International Conference on Systems, Science, Control, Communication, Engineering and Technology 2015 [ICSSCCET 2015]

ISBN	978-81-929866-1-6
Website	icssccet.org
Received	10 - July - 2015
Article ID	ICSSCCET051

VOL	01
eMail	icssccet@asdf.res.in
Accepted	31- July - 2015
eAID	ICSSCCET.2015.051

Knowledge study of Anonymity Databases

P.Mayilvel kumarm[1], K.Kalaiselvi[2], R.Saranya[3], J.K.Kiruthika[4]

[1, 2, 4] Faculty CSE.

[3]Final year B.E CSE, Karpagam Institute of Technology, Coimbatore.

Abstract: *Knowledge Discovery in Databases (KDDs) is the process of identifying valid, novel, useful, and understandable patterns from large data sets. Data Mining (DM) is the core of the KDD process, involving algorithms that explore the data, develop models, and discover significant patterns. One way to enable effective data mining while preserving privacy is to anonymize the data set that includes private information about subjects before being released for data mining. Two common manipulation techniques used to achieve k-anonymity of a data set are generalization and suppression. Generalization refers to replacing a value with a less specific but semantically consistent value, while suppression refers to not releasing a value at all. Generalization is more commonly applied in this domain since suppression may dramatically reduce the quality of the data mining results if not properly used.In this project, we propose a new method for achieving k-anonymity named K-anonymity of Classification Trees Using Suppression (kACTUS). In kACTUS, efficient multidimensional suppression is performed.Thus, in kACTUS, we identify attributes that have less influence on the classification of the data records and suppress them if needed in order to comply with k-anonymity. Encouraging results suggest that kACTUS pedictive performance is better than that of existing k-anonymity algorithms. Attackers often have background knowledge, and we show that k-anonymity does not guarantee privacy against attackers using background knowledge. So we propose the novel and powerful privacy definition called L-diversity. L-Diversity provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. The main idea behind L-diversity is the requirement that the values of the sensitive attributes are well-represented in each group.*

Keywords- Anonymity ,privacy preservation, QI

I.INTRODUCTION

Anonymity typically refers [1,5,7,11]to the state of an individual's personal identity, or personally identifiable information, being publicly unknown. There are many reasons why a person might choose to obscure their identity and become anonymous. Several of these reasons are legal, legitimate and socially approved of many acts of charity are performed anonymously, as benefactors do not wish, for whatever reason, to be acknowledged for their action. Someone who feels threatened by someone else might attempt to hide from the threat behind various means of anonymity, a witness to a crime can seek to avoid retribution, for example, by anonymously calling a crime tip line. There are also many reasons to hide behind anonymity. Criminals typically try to keep themselves anonymous either to conceal the fact that a crime has been committed, or to avoid capture. Anonymity may also be created unintentionally, through the loss of identifying information due to the passage of time or a destructive event.

Cite this article as: P.Mayilvel kumarm, K.Kalaiselvi, R.Saranya, J.K.Kiruthika. “Knowledge study of Anonymity Databases.” *International Conference on Systems, Science, Control, Communication, Engineering and Technology (2015)*: 236-240. Print.

II.PRIVACY PRESERVING DATA MINING

Privacy Preserving Data Mining or PPDM , is a research area concerned with the privacy driven from personally identifiable information when considered for data mining. PPDM provide security to protect data. PPDM have different kind of algorithms. PPDM Includes Privacy preserving Association Rule Mining, Privacy Preserving Clustering and Privacy Preserving Classification.

III.QUASI-IDENTIFIERS

Combinations of attributes within the data that can be used to identify individuals. For example, the statistic given is that 87% of the population of the United States can be uniquely identified by gender, date of birth, and 5-digit zip code. Given that three-attribute "quasi-identifier", a dataset that has only one record with any given combination of those fields is clearly not anonymous – most likely it identifies someone. Datasets are "k-anonymous" when for any given quasi-identifier, a record is indistinguishable from k-1 others.

