Summary

Data-Set : Adult Data-Set, UCI Machine Learning Repository. The data set was extracted by Barry Becker from the 1994 Census database. No of Records - 5000

It has the following set of Attributes:

Age, Gender, Race, Education , Marital Status, Native Country, Work Class, Occupation, Salary Class

1. Age : It is a continuous range of values. 67 Values

2. Sex : Male/Female. 2 Values

3. Race : White/Black. 2 Values

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. 16 Values

5. Marital Status : Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. 7 Values

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. 39 Values

7. Work Class : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. 6 Values

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. 14 Values

9. Salary Class : This is divided into two categories. Salary <= 50K and > 50K. 2 Values

Sensitive Attribute= {Occupation}

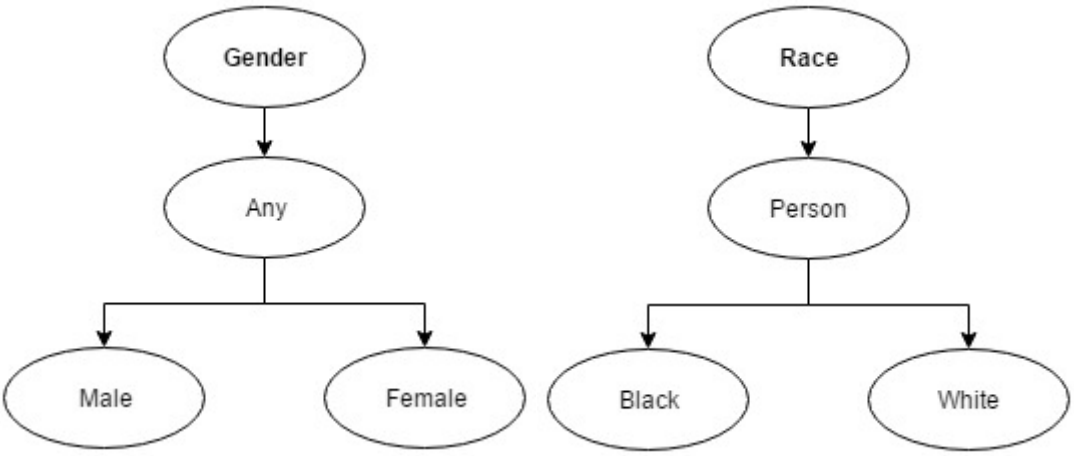
Quasi-Identifier = {Age, Gender, Race, Education, Marital Status, Native Country, Work Class, Salary Class}

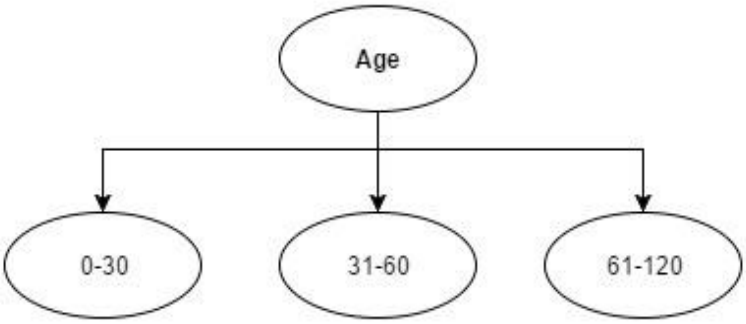
Data Privacy Techniques Used :

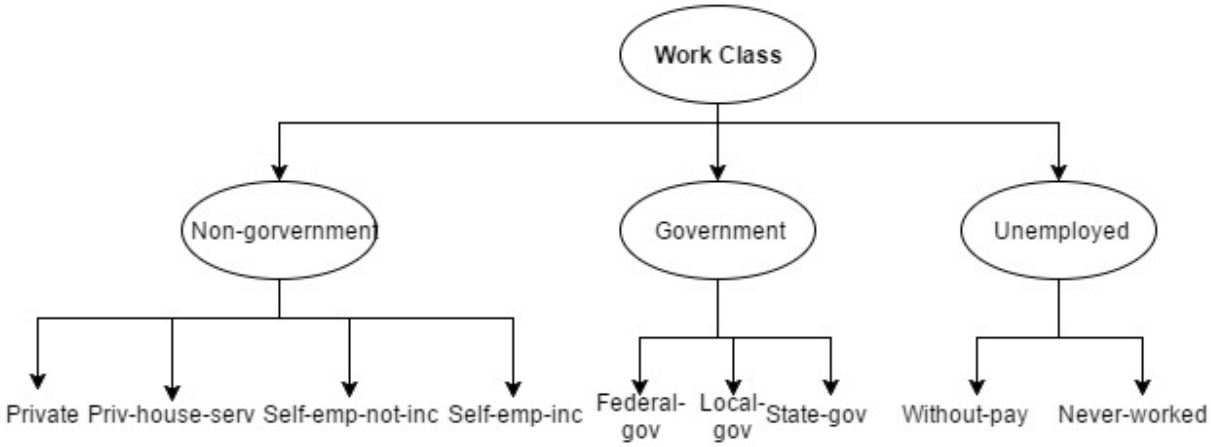
1. K-anonymity (K=2)

