ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 4, April 2015

Privacy Preservation for Microdata by using K-Anonymity Algorthim

Sowmiyaa P[1], Tamilarasu P[2], Kavitha S[3], Rekha A[4], Gayathri R Krishna[5]

PG Scholar, Department of Computer Science and Engineering, Dr.N.G.P. Institute of Technology,

Coimbatore, Tamil Nadu, India[1]

Lecturer, Department of Information Technology, University College of Engineering,

Villupuram, Tamilnadu, India[2,3,4,5]

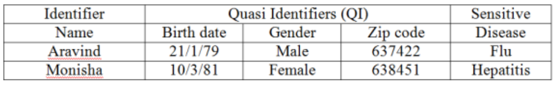
Abstract: Privacy for microdata is common problem in external database and data publishing. K-anonymity is one technique to protect micro data against linkage and identification of records. While in previous k-anonymity algorithms exist for producing k-anonymous data, due to privacy issues, the common data from different sites cannot be shared directly and assumes existence of a public database that can be used to breach privacy. During anonymization process, public database are not utilized. In existing generalization algorithm creates anonymous table by using microdata table and focusing on identity disclosure, and k-anonymity model fails to protect attribute disclosure to some extent. In propose two new privacy protection models called (p, α)-sensitive k-anonymity and (p+, α)-sensitive k-anonymity, respectively. It is different from previous the p-sensitive, these new introduced models allow us to release a lot more information without compromising privacy. Moreover, we prove that the (p, α)-sensitive and (p+, α)-sensitive k-anonymity problems are NP-hard and include testing and heuristic generating algorithms to generate desired micro data table.

Keywords: Microdata, Privacy, k-anonymity, k-join-anonymity,(p, α)-sensitive k-anonymity and (p+,α)-sensitive k-anonymity**.**

I. INTRODUCTION

Many data holders publish their micro data for different reasons. So, they have difficulties in releasing information which does not compromise privacy. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields i.e name, address, etc. Although, joining this de-identified table with a publicly available database (like the voters database) on attributes like race, age, and zip code (usually called quasi-identifier) can be used to identify individuals. k-anonymity model fails to protect attribute disclosure [7]. Several models such as p-sensitive k-anonymity [19], l-diversity [10], (α, k)-anonymity [22] and t-closeness [9] were proposed in the literature in order to deal with the problem of k-anonymity. The work presented in this paper is highly inspired by [19].The main contribution of [19] is to introduce the p-sensitive k-anonymity property, which requires, in addition to k-anonymity, that for each group of tuples with identical combination of quasi-identifier values, the number of distinct sensitive attributes values must be at least p. However, depending on the nature of the sensitive attributes, even p-sensitive property still permits the information to be disclosed. We identify in this paper situations when p-sensitive property is not enough for privacy protection and we propose two solutions to overcome this identified problem: (p, α)-sensitive and (p+,α)-sensitive k-anonymity models and the heuristic algorithms to enforce these properties. We introduce some basic concepts and p-sensitive k-anonymity model. Enhanced p-sensitive k-anonymity models are discussed.

Several concepts have been proposed to achieve privacy preservation. Most database literature has focused on k-anonymity [8], [10]. Specifically, a table T is k-anonymous if each record is indistinguishable from at least k-1 other tuples in T with respect to the QI set. The process of generating a k-anonymous table given the original microdata is called k-anonymization. The most common form of k-anonymization is generalization, which involves replacing specific QI values with more general ones. The output of the generalization is an anonymized table AT containing anonymized groups, each including at least k tuples with identical QI values.



Identifier	Quasi Identifiers (QI)			Sensitive
Name	Birth date	Gender	Zip code	Disease
Aravind	21/1/79	Male	637422	Flu
Monisha	10/3/81	Female	638451	Hepatitis

The concept of K-join-anonymity permits the utilization of existing generalization techniques and creates the join table by using both microdata table and public database. In join table unique IDs and SAs are removed. Protects the microdata against the linkage and identification of records during the data publishing.

