International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013)

International Conference on Information Systems and Computing (ICISC-2013), INDIA.

PRIVACY PRESERVATION OF CONFIDENTIAL DATABASE USING ANONYMIZATION

J. Jesmitha[1], R. Seethalakshmi[2]

[1,2]PG Student, Sri Sairam Engineering College, Chennai – 44.

jesmithaj@yahoo.co.in
seetha219@gmail.com

Abstract

The main objective of the project is to design and develop a system which preserves user's privacy. The anonymization method which provides both privacy protection and knowledge preservation is explained here. Data is anonymized to protect privacy; on the other hand, miners are allowed to discover useful knowledge from anonymized data. For case study, we have taken legal system where the client's data need to be published. This puts personal privacy at risk. To surmount this risk, attributes that clearly identifies individuals, such as Name, Original Suit number, are removed. But this is not enough because such a database can sometimes be joined with other public database on attributes such as Gender, Job, Date of Birth, and Pin code to identify individuals who were supposed to remain anonymous. To protect user privacy, we use anonymization algorithm. The user details should be anonymized using this algorithm before publishing them. Anonymization algorithm used should protect privacy at the same time the utility of data should also be preserved. While considering Legal Privacy System, the sensitive details of client should not be revealed, Anonymization algorithm preserves the sensitive data.

Keywords**-- Anonymization, Quasi identifier, Generalization.**

I. INTRODUCTION

The aim of the project is to design a system which preserves user's privacy. The Anonymization method which provides both privacy protection and knowledge preservation is explained here. We anonymize data to protect privacy; at the same time miners should discover useful knowledge from anonymized data. A recent study reveals approximately 87% of the population of the United States can be uniquely identified on the basis of Gender, Date of Birth, and 5-digit Zip code.

II. RELATED WORKS

2.1 Achieving k-anonymity privacy protection using generalization and suppression

Organizations need to share person-specific records in such a way that the identities of the individuals who are the subjects of the data cannot be determined. This is achieved by releasing records that adhere to *k* anonymity, which means each released record has at least *k*-1 other records in the release whose values are indistinct over those fields that appear in external data. So, *k* anonymity provides privacy protection by guaranteeing that each released record will relate to at least *k* individuals even if the records are directly linked to external information. This paper provides a formal presentation of combining generalization and suppression to achieve *k*-anonymity. Generalization involves replacing a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all.

The Preferred Minimal Generalization Algorithm (MinGen), which is a theoretical algorithm presented here, combines these techniques to provide *k*-anonymity protection with minimal distortion.

Issues: With sufficient background knowledge about the individual it is possible to locate a person uniquely. Data loss occurs because of suppression of details. Thus the data may not provide utility.

2.2 L-diversity: privacy beyond k-anonymity

Publishing data about individuals without revealing sensitive information about them is very important. In previously used *k*-anonymized dataset, each record is indistinguishable from at least $k - 1$ other records with respect to certain identifying attributes. There are two attacks that a *k*-anonymized dataset has, which cause severe privacy problems. First, an attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attribute. Second, attackers often have background knowledge; *k*-anonymity does not guarantee privacy against attackers using background knowledge. A novel and powerful privacy criterion called *l*-diversity that can defend against such attacks is proposed here. A proper formal foundation for *l*-diversity as well experimental evaluation showing that *l*-diversity is practical and can be implemented efficiently is described. This *l*-diversity also handles multiple sensitive attributes and it has methods for handling continuous sensitive attributes.

Issues: The privacy and utility are duals of each other; privacy has been given importance than the utility of a published table. As a result, the concept of utility is not well understood. Therefore knowledge discoverers cannot use the data for inferring useful knowledge from them.

III. PROPOSED SYSTEM

From the literature survey, it is known that there are several drawbacks in the system. To overcome all the problems in existing systems, we proposed a method which anonymizes data by randomly breaking links among attribute values in record. This is based on probabilistic distribution of attributes. In this method, the privacy of the individual as well as the knowledge is preserved. An enhanced algorithm is proposed to tackle situation where user's prior knowledge may cause privacy leakage. The anonymization is done in such a way that the utility of the data published is preserved.