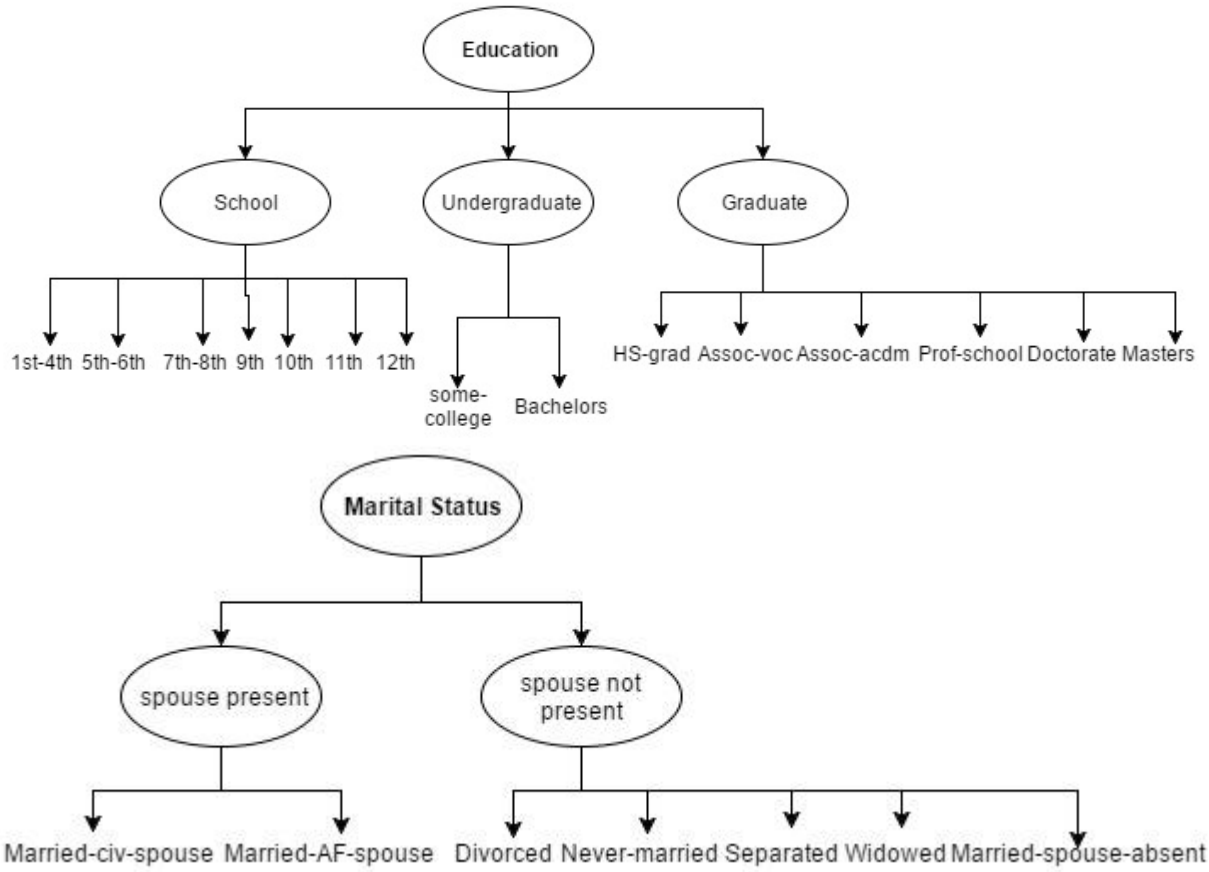
Let $T(A_1,A_2,.....,A_n)$ be a table and QI_T be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if each sequence of values in $T[QI_T]$ appears with at least k-occurrences in $T[QI_T]$.

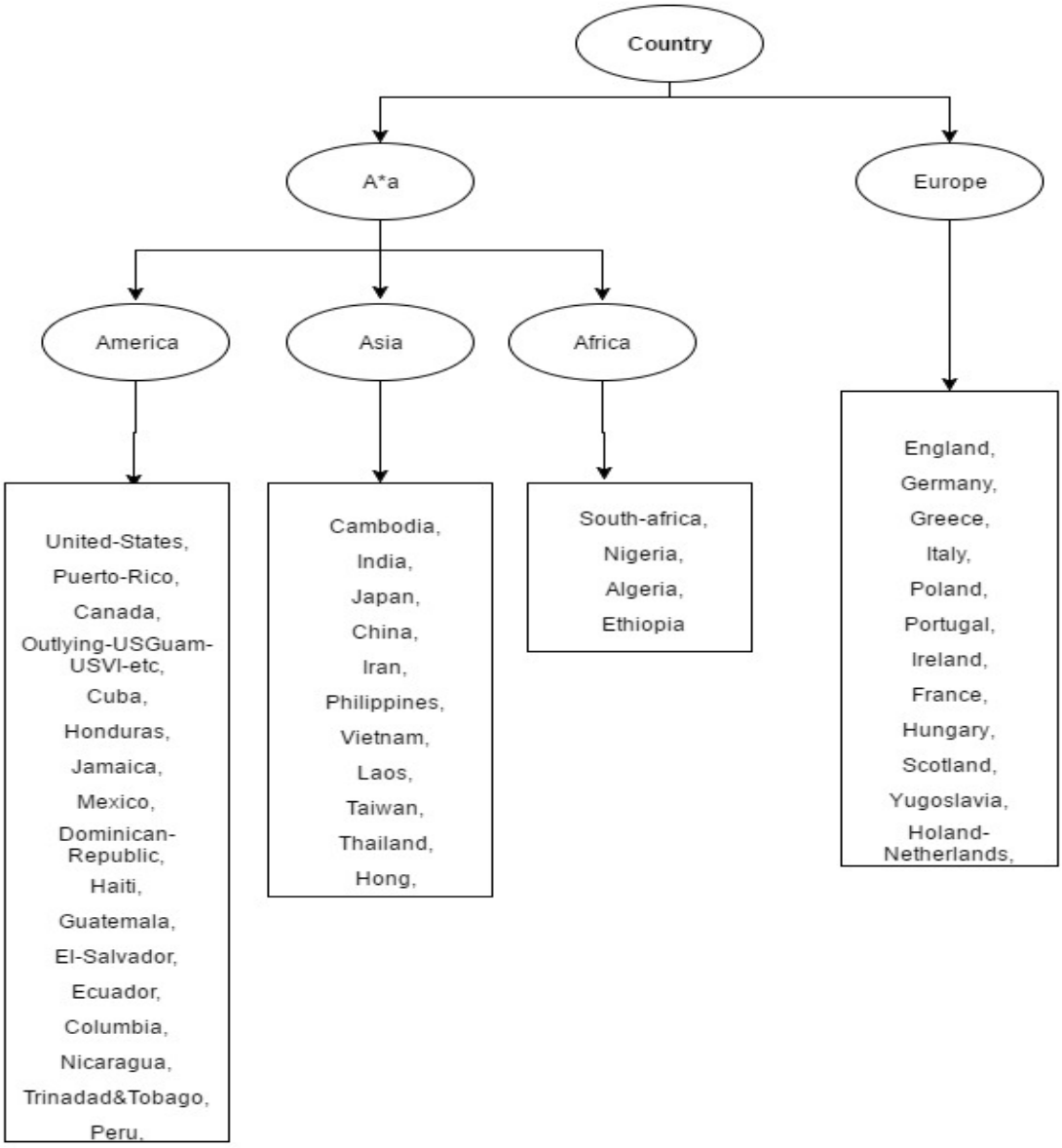
For the data-set, generalization and suppression is done as follows to achieve 2-anonymity:


