Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 11, November 2014, pg.144 – 148

SURVEY ARTICLE

A SURVEY ON DATA ANONYMIZATION TECHNIQUES FOR LARGE DATA SETS

Dhamodran.P[1], Priyadharsini.P[2], Kavitha.M.S[3]

Department of Computer Science and Engineering, India

[1] dhamodranp@gmail.com; [2] priyadharsini.pp@gmail.com; [3] kaviktg@gmail.com

Abstract— Data anonymization is used to remove user specific information from published data sets. Different kinds of anonymization techniques are used to eliminate various types of attacks. Anonymization process modifies the data into human unidentifiable form and it is most efficient than any other privacy preserving techniques like encryption etc. Encryption is costly when compared to anonymization as the time taken to decrypt and obtain the original data is huge. In this paper we discuss different kinds of anonymization techniques to protect the user privacy from published data sets.

Keywords— Anonymization, Encryption, Decrypt

I. INTRODUCTION

The data that is stored in cloud may contain user specific information. This information's must be protected in order to preserve the user's identity. The data contains the following identifiers such as explicit identifier, sensitive identifier, non-sensitive identifier and quasi identifier.

The explicit identifier provides the direct information about the record owner and combination of quasi identifier identifies the user specific information from the data record; sensitive identifier provides the details about record owners salary, health issues etc; All the other set of data's belong to the non-sensitive identifier.

Anonymization technique is used to hide these sensitive data's assuming that these user specific data's must be retained for future analysis. Attacks such as linking attacks can be prevented by anonymizing the quasi identifier before publishing the data. The quasi identifier is modified using anonymization operations on QID (quasi identifier) attributes in the original table. If still the data owner is linked using the modified QID, then the data is mapped to multiple records making the link ambiguous. There are many other types of attacks are possible on the data set which can be fixed using appropriate anonymization techniques.

II. TYPES OF ATTACK MODELS

The linkage of a particular data with the data in a published record can happen by the following attack models namely, table linkage, record linkage and attribute linkage .In all the three cases the adversary has some background knowledge about the victim that his record is present inside the published data or not. A published data is determined to preserve privacy if it can efficiently prevent the adversary from identifying user specific information. Another category of attack is known as the probabilistic attack where the published data along with background knowledge provides additional information about the victim.

A. Record Linkage Model

In record linkage model the victim's qid value on QID is matched with the records of the released table. Now this model identifies the group of records with the same qid value as that of the victim. The adversary by using the background knowledge about the victim has additional chances of identifying the user specific information from the identified group of records. For example: consider this two tables table I and II.

Table I. Original Patient Data

JOB	SEX	AGE	DISEASE
Professor	Male	33	Flu
Professor	Male	35	HIV
Dentist	Male	35	Hepatitis
Engineer	Female	37	Flu
Engineer	Female	37	Hepatitis
Lawyer	Female	38	HIV
Lawyer	Female	38	HIV

Table II. External Record

NAME	JOB	SEX	AGE
Doug	Lawyer	Female	38
Rancal	Engineer	Female	37
Niki	Professor	Male	35
Parker	Lawyer	Female	38
Ben	Engineer	Female	37
Logan	Dentist	Male	35
Micky	Professor	Male	33

Here table I is the published patient record to a research centre. Assume that the research centre has access to the external table II and also know that every patient in the published table has a record in the external table. Comparing the records in both the tables over common attributes like job, sex and age a link can be arrived at to establish the name of a patient affected by a particular disease.

1) K-Anonymity

K-Anonymity is a anonymization technique using which the record linkage through QID can be prevented. In k-anonymization some qid is assigned to a particular record and there will be k-1 other records with the same qid value. It means that the number of records with the same QID values is at least K-A table which satisfies this property is termed to be k-anonymous. In a k-anonymous table each qid is similar to at least k-1 qid's of the remaining records. it becomes difficult for an adversary to map a qid with an external table as the adversary may end up with mapping multiple records with the same value qid. The probability of mapping a victim's qid with the other records qid values is at most 1/k. The qid is a generalized value of the user specific information. For example table III shows the generalized QID values based on age of the patient.

Through these QID values.

Table III. Generalized Table

JOB	SEX	AGE	DISEASE
Professor	Male	30-35	Flu
Professor	Male	30-35	HIV
Dentist	Male	30-35	Hepatitis

Engineer	Female	36-40	Flu
Engineer	Female	36-40	Hepatitis
Lawyer	Female	36-40	HIV
Lawyer	Female	36-40	HIV

The generalization on QID is carried out by using the below taxonomy tree.

