Indian Institute of Information Technology, Allahabad

PROJECT REPORT

DATA PRIVACY

Project Supervisor - Dr. K.P. Singh

Declaration by the Candidates

We, hereby declare that the project titled *Data Privacy* is a record of bonafide project work carried out by us under the guidance of *Dr. K.P. Singh* in partial fulfillment of the 6th semester Mini-Project work for the B.Tech (IT) Course in Indian Institute of Information Technology, Allahabad.

Due acknowledgments have been made in the text to all the references and frameworks used.

Nishit Gupta – IIT2014502
Sahil Prakash – IIT2014504
Sacheendra Mohan Singh – IIT2014506

Certificate

This is to certify that the project report entitled *Data Privacy* submitted to Board No : 4, Department of Information Technology, Indian Institute of Information Technology, Allahabad in partial fulfillment of the 6th semester Mini-roject work, is a record of bonafide work carried out by:

1. Nishit Gupta – IIT2014502
2. Sahil Prakash – IIT2014504
3. Sacheendra Mohan Singh – IIT2014506

under my supervision and guidance.
This report has not been submitted anywhere else for any other purpose.

Submission Date : 09/03/2017

Dr. K.P. Singh
Assistant Professor
Department of Information Technology
IIIT Allahabad
Email: kpsingh@iiita.ac.in
Contact: +915322922226(O)

Contents

Abstract

Data Privacy, also known as Information Privacy, is the aspect of Information Technology, that deals with the ability of an organization to share the information or data in the computer system with the third parties.

This project is divided into two phases.
The first phase discusses the definition of data protection, the need for data protection, how does the data protection work and the three important data protection or data anonymization techniques which are k-anonymity, l-diversity and t-closeness. The k-anonymity privacy requirement for publishing microdata requires that each equivalence class in the data-set contains at least k records. L-diversity requires that each equivalence class in the data-set must have at least t distinct values for the sensitive attribute. The concept of t-closeness requires that the distribution of sensitive attribute in the anonymized data is close to its distribution in the whole data-set by a metric known as Earth Mover Distance.

The second phase of the project discusses whether Machine Learning could be incorporated on the anonymized data to extract the sensitive feature which in turn can reveal the identity of a record in the dataset. Also, the second phase will let us know the possible attacks a third party can try to extract the sensitive feature. This would help us to evaluate the usefulness and accuracy of the current prevailing data anonymization techniques and would help us judge whether the current prevailing data-anonymization methods are worth or not. If they fail, then what could be done to prevent the data from revealing the identity of a person.

Introduction

Data Protection : Individuals as consumers, citizens, patients, employees etc. have the right to privacy and protect themselves and their information from abuse. Data protection is about safeguarding the fundamental right to privacy and preventing this sensitive information from being revealed to the outer world.
The personal information is collected, processed and stored by “automated” means which is intended to be a part of the filing system.
Privacy concerns exist wherever personally identifiable information is collected, stored, used and destroyed knowingly or unknowingly.
The sources of data privacy issues can be:

1. Healthcare records
2. Social Media records
3. Financial records and transactions
4. Citizen records of a country
5. Criminal records

6. Location based service and geolocation
7. Taxi service records
8. Biological records etc.

Need for Data Protection : Data protection is important to safeguard the personal information stored in a database. The data available is prone to revealing of identities of the individuals which can pose serious threats to a person's private life.

Hence, there is a need to protect the data from attackers and information gatherers who could misuse the data for their own personal gains thereby causing a threat to other person's life.

Therefore, the database must be converted to a form which could be released to seekers thereby not allowing them to reveal the identity of a person.

There are a number of ways which could help in this : data generalization, suppression, methods like k-anonymity, l-diversity, t-closeness etc.

Problem Statement and Objective

Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful.

Also, can we train the system on the basis of test-data by using some Machine Learning approach to reveal the identity of a person from the converted data (i.e data in anonymised form).

Literature Survey

S No.	Author	Paper Title	Year	Crux	Venue
1.	Latanya Sweeney	*K-Anonymity: A Model for protecting privacy.*[1]	2002	*This Paper introduces a formal protection model called k-anonymity and a set of accompanying policies for deployment . A released table provides k-anonymity protection if information for each person contained cannot be distinguished from at least k-1 individuals whose information is also in the dataset.*	*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems October 5,2002*
2.	A.Machanavajjhala J. Gehrke D.Kifer M.Venkitasubramaniam	*l-Diversity: Privacy Beyond K-Anonymity.*[2]	2006	*In this paper it is shown that k-anonymity model does not guarantee privacy against attackers using two simple attacks, first, lack of diversity in sensitive attributes, second, using background knowledge . It proposes a powerful privacy definition called l - diversity which requires that each equivalent class has at least l well represented values for sensitive attribute.*	Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on 3-7 April 2006
3.	Ninghui Li Tiancheng Li Suresh Venkatasubramanian	*T-closeness: Privacy Beyond K-Anonymity and l-Diversity.*[3]	2007	*This Paper shows limitations of l-diversity and then proposes another privacy notion called t-closeness , which requires that the distribution of sensitive attribute in any equivalence class is close to the distribution of the attribute in overall table. EMD is proposed as a distance measure, to evaluate the closeness.*	Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 15-20 April 2007

Dataset Description

The work carried out in this project is performed on the Adult Data Set downloaded from the UCI Repository.

Adult Data Set[4]:

The data set was extracted by Barry Becker from the 1994 Census database.
It has the following set of attributes :

1. Age : It is a continuous range of values.

2. Sex : Male/Female.

3. Race : White/Black

4. Education : Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

5. Marital Status : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

6. Native Country : United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

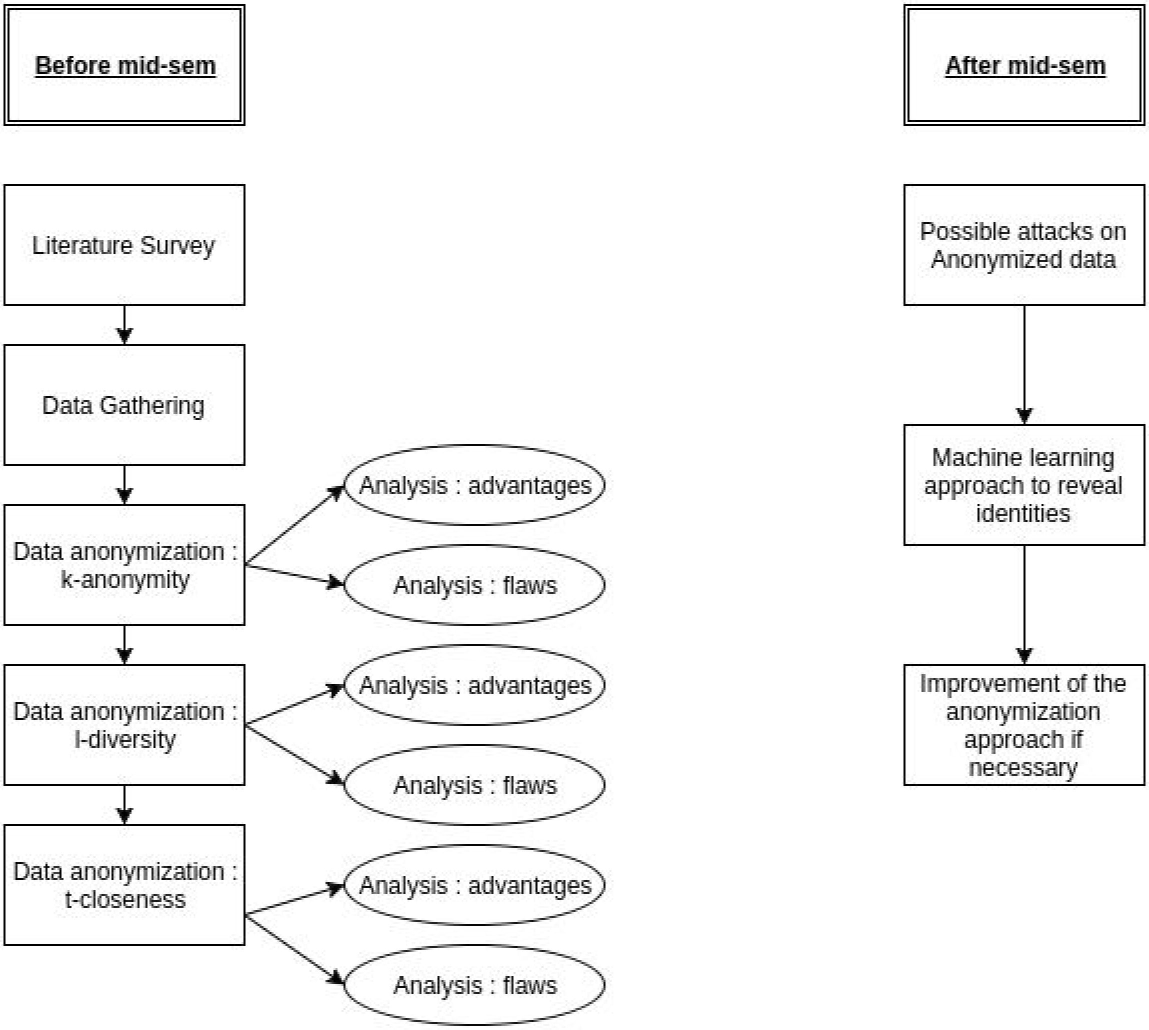
7. Work Class : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

