iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: III Month of publication: March 2018

DOI: http://doi.org/10.22214/ijraset.2018.3493

B-Anonymization: Privacy beyond k-Anonymization and l-Diversity

B. Prakash[1], S. Kranthi Reddy[2], Daljit Singh[3], VBV Phani Sai Yeshwanth[4], Mutukulloju Sai Kumar[5]
[1]*Professor, Head of Department, Vignan Institute of Technology & Science*
[2]*Assistant Professor, Vignan Institute of Technology & Science*
[3, 4, 5] *CSE, Vignan Institute of Technology & Science, Hyderabad*

Abstract: Privacy is very important for both users and enterprises. Research is being done on various aspects of privacy preserving in data management systems. Many algorithms like k-anonymization, l-diversity and t-closeness have been proposed, but each of them has their own advantages and disadvantages. For example taking into account k-anonymization, different attacks such as Background Knowledge and Homogeneity attack can be done. Also lot of time is used in dividing the whole database into equivalence classes by comparing records. So, this algorithm is a slight improvement over k-anonymization in term of privacy and efficiency.

Keyword: Privacy, b-anonymization, k-anonymization, l-diversity, t-closeness

I. INTRODUCTION

In today’s world of internet and inexpensive computing power, society has developed an insatiable appetite for information. Most action performed by a user in daily life are recorded somewhere in computer. This information in turn is shared with others, exchanged or sold. A normal user may not care about the fact that local grocery keeps track of which items they purchase, how frequently they purchase, but this information may be quite sensitive or damaging to individuals or big organizations. Medical information, financial information or matters of national security can have alarming result because of improper disclosure. The objective is to release the information freely but to do so in a way that the identity of any individual contained in the dataset cannot be recognized in any way.

Shockingly, there remains a common belief among people in society that if data looks anonymous, it is anonymous. Data holders, including government agencies, often remove all explicit identifiers, such as name, address and phone number from data so that other information in the database can be shared, believing that the identities of the person who’s data is contained in the database cannot be identified. On the contrary, de-identifying data provides no guarantee of anonymity.

Release information (i.e. that is available for public) often contain other data such as birth date, gender and zip code that in combination can be linked with other publicly available information such as voter id details. Most municipalities sell population registers that include the identities of individuals along with basic demographics; examples include local census data, voter lists, city directories, and information from motor vehicle agencies, tax assessors, and real estate agencies.

For example, an electronic version of a city's voter list was purchased for twenty dollars and used to show the ease of re-identifying medical records. In addition to names and addresses, the voter list included the birth dates and genders of 54,805 voters. Of these, 12% had unique birth dates, 29% were unique with respect to birth date and gender, 69% with respect to birth date and a 5-digit ZIP code, and, 97% were identical with just the full postal code and birth date.

These results reveal how uniquely identifying combinations of basic demographic attributes, such as ZIP code, date of birth, ethnicity, gender and marital status, can be[1].

To illustrate this problem, table of released medical data de-identified by suppressing names and Social Security Numbers (SSNs) so as not to disclose the identities of individuals to whom the data refer. However, values of other released attributes, such as ZIP, Date of Birth, Ethnicity, Sex, and Marital Status can also appear in some external table jointly with the individual identity and can therefore allow it to be tracked. ZIP, Date of Birth, and Sex can be linked to the Voter List to reveal the Name, Address, and City. Likewise, Ethnicity and Marital Status can be linked to other publicly available population registers.

Several protection techniques have been developed with respect to statistical databases, such as scrambling and swapping values and adding noise to the data in such a way as to maintain an overall statistical property of the result.

However, many new uses of data, including data mining, cost analysis and retrospective research, often need accurate information within the tuple itself.

II. K- ANONYMIZATION

A. Understanding the algorithm

It is very easy to understand the protection provided by the *k*-anonymization. If a table satisfies *k*-anonymity for some value *k*, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/*k* [3]. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. Two attacks that can be done on database after anonymization are: the homogeneity attack and the background knowledge attack [8].