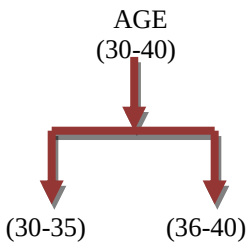

Fig. 1 Taxonomy tree

Now the adversary will end up with multiple records matching the victim's qid value through record linkage. The major disadvantage of this technique is that the k-anonymous property of the table will not hold in certain conditions like, a patient can contain more than one record if he is affected by more than a disease. Here the k-anonymous property of the table is violated.

2) (X-Y) Anonymity

The limitations of the k-anonymity is removed by X-Y anonymity, where X,Y are two disjoint sets.

Definition: Let x be the attribute in the set X. The Y(x) be the anonymity of x with respect to Y. Y(x) also denotes the number of distinct attributes that co-occur with the value x. For a given table T, if Y be the key value then Y(x) will be equal to the number of records containing the value x.

In this form of anonymization each X value is mapped to at least k different values on Y. The disadvantage of k-anonymity is the x value denotes the QID value and Y acts as the key in table T. It uniquely identifies the record owner through these QID values. X-Y anonymity is a special case was it provides a flexible way to denote different kinds of privacy requirements. If X represents the set of record owners and Y represents the sensitive values then each set of X-Y combination is mapped to diverse set of values. It becomes difficult to arrive at a user specific sensitive set of values.

B. Attribute Linkage Model

In attribute linkage model the adversary does not identify the record of the victim but he/she precisely identifies the sensitive values from the published record. This identification is made by associating the sensitive values of the victim with the group. The attribute linkage model can be eliminated using anonymization techniques like l-diversity etc.

1) l –Diversity

This model requires every qid group to have well represented sensitive values. A well represented sensitive value in a qid group should be associated with l distinct values. The disadvantage with this l -diversity model is it cannot eliminate probabilistic attack because some sensitive values occur frequently.

2) X-Y Linkability

(X, Y)-anonymity states that each group on X has at least k distinct values on Y. However, being linked to k persons does not imply that the probability of being linked to any of them is 1/k. If some Y values occur more frequently than others, the probability of inferring a particular Y value can be higher than 1/k. The (X, Y) - linkability below addresses this issue.

Definition: (X, Y)-linkability [8] Let x be a value on x and y be a value on Y. The linkability of x to y, denoted by $l_y(x)$, is the percentage of the records that contain both x and y among those that contain x, i.e., a(y, x)/a(x), where a(x) denotes the number of records containing x and a(y, x) is number of records containing both y and x. Let $L_y(X)= \max\{l_y(x) \mid x \in X\}$ and $L_Y(X)=\max\{L_y(X) \mid y \in Y\}$. We say that T satisfies the (X, Y)-linkability for some specified real $0<k\leq 1$ if $L_Y(X)\leq h$.

In words, (X, Y)-linkability limits the confidence of inferring a value on Y from a value on X. With X and Y describing individuals and sensitive properties, any such inference with a high confidence is a privacy breach. Often, not all but some values y on Y are sensitive, in which case Y can be replaced with a subset of y_i values on Y, written $Y =\{y_1,...,y_p\}$, and a different threshold h can be specified for each y_i. More generally, we can allow multiple Y_i, each representing a subset of values

on a different set of attributes, with Y being the union of all Y_i.Such a "value-level" specification provides a great flexibility essential for minimizing the data distortion.

3) (X, Y)-Privacy

Wang and Fung [8] propose a general privacy model, called (X, Y)-privacy, which combines both (X, Y)-anonymity and (X, Y)-linkability. The general idea is to require each group x on X to contain at least k records and the confidence of inferring any $y \in Y$ from any $x \in X$ is limited to a maximum confidence threshold h. Note, the notion of (X, Y)-privacy is not only applicable to a single table, but is also applicable to the scenario of multiple releases.

4) (α, k)-Anonymity

Wong et al. [9] propose a similar integrated privacy model called (α, k)-anonymity, requiring every qid in a Table T to be shared by at least k records and $conf(qid \rightarrow s) \leq \alpha$ for any sensitive value s, where k and α are data holder-specified thresholds. Nonetheless, both (X, Y)-Privacy and (α, k) -anonymity may result in high distortion if the sensitive values are skewed.