8. Occupation : Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

9. Salary Class : This is divided into two categories. Salary <= 50K and > 50K

For our work, we downloaded the Adult Data Set consisting of over 30,000 records and we selected 5,000 records out of those 30,000 records, randomly

Proposed Approach


Before mid-sem
Literature Survey
Data Gathering
Data anonymization : k-anonymity
Analysis : advantages
Analysis : flaws
Data anonymization : l-diversity
Analysis : advantages
Analysis : flaws
Data anonymization : t-closeness
Analysis : advantages
Analysis : flaws
After mid-sem
Possible attacks on Anonymized data
Machine learning approach to reveal identities
Improvement of the anonymization approach if necessary

Some Terminologies:

1. **Quasi-Identifier** : Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier.

2. **Suppression** : In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. This means that all the information that allows the inference of sensitive information is simply not released.

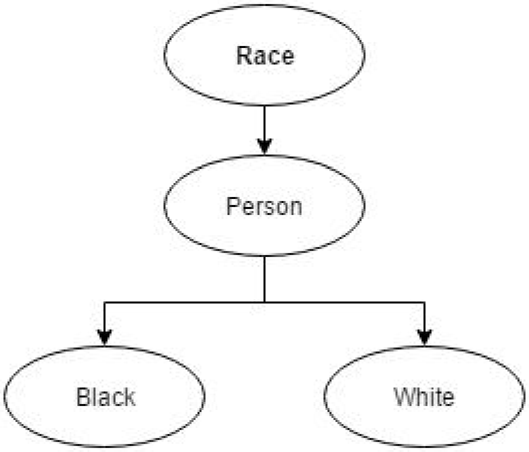
3. **Generalization** : In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' etc.

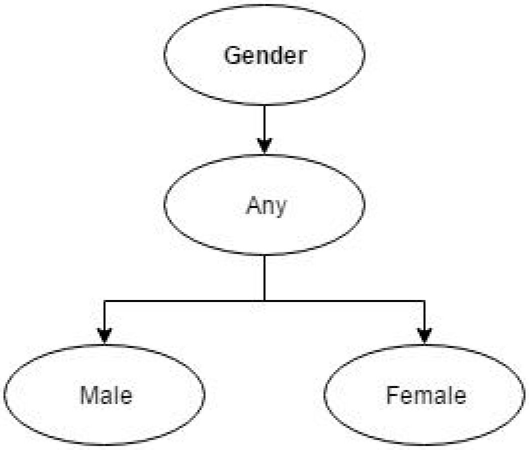
Anonymization Methods

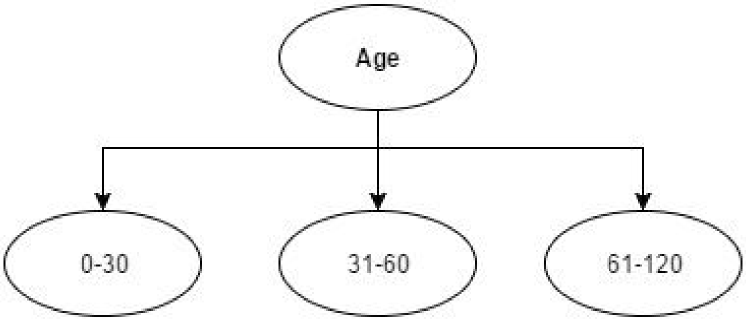
1. K-anonymity

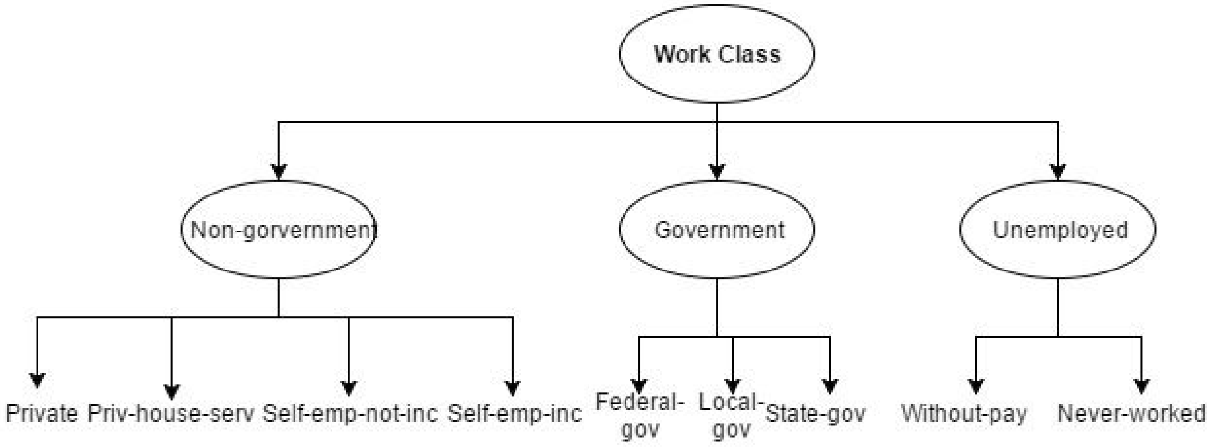
Let T(A_1,A_2,.........,A_n) be a table and QI_T be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if each sequence of values in T[QI_T] appears with **at least** k-occurrences in T[QI_T].