IV.MULTIDIMENSIONAL SUPPRESSION FOR PRIVACY PRESERVATION

Suppression refers to removing a certain attribute value and replacing occurrences of the value with a special value "*" indicating that any value can be placed instead [6]. The original ZIP codes {06148, 06149} can be generalized to 0614*, thereby stripping the rightmost digit and semantically indicating a larger geographical area. The domains in databases are used to describe the set of values that attributes assume. For example, there might be a ZIP domain, a number domain and a string domain. In the original database, where every value is as specific as possible, every attribute is considered to be in a ground domain. For example, 06148 and 06149 are in the ground ZIP domain, Z0. In order to achieve anonymity the ZIP codes should be less informative. This can be done by making the domain of them at higher level Z1 in which the last digit has been replaced by '*'.

V.K-Anonymity

If the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release. If you try to identify a man from a release, but the only information you have is his birth date and gender. There are k people meet the requirement. This is k-Anonymity[1,4,9]. Each released record should be indistinguishable from at least(k-1) others on its QI attributes. Alternatively, cardinality of any query on released data should be atleast k. k-anonymity is(the first) one of many privacy definitions in this line of work.

VI. Complementary Release Attack

Different releases can be linked together to compromise k- anonymity.

VI a. Solution

Consider all of the released tables before release the new one, and try to avoid linking. Other data holders may release some data that can be used in this kind of attack. Generally, this kind of attack is hard to be prohibited completely.

VI b. L Diversity

The l-Diversity principle advocates ensuring well represented values for sensitive attributes but does not define what well represented values mean. In a l-diverse q*-block an attacker would need l-1 pieces of background knowledge to eliminate l-1 sensitive values to infer positive disclosure. A q*-block is l-diverse if contains atleast l "well-represented" values for the sensitive attributes S.A table is l-diverse[11] if every q*-block is l-diverse. This implies that for a table to be entropy l-Diverse, the entropy of the entire table must be at least log(l). Therefore, entropy l-Diversity may be too restrictive to be practical. Less restrictive than entropy l-diversity .Let s1, …, sm be the possible values of sensitive attribute S in a q*-block. Assume we sort the counts n(q*,s1), …, n(q*,sm) in descending order with the resulting sequence r1, …, rm. We can say a q*-block is recursive (c,l)-diverse if r1 < c(r2+ …. +rm) for a specified constant c.

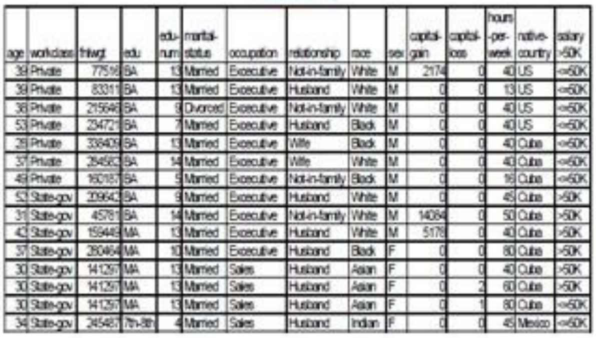
VI c. OVERVIEW OF THE PROJECT

To protect respondents' identity when releasing microdata, data holders often remove or encrypt explicit identifiers, such as names and social security numbers. De-identifying data, however, provide no guarantee of anonymity. Released information often contains other data, such as race, birth date, sex, and ZIP code, that can be linked to publicly available information to re-identify respondents and to infer information that was not intended for release. One of the emerging concept in microdata protection is k-anonymity,which has been recently proposed as a property that captures the protection of a microdata table with respect to possible re-identification of the re-spondents to which the data refer. k-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than k respondents. One of the interesting aspect of k-anonymity is its association with protection techniques that preserve the truthfulness of the data. In this chapter we discuss the concept of k-anonymity, from its original proposal illustrating its enforcement via generalization and suppression. We then survey and discuss research results on k-anonymity in particular with respect to algorithms for its enforcement. We also discuss different ways in which generalization and suppressions can be applied to satisfy k-anonymity and, based on them, introduce a taxonomy of k-anonymity solutions. k-anonymity requirement- Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k respondents.Since it seems impossible, or highly impractical and limiting, to make assumptions on the datasets available for linking to external attackers or curious data recipients, essentially k-anonymity takes a safe approach requiring that,in the released table itself, the respondents be indistinguishable (within a given set) with respect to the set of attributes. To guarantee the k-anonymity requirement, k-anonymity

```
1.  Read Dataset
2.  Preprocessing the dataset
    If any missing values then
            Remove the column
    Elseif check validated data then
            Validate the dataset
    End
3.  Built classification tree when threshold accuracy reached
4.  Calculate quasi identifier for classification tree
5.  Apply k anonymity process for training dataset
    Calculate accuracy
6.  Apply k anonymity process for training dataset
    Calculate accuracy
7.  Anonimize the datset based on higher accuracy
8.  Apply L diversity based clustering
9.  Anonimize the dataset based on clustered values.
10. Anonimized data.
```