Probabilistic anonymity: Suppose that a data set D is anonymized to D_0. Let r be a record in D and r0 belongs to D0 be its anonymized form. Denote r(QI) as the value combination of the quasi-identifier in r. The probabilistic anonymity of data set D_0 is defined by $1/P(r(QI)/ r_0 (QI))$, where $P(r(QI)/ r_0 (QI))$ is the probability that r(QI) (for all r belongs to D) may be inferred given $r_0(QI)$.The probabilistic anonymity gives a measurement of how unlikely the user can infer original associations. The greater the probabilistic anonymity, the less probable the user can guess the original data. The Algorithm should maximize the anonymity of the data set produced.

Quasi identifier: Given a data set $D(A_1, A_2; . . . , A_m)$ and an external table D_E. For all records r_i belongs to D, if the value combination $r_i(A_j, . . . ,A_k)$, j, k < m,{ $A_j, . . . ,A_k$} contains no identifiers, can be uniquely located in D_E, we call the set of attributes {$A_j, . . . ,A_k$} as quasi-identifier. For example, in the data set the external table D_E would consist of Name, Age, Job, and Country and a quasi-identifier would be {Age, Job, Country}.

Generalization: Suppose that a domain M consists of disjoint partitions {Pi}, i=1 . . . n, and UPi =M. On a given value combination v, we call the generalization process as returning the only partition Pi containing v. By generalizing the quasi-identifier, each individual in a k anonymous table is identical to at least k - 1 other ones with respect to the quasi-identifier.

K-anonymous: Given a data set $D(A_1, A_2; . . . , A_m)$ and its quasi-identifier QI. If for any subset C subset of QI and for any record r_i belongs to D, there exist at least k -1 other records sharing the same values with r_i on the attribute set C, then data set D is k-anonymous.

When the data set is k-anonymous, we can group together the records with the same value combinations of the quasi-identifier.

3.1 System architechture

The System architecture diagram provides a top-down description of the structure of the System. The diagram (Fig: 1) shown below describes the architecture of the system developed.