C. Table Linkage Model

Both record linkage and attribute linkage assume that the adversary already knows the victim's record. However, in some cases, the presence (or the absence) of the victim's record. Already reveals the victim's sensitive information. Suppose a hospital releases a data table with a particular type of disease. Identifying the presence of the victim's record in the table is already damaging. A table linkage occurs if an adversary can confidently infer the presence or the absence of the victim's record in the released table. The following example illustrates the privacy threat of a table linkage.

D. Probabilistic Model

There is another family of privacy models that does not focus on exactly what records, attribute values, and tables the adversary can link to a target victim, but focuses on how the adversary would change his/her probabilistic belief on the sensitive information of a victim after accessing the published data. In general, this group of privacy models aims at achieving the uninformative principle, whose goal is to ensure that the difference between the prior and posterior beliefs is small.

1) (c, t)-Isolation

Chawla et al. [5] suggest that having access to the published anonymous data table should not enhance an adversary's power of isolating any record owner. Consequently, they proposed a privacy model to prevent (c, t)-isolation in a statistical database. Suppose p is a data point of a target victim v in a data table, and q is the adversary's inferred data point of v by using the published data and the background information. Let δ_p be the distance between p and q. We say that point q(c, t)-isolates point p if $B(q, c\delta_p)$ contains fewer than t points in the table, where $B(q, c\delta_p)$ is a ball of radius $c\delta_p$ centered at point q. Preventing (c, t)-isolation can be viewed as preventing record linkages. Their model considers distances among data records and, therefore, is more suitable for numerical attributes in statistical databases.

2) Distributional Privacy

Motivated by the learning theory, Blum et al. [6] present a privacy model called distributional privacy for a non-interactive query model. The key idea is that when a data table is drawn from a distribution, the table should reveal only information about the underlying distribution, and nothing else. Distributional privacy is a strictly stronger privacy notion than differential privacy, and can answer all queries over a discretized domain in a concept class of polynomial VC-dimension, where Vapnik-Chervonenkis (VC) dimension is a measure of the capacity of a statistical classification algorithm. Yet, the algorithm has high computational cost. Blum et al. [6] present an efficient algorithm specifically for simple interval queries with limited constraints. The problems of developing efficient algorithms for more complicated queries remain open.

III. ISSUES IN ANONYMIZATION

Due to advancement and innovations in technologies the amount of data being generated is huge. it is a challenge for existing anonymization approaches to achieve privacy preservation on privacy-sensitive large-scale data sets due to their insufficiency of scalability.

IV. PROPOSED SOLUTION

To achieve improvement in anonymization over large scale data sets, anonymization techniques are used over Hadoop distributed environment using MapReduce computing model. This can efficiently anonymize large data sets in quick time.

V. CONCLUSION

In this paper we have discussed various anonymization methods which can efficiently eliminate different types of attack models on published data sets. The scalability issue in anonymizing large scale data sets is analysed and a solution to overcome this limitation over large data sets is also proposed .

REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. InProc. of the 10th International Conference on Database Theory (ICDT), pages 246–258, Edinburgh, UK, January 2005.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. InProc. Of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), Chicago, IL, June 2006.

[3] G. Aggarwal, N. Mishra, and B. Pinkas. Secure computation of the kth-ranked element. InProc. of the Eurocyrpt, 2004.

[4] E. Agichtein, L. Gravano, J. Pavel, V. Sokolova, and A. Voskoboynik. Snowball: A prototype system for extracting relations from large text collections.ACM SIGMOD Record, 30(2):612, 2001.

[5] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. Toward privacy in public databases. InProc. of Theory of Cryptography Con-ference (TCC), pages 363–385, Cambridge, MA, February 2005.

[6] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. InProc. of the 40th annual ACM Symposium on Theory of Computing (STOC), pages 609–618, Victoria,Canada, 2008.

[7] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. InProc. of ACM International Conference on Management of Data (SIGMOD), 1993.

[8] K. Wang and B. C. M. Fung. Anonymizing sequential releases. InProc.of the 12th ACM SIGKDD International Conference on Knowledge Dis-covery and Data Mining (SIGKDD), pages 414–423, Philadelphia, PA,August 2006.

[9] R. C. W. Wong, J. Li., A. W. C. Fu, and K. Wang. (α,k)-anonymity: An enhancedk-anonymity model for privacy preserving data publish-ing. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 754–759, Philadelphia,PA, 2006.