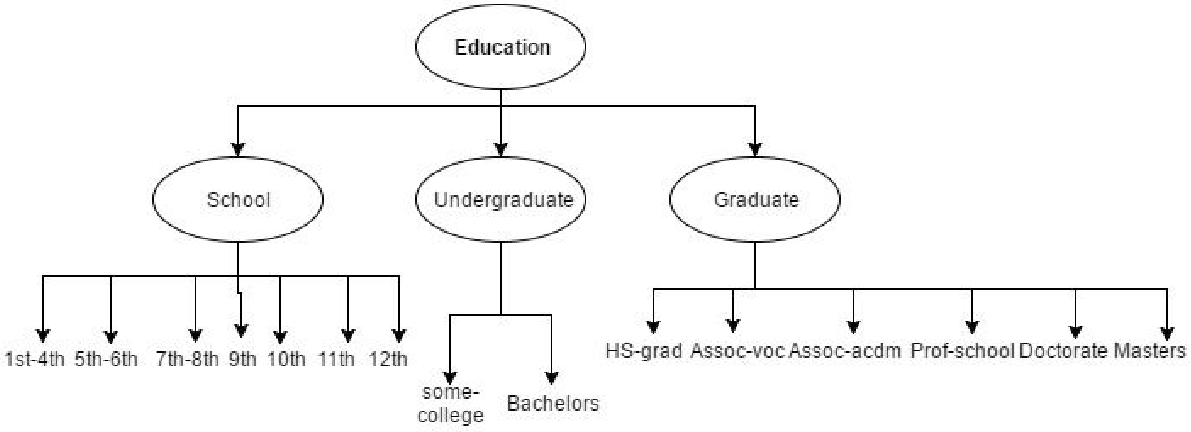
For the data-set, generalization and suppression is done as follows to achieve **2-anonymity:**

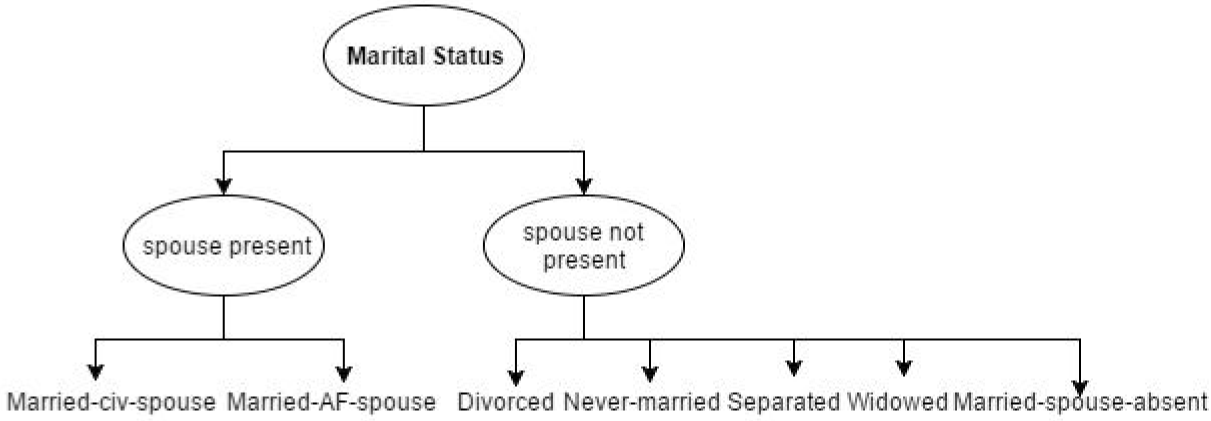


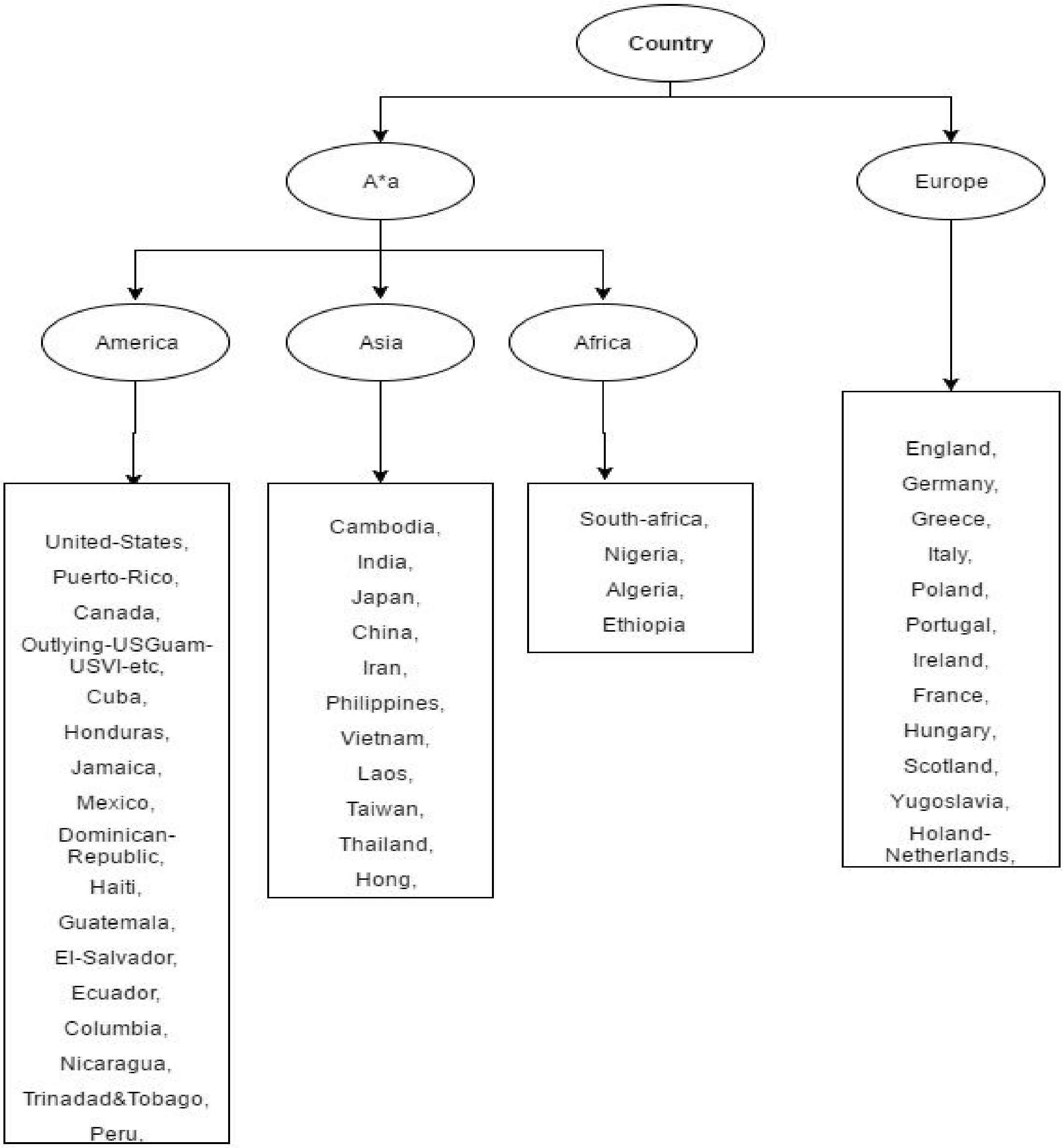





Work Class
Non-gorvernment
Government
Unemployed
Private
Priv-house-serv
Self-emp-not-inc
Self-emp-inc
Federal-gov
Local-gov
State-gov
Without-pay
Never-worked


Education
School
Undergraduate
Graduate
1st-4th
5th-6th
7th-8th
9th
10th
11th
12th
some-college
Bachelors
HS-grad
Assoc-voc
Assoc-acdm
Prof-school
Doctorate
Masters


Marital Status
spouse present
spouse not present
Married-civ-spouse
Married-AF-spouse
Divorced
Never-married
Separated
Widowed
Married-spouse-absent