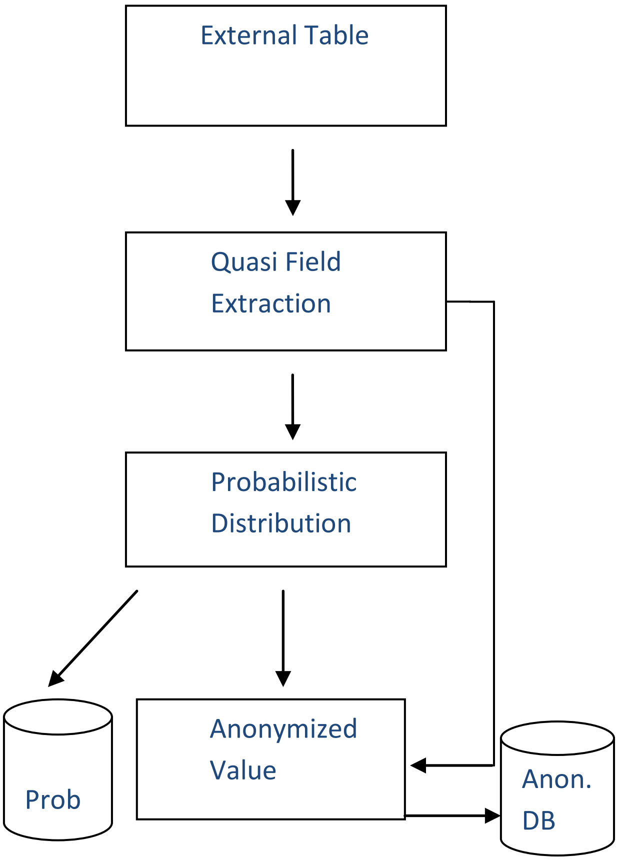


Figure 1: System Architecture

3.2 Context diagram

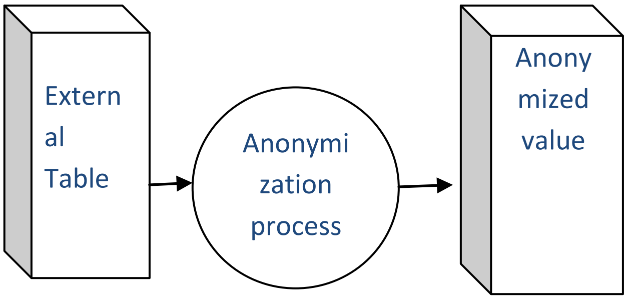


Figure 2: Context Diagram

Figure 2 describes the context diagram of the system, which is also called as zero level DFD. Context diagram describes the input, output and the process.

IV. MODULES

A module is a self-contained component of a system which has a well-defined interface to other components of the system. Here the Development module is split up into three modules, they are: Quasi Identification, Probability Distribution and Anonymization. Each module and its function are described in detail.

4.1 Quasi field identification

From the information provided by client, the fields which act as quasi are identified. Combination of two or more field may reveal the sensitive data; those combinations of field are called quasi identifiers. Such an identity-leaking attribute combination is called as a quasi-identifier. In the first module we find which attributes act as a strong quasi identifier.

In the first test set we will take **Gender, Date of Birth and Zip code**

The number of distinct value for Gender (C1) is d1=2

The number of distinct values of column Date of Birth (C2) is d2=60*365=2*(10^4)

The number of distinct values for Zip code (C3) is d3=10^5

D=d1*d2*d3

=2*(2*(10^4))*10^5

=4*(10^9)

In the second test set we will take **Occupation, Date of Birth and Zip code**

The number of distinct value for Occupation (C1) is roughly d1=100

The number of distinct values of column Date of Birth (C2) is d2=60*365=2*(10^4)

The number of distinct values for Zip code (C3) is d3=10^5

D=d1*d2*d3

=100*(2*(10^4))*10^5

=4*(50(10^9))

Table 1
Quasi Identifiers

Date Of Birth	Job	Pin code
05/05/1975	Doctor	600040
08/05/1989	Lawyer	600030
05/05/1975	Doctor	600040
05/04/1989	Doctor	600116
08/07/1982	Vendor	566857

Table 1 contains a sample of quasi identifiers Date of Birth, Job, Pin code and their corresponding values.

4.2 Probabilistic distribution

From Quasi identification module, we find the fields: Date of Birth, Zip code and Occupation act as strong quasi identifiers. In Probabilistic distribution module, we find the probability of each field identified.

Algorithm

```
For each column do
    Group the similar values, let the value be s1
Loop begin
        Calculate total number of records, let the value be t1
        Probability distribution for each attribute p= s1/t1;
Loop end
If no similar value
For each distinct attribute d1
        Calculate total number of records, Let the value be t1
        Probability distribution for each attribute=d1/t1;
End
```

Table 2
Probability Distribution of Job

Doctor	0.6
Lawyer	0.2
Vendor	0.2

Table 2 contains the probability value of quasi identifier job. Similarly probability value is calculated for all Quasi fields.

4.3 Anonymization module

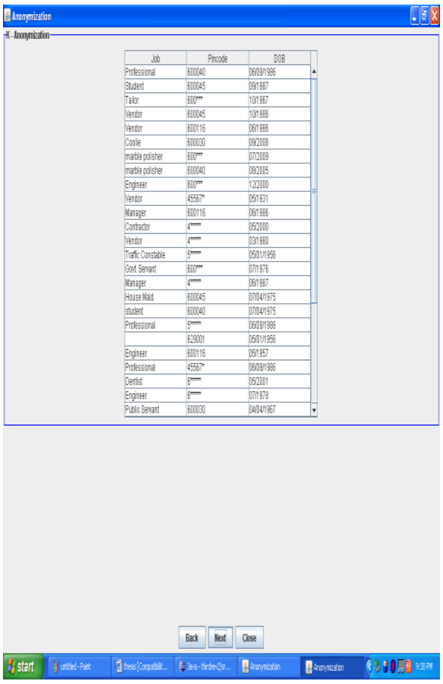
In anonymization module we anonymize the values which are selected as quasi identifiers. Anonymization of values should be done in such a way miners can discover useful knowledge from the data. So permuting the data or removing the values does not hold good. The data should be given to the miners as well the individual's identity should be preserved. So we anonymize values depending on probability value. The attributes with lesser probability should be anonymized other values can be presented as such. The principle used here is, if an attribute has high probability of occurrence, the attribute cannot be used to identify an individual among the crowd because of the large search space. On the other hand if an attribute occur as singleton or if it has lesser probability, the value should be anonymized otherwise it may cause privacy leakage.