Gender
Any
Male
Female
Race
Person
Black
White


Age
0-30
31-60
61-120


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



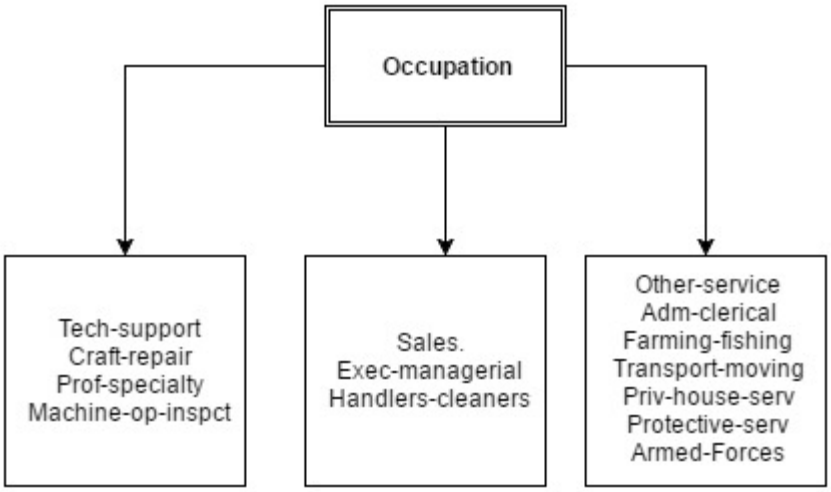
2. L-diversity (L=3)

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.

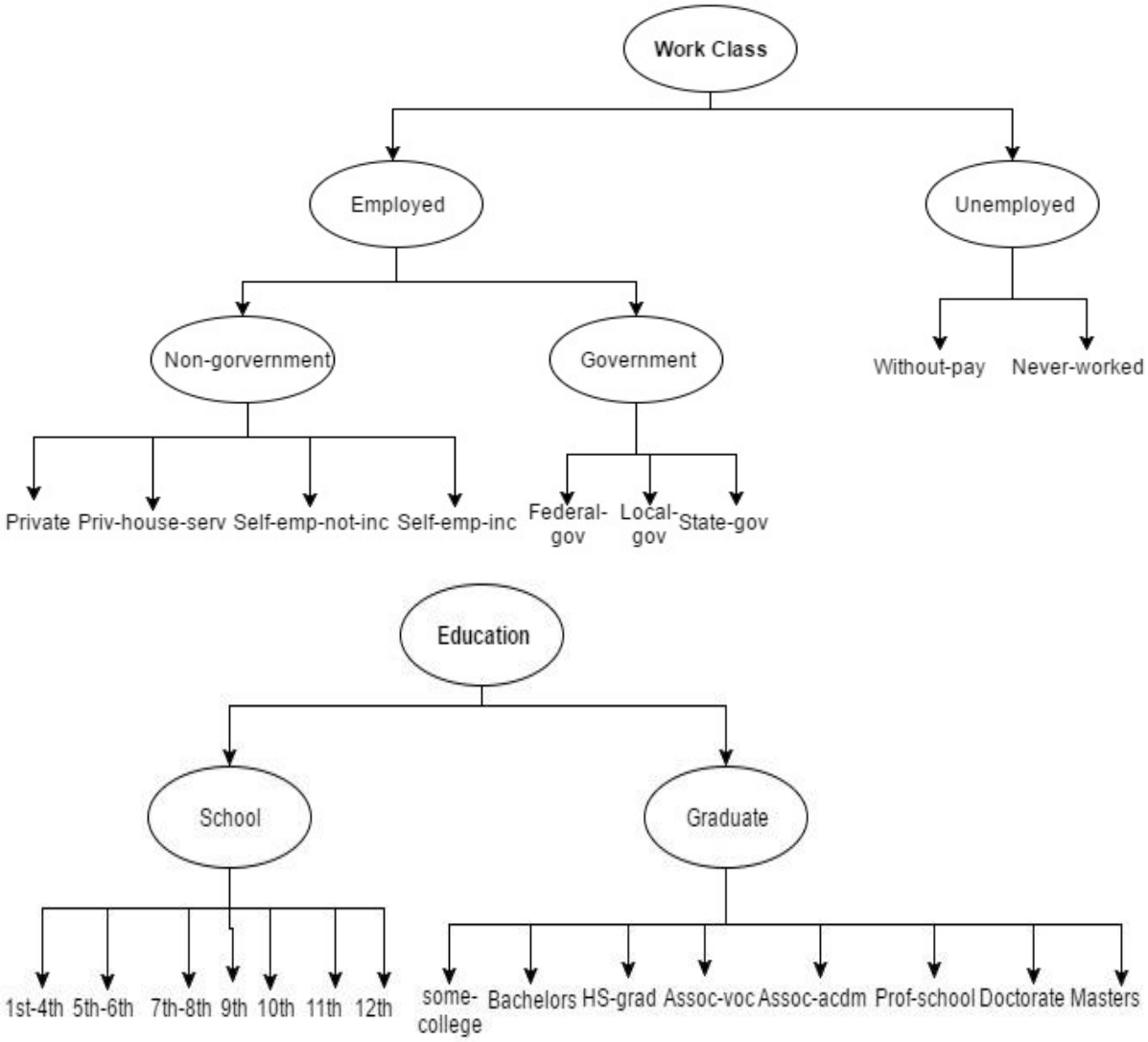
Distinct l- Diversity

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.

