attribute revelation, *K*-anonymity is not sufficient because of two types of attacks, one is called similarity attack and other is called background knowledge attack. To solve this problem, several models such *l*-diversity, enhance version of *l*-diversity, p-sensitive k-anonymity and (*p*, *α*)-sensitive *k*-anonymity were proposed [2, 6, 8, 9]. But these improved versions of k-anonymity still allow the sensitive values to be exposed or contain several limitations. Following this, this study attempts to prevent the sensitive attributes disclosure against individuals. For this purpose a new technique called enhanced (*p*, *α*)-sensitive *k*-anonymity model has been proposed.

# CHAPTER 4: PROPOSED TECHNIQUE

To secure sensitive attributes, enhance privacy and reduce similarity attack, a specific category in (*p*, *α*)-sensitive *k*-anonymity model [9] has been used, instead of specific value we called it enhanced (*p*, *α*)-sensitive *k*-anonymity model. For proposed algorithm, incognito algorithm is extended [12], which is a global-recoding based algorithm and may produce needless data loss to the dataset. Here a local-recoding based algorithm has been proposed, called top-down local recoding algorithm

## 4.1 Enhanced (*p, α*)-sensitive *k*-anonymity

A released table satisfies enhanced (*p*, *α*)-sensitive *k*-anonymity, if every group of quasi-identifier consists of at least p different sensitive categories with its total weight at least α and also it satisfies k-anonymity.

For the protection of sensitive attribute, values of sensitive attribute *H* are sorted based on their sensitivity. An ordered value domain *D* are formed by the arrangement of *H*. The sensitive attribute is partitioned into *x*-categories $(H_1, H_2, \ldots, H_x)$, such that such that $H = H_{i=1}^{x} H_i$ , $H_i \cap H_j = \emptyset$ for $(i \neq j)$, $H_i \leq H_j$ means that $H_i$ is more sensitive than $H_j$ (for $i \leq j \leq x$).

For more explanation consider Health_condition *H* = {Cancer, HIV, Hepatitis, Phthisis, Asthma, Obesity, Indigestion, Flu} in below table 4.1

Table4.1: Raw data

| Zip code | Age | Country | Health condition |
|---|---|---|---|
| 25359 | 25 | Denmark | Flu |
| 25308 | 29 | France | Asthma |
| 25305 | 23 | Germany | Flu |
| 25308 | 26 | France | Indigestion |
| 24064 | 42 | Japan | Hepatitis |
| 24085 | 49 | China | Obesity |
| 24075 | 44 | Pakistan | Flu |
| 24073 | 41 | Pakistan | Phthisis |
| 25306 | 35 | Canada | HIV |
| 25305 | 39 | USA | Cancer |
| 25306 | 32 | Canada | Cancer |
| 25359 | 31 | Canada | HIV |

According to the sensitivity of the health condition it has been partitioned into four categories according to the table 4.2 below, where $H_1$ shows most secret where as $H_4$ is non-secret and shows the minimum level of secrecy.

Table4.2: Categories of Health_condition

| **Category #** | **Health condition** | **Sensitivity** |
|---|---|---|
| 1 | Cancer, HIV | Most secret |
| 2 | Hepatitis, Phthisis | Secret |
| 3 | Asthma, Obesity | Less secret |
| 4 | Indigestion, Flu | Non secret |

For enhanced (*p*, *α*)-sensitive *k*-anonymity, ordinal weight has been proposed for each category to show the level of each sensitive value belong to the quasi-identifier group.

For an attribute $H$, let $D(H) = \{H_1, H_2; H_3, \ldots., H_x\}$ represent a separation of categorical domain and $Weight\ (H_i)$ represent the weight of category $(H_i)$. Then

$$\left.\begin{aligned} &Weight\ (H_i) = (i-1)/(x-1);\quad 1 \le i < x \\ &weight\ (H_x) = 1 \end{aligned}\right\} \longrightarrow \textcircled{1}$$

According to above formula, sensitive attributes has been partition as shown in table 4.2

$$weight\ (S_1) = (1-1)/(4-1) = 0$$
$$weight\ (S_2\ ) = (2-1)/(4-1) = 1/3$$
$$weight\ (S_3\ ) = (3-1)/(4-1) = 2/3$$
$$weight\ (S_4\ \ ) = (4-1)/(4-1) = 1$$

So it means that weight of the category is equal to the weight of the sensitive value that belongs to the category. The total weight of each sensitive value that the quasi-identifier group contains is the weight of the quasi-identifier group. As shown in table 4.2, four values set *A*= {Indigestion, Obesity, Hepatitis, HIV}.

According to formula (1), the total weight of A is 1 + 2/3 + 1/3 + 0 = 2

The distance between HIV $(H_1)$ and Indigestion $(H_4)$ is 3/3=1, while the distance between Hepatitis $(H_2)$ and Obesity $(H_3)$ is 1/3.

## 4.2 The Anonymization Algorithms

### 4.3.1 Global Recoding

Incognito algorithm is a global-recoding based algorithm which is extended [12] for enhanced (*p*, *α*)-sensitive *k*-anonymity model.

#### 4.2.1.1 Incognito Algorithm

For the *k*-anonymity, incognito algorithm is an optimum global-recoding based algorithm; incognito algorithm produces all probable *k*-anonymous full-domain generalizations of *T*, alongwith elective suppression of tuples. According to subset property of incognito algorithm, it starts from subsets of the quasi-identifier by checking single attribute, and then k-anonymity is checking in iterations with respect to gradually large subsets.

Each iteration of incognito algorithm consists of two main parts:

1- Every iteration considers all the nodes in a set *S* constructed from subsets of the quasi-identifier of size i. Taking advantage of the generalization and rollup property it goes through these nodes in a bottom-up breadth first search.

2- Next the incognito algorithm builds the set of candidate nodes *S* with quasi-identifier of size i + 1 and taking advantage of the subset property by avoiding the nodes that cannot be solved, when the set of attributes is larger.

This summarizes that incognito algorithm using search of bottom-up breadth first on generalization hierarchy and checking the attributes in iteration

For example, for quasi-identifier it checks *k*-anonymity for each single attribute in iteration 1, and removes those generalizations that do not fulfill *k*-anonymity. Then in iteration 2, the remaining generalizations are combined in pair and performing the similar process on pair of attributes and so on until the whole set of attribute is complete.

To more explain, consider below table with quasi-identifier = {Zip code, Marital_Status, Gender} and assume that *k* = 3 and *MaxSup*=2

Table4.3: Raw data

| Zip code | Marital_status | Gender | Health_condition |
|---|---|---|---|
| 22030 | Married | Female | Hypertension |
| 22030 | Married | Female | Hypertension |
| 22030 | Single | Male | Obesity |
| 22032 | Single | Male | HIV |
| 22032 | Divorced | Female | Obesity |
| 22032 | Divorced | Female | Hypertension |
| 22045 | Divorced | Male | Obesity |
| 22047 | Widow | Male | HIV |
| 22047 | Widow | Male | HIV |
| 22047 | Single | Female | Obesity |

In below figure 4.1, the complete value generalization hierarchies of quasi-identifier of all the subsets are shown on the left side, while the sub-hierarchies performed by incognito algorithm at every iteration are shown on the right side for the above table 4.3.

In the hierarchy, Zip code is denoted by Z, Marital_status is denoted by M and Gender is represented by G. Also the different values of QI assigned to hierarchy are mention below

$Z_0$= {22030, 22032, 22045, 22047}, $Z_1$= {2203*, 2204*}, $Z_2$ = {220**}.

$M_0$= {Widow, Divorced, Married, Single}, $M_1$= {Been Married, Never Married}, $M_2$ = {*}.

$G_0$= {Male, Female}, $G_1$= {Gender}.