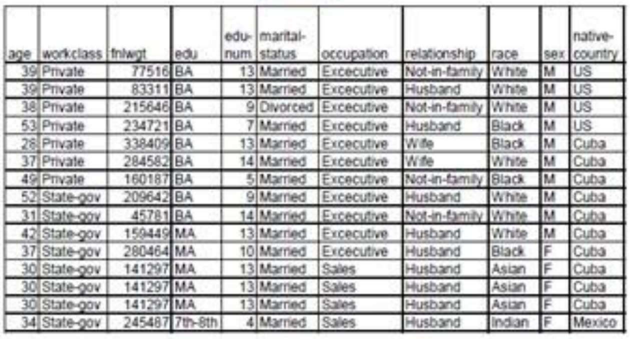
VI d.DATABASE DESIGN

Table 1
DataSet



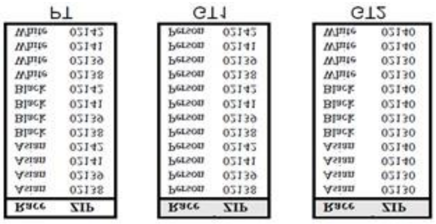
age	workclass	fnlwgt	edu	edu-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary >50K
39	Private	77516	BA	13	Married	Excecutive	Not-in-family	White	M	2174	0	40	US	<=50K
39	Private	83311	BA	13	Married	Excecutive	Husband	White	M	0	0	13	US	<=50K
38	Private	215646	BA	9	Divorced	Excecutive	Not-in-family	White	M	0	0	40	US	<=50K
53	Private	234721	BA	7	Married	Excecutive	Husband	Black	M	0	0	40	US	<=50K
28	Private	338409	BA	13	Married	Excecutive	Wife	Black	M	0	0	40	Cuba	<=50K
37	Private	284582	BA	14	Married	Excecutive	Wife	White	M	0	0	40	Cuba	<=50K
49	Private	160187	BA	5	Married	Excecutive	Not-in-family	Black	M	0	0	16	Cuba	<=50K
52	State-gov	209642	BA	9	Married	Excecutive	Husband	White	M	0	0	45	Cuba	>50K
31	State-gov	45781	BA	14	Married	Excecutive	Not-in-family	White	M	14084	0	50	Cuba	>50K
42	State-gov	159449	MA	13	Married	Excecutive	Husband	White	M	5178	0	40	Cuba	>50K
37	State-gov	280464	MA	10	Married	Excecutive	Husband	Black	F	0	0	80	Cuba	>50K
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	0	0	40	Cuba	>50K
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	0	2	60	Cuba	>50K
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	0	1	80	Cuba	<=50K
34	State-gov	245487	7th-8th	4	Married	Sales	Husband	Indian	F	0	0	45	Mexico	<=50K

Table 2
DataSet



age	workclass	fnlwgt	edu	edu-num	marital-status	occupation	relationship	race	sex	native-country
39	Private	77516	BA	13	Married	Excecutive	Not-in-family	White	M	US
39	Private	83311	BA	13	Married	Excecutive	Husband	White	M	US
38	Private	215646	BA	9	Divorced	Excecutive	Not-in-family	White	M	US
53	Private	234721	BA	7	Married	Excecutive	Husband	Black	M	US
28	Private	338409	BA	13	Married	Excecutive	Wife	Black	M	Cuba
37	Private	284582	BA	14	Married	Excecutive	Wife	White	M	Cuba
49	Private	160187	BA	5	Married	Excecutive	Not-in-family	Black	M	Cuba
52	State-gov	209642	BA	9	Married	Excecutive	Husband	White	M	Cuba
31	State-gov	45781	BA	14	Married	Excecutive	Not-in-family	White	M	Cuba
42	State-gov	159449	MA	13	Married	Excecutive	Husband	White	M	Cuba
37	State-gov	280464	MA	10	Married	Excecutive	Husband	Black	F	Cuba
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	Cuba
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	Cuba
30	State-gov	141297	MA	13	Married	Sales	Husband	Asian	F	Cuba
34	State-gov	245487	7th-8th	4	Married	Sales	Husband	Indian	F	Mexico