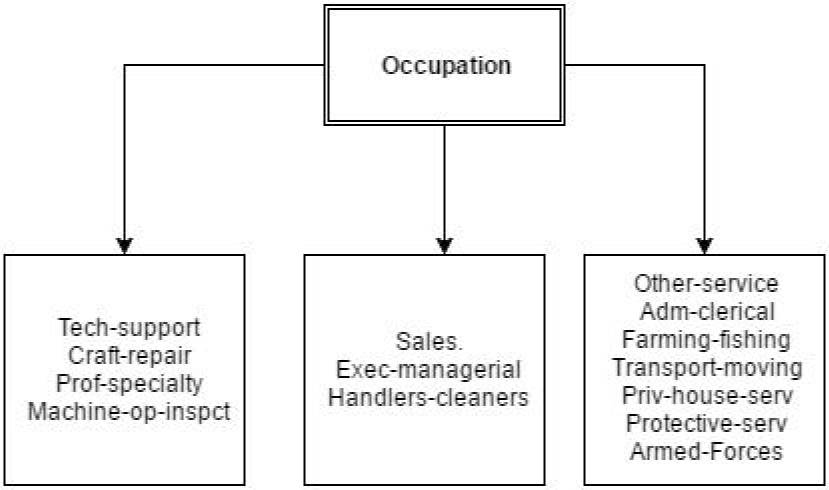
2. L-diversity

An equivalence class is l-diverse if it contains at least ‘l’ well represented values for sensitive attribute S. A table is l-diverse if every equivalence class is l-diverse.

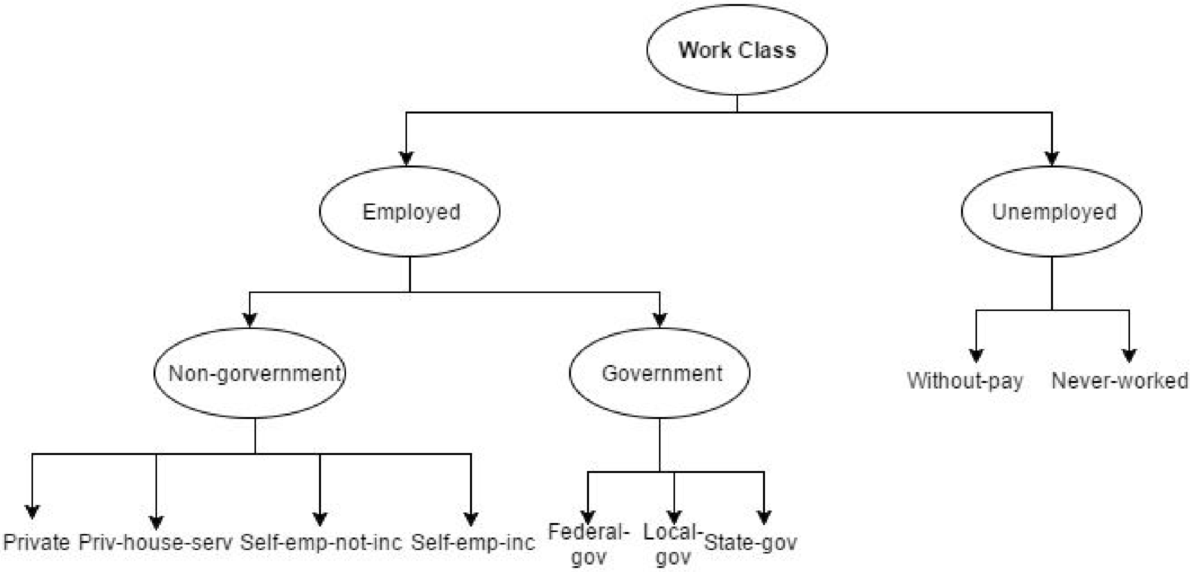
Distinct l- Diversity

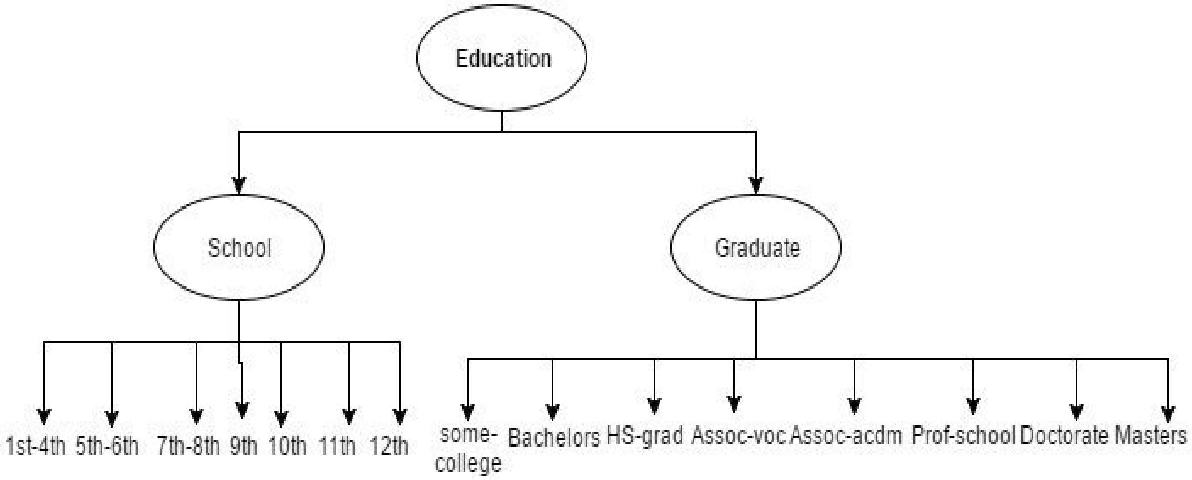
An equivalence class has distinct l-diversity if it has at least ‘l’ well - defined sensitive values. When each equivalence class in the table has distinct l-diversity, the table is said to be having distinct l-diversity.

For our data-set, we first divided the tuples into three categories on basis of occupation as defined below, so that each divided group has almost equal number of records:

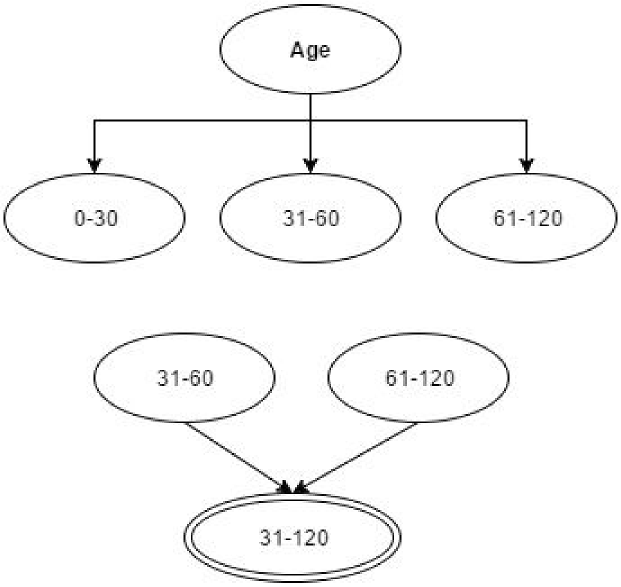


To make the data-set **3-diverse** the following changes to generalizations were added to the generalizations used in k-anonymity, as defined previously:





For some of the records, age generalization was done as follows:



3. T-closeness

An equivalence class has t-closeness, if the distance between distribution of sensitive attribute in that class and the distribution of that attribute in the overall table is not more than a threshold ‘t’. It minimizes the amount of knowledge the observer gains by looking at the released table as compared to the knowledge he/she gains by looking at the complete table.