Complete Hierarchies | Sub Hierarchies

Iteration 1:

Complete Hierarchies:

$$\begin{tikzcd}
(Z_2) & (M_2) & \\
(Z_1) \arrow[u] & (M_1) \arrow[u] & (G_1) \\
(Z_0) \arrow[u] & (M_0) \arrow[u] & (G_0) \arrow[u]
\end{tikzcd}$$

Sub Hierarchies:

$$\begin{tikzcd}
(Z_2) & (M_2) & \\
(Z_1) \arrow[u] & (M_1) \arrow[u] & (G_1) \\
(Z_0) \arrow[u] & & (G_0) \arrow[u]
\end{tikzcd}$$

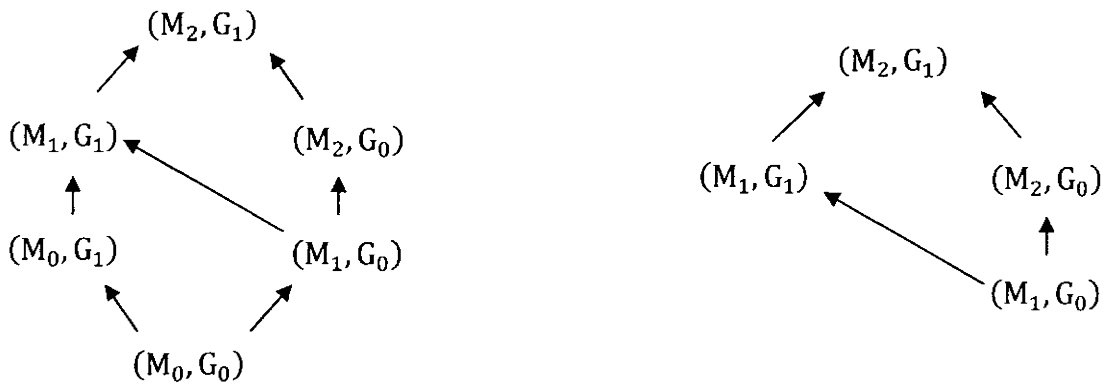
Iteration 2:

$$\begin{tikzcd}
 & (Z_2, M_2) & \\
(Z_1, M_2) \arrow[ur] & & (Z_2, M_1) \arrow[ul] \\
(Z_0, M_2) \arrow[u] & (Z_1, M_1) \arrow[ul] \arrow[ur] & (Z_2, M_0) \arrow[u] \\
(Z_0, M_1) \arrow[u] \arrow[ur] & & (Z_1, M_0) \arrow[ul] \arrow[u] \\
 & (Z_0, M_0) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

$$\begin{tikzcd}
 & (Z_2, M_2) & \\
(Z_1, M_2) \arrow[ur] & & (Z_2, M_1) \arrow[ul] \\
(Z_0, M_2) \arrow[u] & (Z_1, M_1) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

$$\begin{tikzcd}
 & (Z_2, G_1) & \\
(Z_1, G_1) \arrow[ur] & & (Z_2, G_0) \arrow[ul] \\
(Z_0, G_1) \arrow[u] & & (Z_1, G_0) \arrow[ull] \arrow[u] \\
 & (Z_0, G_0) \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

$$\begin{tikzcd}
 & (Z_2, G_1) & \\
(Z_1, G_1) \arrow[ur] & & (Z_2, G_0) \arrow[ul] \\
(Z_0, G_1) \arrow[u] & & (Z_1, G_0) \arrow[ull] \arrow[u]
\end{tikzcd}$$



Iteration 3:

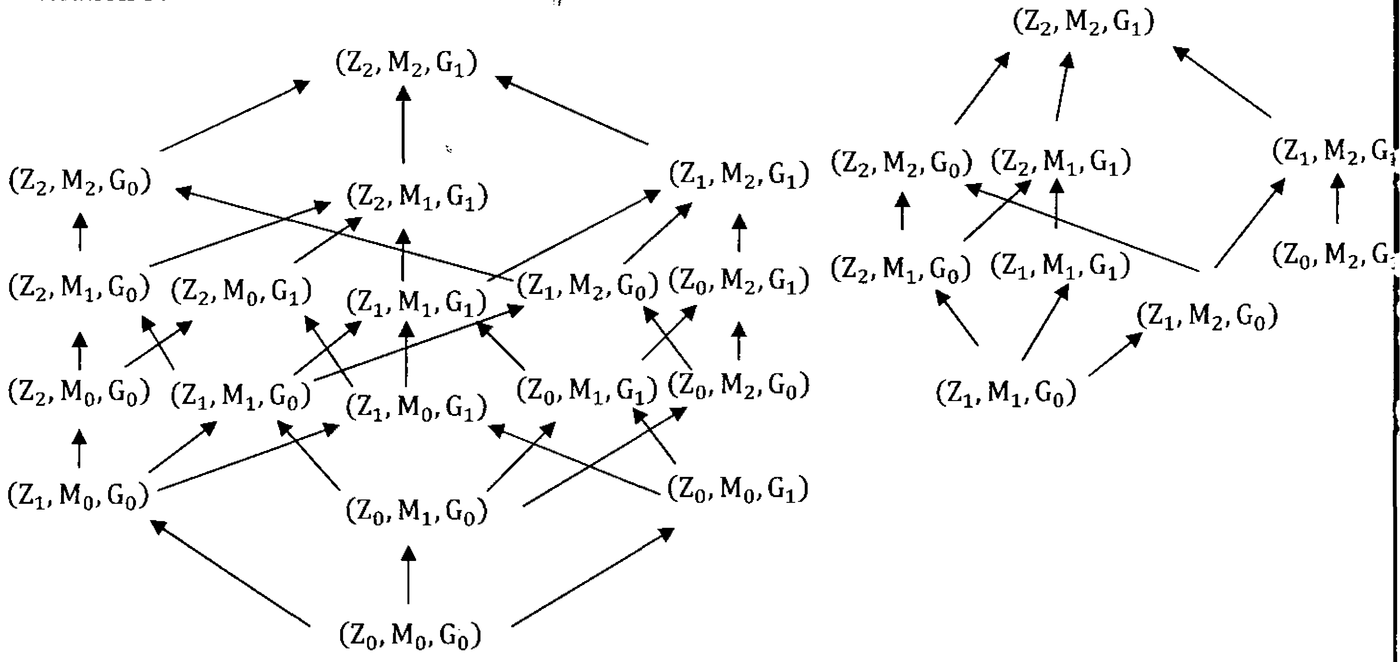


Figure4.1: Sub-hierarchies computed by Incognito algorithm according to table 4.3

Explanations of above figure 4.1 are mentioned below in different iterations

Iteration 1:

$DGH_{(Z_0)}$: Vertices $(Z_0)$, $(Z_1)$ and $(Z_2)$ are noticeable true, since table $T_{(Z_0)}$ satisfies 3-anonymity by suppressing a number of records lower than MaxSup.

$DGH_{(M_0)}$: Vertices $(M_0)$, is marked false, since to satisfy 3-anonymity, in table $T_{(M_0)}$ we need to suppress more than 3 tuples. Vertex $(M_1)$ and vertex $(M_2)$ are marked true since table $T_{(M_1)}$ satisfies 3-anonymity by suppressing a number of records lower than MaxSup.

$DGH_{(Z_0)}$: Vertices $(G_0)$ and $(G_1)$ are marked true, since table $T_{(G_0)}$ satisfies 3-anonymity by suppressing a number of records lower than MaxSup.

Iteration 2:

$DGH_{(Z_0,M_0)}$: Since $(M_0)$, has been false in the previous section, this hierarchy does not include vertices $(Z_0, M_0)$, $(Z_1, M_0)$ and $(Z_2, M_0)$. Vertex $(Z_0, M_1)$ is noticeable false, since $T_{(Z_0,M_1)}$ satisfy 3-anonymity only if more than 3 tuples are suppressed. Vertices $(Z_0, M_2)$, $(Z_1, M_2)$, $(Z_2, M_2)$, $(Z_1, M_1)$ and $(Z_2, M_1)$ are marked true, since table $T_{(Z_0,M_2)}$ and $T_{(Z_1,M_1)}$ satisfy 3-anonymity by suppressing a number of records lower than MaxSup.

