ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 10, December 2012

A Survey on Anonymous Publication of Data

Jyoti Gajendra[1]
M. Tech. Scholar, Department of Computer Science and Engg. CSIT, Durg (CG) INDIA
Mr. Khom Lal Sinha[2]
Assistant Professor, Department of Computer Science and Engg. CSIT, Durg (CG) INDIA

***Abstract*— Data publication is one of major issue for organizations such as hospitals that publish detailed data about individual's e. g. medical records for research or analysis purpose. However, sensitive personal information may be disclosed in this process, due to the existence of data such as zip code, age etc. so in the data publishing privacy will be provided for data at different levels. There are different approaches that are used in data publishing. Data anonymization is one of the approach that modified original data before being publishes. In this paper, we analyze data publishing, anonymization approaches and operations that is used for publication of data for data mining task.**

***Index Terms*— Privacy, Anonymity, Sensitive- Information, Quasi Identifier.**

I.INTRODUCTION

The collection of information by governments, corporations and individuals has created many opportunities for knowledge-based decision making that require certain data to be exchange and published among various parties.

Data publishing is major factor because the original form of person-specific data contents sensitive information about individuals and publishing such data violates individuals privacy [1]. AOL published releases of query logs but quickly removed is due to the re-identification of a researcher [2].

The current practice release on guidance and policies to restrict publishable data types and use on agreements and storage of sensitive data but its limitation is that it requires excessively trust level that is impractically high in data sharing. This survey aims to achieve privacy preservation in data publishing and discussed different anonymized techniques to provide privacy in data publishing.

A. Data collection and data publishing

In data publishing, there is mainly two phase one is data collection and another is data publishing phase. In the data collection phase, the data publisher collect data from different record owners (e. g. Alice, Bob). In the data publishing phase, the data publisher releases the collected data for research or analysis purposes to the public that is called data recipient [1]. For example, a hospital collects data from patient and publishes the patient records to an external medical centre. In this example, the hospital is data publisher, patients are record owner and medical center is the data recipient. The data collection and data publishing is described in Fig.1.

B. Data publisher model

There are mainly two models of data publisher [3] that are trusted model and un-trusted model.

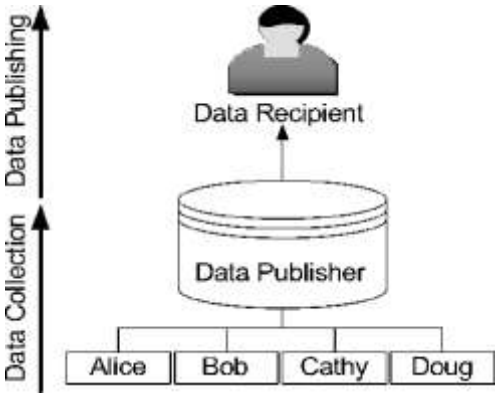

Fig.1: Data Collection and Data Publishing

1. Trusted model- In the trusted model, record owner to provide their personal information to the data publisher and the data publisher is trustworthy.

2. Un-trusted model- In the un-trusted model data publisher is not trusted and may attempt to identify record owner sensitive information.

C. Data anonymization

Data anonymization is modification of original data before being publishes. The terms anonymity means without name or nameless data that is used for privacy purpose. The original data table does not satisfy a specified privacy requirement and the data must be modified or anonymize before being published.

II.THE DATA ANONYMIZATION APPROCH

The data publisher has published data in the most basic form of table-

D(Explicit_Identifier,Quasi_Identifier, Sensitive_Attributes, Non-Sensitive_Attributes)

Where Explicit Identifier is a set of attributes, such as name and Security Number (SN), containing information that explicitly identifies record owners. Quasi Identifier (QID) is a set of attributes that could potentially identify record owners. Sensitive Attributes consists of sensitive person-specific information such as disease, salary, and disability status and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories [4].

Anonymization [5],[6] refers to the data publishing in privacy preserving approach that seeks to hide

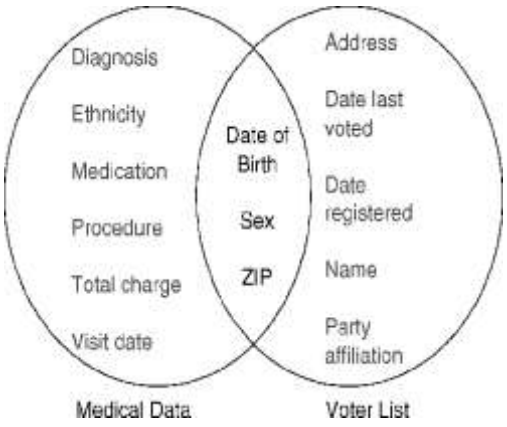

Fig.2: Linking to Re-Identify Record Owner

the identity and the sensitive data of record owners, assuming that sensitive data must be retained for data analysis and Explicit identifiers of record owners must be removed. In given example [7] a public voter list individual name was linked his record in a published medical database through the combination of zip code, date of birth and sex as shown in Fig. 2.