Metrics for calculating the distance: **Earth Mover Distance (EMD)**

EMD : minimal amount of work needed to transform one distribution to another. Let one distribution be assumed as mass of earth lying in space. The second distribution is assumed as holes in space. We need to fill masses in holes. This is EMD.

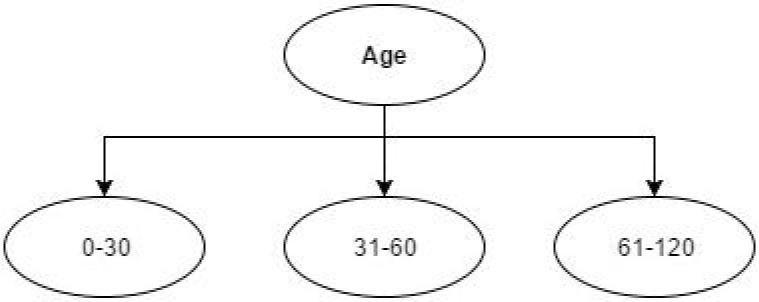
EMD between distributions $P = (p_1,p_2,p_3,...)$ and $Q = (q_1,q_2,q_3,...)$

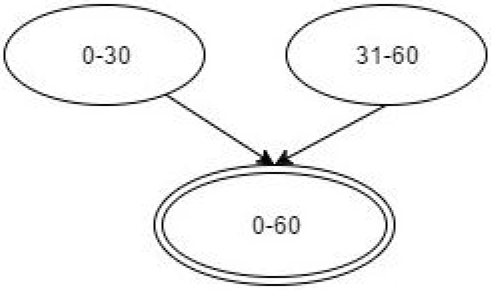
$f_{i,j}$ = flow of mass from element i of P to element j of Q
$d_{i,j}$ = ground distance between element i of P and element j of Q
$f_{i,j} >= 0$
$p_i - \sum f_{i,j} + \sum f_{j,i} = q_i$
$\sum \sum f_{i,j} = \sum p_i = \sum q_i = 1$

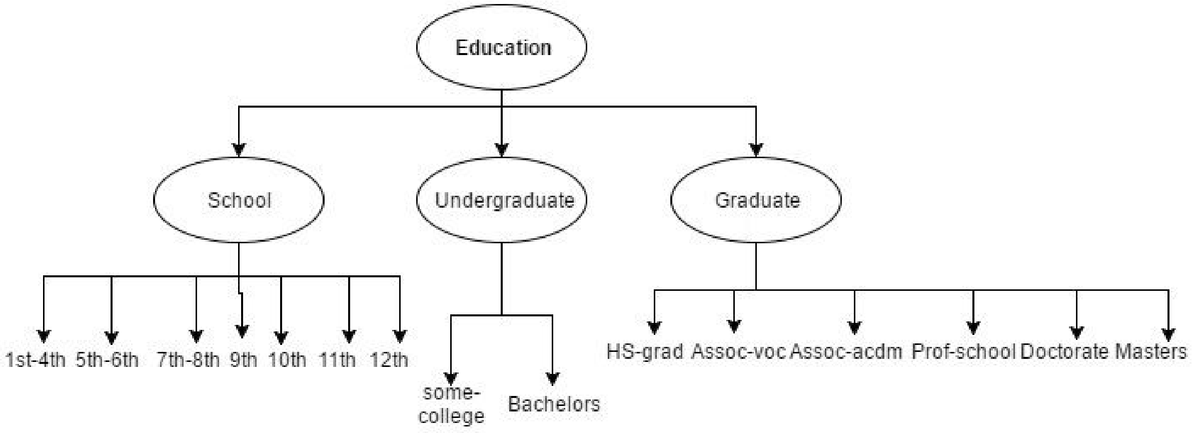
EMD : $D[P,Q] = \sum \sum d_{i,j} f_{i,j}$

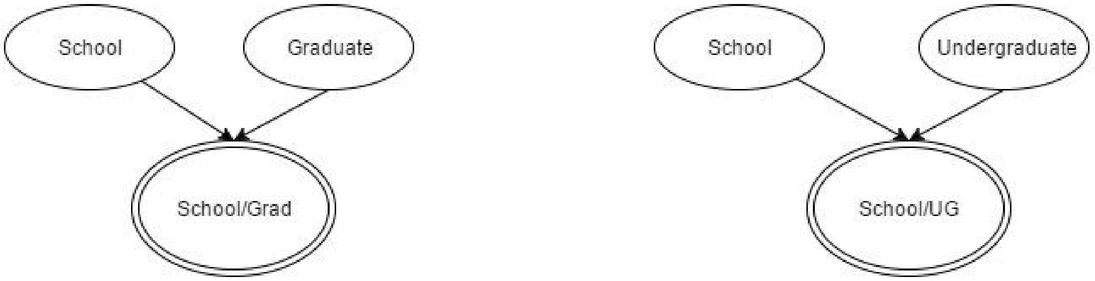
We chose value of t to be 0.2 to implement t-closeness on the data-set.
Hence, data was anonymized in such a way that EMD between the generalized data of an equivalence class and original data-set was always less than or equal to 0.2

Generalizations were done as follows for some records in addition to generalization done for k-anonymity:









The Need for Different Techniques

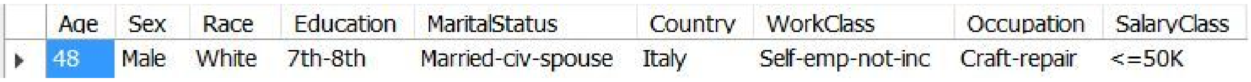
Flaw of K-Anonymity

Though K-anonymity generalizes the data to certain extent, however there is still a possibility of knowing the sensitive feature.

Query

SELECT * FROM dataset WHERE Age=48 AND Sex='Male' AND RACE='White' AND Education='7th-8th' AND MaritalStatus='Married-civ-spouse' AND Country='Italy' AND WorkClass='Self-emp-not-inc' AND SalaryClass='<=50K';

The Result of Data-set is:

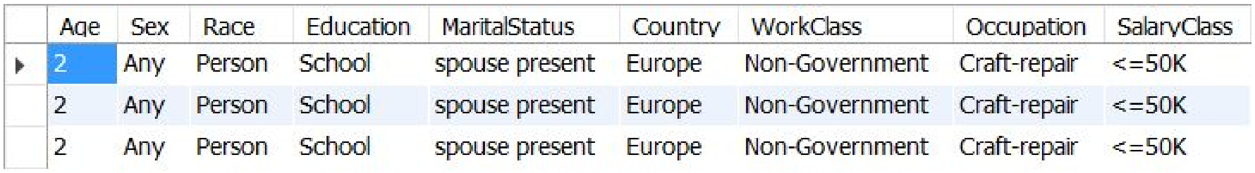


Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
48	Male	White	7th-8th	Married-civ-spouse	Italy	Self-emp-not-inc	Craft-repair	<=50K

Modified Query for K-anonymity:

SELECT * FROM k_anonymous WHERE Age=2 AND Sex="Any" AND Race="Person" AND Education = "School" AND MaritalStatus = "spouse present" AND Country = "Europe" AND WorkClass = "Non-Government" AND SalaryClass = "<=50K";

The Result of K-anonymity is :