$DGH_{(Z_0,G_0)}$: Vertex $(Z_0, G_0)$ is noticeable false, since $T_{(Z_0,G_0)}$ satisfy 3-anonymity only if more than MaxSup tuples are suppressed. Vertices $(Z_0, G_1)$, $(Z_1, G_1)$, $(Z_2, G_1)$, $(Z_1, G_0)$ and $(Z_2, G_0)$ are marked true, since table $T_{(Z_0,G_1)}$ and $T_{(Z_1,G_0)}$ satisfy 3-anonymity by suppressing a number of records lower than MaxSup.

$DGH_{(M_0,S_0)}$: Since $(M_0)$, has been false in the previous section, this hierarchy does not include vertices $(M_0, S_0)$ and $(M_0, S_1)$. All the other vertices in the hierarchy are marked true, since $T_{(M_1,S_0)}$ satisfy 3-anonymity by suppressing a number of records lower than MaxSup.

Iteration 3:

$DGH_{(Z_0,M_0,G_0)}$: Since $DGH_{(Z_0,M_0)}$ does not contain Vertices $(Z_0, M_0)$, $(Z_1, M_0)$ and $(Z_2, M_0)$ and vertex $(Z_0, M_1)$ has been marked false, this hierarchy does not contain vertices $(Z_0, M_0, G_0)$, $(Z_1, M_0, G_0)$, $(Z_2, M_0, G_0)$, $(Z_0, M_0, G_1)$, $(Z_1, M_0, G_1)$, $(Z_2, M_0, G_1)$, $(Z_0, M_1, G_0)$ and $(Z_0, M_1, G_1)$. Similarly since vertex $(Z_0, G_0)$ has been marked false in $DGH_{(Z_0,G_0)}$, this hierarchy does not contain vertex $(Z_0, M_2, G_0)$. Vertices $(Z_1, M_1, G_0)$, $(Z_1, M_1, G_1)$, $(Z_1, M_2, G_0)$, $(Z_1, M_2, G_1)$, $(Z_2, M_1, G_0)$, $(Z_2, M_1, G_1)$, $(Z_2, M_2, G_0)$ and $(Z_2, M_2, G_1)$ are marked true, since table $T_{(Z_1,M_1,G_0)}$ satisfy 3-anonymity by suppressing a number of records lower than *MaxSup*. Similarly, Vertex $(Z_0, M_2, G_1)$ is marked true, since table $T_{(Z_0,M_2,G_1)}$ satisfy 3-anonymity by suppressing a number of records lower than *MaxSup*. Incognito algorithm has been widely used in the research of *k*-anonymity, similarly for the research of *p*-sensitive *k*-anonymity [8] and (*p*, *α*)-sensitive *k*-anonymity [9], incognito algorithm has also been used.

### 4.2.2 Local Recoding

An extended global-recoding based algorithm called incognito algorithm [12], which is not capable and may produce needless data loss to the dataset

A capable local-recoding based algorithm has been proposed here. The algorithms for enhanced (*p*, *α*)-sensitive *k*-anonymity are like to incognito and (*p*, *α*)-sensitive *k*-anonymity [12, 9] but the testing criteria of every node in the solution space is difference.

#### 4.2.2.1 Top down Local-recoding Algorithm

```
Step1:  All tuples should generalize fully.
Step2:  Let $A$ be a set having all these generalized tuples
Step3:  $H \leftarrow \{A\}; 0 \leftarrow \emptyset$
Step4:  Repeat
Step5:    $H' \leftarrow \emptyset$
Step6:    For all $A \in H$ do
Step7:      All tuples of $A$ should specialize one level down in the generalization hierarchy
            forming a number of specialized child nodes
Step8:      The nodes which do not satisfy enhanced ($p$, $\alpha$)-sensitive $k$-anonymity un-
            specialize by moving the tuples back to the parent node.
Step9:      If the parent node $A$ does not satisfy enhanced ($p$, $\alpha$)-sensitive $k$-anonymity
            Then
Step10:       Some tuples in the remaining child nodes un-specialize, so that the parent
              node $A$ satisfies enhanced ($p$, $\alpha$)-sensitive k-anonymity
Step11:     End of if
Step12:       For all un-empty branches $B$ of $A$, do $H' \leftarrow H' \cup \{B\}$
Step13:       $H \leftarrow H'$
Step14:       If $A$ is un-empty then $0 \leftarrow 0 \cup \{A\}$
Step15:     End of for
Step16: Until $H = \emptyset$
```

Step17: **Return** 0.

The concept of this algorithm is that in initial stage it completely generalizes all tuples. Then, in iterations tuples are specialized one level down forming child nodes. Throughout the specialization, enhanced (*p*, *α*)-sensitive *k*-anonymity must be maintained and the process will continue until the tuples cannot be specialized further. For enhanced (*p*, *α*)-sensitive *k* -anonymity the pseudo code is depicted in above algorithm. Consider a diagram, to initially illustrate how the algorithm works for quasi-identifier of size 1. Then, the technique is extended for the size of quasi-identifier greater than 1.

For example, consider a sample data, where only one quasi-identifier, that is, Zip code

Table4.1(a): Raw data

| Zip code | Gender | Health condition |
|---|---|---|
| 73456 | Male | HIV |
| 73456 | Male | Indigestion |
| 73456 | Female | Flu |
| 73455 | Female | Cancer |

Table 4.4(b): Projected Table

| S# | Zip code | Health condition |
|---|---|---|
| 1 | 73456 | HIV |
| 2 | 73456 | Indigestion |
| 3 | 73456 | Flu |
| 4 | 73455 | Cancer |

Table4.4(c): Generalized Table

| S# | Zip code | Health condition |
|---|---|---|
| 1 | 73456 | HIV |
| 2 | 73456 | Indigestion |
| 3 | 7345* | Flu |
| 4 | 7345* | Cancer |

As in above table 4.4(a), there are only two sensitive values, that is, HIV and Cancer, we suppose that *α*= 1, *p*= 2, *k* = 2. Initially, totally generalize all four tuples to a mainly generalized value, such that, Zip code=***** as shown in below figure 4.2(a). Then, in the generalization hierarchy every tuple should specialize one level down forming child nodes. In figure 4.2(b) the branch with Zipcode = 7**** is obtained. In the next iterations, the branch with Zipcode = 73*** in figure 4.2(c) and the branch with Zip code = 734** and with Zipcode = 7345* in figure 4.2(d) and figure 4.2(e) respectively is obtained. Next, two branches are obtained by further specialization of tuples as shown Figure 4.2(f). Thus processing of the specialization is view in the form of growth of a tree.

The specialization will be successful, if every leaf node fulfills the criteria of enhanced (*p*, *α*)-sensitive *k*-anonymity. However a number of problematic leaf nodes that are not satisfied enhanced (*p*, *α*)-sensitive k-anonymity may encounter. In the generalization hierarchy all those tuples that are not specialized will be pushed back to parent node and should keep unspecialized in this process. For example, the leaf node in figure 4.2(f), with Zip code = 73455 has only one tuple, which does not satisfy enhanced (*p*, *α*)-sensitive *k*-anonymity. Thus, this tuple has to be pushed upward with Zip code =7345*, shown in below figure 4.2(g).