Table 1 contains the original data where as the Table 2 also called as release table contain 3-anonymized data. Here, the attribute 'disease' is called as the sensitive attribute. Sensitive attributes are the one which cannot be modified. Other attributes such as Zip code and Age are called quasi identifiers. These attributes are the one where anonymization takes place.

	Zip code	Age	Disease
1	501963	26	Arthritis
2	501978	24	Arthritis
3	501966	22	HIV
4	501936	23	HIV
5	501590	49	Ulcer
6	501593	59	Arthritis
7	501596	41	HIV
8	501598	51	HIV
9	501106	31	Ulcer
10	501119	36	Ulcer
11	501199	37	Ulcer
12	501153	35	Ulcer

Table 1: Original Patient Table

	Zip code	Age	Disease
1	5019**	<30	Arthritis
2	5019**	<30	Arthritis
3	5019**	<30	HIV
4	5019**	<30	HIV
5	50159*	>=40	Ulcer
6	50159*	>=40	Arthritis
7	50159*	>=40	HIV
8	50159*	>=40	HIV
9	5011**	3*	Ulcer
10	5011**	3*	Ulcer
11	5011**	3*	Ulcer
12	5011**	3*	Ulcer

Table 2: 3-Anonymous Table

B. Attacks

1) *Homogeneity Attack:* Suppose Srikar knows that Rishi is 31 year old, lives in Zip code 501106 and was admitted to hospital due to some disease. After the anonymized data is published on the hospital website, he checks and find out that Rishi's record belongs to the third equivalence class. As all the disease in third equivalence class is same, he concludes that Rishi was suffering from Ulcers.
2) *Background Knowledge attack:* Suppose Akhil is Saketh's neighbour and know his zip code and age and concludes that saketh's record belongs to first equivalence class. Also akhil know additional information that saketh has very low risk of HIV. Hence akhil conclude that saketh have Arthritis.

Though *k*-anonymization provides some level of anonymization, attack can be done on it and information can be inferred.

III. L-DIVERSITY

A. Introduction

The *l* –diversity principle states that “A q*-block is *l*-diverse if contains at least *l* well represented values for the sensitive attributes S. A table is *l*-diverse if every q*-block is *l*-diverse [4]. In simple words it means that if the value of *l* is 3, then there must be minimum of three distinct sensitive attribute in each equivalence class. This ensures that the Homogeneity and Background Knowledge attacks cannot be done on it. To understand this algorithm lets go back to the previous example and modify it according to *l*-diverse principle.

	Zip code	Age	Disease
1	501***	<=40	Arthritis
2	501***	<=40	HIV
3	501***	<=40	Ulcer
4	501***	<=40	Ulcer
5	50159*	>=40	Ulcer
6	50159*	>=40	Arthritis
7	50159*	>=40	HIV
8	50159*	>=40	HIV
9	501***	<=40	Arthritis
10	501***	<=40	HIV
11	501***	<=40	Ulcer
12	501***	<=40	Ulcer

Table 3: 3 –diverse Table

B. Similarity attack on l-Diversity

Though *l*-diversity prevents Homogeneity and Background knowledge attack, but it fails to address similarity attack. When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information [4]. Let’s look at the following example:

	Zip Code	Age	Disease
1	505**	3*	Pneumonia
2	505**	3*	Ulcer
3	505**	3*	Flu
4	505**	3*	Flu
5	5078*	<=20	Tuberculosis
6	5078*	<=20	Ulcer
7	5078*	<=20	Flu
8	5078*	<=20	Pneumonia
9	5093*	4*	Pneumonia
10	5093*	4*	Tuberculosis
11	5093*	4*	Lung abscess
12	5093*	4*	Baritosis

Table 4: 3-diverse Table

Suppose Krishna knows that Subheem record belongs to the 3rd equivalence class. Though the table is 3-diverse Krishna can conclude that Subheem has some Lung related disease as all diseases in 3rd equivalence class are Lung related. Hence if the attributes are semantically similar then problem can arise in *l*-diversity algorithm.