II. PROBLEM DEFINITION

All the previous k-anonymity techniques are not utilizing the existence of a Public Database during the anonymization process. This omission leads to unnecessarily high information loss. Grouping the fields that contain tuples with different quasi identifiers values. Publicly available databases (voter lists, city directories) can reveal the "hidden" identity [2], [3]. Attacker can re-

DOI 10.17148/IJARCCE.2015.4483

identify the sensitive information by using background knowledge. By using micro data table alone it's create the anonymous table. Not utilizes the Public Database for generalization algorithm [5]. Sensitive attributes are not consider while data publishing. Attackers have the Background Knowledge about microdata by using public database (example: voter list, city directory). Unnecessarily high information loss.

In the example 2 attacker use the public database (voter list) to know the details of the microdata like medical data. By using background knowledge attacker know the privacy data. In the existing system there is possible for information leakage and data linking and identification problem. In this anonymous table created for microdata table alone. To avoid the identification of records in microdata, uniquely identifying information like names and social security numbers are removed. However, this first sanitization still does not ensure the privacy of individuals in the data. A recent study estimated that 87% of the population of the United States can be uniquely identified using the seemingly innocuous attributes gender, date of birth, and zip code [9]. In fact, those three attributes were used to link voter registration records are name, gender, zip code, and date of birth to anonymized medical data are included gender, zip code, etc. This "linking attack" managed to uniquely identify the medical records of the individual [10].

III. K-JOIN-ANONYMITY

Definition 1. (Quasi-identifier). A set of nonsensitive attributes {Q1,…,Qw} of a table is called a quasi-identifier if these attributes can be linked with external data to uniquely identify at least one individual in the general population.

Definition 2. The schema of a microdata table (MT) consists of the unique ID, QI and sensitive attributes.

Definition 3. The schema of a public database (PD) consists of the unique ID and all QI attributes appearing in MT. Using PD, the attacker identifies the QI values of an individual.

The concept of K-anonymity utilization of existing generalization techniques and protects the microdata against the linkage and identification of records during the data publishing. Anonymized table (AT) is created by using microdata table and not utilizing public database. Identifiers and sensitive information are removed and generalization is performed in the AT. Sets of attributes are gender, date of birth, etc., that can be linked with external data to uniquely identify individuals information are called quasi-identifiers. The counter linking attacks using quasi-identifiers a table satisfies k-anonymity if every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table.

It proposes the privacy is common problem while data publishing and K-anonymity methods of Privacy Protection have great influence on the data precision [6]. Experimental results show that the improved algorithm of K-anonymity model increases the data precision effectively. It proposes and evaluates an optimization algorithm for de-identification of data is known as k-anonymization [7]. All the previous work has shown the necessity of considering an attacker's background knowledge when reasoning about privacy in data publishing [3].

K-join-anonymity permits the utilization of existing generalization techniques and protects the micro data against the linkage and identification of records during the data publishing. Join table is created by using both micro data table and public database. Identifiers and sensitive information are removed from the join table. Reduce the loss of information and provide privacy for micro data by utilizing the public database. The goal of k-join-anonymity is to provide the same privacy guarantees with k-anonymity incurring, however, less information loss. To achieve this, it shrinks the G-boxes using public knowledge about universe (U) tuples. In some applications, the entire U is available to the publisher, e.g., company payroll. First generalizes the combination of microdata table and public database under the constraint that each group should contain at least one tuple of microdata table. Second anonymized microdata table, and then refines the resulting groups using public database.

Definition 4. (k-join-Anonymity) A table T satisfies k-join-anonymity if for every tuple tϵT there exist k - 1 other tuples$t_{i1}, t_{i2}, \ldots\ldots\ldots, t_{ik-1} \in T$such that$t[c] = t_{i1}[c] = t_{i2}[c] = \cdots = t_{ik-1}[c]$for all$C \in QI$.

The Anonymized Table T*. Since the quasi-identifiers might uniquely identify tuples in T, the table T is not published; it is subjected to an anonymization procedure and the resulting table T*is published instead.

Definition 5. Ananonymized table AT of join table is k-join-anonymous if the mapping of each record in join table is indistinguishable among the mapping of at least k-1 other join table tuples.

Definition 6. (**Distance between two numeric values**) Let D be a finite numeric domain. Then the normalized distance between two values v_i, $v_j \in D$ *is defined as:*
$\delta_N (v_1, v_2) = |v_1 - v_2| \ / \ |D|$,where $|D|$ is the domain size measured by the difference between the maximum and minimum values in D.