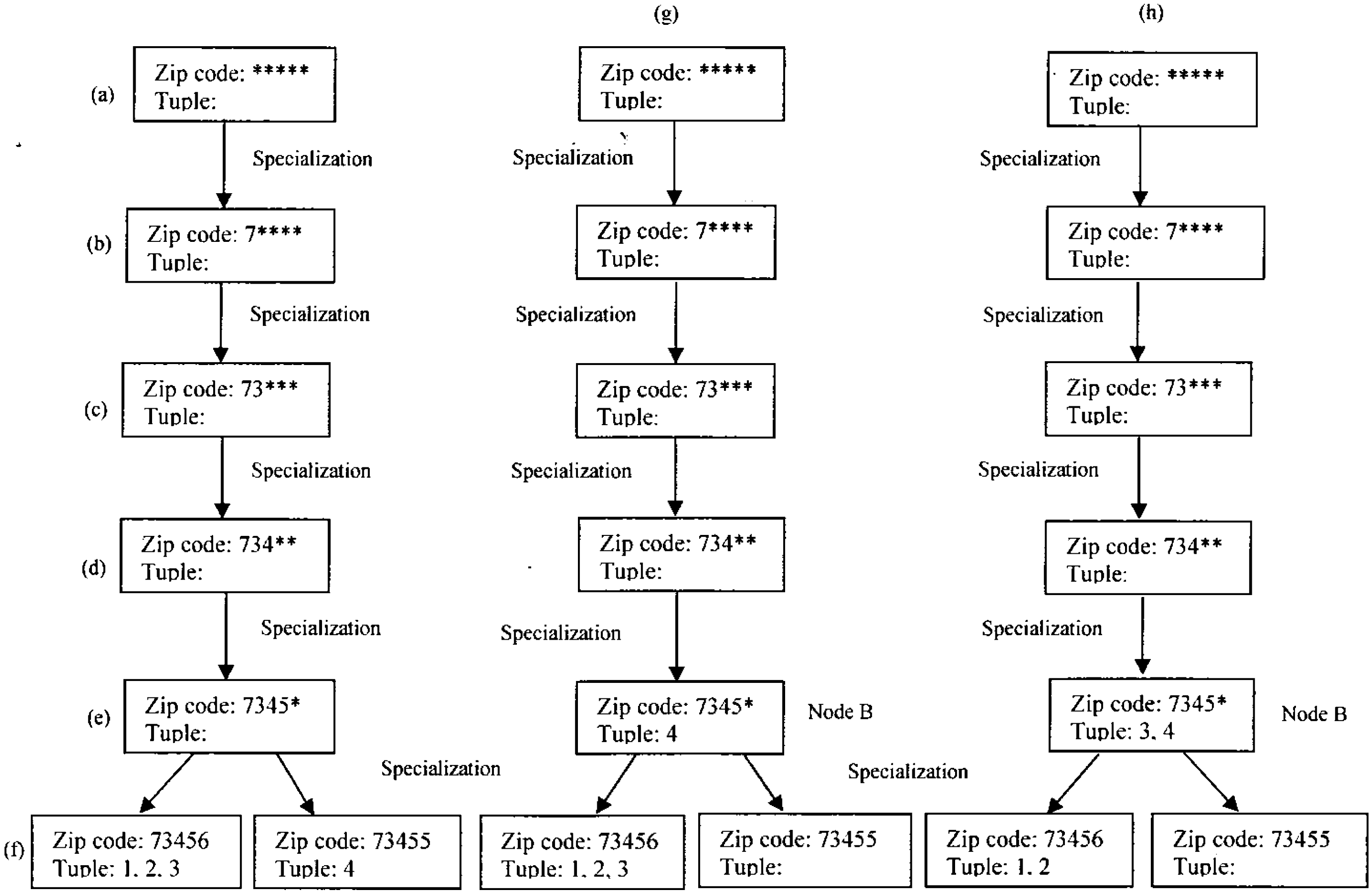


Figure4.2: Diagram for quasi-identifier = 1 (Zipcode) [9]

After that, a number of problematic leaf nodes that are not satisfied enhanced (*p*, *α*)-sensitive k-anonymity pushed back to parent node. However, in the parent node all the tuples that do not fulfill the condition of enhanced (*p*, *α*)-sensitive *k*-anonymity, several tuples from leaf nodes L are further moved to the parent node. Such that the leaf nodes L and parent node can maintain enhanced (*p*, *α*)-sensitive *k* -anonymity. For example, with Zip code = 7345* in figure 4.2(g), the parent node is not satisfied enhanced (*p*, *α*)-sensitive *k* -anonymity. Thus, in the node *B* with Zipcode = 73455 should move one tuple back to parent node (which satisfies enhanced (*p*, *α*)-sensitive *k*-anonymity).

Lastly, a dataset is obtained, as shown in figure 4.2(h), where tuples 3 and 4 of the Zip code are generalized to 7345* and tuples 1 and 2 of the Zip code remains 73456. After the specialization, final allocation of tuples is shown in figure 4.2(h) and the resulted table can be seen in Table 4.4 (c).

In step10 of the above top-down algorithm, some tuples are un-specialized which have fulfill condition of enhanced (p, α)-sensitive k-anonymity already. So what criterion is applied, which selects tuples in such away to create a generalized dataset with less data loss? The following extra steps are applied to handle this problem.

All tuples are further specialized in all candidate nodes and specialization procedure is repeatedly performed until the tuples do not specialize anymore. Then, the numbers of times of specialization for every tuple are recorded. If the specializations of tuple require less time, then it should be assumed as an excellent option for un-specialization because it cannot be specialized deeply in later steps.

Next the top-down local recoding algorithm is extended to grip the situation where the size of quasi-identifier has more than one.

**More Than 1 Size of Quasi-identifier:**

In the first step, generalize fully all attributes of the tuples. Then, the "best" attribute for specialization for every iteration, is find out and do the specialization for the "best" attribute. The iteration performs until no more specialization is needed.

Suppose a group G, for choosing the criteria of "best" attributes.

**Criterion 1 (Maximum No of Specialized Tuples):** Final sharing of the tuples is obtained throughout specialization of G. A number of tuples are specialized and several may still stay in G. The "best" specialization will give the greatest number of tuples to be specialized because that corresponds to the least entire distortion.

For example, below figure 4.3 (a) and 4.3 (b) shows the final distribution of tuples of the specialization with attributes Zipcode and Age, respectively. If the dataset has these two quasi-identifiers only, attribute Zipcode for specialization should be chosen because it gives the greatest number of tuples to be specialized.

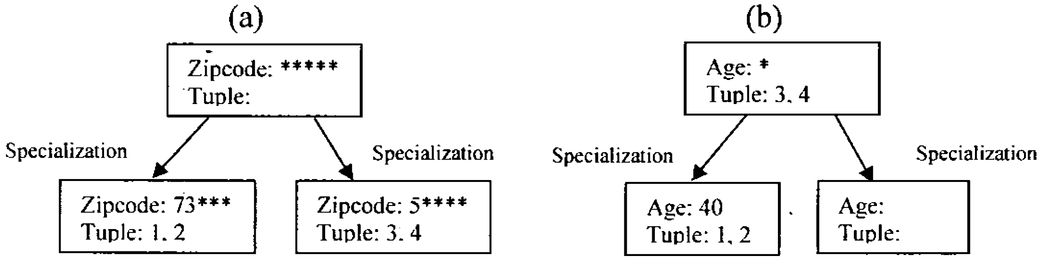


Figure4.3 (a, b): illustration for criteria of choosing the "Best" attribute

**Criterion 2 (Specialize Smallest No of Branches):** When considering the first criterion, we will think the more number of branches to be specialized (i.e. un-empty branches); in situation there is a tie. The "best" specialization gives to specialized the least number of branches. A pointer of further generalized domain indicates the smallest number of branches and compared to a fewer generalized domain it is a better option.

For example, figure 4.4(a) and 4.4(b) shows the final sharing of specialization of tuples with attribute Zipcode and Age, respectively. If the dataset contains these two attribute only, then for specialization Age is chosen, because the specialization of Age gives the smallest number of branches.

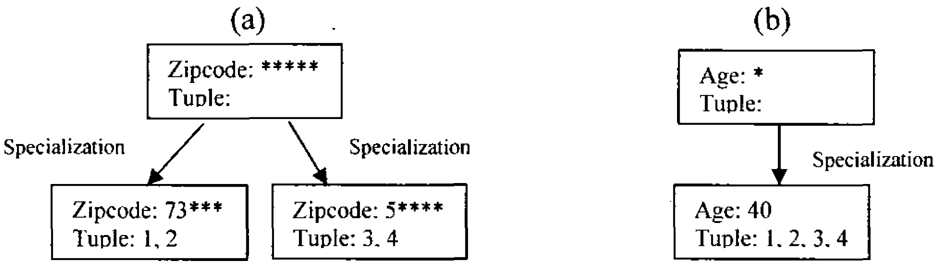


Figure4.4 (a, b): diagram for criterion of selecting the "Best" attribute

## 4.3Summary

In this chapter, the proposed technique called enhanced ($p$, $\alpha$)-sensitive $k$-anonymity has been discussed. The proposed algorithm extends Incognito algorithm [12]. The Incognito algorithm is a fully global recoding algorithm and may produce unnecessary distortions. A local-recoding algorithm has been proposed, called top-down local recoding algorithm.