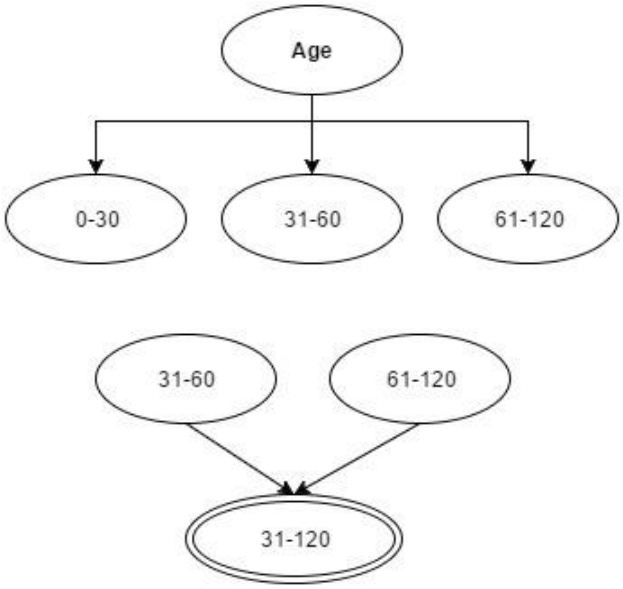
For our data-set, we first divided the tuples into three categories on basis of occupation as defined below, so that each divided group has almost equal number of records:



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 (T=0.2)

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.

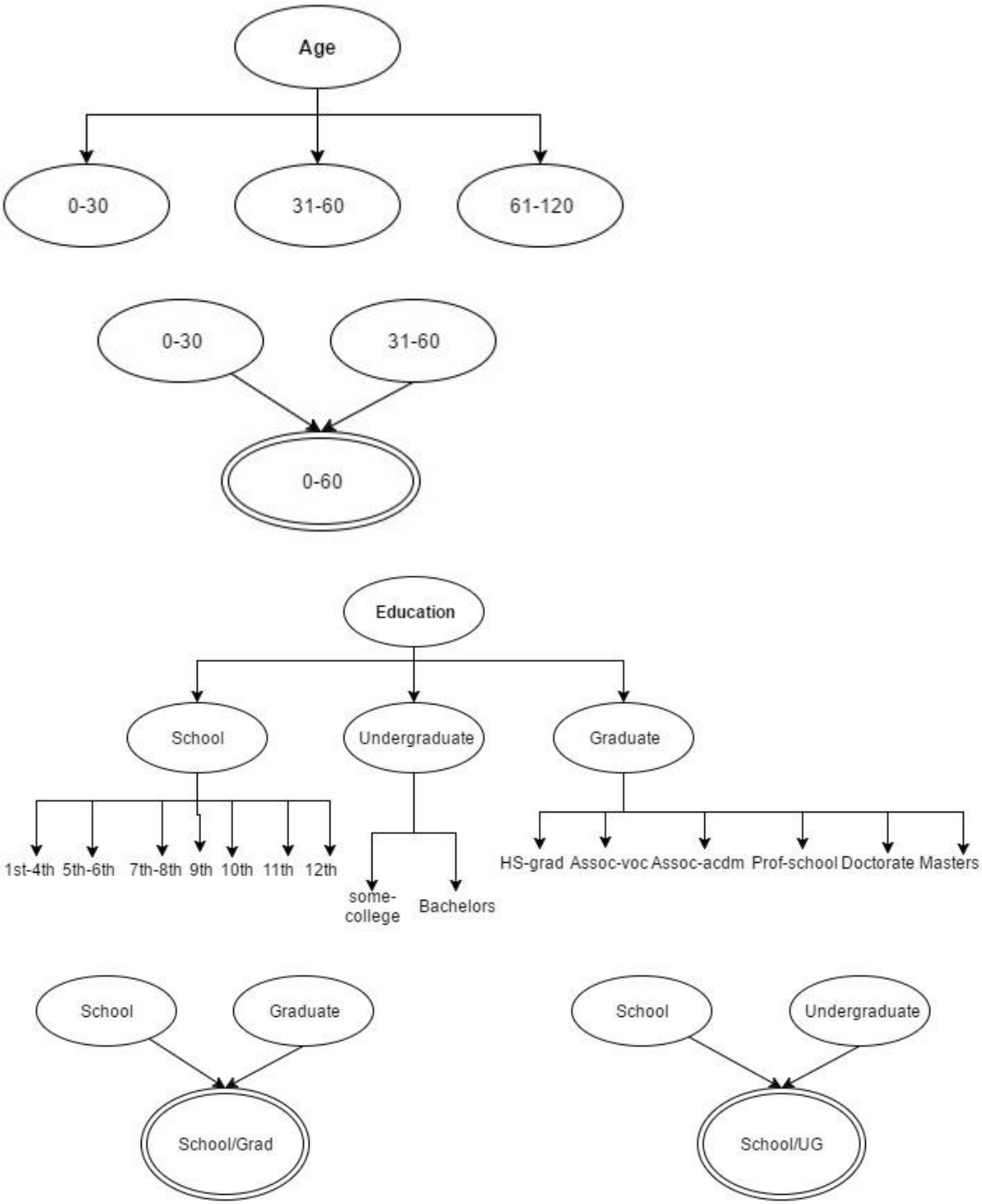
EMD between distributions P = $(p_1,p_2,p_3,...)$ and Q = $(q_1,q_2,q_3,...)$

$f_{i,j}$ = 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:



Conclusion of Anonymization

a. 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.
b. 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.

c. 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.

