Privacy Measure for Publishing the Data- A Case Study

Chinta Someswara Rao, Bhadri Raju MSVS

Abstract— Privacy-maintaining data release is one of the most important challenges in an information system because of the wide collection of sensitive information on the World Wide Web (WWW). Many solutions have been proposed by several researchers for privacy-maintaining data release. This paper provides an inspection of the state-of-the-art methods for privacy protection. The paper discusses novel and powerful privacy definitions which can be categorized into micro data anonymity methods and differential privacy methods for privacy-maintaining data release. The methods include K-anonymity, L-diversity, T-closeness and JS-reduce defense. This paper proposes a study which will provide sequential background knowledge and provides some anonymization.

Index Terms— WWW, Privacy preserving, K-Anonymity; L-Diversity; T-Closeness; JS-Reduce.

I. INTRODUCTION

The collection of information by governments and corporations has created massive opportunities for knowledge-based decision making. Driven by mutual benefits, or by regulations that require certain data to be published, there is a demand for the exchange and publication of data among various parties. For example, Netflix, the world's largest online DVD rental service, recently published a data set contains 100M ratings of 17K movies by 500K users, announced the 1-million Netflix Prize in a drive to improve the quality of movie recommendations based on user preferences [1]. Releasing data to the public or other parties for research is an inevitable trend and has substantial benefits to the company and the society. However, such activities have been strongly opposed by their users since the released data often contain their sensitive information and by publishing data directly, will violate users' privacy. Hence users argue that their safety of integrity would be intruded and the privacy issue has been raised with increasing importance today. This undertaking is in the scope of privacy preserving data publishing [2].

A typical privacy preserving data publishing scenario is described in the figure 1. Assume there is a centralized trusted server, called data publisher, who has a collection of data from users and wants to release the collected data to a data miner or to the public for research or other purposes. A task of the utmost importance here for the data publisher is to anonymize data before it being published such that the data recipient cannot learn the privacy information about users while still can get meaningful data and perform data mining activities in a decent accuracy. One trivial anonymization method is that before dataset to be released, user names and IDs are replaced with random numbers or simply removed.

Manuscript received March 2014.

China Someswararao, belong to Department of CSE, SRKR Engineering College, Bhimavaram, West Godavari, AP, India.

Dr. Bhadri Raju MSVS, belong to Department of CSE, SRKR Engineering College, Bhimavaram, West Godavari, AP, India.

However, this kind of trivial anonymization is not good enough to protect users' privacy. Private or sensitive user information can still be mined from the remaining user data, so called re-identification [3]. For example, Netflix disclosed what it considered was anonymize user data to those trying to come up with solutions. This, however, led to a lawsuit by a mother who argued that Netflix had not sufficiently anonymize the information and that she (among others) could be easily outed according to her own rental history. Indeed, within weeks of the data being released, researchers like Narayanan and Shmatikov had found a way to use an external data source (e.g. IMDb) to decode an individual's viewing history with surprising accuracy [4].

II.PRIVACY MODELS INVESTIGATION

In this section, we investigate different privacy models for privacy preserving data publishing as well as a brief discussion on pros and cons of each method. Furthermore, we state theoretical challenges in high dimensionality of data along with related work for overcoming it. Publishing personal micro-data for analysis or seeking better data mining performances, while maintaining individual privacy, is a problem of increasing importance today. Privacy preserving data publishing (PPDP) techniques have been deserved serious thinking and widely studied in recent years. Several methods and models have been developed toward PPDP, such as randomization and *k*-anonymity. Furthermore, the problem has been discussed in several communities such as the database community, the statistical disclosure control (SDC) community and the cryptography community. A survey on some of the techniques used for privacy preserving data mining (publishing) may be found in [5,6]. In [6], authors classified privacy preserving techniques based on five dimensions: data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation.