# CHAPTER 5: RESULTS AND ANALYSIS

This chapter will highlight the dataset that is; Adult dataset which are used in experimentation, the experimental results will calculate the similarity attacks and will measure performance in term of distortion ratio and running time.

## 5.1Dataset

The proposed algorithm has been implemented on the standard database called adult dataset from UCI Machine Learning Warehouse [10] with 30169 records. The Adult dataset contains categorical as well as numerical attributes which is suitable for generalization required in the experiment. In 1994, The Adult dataset was taken out by Ronny Kohavi and Barry Becker from the database of census bureau. The Adult dataset is publicly available dataset, at the UCI Machine Learning Warehouse [10], which has been selected by [6, 12, 29] and become the standard of this field.

Below table 5.1 presents a short explanation of the adult dataset. The table shows name and type of each attributes, it also shows height of the generalization hierarchy and the number of distinct values for every attribute.

Table5.1: Brief Description of Adult Data Set [10]

| Name of attribute | Type of attribute | Distinct values | Height |
|---|---|---|---|
| Age | Numeric | 72 | 4 |
| Workclass | Categorical | 14 | 3 |
| Marital_Status | Categorical | 7 | 3 |
| Race | Categorical | 5 | 3 |
| Gender | Categorical | 2 | 2 |
| Education | Categorical | 16 | 4 |
| Country | Categorical | 41 | 3 |
| Health _condition | Sensitive | 8 | 1 |

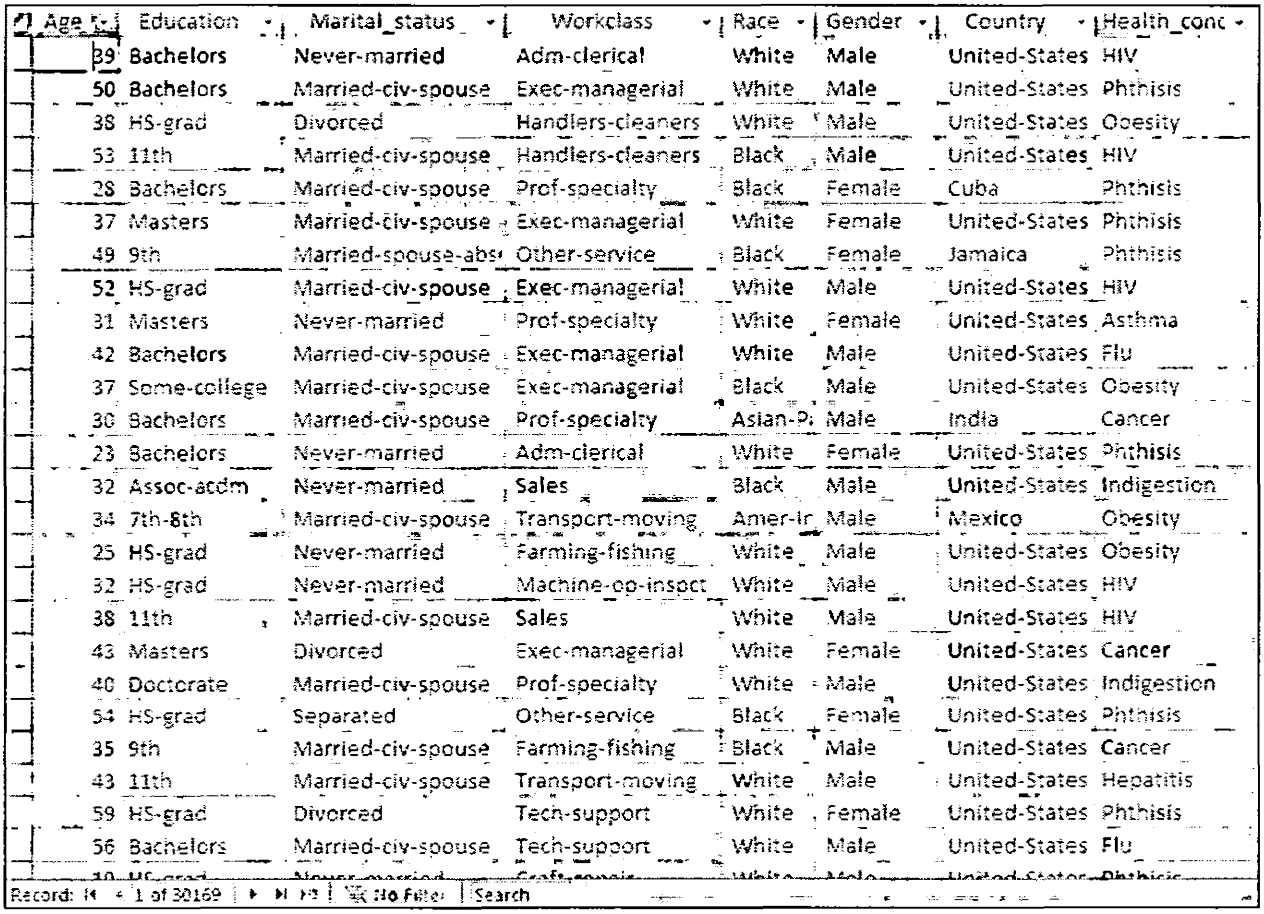
Furthermore, below table 5.2 shows the list of distinct attributes contains in Adult dataset.

Table5.2: List of distinct attribute used in Adult dataset [10]

| **Attribute name** | **Total** | **Distinct value** |
|---|---|---|
| Marital_status | 7 | Divorced, Never-married, Separated, Widowed, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, |
| Age | 72 | 117, 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, 82, 83, 84, 85, 86, 88, 90 |
| Race | 5 | Black, Other, White, Amer-Indian-Eskimo, Asian-Pac-Islander |
| Education | 16 | Preschool, Prof-school,1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Assoc-acdm, Assoc-voc, Some-college, HS-grad, Bachelors, |

| | | |
|---|---|---|
| | | Masters, Doctorate |
| Workclass | 14 | Armed-Forces, Craft-repair, Adm-clerical, Exec-managerial, Farming-fishing, Handlers-cleaners, Priv-house-serv, Tech-support, Transport-moving,Prof-specialty, Protective-serv, Sales, Machine-op-inspct, Other-service |
| Country | 41 | Outlying-US(Guam-USVI-etc), Peru, Scotland, South, Taiwan, Philippines, Poland, Portugal, Puerto-Rico, Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holland-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Thailand, Trinadad&Tobago, United-States, Vietnam, Yugoslavia |
| Gender | 2 | Male, Female |
| Health_condition | 8 | HIV, Obesity, Flu, Cancer, Phthisis, Indigestion, Hepatitis, Asthma |