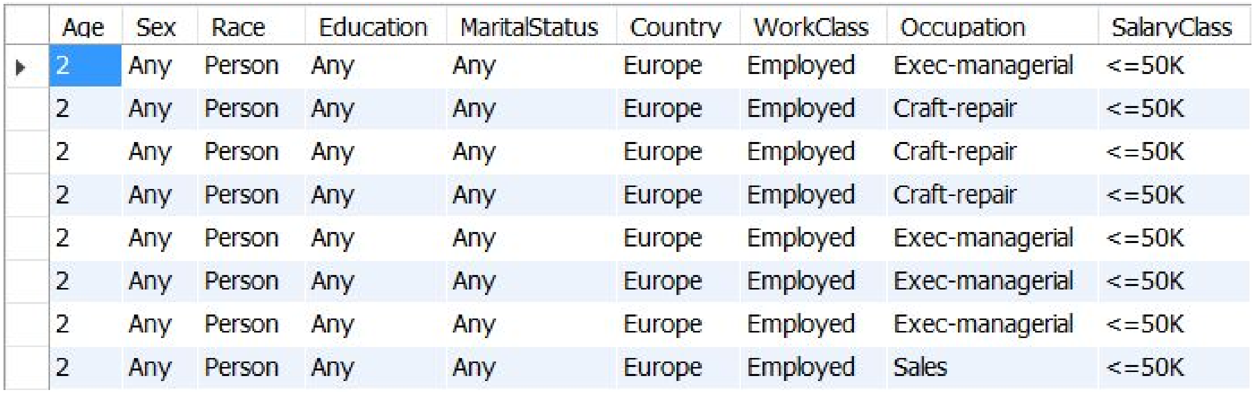
	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K
	2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K
	2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K

From the example, shown above we can see that although we are getting more than one record in the output, however we still have only one Occupation class. We are able to identify the sensitive attribute for the queried tuple. This is what we call, Attribute disclosure.
Hence, there is a need to generalize in such a way that occupation should have more than one class in the output so that it becomes difficult for the attacker to gain some useful attribute information.

Advantage of using L-diversity
So, L-diversity comes into picture.

Modified Query for L-diversity:
SELECT * FROM l_diverse WHERE Age=2 AND Sex="Any" AND Race="Person" AND (Education = "School" OR Education = "Any" OR Education = "***") AND (MaritalStatus = "spouse present" OR MaritalStatus = "Any") AND Country = "Europe" AND WorkClass = "Employed" AND SalaryClass = "<=50K";

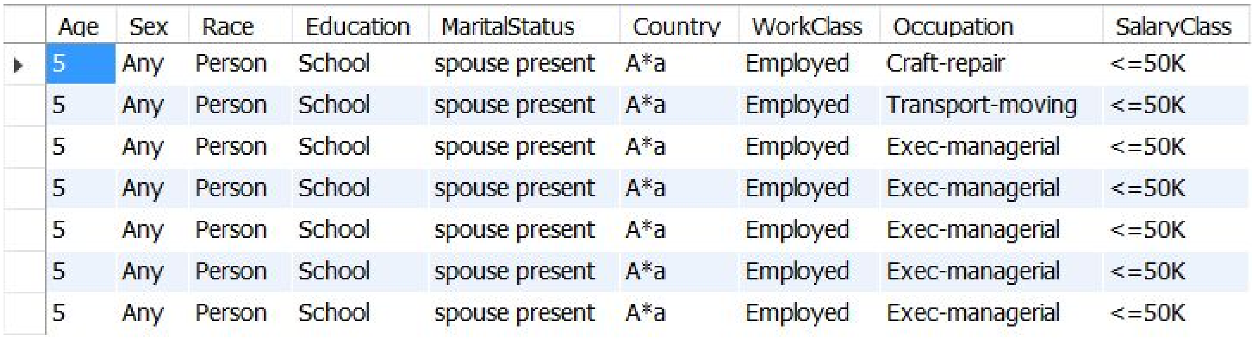
The Result of L-diversity is :



	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
	2	Any	Person	Any	Any	Europe	Employed	Sales	<=50K

Here, we see that we are now getting 8 tuples as output and there are 3 distinct values for Occupation Class as the data is 3-diverse. So now attacker can not identify the occupation easily. Attribute disclosure is prevented in this approach.

Flaws of L-diversity
1. The data quality is reduced as we have seen in the example shown earlier.
2. For the following equivalence class :

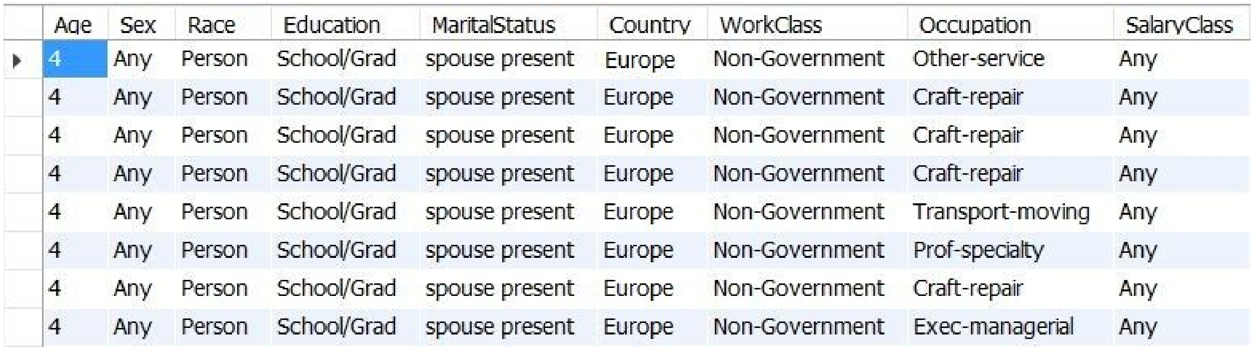


Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
5	Any	Person	School	spouse present	A*a	Employed	Craft-repair	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Transport-moving	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K

We have 5 records out of 7 which have Occupation as Exec-managerial (71.4 %).
In the original dataset, there were 663 out of 5000 records with Occupation as Exec-managerial (13.26 %). (Skewness attack)

Advantages of T-closeness

We have discussed that t-closeness brings the distribution of the sensitive attribute in an equivalence class much close to that in the original data so skewness problem of l-diversity is eliminated using T-closeness.



Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Other-service	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Transport-moving	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Prof-specialty	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Exec-managerial	Any