IV. PROPOSED ALGORITHM (b-ANONYMIZATION)

To address these limitations of *k*-anonymization and *l*-diversity, *b*-anonymization algorithm is introduced. *b*-anonymization algorithm make use of a simple technique where in the whole database is sorted in accordance with age. After sorting the database middle record is found out in following ways:
Consider the total number of records in the table to be t. Then,

A. For odd number of records

$$Mid = (t+1)/2$$

B. For even number of records

$$Mid = t/2$$

From the middle record we take the age. Now records having age less than and equal to mid is replace by "<=mid" and records with age greater than mid is replaced by ">mid". Also other techniques of *k*-anonymization such as generalization and suppression are used to anonymize other quasi identifiers.

1) *Generalization:* In this method, individual value of attribute is replaced with a broader category [2]. For example the value '51' is replaced by '>=40' in Table 2.
2) *Suppression:* In this method, certain values of the attributes are replaced by an asterisk '*'[2]. All or some values of a column may be replaced by '*'. For example zip code value '135136' is replaced by '1351**'.

	Zip code	Age	Disease
1	501***	<=35	HIV
2	501***	<=35	HIV
3	501***	<=35	Arthritis
4	501***	<=35	Arthritis
5	501***	<=35	Ulcer
6	501***	<=35	Ulcer
7	501***	>35	Ulcer
8	501***	>35	Ulcer
9	501***	>35	HIV
10	501***	>35	Ulcer
11	501***	>35	HIV
12	501***	>35	Arthritis

Table 5: Binary Anonymized Table

From Table 3, it can be seen that attacks such as Homogeneity, Background Knowledge and Similarity attacks are not possible because of the database is divided into two sections and each section have more than two different diseases. Here zip code is suppressed and age is generalized.

V. CONCLUSION

"b- Anonymization" can be used to anonymize the relational databases. This technique is more efficient than k-anonymity and has higher degree of anonymization. *k*-anonymity takes more time as it has to compare records with each other in order to form equivalence classes.

This is the step where most of the time is wasted in *k*-anonymization. Whereas, in binary anonymization technique simple sorting is done with respect to 'age' and mid element is found.

Also the degree of anonymization increase with the increase in number records in the database. This can be seen from the fact that the whole database is divided into two classes based on age and each class contain 'n' different types of diseases.

VI. FUTURE WORK

The scope of *b*- anonymization is not limited. This algorithm can also be applied to databases containing more than one sensitive attributes. Also we can take more number of points in 'age' column so as to preserve more information which can be used for research purposes.

REFERENCES

[1] Privacy preserving data publishing and data anonymization approaches: A review **DOI:** 10.1109/CCAA.2017.8229787,IEE
[2] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertain. Fuzz., 10(6):571–588, 2002
[3] L. Sweeney. K-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzz., 10(5):557–570, 2002, IEE
[4] (a, d)-Diversity: Privacy Protection Based on l-Diversity DOI: 10.1109/WCSE.2009.362, IEE
[5] t-Closeness: Privacy Beyond k-Anonymity and l-Diversity DOI: 10.1109/ICDE.2007.367856,IEE
[6] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. ACM Comput. Surv., 21(4):515–556, 1989.
[7] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In EDBT, pages 183–199, 2004.
[8] Analysis of privacy preserving K-anonymity methods and techniques INSPEC Accession Number: **11887481**, IEEE
[9] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In VLDB, 1994.
[10] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. A.I., 87(1-2), 1996.
[11] R. J. Bayardo and R. Agrawal. Data privacy through optimal kanonymization. In ICDE-2005, 2005.
[12] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. In TCC, 2005.
[13] . H. Cox. Suppression, methodology and statistical disclosure control. Journal of the American Statistical Association, 75, 1980.
[14] T. Dalenius and S. Reiss. Data swapping: A technique for disclosure control. Journal of Statistical Planning and Inference, 6, 1982.
[15] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. Annals of Statistics, 1:363–397, 1998.