III. DATA ANONYMIZATION OPERATIONS

The modification to the table is done by applying a sequence of anonymization operations because the original table does not satisfy a specified privacy requirement and the table must be modified before being published. There are many anonymization operations that are as follows:

A. Generalization

Each generalization operation replace values of specific description, typically the QID attributes, with less specific attributes. It hides some details in QID. In Fig. 3 taxonomy tree for job is represented. In this taxonomy the parent node professional is more general than the child node doctor and lawyer and the root node, JOB_ANY represents the most general value in Job.

There are mainly five generalization schemes that are:

1. In full domain generalization scheme, all values in an Attribute are generalized to the same level of the taxonomy tree. For example, in Fig 3, if doctor and lawyer are generalized to professional, then it also requires generalizing dancer and painter to artist. The search space is minimum as compare to other scheme but the data distortion is largest.

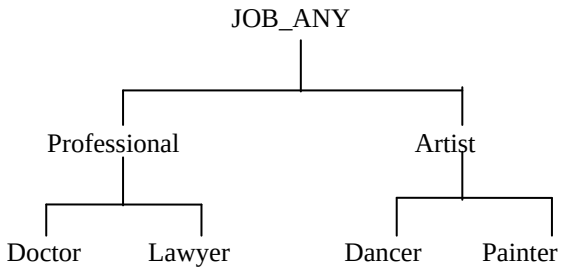

Fig. 3: Taxonomy Tree of Job

2. In sub-tree generalization scheme, at a non-leaf node either all child values or none are generalized. For example, in Fig 3, if Doctor is generalized to Professional, this scheme also requires the other child node, lawyer to be generalized to professional, but Dancer and Painter, which are child node of Artist.

3. Sibling generalization scheme is similar to the sub-tree generalization, except that some sibling may remain un-generalized.

4. In cell generalization scheme, if a value is generalized, all its instances are generalized.

B. Suppression

Suppression replaces some values with a special value, indicating that the replaced values are not disclosed. The reverse operation of suppression is called disclosure. There are also different suppression schemes. In Record suppression, it suppressing entire record. Value suppression, it suppressing every instance of a given value in table, and Cell suppression, it suppressing some instances of a given value in a table.

C. Anatomization

Anatomization [8] does not modify the quasi-identifier or the sensitive attribute, but de-associates the relationship between the two. Precisely, the method releases the data in two separate tables on QID and sensitive attribute. A quasi-identifier table (QIT) contains the QID attributes and a sensitive table (ST) contains the sensitive attributes and both QIT and ST have one common attribute, group ID. All records in the same group will have the same value on group id in both tables and therefore are linked to the sensitive values in the group in the exact same way. The anatomized tables can more accurately answer aggregate queries involving domain values of the QID and Sensitive Attributes as compare to generalization approach because the data in QIT and ST is unmodified.

D. Permutation

Permutation [9] is same spirit of anatomization. The basic idea is to de-associate a quasi-identifier and a numerical sensitive attribute relationship by partitioning a set of data records into groups and shuffling their sensitive values with in each group.

E. Perturbation

Perturbation is used to preserve statistical information due to its simplicity, efficiency and ability and it also used in statistical disclosure control [10]. The basic idea is to replace the original data values with same synthetic data values, so that statistical information computed from the statistical information computed from the original data. The perturbation approach limitation is that the published records are synthetic in that they do not correspond to the real-world entities represented by the original data.

IV. INFORMATION METRICS

Information metric is used to measure the utility of an anonymous table. A data metric measures the data quality in the entire anonymous table with respect to the data quality in the original table. A search metric guide each step of an anonymization algorithm to identify an anonymous table with maximum information. Information metric can be categorized by its information purposes, including general purpose, special purpose, or trade-off purpose.

ILoss is a data metric proposed [11] to capture the information loss of generalizing a specific value to a general value *vg*:

$$ILoss(vg) = |vg|-1/|DA| \quad (1)$$

Where $|vg|$ is the number of domain values that are descendants of *vg*, and $|DA|$ is the number of domain values in the attribute *A* of *vg*. This data metric requires all original data values to be at the leaves in the taxonomy.

If $ILoss(vg) = 0$ then *vg* is an original data value in the table. In words, $ILoss(vg)$ measures the fraction of domain values generalized by *vg*. For example, generalizing one instance of Dancer to Artist in Fig.3 has $ILoss(\text{Artist}) = 2-1/4 = 0.25$.