Software & Hardware Requirements

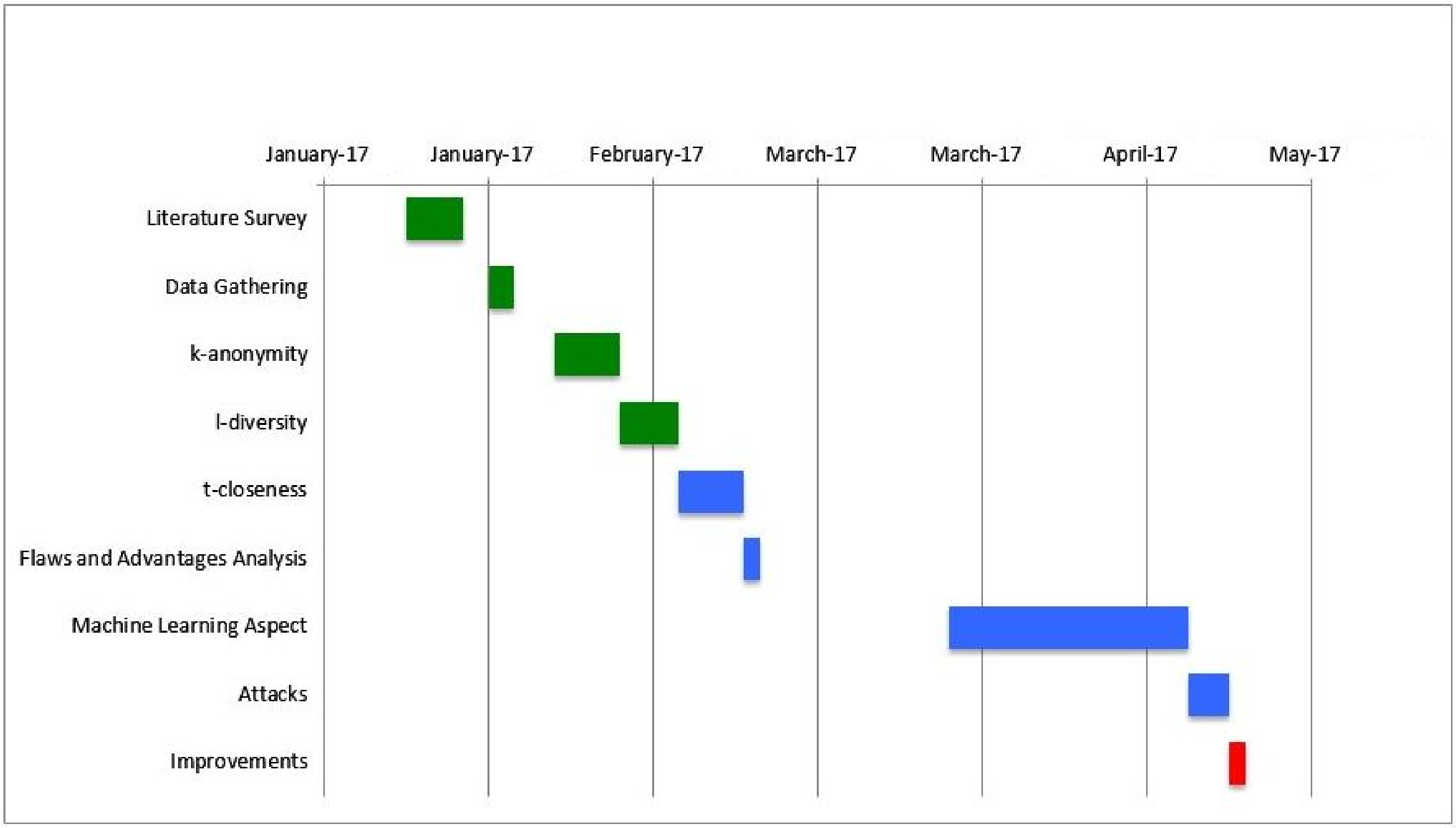
Software Requirements:

1. MATLAB R2012A or above
2. G++ compiler for C++ 11 or above
3. Microsoft Office (MS Excel)
4. MySQL
5. Windows/Linux
6. Suitable Text Editors like Sublime, Notepad++ etc.

Hardware Requirements:
Testing Platform : Laptop HP ENVY-J106TX, running under Linux (Ubuntu 16.04LTS) and Windows 10

Hardware/Software Specifications:
Architecture : 64 bit system
Operating System : Windows 10
Microprocessor : 2.8 GHz Intel Core i5-6610U
Memory : 12 GB 1600 MHz DDR4L SDRAM (1 x 4 GB + 1 x 8 GB)

Activity Chart



Work Completed Pre-Mid Semester

Pre-Mid Semester, the following work have been completed :

1. Literature survey completed successfully
2. Adult dataset successfully gathered from UIC datastore.
3. Anonymization using k-anonymity achieved on the dataset and advantages and flaws were noted.
4. Anonymization using l-diversity achieved and advantages and flaws were noted.
5. Data was made anonymous as per the t-closeness principle.

Work to be Completed Post-Mid Semester

Post Mid-semester, Machine Learning aspect has to be implemented to learn about the dataset using samples (or clustering) and hence to provide the output class label ‘Occupation’ on the basis of Quasi-Identifier as input.

Also, we have to see if the current anonymization methodologies are worth or not i.e if the data can still be attacked to reveal the personal information.

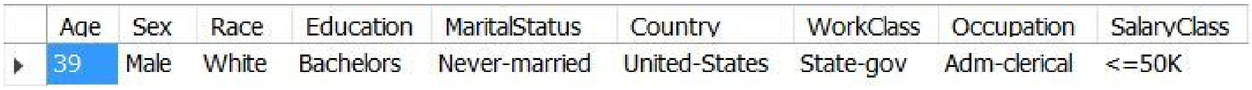
If Machine Learning aspect does not give satisfactory results in favour of data privacy, then how could we improve those previously discussed methods.

Results

Query

SELECT * FROM dataset WHERE Age=39 AND Sex='Male' AND RACE='White' AND Education='Bachelors' AND MaritalStatus='Never-married' AND Country='United-States' AND WorkClass='State-gov' AND SalaryClass='<=50K';

Output



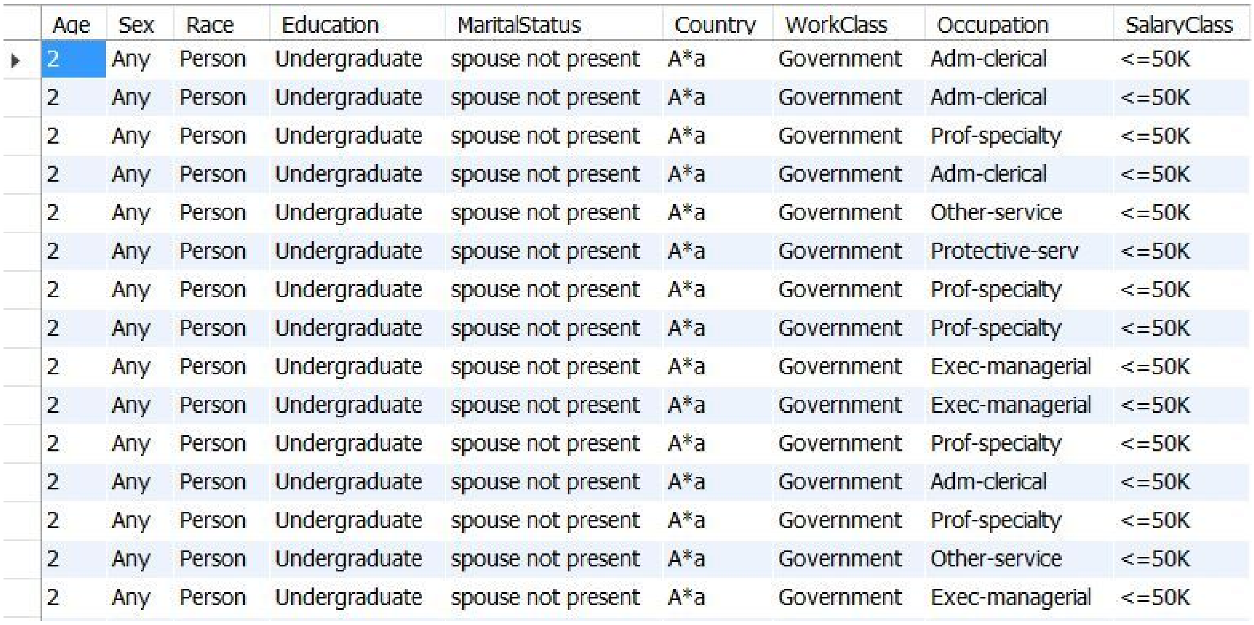
	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	39	Male	White	Bachelors	Never-married	United-States	State-gov	Adm-clerical	<=50K

K-anonymity:

Modified Query

SELECT * FROM k_anonymous WHERE Sex="Any" AND Race="Person" AND Age = 2 AND Education = "Undergraduate" AND MaritalStatus = "spouse not present" AND Country = "A*a" AND WorkClass = "Government" AND SalaryClass = "<=50K" ;