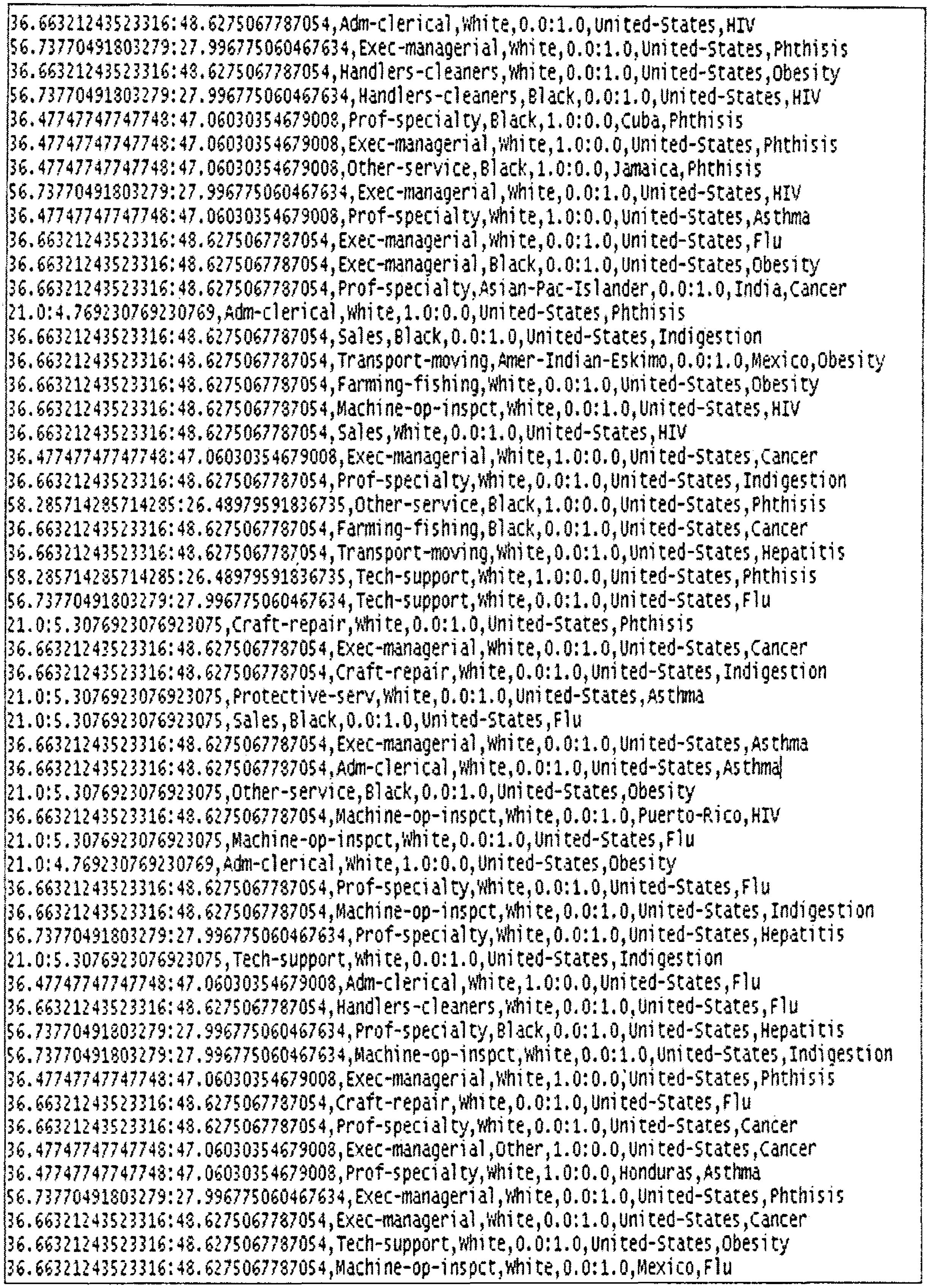
Below figure 5.1 shows a sample of the adult database that is used for conducting the experiments.



| 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 |
| 49 | 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 |
| [illegible] | HS-grad | Never-married | Craft-repair | White | Male | United-States | Phthisis |

Figure5.1: Adult Dataset from UCI Repository

Below figure 5.2 shows a sample of Anonymization through Incognito algorithm with k=3



```
36.66321243523316:48.6275067787054,Adm-clerical,White,0.0:1.0,United-States,HIV
56.73770491803279:27.996775060467634,Exec-managerial,White,0.0:1.0,United-States,Phthisis
36.66321243523316:48.6275067787054,Handlers-cleaners,White,0.0:1.0,United-States,Obesity
56.73770491803279:27.996775060467634,Handlers-cleaners,Black,0.0:1.0,United-States,HIV
36.47747747747748:47.06030354679008,Prof-specialty,Black,1.0:0.0,Cuba,Phthisis
36.47747747747748:47.06030354679008,Exec-managerial,White,1.0:0.0,United-States,Phthisis
36.47747747747748:47.06030354679008,Other-service,Black,1.0:0.0,Jamaica,Phthisis
56.73770491803279:27.996775060467634,Exec-managerial,White,0.0:1.0,United-States,HIV
36.47747747747748:47.06030354679008,Prof-specialty,White,1.0:0.0,United-States,Asthma
36.66321243523316:48.6275067787054,Exec-managerial,White,0.0:1.0,United-States,Flu
36.66321243523316:48.6275067787054,Exec-managerial,Black,0.0:1.0,United-States,Obesity
36.66321243523316:48.6275067787054,Prof-specialty,Asian-Pac-Islander,0.0:1.0,India,Cancer
21.0:4.769230769230769,Adm-clerical,White,1.0:0.0,United-States,Phthisis
36.66321243523316:48.6275067787054,Sales,Black,0.0:1.0,United-States,Indigestion
36.66321243523316:48.6275067787054,Transport-moving,Amer-Indian-Eskimo,0.0:1.0,Mexico,Obesity
36.66321243523316:48.6275067787054,Farming-fishing,White,0.0:1.0,United-States,Obesity
36.66321243523316:48.6275067787054,Machine-op-inspct,White,0.0:1.0,United-States,HIV
36.66321243523316:48.6275067787054,Sales,White,0.0:1.0,United-States,HIV
36.47747747747748:47.06030354679008,Exec-managerial,White,1.0:0.0,United-States,Cancer
36.66321243523316:48.6275067787054,Prof-specialty,White,0.0:1.0,United-States,Indigestion
58.285714285714285:26.48979591836735,Other-service,Black,1.0:0.0,United-States,Phthisis
36.66321243523316:48.6275067787054,Farming-fishing,Black,0.0:1.0,United-States,Cancer
36.66321243523316:48.6275067787054,Transport-moving,White,0.0:1.0,United-States,Hepatitis
58.285714285714285:26.48979591836735,Tech-support,White,1.0:0.0,United-States,Phthisis
56.73770491803279:27.996775060467634,Tech-support,White,0.0:1.0,United-States,Flu
21.0:5.3076923076923075,Craft-repair,White,0.0:1.0,United-States,Phthisis
36.66321243523316:48.6275067787054,Exec-managerial,White,0.0:1.0,United-States,Cancer
36.66321243523316:48.6275067787054,Craft-repair,White,0.0:1.0,United-States,Indigestion
21.0:5.3076923076923075,Protective-serv,White,0.0:1.0,United-States,Asthma
21.0:5.3076923076923075,Sales,Black,0.0:1.0,United-States,Flu
36.66321243523316:48.6275067787054,Exec-managerial,White,0.0:1.0,United-States,Asthma
36.66321243523316:48.6275067787054,Adm-clerical,White,0.0:1.0,United-States,Asthma
21.0:5.3076923076923075,Other-service,Black,0.0:1.0,United-States,Obesity
36.66321243523316:48.6275067787054,Machine-op-inspct,White,0.0:1.0,Puerto-Rico,HIV
21.0:5.3076923076923075,Machine-op-inspct,White,0.0:1.0,United-States,Flu
21.0:4.769230769230769,Adm-clerical,White,1.0:0.0,United-States,Obesity
36.66321243523316:48.6275067787054,Prof-specialty,White,0.0:1.0,United-States,Flu
36.66321243523316:48.6275067787054,Machine-op-inspct,White,0.0:1.0,United-States,Indigestion
56.73770491803279:27.996775060467634,Prof-specialty,White,0.0:1.0,United-States,Hepatitis
21.0:5.3076923076923075,Tech-support,White,0.0:1.0,United-States,Indigestion
36.47747747747748:47.06030354679008,Adm-clerical,White,1.0:0.0,United-States,Flu
36.66321243523316:48.6275067787054,Handlers-cleaners,White,0.0:1.0,United-States,Flu
56.73770491803279:27.996775060467634,Prof-specialty,Black,0.0:1.0,United-States,Hepatitis
56.73770491803279:27.996775060467634,Machine-op-inspct,White,0.0:1.0,United-States,Indigestion
36.47747747747748:47.06030354679008,Exec-managerial,White,1.0:0.0;United-States,Phthisis
36.66321243523316:48.6275067787054,Craft-repair,White,0.0:1.0,United-States,Flu
36.66321243523316:48.6275067787054,Prof-specialty,White,0.0:1.0,United-States,Cancer
36.47747747747748:47.06030354679008,Exec-managerial,Other,1.0:0.0,United-States,Cancer
36.47747747747748:47.06030354679008,Prof-specialty,White,1.0:0.0,Honduras,Asthma
56.73770491803279:27.996775060467634,Exec-managerial,White,0.0:1.0,United-States,Phthisis
36.66321243523316:48.6275067787054,Exec-managerial,White,0.0:1.0,United-States,Cancer
36.66321243523316:48.6275067787054,Tech-support,White,0.0:1.0,United-States,Obesity
36.66321243523316:48.6275067787054,Machine-op-inspct,White,0.0:1.0,Mexico,Flu
```