Clustering (Use of Machine Learning to Reveal the Sensitive Attribute)

Algorithm Used: k-Modes Algorithm for Mixed Type of Data (Numerical + Nominal)
The K-means cluster technique cannot cluster categorical data since of the different measure it using. The K-modes cluster algorithms is base on K-mean pattern other than remove the numeric data limitation even as preserve its effectiveness.
This K-mode technique extend K-mean pattern to cluster categorical data through eliminate the limitation forced by Kmeans follow modification:

1. Using simple match dissimilar evaluate or hamming distance used for categorical data object
2. change means of cluster by modes

$$d(x, y) = \sum_{i=1}^{f} \delta(X_j, Y_j) \ldots\ldots\ldots\ldots(1)$$

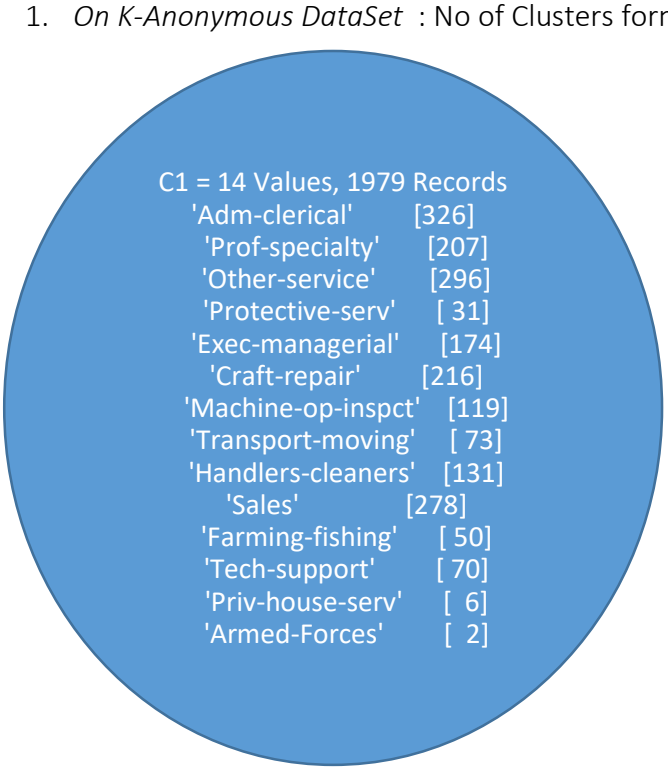
d (x, y) gives equal significance to every kind of an attribute. Let Z be a set of categorical data objects described by categorical attributes, A1, A2 . . . . . Am. while the above is used because the dissimilarity determine for categorical data objects, the cost function become

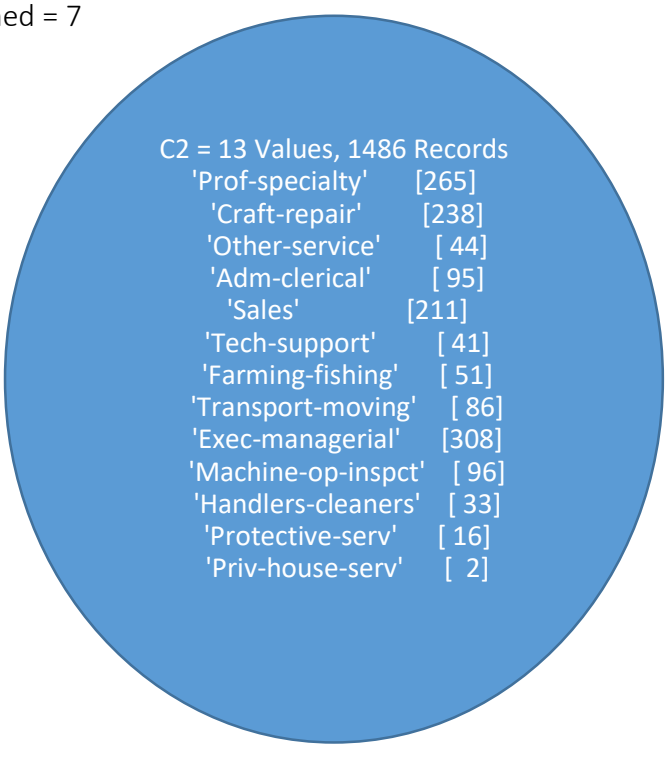
$$C(Q) = \sum_{i=1}^{n} d(Z_i, Q_i) \ldots\ldots\ldots\ldots(2)$$