The loss of a generalized record *r* is given by

$$ILoss(r) = \sum_{v_g \in r}(wi \times ILoss(v_g)) \quad (2)$$

Where *wi* is a positive constant specifying the penalty weight of attribute *Ai* of *vg*.
The overall loss of a generalized table *T* is given by

$$ILoss(T) = \sum_{r \in T} ILoss(r) \quad (3)$$

A search metric [12], [13] based on the principle of information/ privacy trade-off. Suppose that the anonymous table is searched by iteratively specializing a general value into child values. Each specialization operation splits each group containing the general value into a number of groups, one for each child value. Each specialization operation *s* gains some information, denoted $IG(s)$, and loses some privacy, $PL(s)$. This search metric prefers the specialization *s* that maximizes the information gained per each loss of privacy:

$$IGPL(s) = IG(s)/PL(s) + 1 \quad (4)$$

The choice of $IG(s)$ and $PL(s)$ depends on the information metric and privacy model. For *k*-anonymity [12],[13] measured the privacy loss $PL(s)$ by the average decrease of anonymity over all *QIDj* that contain the attribute of *s*, that is,

$$PL(s) = avg\{A(QIDj) - As(QIDj)\} \quad (5)$$

Where $A(QIDj)$ and $As(QIDj)$ denote the anonymity of *QIDj* before and after the specialization. One variant is to maximize the gain of information by setting $PL(s)$ to zero. The principle of information/privacy trade-off can also be used to select a generalization *g*, in which case it will minimize

$$ILPG(g) = IL(g)/PG(g) + 1 \quad (6)$$

Where $IL(g)$ denotes the information loss and $PG(g)$ denotes the privacy gain by performing *g*.

V. ANONYMIZING DIFFERENT TYPES OF DATA

Anonymizing relational and statistical data is result in privacy and sensitive information leakages and non relational data that means other types of data like publishing high transaction data, moving object data, and textual data may also result in privacy threats and sensitive information leakages.

A. High-Dimensional Transaction Data

Publishing high-dimensional data is part of the daily operations in public and commercial activity. An example of high-dimensional data is transaction databases. Each transaction corresponds to a record owner and consists of a set of items selected from a large universe. Examples of transactions are web queries, click streams, e-mails, market baskets, and medical notes. Such data often contains rich information and is an excellent source for data mining. Detailed transaction data provides an electronic image of a record owner's life, possibly containing sensitive Information.

A recent case demonstrates the privacy threats caused by publishing transaction data: AOL released a database of query logs to the public for research purposes [2]. However, by examining query terms, AOL user No. 4417749 was traced back to Ms. Thelma Arnold, a 62-year-old widow who lives in Lilburn. Even. If a query does not contain an address or name, a record owner (the AOL user in this example) may still be re-identified from combinations of query terms that are adequately unique to the record owner. This scandal led not only to the disclosure of private information of AOL users, but also damaged data publishers' enthusiasm in offering anonymized transaction data for research purposes.

B. Moving Object Data

Moving object data have new challenges to traditional database, data mining, and privacy-preserving technologies due to its characteristics time-dependent, location-dependent, and is generated in large volumes of high-dimensional stream data. Location-based services (LBS) are information services provided to mobile subscribers based on their specific physical locations. Although the advancement of telecommunication technology has improved our quality of life, research has shown that 24% of potential LBS users are seriously concerned about the privacy implications of disclosing their locations in conjunction with other personal data [14].

Pid	Path	Disease
1	(a1- d2-b3-e4-f6-c7)	HIV
2	(b3-e4-f6-e8)	Flu
3	(b3-c7-e8)	Flu
4	(d2-c5-f6-c7)	Allergy
5	(c5-f6-e9)	HIV
6	(f6-c7-e9)	Fever

Table.1: Patient-Specific Path Table

The following example shows the privacy threats caused by publishing moving object data. For example, a hospital wants to release the patient-specific path table, Table 1.to a third party for data analysis. Explicit identifiers, such as patient names and Pid, have been removed. Each record contains the moving path of a patient in the hospital and some patient-specific (sensitive) information, for example, contracted diseases. A moving path contains a sequence of pairs (loci) indicating the patient's visited location loci at timestamp ti. For example, Pid#3 has a path b3 - c7 - e8, meaning that the patient has visited locations b, c, and e at timestamps 3, 7, and 8, respectively. An attacker seeks to perform record and/or attribute linkages by using the moving path as QID for matching.

1. Record linkage-

Suppose the attacker knows that the target victim, Alice, has visited e and c at timestamps 4 and 7, respectively. Alice's record, together with her sensitive value (HIV in this case), can be uniquely identified because Pid#1 is the only record that contains e4 and c7.