Top down Greedy Algorithm

```
1. IF |T| ≤ k THEN
2. RETURN;
3. ELSE   {
4. Partition T into two exclusive subsets T1
and T2 such that T1 and T2;
5. IF |T1| > k THEN
```

DOI 10.17148/IJARCCE.2015.4483

```
6. recursively partition T1;
7. IF |T2| > k THEN
8. recursively partition T2;}
9. Adjust the groups so that each group has at least k tuples;
```

K-Join-Anonymous Algorithm

```
1.  read quasi-identifier from MT, RT and JT is empty.
2.  read quasi-identifier from PT, RT and JT is empty.
3.  FOR i=1 to n DO
4.  JT= (MT,PT);
5.  FOR i=1 to m DO
      a) marked 0 on the Tuple of table T;
      b) read into an Tuple;
      c) FOR j=1 TO m DO to find the Tuple which contain
         the attribute most close to other tuple;
      d)The Tuple of the smallest mark do with a
         generalization, and be integrated into the RT;
      e) Repeat the step 4 until all tuples of JT were
         generalized;
```

6. Output the table of RT.

On a data table, replaced the original value of attribute with another value that can indicate a larger geographical area and have the same semantic, this process is known as a generalization [6]. For example, zip= 637408 can become zip= 63740* and zip=63740* and zip=63742* can become zip=6374**, the generalization value and the original value maintain the right consistency and expand the area represented by the attribute. In relational database system, a domain are used to present of attribute a set of value that attributes assume, in order to facilitate the description of generalization on the attribute, there is need to expand the concept of attribute domain. The original table of data is as specific as possible, but in order to achieve K anonymous, it is necessary to generalize the original data, so reached the level of a more wide. After a generalization, a set of attribute value become a high-level domain. for example, in Figure 2 the zip code 637408 is located in the bottom of the domain Z0, generalization of the zip is refers to more widely domain, with Z1 instead of Z0, the operation can be considered from Z0 to Z1 mapping637408 → 63740*.

IV. CONCLUSION

In existing generalization algorithm p-sensitive satisfied by micro data sets, can help increase the privacy of the respondents whose data is being used but this property is not enough for protecting sensitive attributes. So we introduced new concept against Similarity Attack, namely (p, α)-and (p+,α)-sensitive k-anonymity models. Our results show that our proposed models could significantly reduce the possibility of Similarity Attack and incur less distortion ratio compared with previous p-sensitive k-anonymity model.

REFERENCES

[1]. J. Domingo-Ferrer and V. Torra, Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery. v11. 195-212.

[2]. B. Fung, K. Wang, P. Yu. Top-down specialization for information and privacy preservation. In Proc. of the 21st International Conference on Data Engineering (ICDE05), Tokyo, Japan.

[3]. Xiaoxun Sun Hua Wang Lili Sun, "Extended K-Anonymity Models Against Attribute Disclosure", 2009 Third International Conference on Network and System Security.

[4]. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas and A. Zhu. Anonymizing tables. In Proc. of the 10th International Conference on Database Theory (ICDT05), pp. 246-258, Edinburgh, Scotland.

[5]. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu. Approximation algorithms for k-anonymity. Journal of Privacy Technology, paper number 20051120001.

[6]. D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y.Halpern, "Worst-Case Background Knowledge for Privacy-Preserving Data Publishing," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[7]. Mohammad Reza ZareMirakabad, AmanJantan, "Diversity versus Anonymity for Privacy Preservation".Proc. IEEE, pp. 978-1-4244-2328-6, 2008.

[8]. JianXu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, Ada Wai-Chee Fu " Utility-Based Anonymization Using Local Recoding", KDD'06, August 20--23, 2006.

[9]. Song Ren-jie, Lei Zhong-yue and Feng Liang-tao, "An Improved K-anonymity Algorithm Model", The 1st International Conference on Information Science and Engineering (ICISE2009)"

[10]. R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proceedings of the 21st International conference on Data Engineering (ICDE 2005) 1084-4627/05."

[11]. P. Samarati. Protecting respondents' Identities in Microdata Release," In IEEE Transactions on Knowledge and Data Engineering, 2001.

[12]. L. Sweeney. Uniqueness of simple demographics in the u.s. population. Technical report, Carnegie Mellon University,2000.

[13]. L. Sweeney. k-anonymity: A model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557–570, 2002.