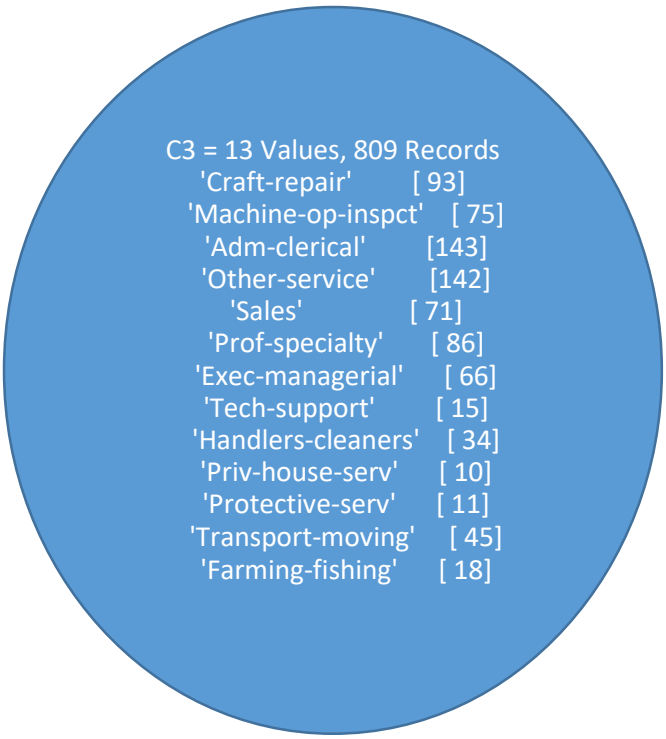
Where Zi is the ith element and Qi is the near cluster centre of Zi.
The K-modes technique minimizes the cost Function defined in Equation 2.
The K-modes assumes that the information of number of probable group of data (i.e. K) is accessible and consists of the following steps : -
1.Generate K clusters by arbitrarily selecting data objects and choose K initial cluster center, one for every of the cluster.
2.Assign data object to the cluster whose cluster center is near toward it according to equation 3.2.
3.Update the K cluster base on allocation of data objects plus Calculate K latest modes of every one clusters.
4.Repeat step 2 to 3 awaiting no data object has changed cluster relationship otherwise some additional predefined criterion is fulfill.

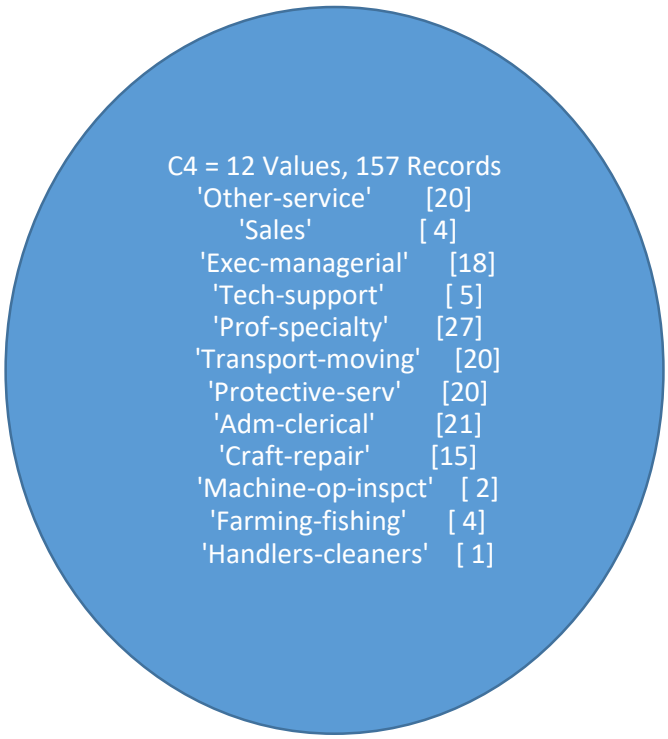
Results

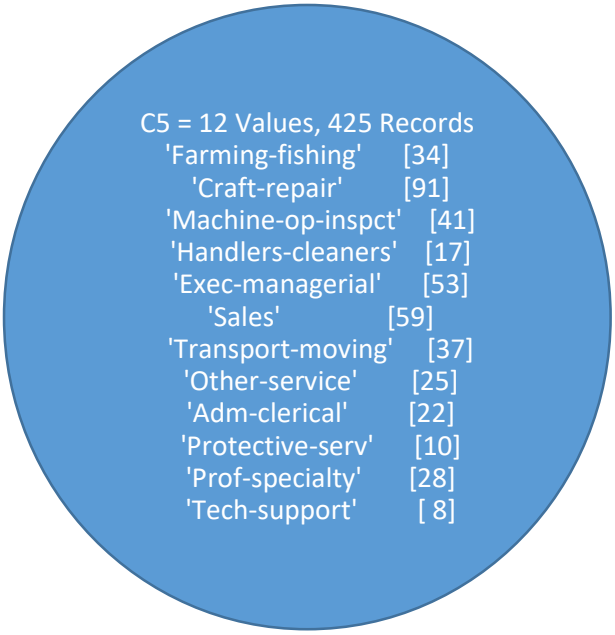
1. *On K-Anonymous DataSet* : No of Clusters formed = 7