Figure5.2: Anonymization through Incognito algorithm with k=3

For experiment Intel Core2 Duo CPU with 2 GM RAM and 1.8 GHz Processor has been used and the algorithm is implemented in C/C++. Similar configuration is used to l-diversity [6] and incognito [12]. The tuples containing unknown values are eliminated and the final dataset has 31069 tuples. In the dataset, seven of the attributes for quasi-identifier were selected. The attribute of sensitive values containing {Cancer, Flu, Indigestion, HIV, Phthisis, Obesity, Hepatitis, Asthma} called "Health _condition" has been added to the dataset. In the dataset, sensitive values are given randomly to every.record in the following manner. To each sensitive attribute, assign a number initially. i.e., {1: Cancer, 2: HIV, 3: Hepatitis, 4: Phthisis, 5: Flu, 6: Indigestion, 7: Asthma, 8: Obesity}. Then a random number for each record is created from1 to 8, and equivalent sensitive value has been given to every tuple according to number. For example, if the first number in the dataset is 2, then the tuple contain sensitive value "HIV", if the second record is 7, then this tuple contain sensitive value "Asthma".

## 5.2 Performance Measure

The proposed algorithm depicts the performance measure in term of similarity attack, distortion ratio and running time.

- **Similarity attack**

  An equivalence class, all the sensitive values are falling in one category, similar or distinct but similar meaning. The quasi-identifier group is exposed to the similarity attack and the attacker can easily get the important information and sensitive values are supposed to be disclosed and such situation is called similarity attack.

- **Distortion ratio**

  Distortion ratio is used to calculate how much data in the resultant table differs from the original table after generalization, that is, how much information is lost?

- **Running time**

  Running time is used to calculate the efficiency of the algorithm. That is, how much time is taken by this algorithm to perform the desire task?

### Scenario 1: Comparison based on Similarity Attack

For similarity attack, last attribute Health_condition in table 5.1 is used as sensitive attribute and the first seven attributes is used as the quasi-identifier. Based on confidentiality of the values according to table 4.2, the eight values of the attribute Health_condition are divided into four categories.

P-sensitive k-anonymity algorithm [8] has been used to generate p-sensitive k-anonymous (that is, 2-sensitive 4-anonymous) tables. There are 21 minimal tables generated and similarity attack is seen in 13 tables (13/21 = 62%). Total of 916 tuples in one table can be deducted their sensitive value.

(p, α)-sensitive k-anonymity algorithm [9] has been used and apply p = 2, k = 4, α = 2. It generate total 30 tables, and similarity attack is experienced in 7 of them (7/30 = 23%).

Then enhanced (p, α)-sensitive k-anonymity algorithm has been used and applied p = 2, k = 4, α = 2. It generate 28 minimal tables and experience that 3 of them are exposed to the similarity attack (3/28 = 11%). Below table 5.3 shows the comparison based on similarity attack.

Table 5.3: Comparison based on similarity attack

| **Algorithm** | **Level of Anonymization** | **Total tables generated** | **Suffer from similarity attack** | **Ratio** |
|---|---|---|---|---|
| p-sensitive k-anonymity | k=4, p= 2 | 2 1 tables | 13 tables | 13/21=62% |
| (p, α)-sensitive k-anonymity | k=4, p= 2, α = 2 | 30 tables | 7 tables | 7/30 = 23% |
| Enhanced (p, α)-sensitive k-anonymity | k=4, p= 2, α = 2 | 28 tables | 3 tables | 3/8= 11% |

It is clear from above observations that enhanced (p, α)-sensitive k-anonymity model considerably decreases the possibility of similarity attacks.

## Scenario 2: Distortion Ratio

Distortion measures are used to calculate how much data in the resultant table differs from the original table after generalization, that is, how much information is lost?

In the derived dataset, the rate of recoding is known by the distortion ratio. In terms of height the distortion of the generalize value is defined. There will be no distortion, if the attribute of tuple has not been generalized and its height is equal to 0; however there is distortion, if the attribute of a tuple is generalized to a further general value. In the taxonomy, if the value has been generalized one level up, its height is equal to 1. For the attribute $x_i$ of the tuple $t_j$, suppose $h_{i,j}$ be the height of the generalized value. In the generalized dataset, the sum of the distortions of all values is equal to distortion of the whole dataset. Such that, distortion $D = \sum_{i,j} h_{i,j}$.

Distortion ratio can be calculated by

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

Where, fully generalized dataset means that, in the taxonomy tree all values of the attributes are generalized to the root.

The distortion ratio depend on the size of quasi-identifier, ratio will greater when the quasi-identifier has more attributes, because there is more possibility of the generalization of tuples. In below figure 5.3, the ratio of distortion decreases when the value for α increases. Obviously, if the value of α is greater there is minimum requirement of calculating α, so in the resulting dataset generalization of the values is needed less operations. Thus the ratio of distortion for enhanced (p, α)-sensitive k-anonymity is smaller than that of p-sensitive k-anonymity and (p, α)-sensitive k-anonymity model.

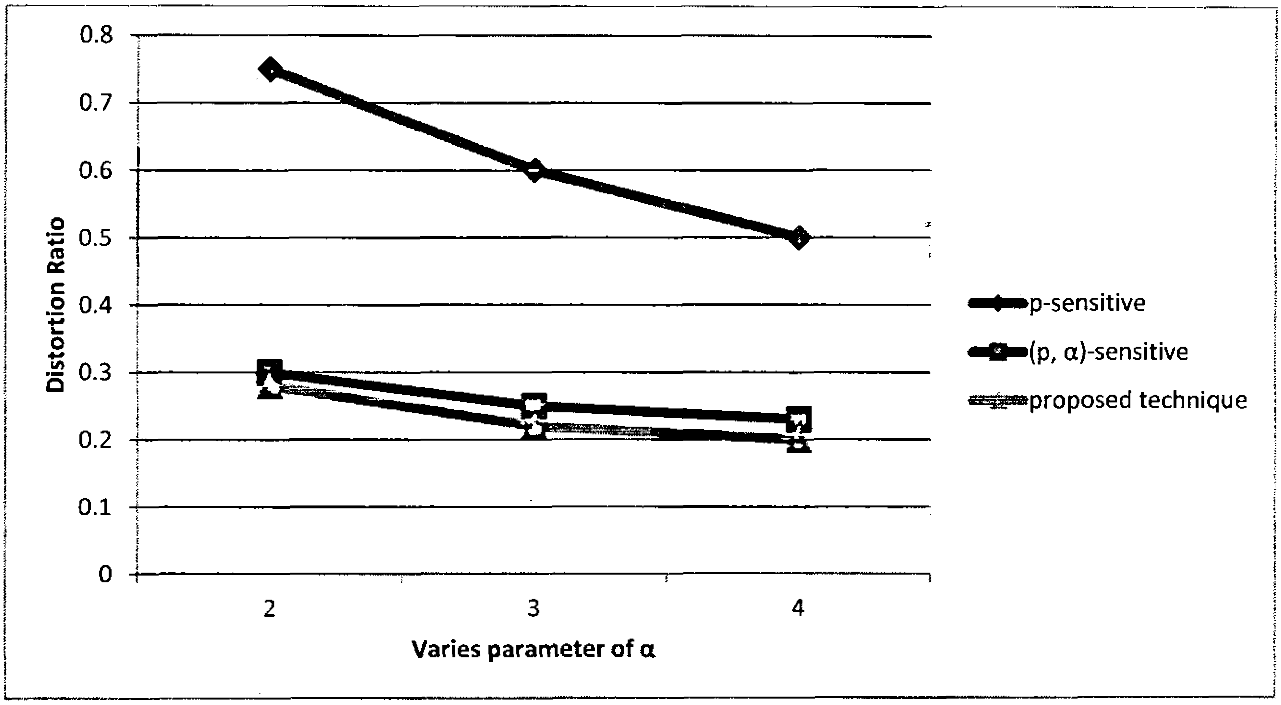


Figure5.3: Comparison of Distortion ratio of the proposed algorithm with variant parameter of p and α with p=2, k=3