For example in a certain data set of 1000 records, if there are 50 Doctors in quasi field occupation, the value can be left as it is because the search space is large. It is difficult to locate a single doctor. On the other hand in the same data set if there is 1 chartered account the individual can be easily identified by using other unanonymized values.

Table 3
Anonymized Table

Date Of Birth	Job	Pin code
05/05/1975	Doctor	600040
*/05/1989	Professional	6000**
05/05/1975	Doctor	600040
*/04/1989	Doctor	600***
08/07/1982	Self Employed	5*****

Table 3 contains the anonymized value of quasi identifier. The Anonymization is done based on the above algorithm. In Date of Birth with lesser probability, date is replaced by a *, which increases the search space. For Job specific values are replaced by more generic value. For example lawyer having less probability is replaced by Professional. In Pin code 600030, if 30 is discarded it has higher probability. Pin code 566857 being a unique number all the digits following 5 should be replaced.



Job	Pincode	DOB
Professional	600040	06/09/1986
Student	600045	09/1987
Tailor	600***	10/1967
Vendor	600045	10/1986
Vendor	600116	06/1986
Coolie	600030	09/2008
marble polisher	600***	07/2009
marble polisher	600040	08/2005
Engineer	600***	12/2000
Vendor	45567*	05/1931
Manager	600116	06/1986
Contractor	4*****	05/2000
Vendor	4*****	03/1980
Traffic Constable	5*****	05/01/1956
Govt. Servant	600***	07/1976
Manager	4*****	06/1987
House Maid	600045	07/04/1975
student	600040	07/04/1975
Professional	5*****	06/09/1986
	629001	05/01/1956
Engineer	600116	05/1957
Professional	45567*	06/09/1986
Dentist	6*****	05/2001
Engineer	6*****	07/1978
Public Servant	600030	04/04/1967

Figure 3 Anonymization Table

Figure 3 contains the anonymized value of quasi identifies.

V. CONCLUSION

I described a method for preserving sensitive details of the user. In my method even if the data is published, privacy breach does not occur because the data is published only after anonymizing using algorithm. In this project, I introduce a novel data anonymization method. Different from the methods such as the k-anonymization methods, I regard the data privacy as the links between the Q-I and sensitive values. By replacing part of the values in each record while maintaining statistical relations in whole data set, my method not only achieves a higher level of the privacy protection but also preserves more non-sensitive knowledge than the other anonymization methods.

Moreover, the useful associations which are less sensitive can be discovered more accurately than the sensitive ones. By comparing the table 1 and 3the data before and after anonymization can be observed. After anonymization sensitive data is not revealed. Only value with high probability of occurrence occurs as it is. All other singleton value and less probability value are replaced. Data anonymization is a popular direction in the research of privacy preserving data mining.

REFERENCES

[1] Ashwin Machaavajjhalal, Daniel Kifer ,"l-Diversity: Privacy Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data, 2007

[2] B.C.M.Fung, K.Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. 21st International Conference on Data Eng, 2005.

[3] K. Wang, P. Yu, and S.Chakraborty, "Bottom-Up Generalization: A Data Mining Solution to Privacy Protection," Proc. Fourth IEEE International Conference on Data Mining, 2004

[4] Sweeney.L, "Achieving k anonymity privacy protection using generalization and Suppression", International journal of uncertainity, 2002.