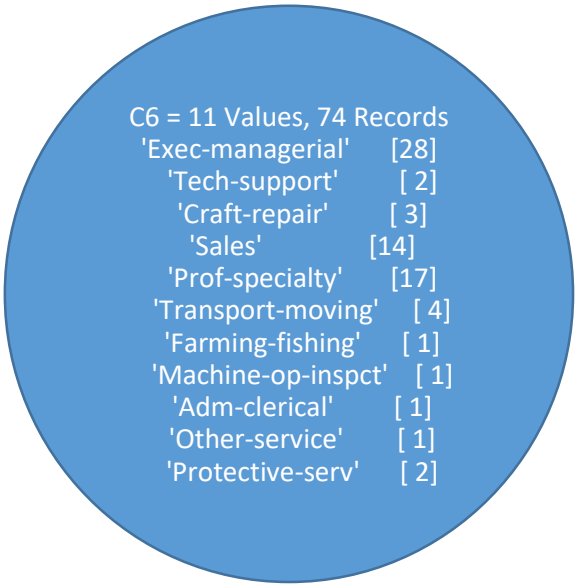


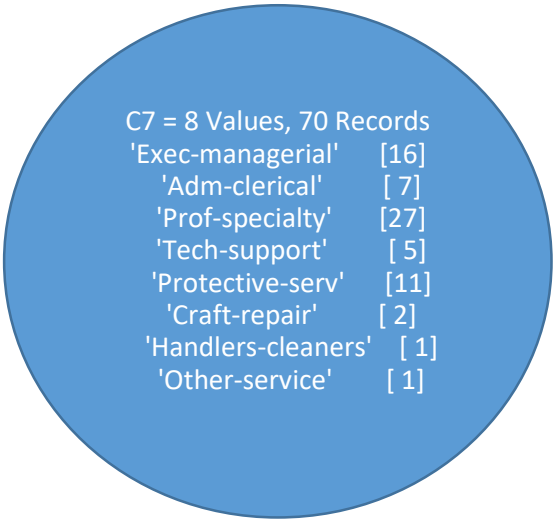



C3 = 13 Values, 809 Records
'Craft-repair' [93]
'Machine-op-inspct' [75]
'Adm-clerical' [143]
'Other-service' [142]
'Sales' [71]
'Prof-specialty' [86]
'Exec-managerial' [66]
'Tech-support' [15]
'Handlers-cleaners' [34]
'Priv-house-serv' [10]
'Protective-serv' [11]
'Transport-moving' [45]
'Farming-fishing' [18]

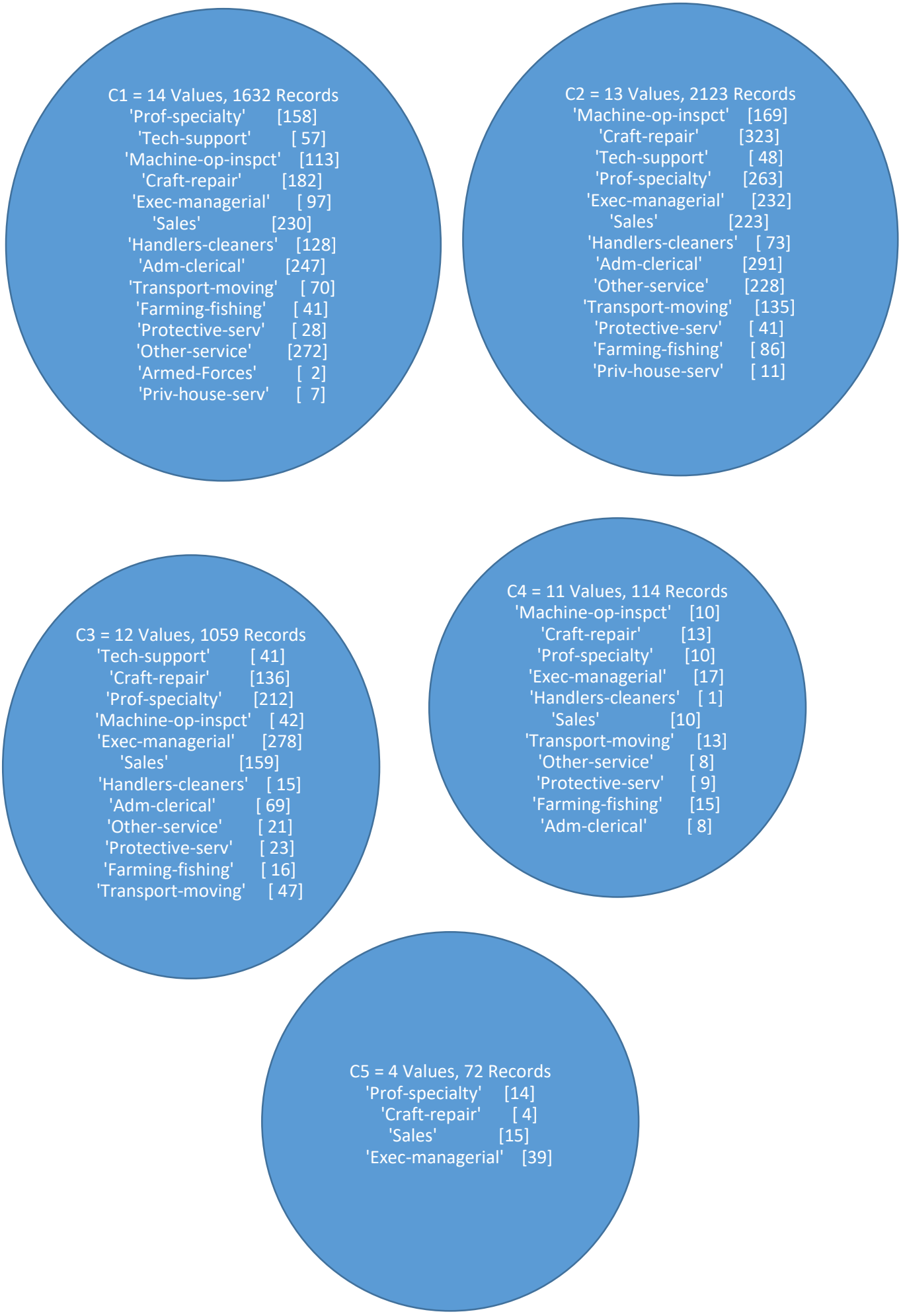

C4 = 12 Values, 157 Records
'Other-service' [20]
'Sales' [4]
'Exec-managerial' [18]
'Tech-support' [5]
'Prof-specialty' [27]
'Transport-moving' [20]
'Protective-serv' [20]
'Adm-clerical' [21]
'Craft-repair' [15]
'Machine-op-inspct' [2]
'Farming-fishing' [4]
'Handlers-cleaners' [1]


C5 = 12 Values, 425 Records
'Farming-fishing' [34]
'Craft-repair' [91]
'Machine-op-inspct' [41]
'Handlers-cleaners' [17]
'Exec-managerial' [53]
'Sales' [59]
'Transport-moving' [37]
'Other-service' [25]
'Adm-clerical' [22]
'Protective-serv' [10]
'Prof-specialty' [28]
'Tech-support' [8]