Table 3
K-ANONYMITY TABLES BASED ON PRIVATE TABLE:



PT

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

GT1

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

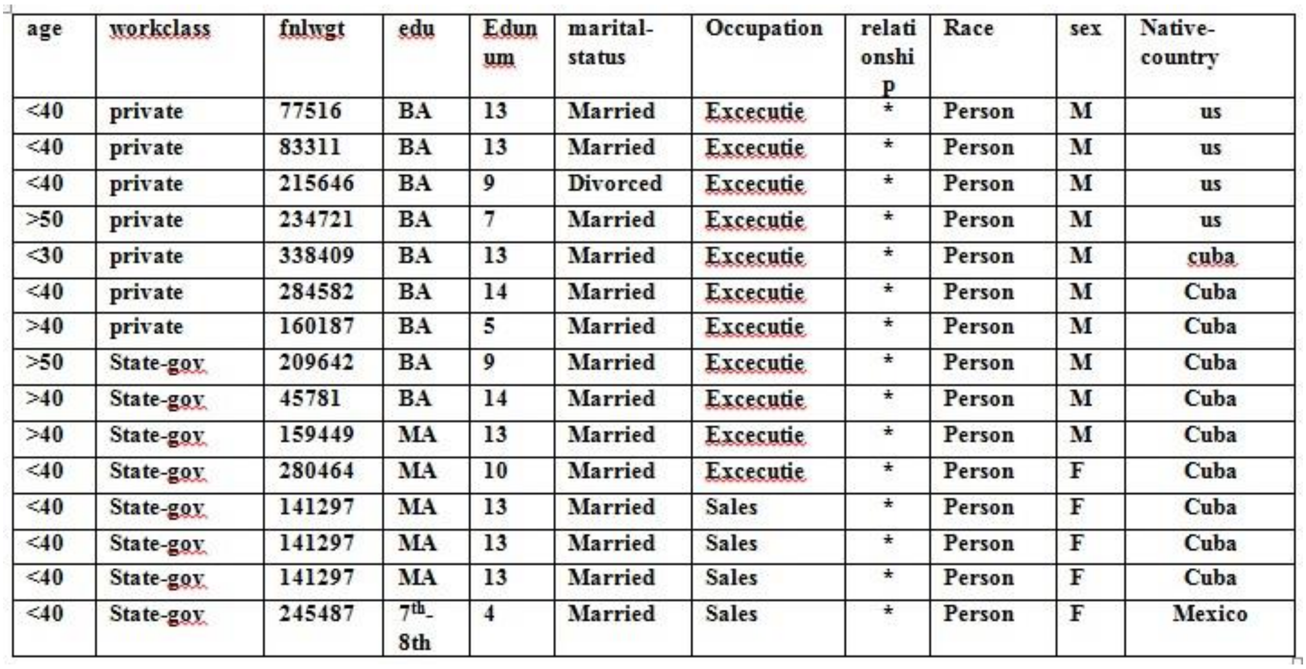
GT2

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

Cite this article as: P.Mayilvel kumarm, K.Kalaiselvi, R.Saranya, J.K.Kiruthika. “Knowledge study of Anonymity Databases.” *International Conference on Systems, Science, Control, Communication, Engineering and Technology (2015)*: 236-240. Print.

TABLE 4

ANONYMIZED



age	workclass	fnlwgt	edu	Edunum	marital-status	Occupation	relationship	Race	sex	Native-country
<40	private	77516	BA	13	Married	Excecutie	*	Person	M	us
<40	private	83311	BA	13	Married	Excecutie	*	Person	M	us
<40	private	215646	BA	9	Divorced	Excecutie	*	Person	M	us
>50	private	234721	BA	7	Married	Excecutie	*	Person	M	us
<30	private	338409	BA	13	Married	Excecutie	*	Person	M	cuba
<40	private	284582	BA	14	Married	Excecutie	*	Person	M	Cuba
>40	private	160187	BA	5	Married	Excecutie	*	Person	M	Cuba
>50	State-gov	209642	BA	9	Married	Excecutie	*	Person	M	Cuba
>40	State-gov	45781	BA	14	Married	Excecutie	*	Person	M	Cuba
>40	State-gov	159449	MA	13	Married	Excecutie	*	Person	M	Cuba
<40	State-gov	280464	MA	10	Married	Excecutie	*	Person	F	Cuba
<40	State-gov	141297	MA	13	Married	Sales	*	Person	F	Cuba
<40	State-gov	141297	MA	13	Married	Sales	*	Person	F	Cuba
<40	State-gov	141297	MA	13	Married	Sales	*	Person	F	Cuba
<40	State-gov	245487	7th-8th	4	Married	Sales	*	Person	F	Mexico