### Scenario 3: Running time

The efficiency in term of running time of the proposed algorithm has been compared with previous technique, that is, with p -sensitive k-anonymity and (p, α)-sensitive k-anonymity. Figure 5.4 shows the running times of

1) p -sensitive k-anonymity model
2) (p, α)-sensitive k-anonymity model and
3) Proposed technique

The execution time of above three properties are shown with k=4, p=4, α=2, and varies size s of quasi-identifier, where the size s of quasi-identifier is from 2 to 7. From below figure it is clear that proposed technique, that is, Enhanced (p, α)-sensitive k-anonymity model is runs slower than both the previous models, due to finding the suitable sensitive value for each category according to calculating weight of α

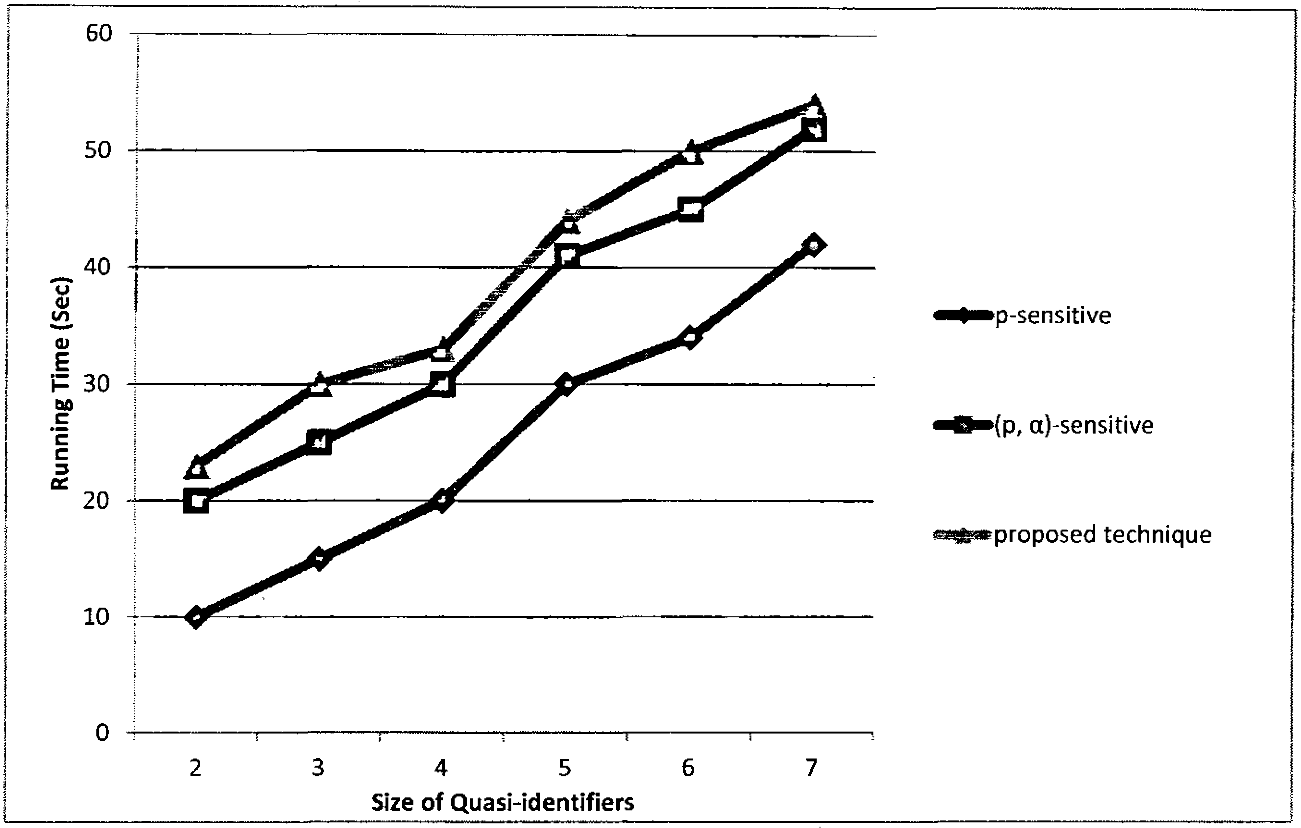


Figure5.4: Comparison of running time of the proposed algorithm with variant QI size with p=4, k=4, α=2

## 5.3 Summary

This chapter discussed the dataset called Adult dataset which are used in experimentation and then depicts the experimental result based on well known performance measures which include similarity attack, distortion ratio and running time. Simulation result shows that enhanced (p, α)-sensitive k-anonymity gives superior results in term of similarity attack and distortion ratio; where as its running time is slightly higher than the existing approaches.

# CHAPTER 6: CONCLUSIONS & FUTURE WORK

This chapter depicts the future work and conclusion. Conclusion shows that what is the purpose of this research and future work highlight the points on the basis of which further work may be done to improve the current technique.

## 6.1 Conclusions

*K*-anonymity is a model which protects the individual's privacy. In k-anonymity the data is shown in such a manner that there are at least *k* identical kinds of tuples in the microdata for every single tuple. But it is not sufficient to protect revelation of attribute due to two type of attack occur in *k*-anonymity; one is called homogeneity attack and other is called background knowledge attack.

Several models were proposed to solve the complication of k-anonymity. But these enhanced properties have some restrictions which still allow the information to be disclosed.

(p, α)-sensitive k-anonymity which is advancement of k-anonymity is a narrative property that satisfies the privacy of the respondents and the data of whose is being used for research or some other purpose, but (p, α)-sensitive k-anonymity model is still not sufficient for the protection of sensitive attributes. To enhance privacy and overcome the deficiency of (p, α)-sensitive k-anonymity, another technique has been proposed called enhanced (p, α)-sensitive k-anonymity model. This technique says that at least its total weight α and p different sensitive attribute categories for every group of quasi-identifier. The proposed technique uses a local recoding based algorithm called top-down algorithm. The concept of top-down local recoding algorithm is that in initial step all tuples are generalized into one quasi-identifier group completely. Then, in every iteration tuples are specialized and enhanced (*p*, *α*)-sensitive *k*-anonymity has been maintain during specialization. The proposed algorithm has been implemented on well known data set called Adult Dataset [10].

This algorithm measures similarity attack, distortion ratio and running time. On the basis of conducted experiment, it is concluded that compared with earlier models, that is, p-sensitive k-anonymity and (p, α)-sensitive k-anonymity the proposed technique reduces rátio of distortion and similarity attack. The proposed algorithm only reduces but not fully eliminates the similarity attack.

## 6.2 Future work

To enhance the privacy and reduce the similarity attack, Enhanced (p, α)-sensitive k-anonymity has been used in this research; this method can further be improved with collaboration of other principle of privacy like t-closeness etc. This method can also be used with advanced technique of data mining, to protect the sensitive attribute and respondent identities. Another technique called slicing may also use instead of generalization and suppression for further research.

# 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. Machanvajjhala 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/~mlearn/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

31. Y. Rubner et al, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, 2000