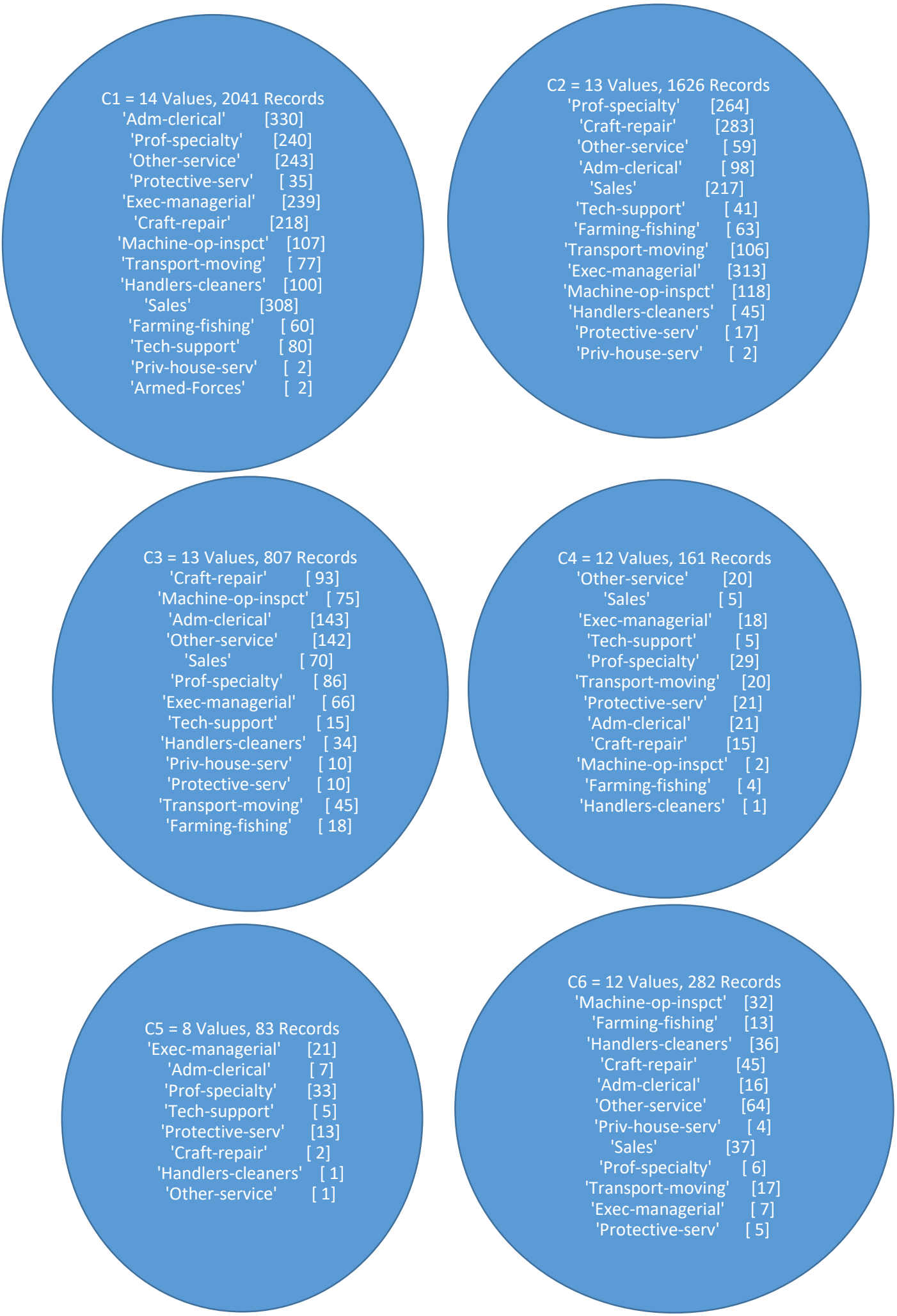

C6 = 11 Values, 74 Records
'Exec-managerial' [28]
'Tech-support' [2]
'Craft-repair' [3]
'Sales' [14]
'Prof-specialty' [17]
'Transport-moving' [4]
'Farming-fishing' [1]
'Machine-op-inspct' [1]
'Adm-clerical' [1]
'Other-service' [1]
'Protective-serv' [2]


C7 = 8 Values, 70 Records
'Exec-managerial' [16]
'Adm-clerical' [7]
'Prof-specialty' [27]
'Tech-support' [5]
'Protective-serv' [11]
'Craft-repair' [2]
'Handlers-cleaners' [1]
'Other-service' [1]

2. *On L-Diverse Data* : No of Clusters Formed = 5



3. *On T-Close Data :* No of Clusters Formed = 6



Utility Measure

We compute the utility of a dataset, when the goal is to get knowledge about attribute $A1$, considering the correlated attributes $A2, A3, \ldots, An$ as follows

$$Umax(A1) = Hmax(A1) - H(A1/A2, A3, \ldots, An) \quad (1)$$

To compute the average utility of a dataset considering all possible usage of the dataset, we first compute the maximum utility for all attributes in the dataset, $Umax(A1)$, $Umax(A2)$, $\ldots$, $Umax(An)$. We then assume that the data publisher has a priori distribution $P(A)$ on the possibility of selection of attributes for an application and computes the average utility of a dataset over all possible applications. That is:

$$UtilityT = \Sigma i P(Ai) Umax(Ai)$$

Approach :

For each attribute A, Umax(A) is calculated by the formula given in (1),
Where Hmax(A1) = log2(k), where k is the number of distinct values of the attribute A1.
H(A1/A2,A3,…,An) is the conditional entropy, and is calculated using the formula :

H(Y/X) = H(X,Y) – H(X).

After calculating Umax for each attribute Ai. The average utility of the dataset is calculated by assuming the priori distribution of selection of attributes for an application to be uniform.

Utilities of various measures :

1. For Original Dataset : 2.7911
2. For K-Anonymous Dataset : 0.4638
3. For L-Diverse Dataset : 0.6670
4. For T-Close Dataset : 0.8926
5. For Trivially Sanitized dataset : 0.0469