VII.ALGORITHM

The algorithm is as follows:

```
Create a root node for the tree
If all examples are positive, Return the single-node tree Root, with label = +.
If all examples are negative, Return the single-node tree Root, with label = -.
If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
Otherwise Begin
A = The Attribute that best classifies examples.
Decision Tree attribute for Root = A.
For each possible value, vi, of A,
Add a new tree branch below Root, corresponding to the test A = vi.
Let Examples(vi), be the subset of examples that have the value vi for A
If Examples(vi) is empty
Then below this new branch add a leaf node with label = most common target value in the examples
Else below this new branch add the subtree C4.5 (Examples(vi), Target_Attribute, Attributes – {A})
End
Return Root
```

VIII.CONCLUSION

WE PRESENTED A NEW METHOD FOR PRESERVING THE PRIVACY IN CLASSIFICATION TASKS USING K-ANONYMITY. THE PROPOSED METHOD REQUIRES NO PRIOR KNOWLEDGE REGARDING THE DOMAIN HIERARCHY TAXONOMY AND CAN BE USED BY ANY INDUCER. THE NEW METHOD ALSO SHOWS A HIGHER PREDICTIVE PERFORMANCE WHEN COMPARED TO EXISTING STATE-OF-THE-ART METHODS.

This work is motivated by the observation that although all previous k-anonymity techniques assume the existence of a PD, which can be used to breach privacy, none actually takes PD into account during the anonymization process. This omission leads to unnecessarily high information loss. In Fig. 1, if k ¼ 3, tuple G 2 MT does not require generalization, as PD already contains two other records (G1 and G2Þ with the same QI values. Based on this fact, we introduce the concept of k-join-anonymity (KJA) to reduce the information loss. Briefly, KJA anonymizes a superset of MT, which includes selected records from PD.In most practical anonymization scenarios, there exists public knowledge (e.g., voter registration data) that can be used by an attacker to breach privacy. On the other hand, this knowledge can also be exploited to reduce the information loss in the published data. Motivated by this observation, we introduce the concept of KJA and show how existing generalization algorithms can be adopted to take into account external databases. We demonstrate the effectiveness of KJA through an extensive experimental evaluation, using real and synthetic data sets. An interesting direction for future work is to apply the general concept of exploiting external knowledge to alternative forms of deidentification. For instance, since some k-anonymity algorithms (e.g., Mondrian) can be easily adapted to capture l-diversity, we expect that the availability of external information will also be beneficial in this case. Additionally, we plan to investigate the issue of updates in MT and PD. Assume that after the initial release of AT, the MT is modified and a new AT must be published. Meanwhile, the PD may have also been updated. A challenging issue is to incrementally update the AT, without compromising the privacy of MT or the utility of AT.

REFERENCES

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.

[2] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.

[3] C. Bettini, X.S. Wang, and S. Jajodia, "The Role of Quasi-Identifiers in k-Anonymity Revisited," Technical Report abs/cs/0611035, Computing Research Repository (CoRR), 2006.

[4] R.J. Bayardo, Jr., and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 217-228, 2005.

[5] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, "Utility- Based Anonymization Using Local Recoding," Proc. ACM SIGKDD, pp. 785-790, 2006.

[6] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," Proc. ACM SIGMOD, pp. 49-60, 2005.

[7] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. IEEE Int'l Conf. Data Eng. (ICDE), p. 25, 2006.

Cite this article as: P.Mayilvel kumarm, K.Kalaiselvi, R.Saranya, J.K.Kiruthika. "Knowledge study of Anonymity Databases." *International Conference on Systems, Science, Control, Communication, Engineering and Technology (2015)*: 236-240. Print.