2. Attribute linkage-

Suppose the attacker knows that another target victim, Bob, has visited d2 and f 6, matching (Pid#1,4,5), the attacker can infer that Bob has HIV with 2/3 = 67% confidence.

C. Textual Data

Most previous work focused on anonymizing the structural or semi-structural data and the un-structural data, such as text documents [15] describes implicit and explicit privacy threats in text document repositories. Sanitization of text documents involves removing sensitive information or removing potential linking information that can associate an individual person to the sensitive information in a document. This research direction is in its infancy. A system[16] implemented for automatically anonymizing hospital discharge letters by identifying and deliberately removing all phrases from clinical text that satisfy some predefined types of sensitive entities. The identification phase is achieved by collaborating with underlying generic named entity recognition system.

VII. CONCLUSION

Information sharing has become part of the routine activity of many individuals companies, organizations and government agencies. The data publishing with preservation of privacy is major factor, while preserving individual privacy and protecting sensitive information. The general objective is to transform the original data into some anonymous form to prevent from inferring its record owners sensitive information. In this paper we reviewed existing method in data publishing, anonymization approach, anonymization operations and information metrics.

REFERENCES

[1] B. C. M. Fung, K. Wang And P. S. Yu, "Privacy-Preserving Data Publishing : A Survey of Recent Developments". ACM Computing Surveys 2010.
[2] M. Barbaro and T. Zeller, "A Face is Exposed for AOL Searcher No. 4417749". New York Times 2006.
[3] J. Gehrke, "Models and Methods for Privacy-Preserving Data Publishing and Analysis". Tutorial at
the 12th ACM SIGKDD 2006.
[4] L. Burnett, K. Barlow-Stewart,A. Pros, And H. Aizenberg, "The Gene Trustee: A Universal Identification System that Ensures Privacy and Confidentiality for Human Genetic Databases". J. Law and Medicine 10, 506–513 2003.
[5] L. H. Cox, "Suppression Methodology and Statistical Disclosure Control". J. Am. Statistical Assoc. 75, 370, 377–385 1980.
[6] T. Dalenius, "Finding A Needle in a Haystack - or Identifying Anonymous Census Record". J. Official
Statistics 2, 3, 329–336 1986.
[7] L. Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization And Suppression". Int. J. Uncertainty, Fuzziness, Knowl.-Based Syst. 10, 5, 571–588 2002.
[8] X. Xiao, And Y. Tao, " Anatomy: Simple And Effective Privacy Preservation". In Proceedings of the 32nd Conference on Very Large Data Bases (VLDB) 2006a.
[9] Q. ZHANG, N. KOUDAS, D. SRIVASTAVA, AND T. YU, "Aggregate query answering on anonymized tables". In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE) 2007.
[10] N. R. Adam, And J. C. Wortman, "Security Control Methods For Statistical Databases". ACM Comput. Surv. 21, 4, 515–556 1989..
[11] X. Xiao, And Y. Tao, "Personalized Privacy Preservation". In Proceedings of the ACM SIGMOD Conference. ACM, New York 2006b.
[12] B. C. M. Fung, K. Wang And P. S. Yu.. "Top-Down Specialization for Information and Privacy Preservation". In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE). 205–216 2005.
[13] B. C. M. Fung, K. Wang And P. S. Yu, "Anonymizing Classification Data For Privacy Preservation". IEEE Trans. Knowl. Data Engin. 19, 5, 711–725 2007.
[14] E. Beinat, "Privacy and Location-Based: Stating The Policies Clearly". GeoInformatic 2001s.
[15] Y. Saygin, D. Hakkani-Tur, and G. Tur,. Web and Information Security. IRM Press, 133–148 2006.
[16] D. Kokkinakis, and A. Thurin, "Anonymization Of Swedish Clinical Data". In Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME). 237–241 2007.

Authors

[1[Jyoti Gajendra

She , received her B.E. (Computer Sc.) in year 2010 and in pursuit for M.Tech. (Computer Sc.) from Chhatrapati Shivaji Institute of Technology (CSIT), Durg, Chhattisgarh, India. Her interests are Data Mining, Operating Systems and Network Security.

[2] Mr. Khom Lal Sinha

He received his B.E. (Information Technology) in year 2005 and M.Tech. in year 2012 (Computer Sc.) from Chhatrapati Shivaji Institute of Technology (CSIT), Durg, Chhattisgarh, India. His interests are Digital Image Processing, Operating Systems and Data Mining. Also he is having Life Membership of Indian Society of Technical Education, India (ISTE) and Institutional Member of Computer Society of India (CSI).