New Output



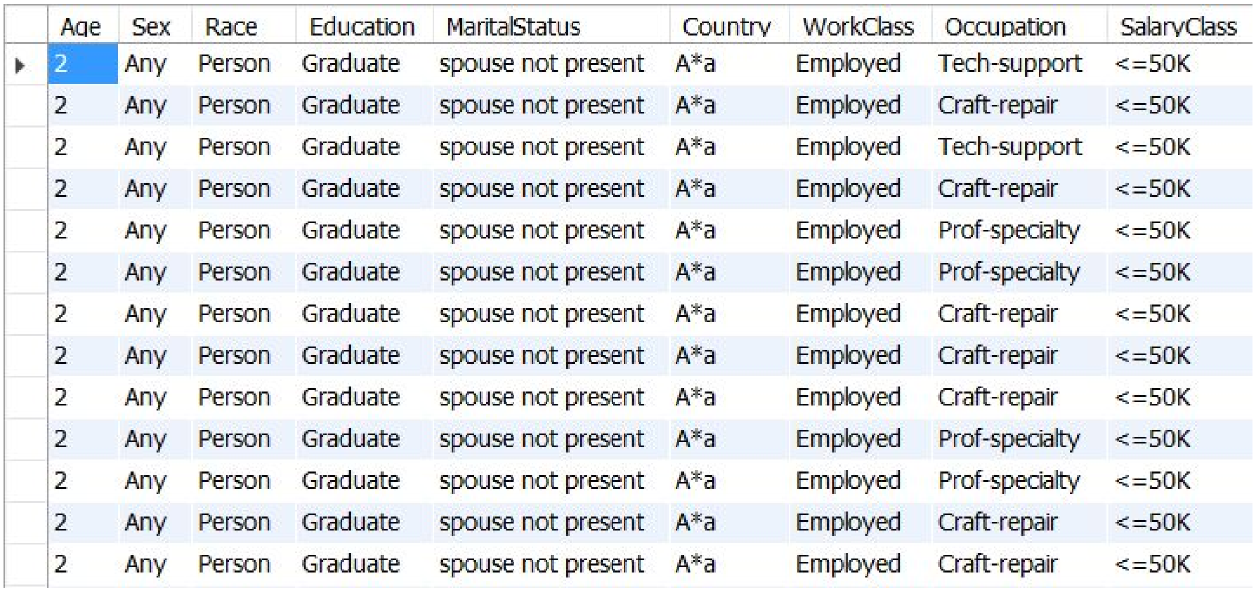
	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Other-service	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Protective-serv	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Exec-managerial	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Exec-managerial	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Other-service	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Exec-managerial	<=50K

L-Diversity:

Modified Query

SELECT * FROM l_diverse WHERE Sex="Any" AND Race="Person" AND (Age = 2 or Age = 4) AND (Education = "Graduate" or Education = "***" or Education = "Any") AND (MaritalStatus = "spouse not present" or MaritalStatus = "Any") AND Country = "A*a" AND WorkClass = "Employed" AND SalaryClass = "<=50K" ;

New Output



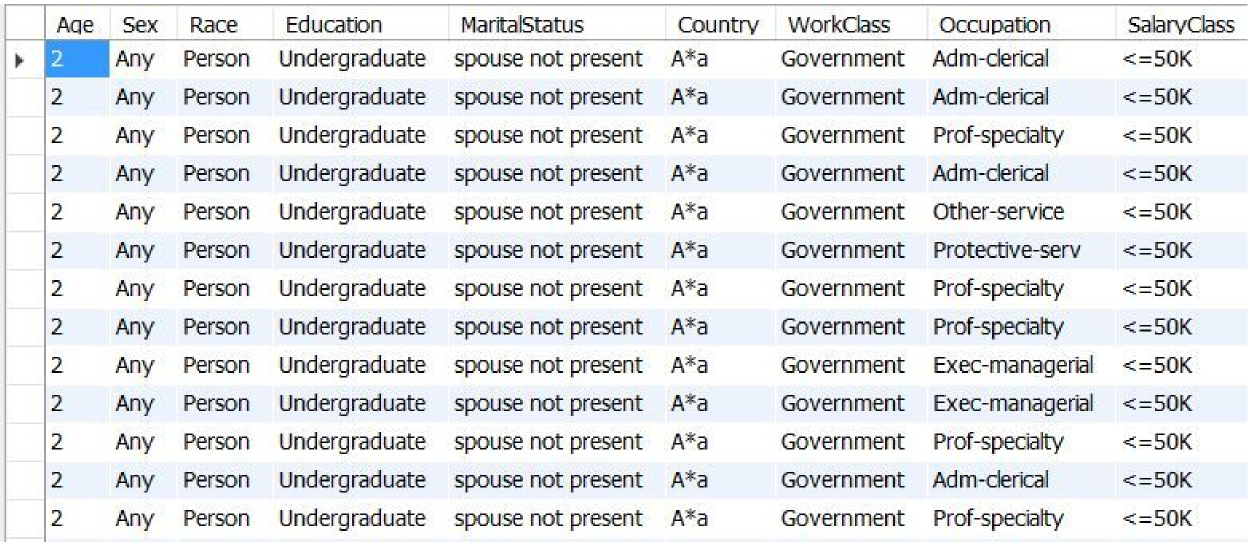
	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	2	Any	Person	Graduate	spouse not present	A*a	Employed	Tech-support	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Tech-support	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Prof-specialty	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Prof-specialty	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Prof-specialty	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Prof-specialty	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K
	2	Any	Person	Graduate	spouse not present	A*a	Employed	Craft-repair	<=50K

T-closeness:

Modified Query

SELECT * FROM t_close WHERE Sex="Any" AND Race="Person" AND (Age = 2 or Age = 4) AND (Education = "Undergraduate" or Education = "School/UG") AND (MaritalStatus = "spouse not present" or MaritalStatus = "Any") AND Country = "A*a" AND WorkClass = "Government" AND SalaryClass = "<=50K" ;

New Output



	Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
▸	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Other-service	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Protective-serv	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Exec-managerial	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Exec-managerial	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Adm-clerical	<=50K
	2	Any	Person	Undergraduate	spouse not present	A*a	Government	Prof-specialty	<=50K

<u>Data Quality measure :</u>

We observe that the data released after applying l-diversity on our data-set has worse data quality as compared to the table released by k-anonymity. Also, the release after applying t-closeness has improved data quality from l-diversity, although it is worse than k-anonymity. Hence, there is a motivation to use t-closeness along with k-anonymity. The cost of reduced data quality is due to the improved protection of the sensitive attribute.

<u>Conclusion</u>

We observe that k-anonymity ensures protection against identity disclose, but it is not sufficient to provide protection against attribute disclosure. We found some equivalent classes where each of the record had same value of the sensitive attribute ‘Occupation’, leading to attribute disclosure.

L-diversity solves this problem by making it necessary to include at least l distinct values of the sensitive attribute. But, it had the limitation that the distribution of sensitive attribute in the equivalent classes was such that it could reveal some general information about that particular class.

T-closeness solves this problem by making sure that the distribution of sensitive attribute in the equivalent class is close to the distribution of the attribute in the overall table.

Future Scope

After successful completion of the project, it can be used to generate a release of data which can be published to third parties. This release of data will be ensuring that data is protected against attacks and privacy remains intact.

References

[1]. L. Sweeney, “K-Anonymity: A model for protecting privacy”, in International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002

[2]. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer and Muthuramakrishnan Venkitasubramaniam, “L-Diversity : Privacy beyond K-Anonymity”, in Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on 3-7 April 2006

[3]. Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, “T-Closeness: Privacy Beyond k-Anonymity and l-Diversity”, in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 15-20 April 2007

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

Suggestions by Board-IV Members