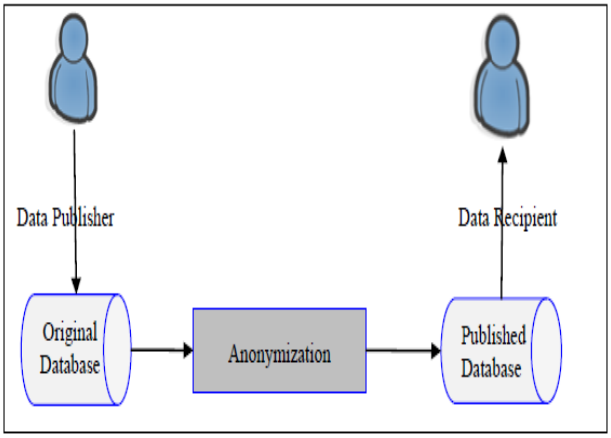

Figure 1 a Simple Model of PPDP

A. K-ANONYMITY

K-anonymity method is proposed by V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. In k-anonymity, data is published in such a way that each record is identical to k-1 other records. That data is published in a group of k records [7]. Two techniques which provide data anonymity are generalization and suppression [8]. The peculiarity of generalization and suppression is that they will maintain truthfulness of the information. Generalization is the method of substituting the given attribute with more general value. For this, the concept of domain, which is the set of values that an attribute can accept, is extended to a set of generalized domains. The original domains along with their generalizations are referred to as Dom. Each generalized domain contains generalized values and mapping between each domain and its generalizations. Another method to obtain k-anonymity is suppression which is applied along with generalization. This will moderate the generalization process when tuples with less number of occurrences undergo a greater amount of generalization. Therefore we can say that generalization is applied to attribute (column) level and suppression is applied to tuple (row) level. The generalization and suppression together provides more general table which provide more privacy to the individuals.

B. L-Diversity

L-Diversity method is proposed by Ashwin Machanavajjhala, Johannes Gehrke and Daniel Kifer. L-diversity Principle is that "a q-block is l-diverse if contains at least l values for the sensitive attribute S. The definition means that each block should contain l different sensitive value. The parameter L can be set depends on how much protection the publisher wants. But the above distinct l-diversity does not prevent probabilistic attack [9]. L-Diversity can be instantiated further as follows.

1. Entropy l-Diversity:

The entropy l-diversity is mathematically represented as follows [9].

$$(q,s)*\log((q,s))s\in S\geq \log(l)$$

where $p(q,s)= n(q,s)\,\Sigma\, n(q,s)\; s \in S$ is the fraction of tuples in the q-block with sensitive attribute value equal to s. this equation represents the entropy of the l-diversity. The entropy gives the average information contained in table. One point that can be infer from above definition of Entropy l-Diversity is that in order to have entropy l-diversity for each equivalence class, the entropy of the entire table must be at least log(l).

2. Recursive (c, l)-Diversity:

Let $s_1 .. . s_m$ be the possible values of the sensitive attribute S in a q-block. Sort the counts $n(q, s_1), . . . , n(q, s_m)$ in descending order and name the elements and results in sequence $r_1, . . . , r_m$. Let T_i denote the number of times the i^{th} most frequent sensitive value appears in that q-block. Given a constant c, the q-block satisfies recursive (c, l)-diversity if $T_1 < c(T_l + T_{r+1} + . . . + T_m)$. That is, q-block satisfies recursive (c, l) - diversity if we can eliminate one possible sensitive value in the q-block and still have a (c, l−1)-diverse block.

C. T-Closeness

T-closeness is proposed by Ninghui L,i Tiancheng Li and Suresh Venkata subramanian. This method proposed as solution to attribute disclosure. T-closeness can be defined as follows. Spreading of sensitive attributes in each quasi-identifier group should resemble to their distribution in whole original database [10]. That is the distance between distribution of the attribute in the whole table and distribution of a sensitive attribute in this class should not be more than a threshold. This method limits the correlation between sensitive attribute and quasi-identifier attribute.

D. JS-Reduce Defense

The methods such as k-anonymity, l-diversity and t-closeness do not consider the sequential background knowledge of the adversary. So another method called JS-Reduce has been proposed which considers adversaries background knowledge in serial microdata release that is background obtained by adversary in serial release of database [11]. The method was proposed by Daniele Riboni, Linda Pareschi and Claudio Bettini. In this method first a model created to find out the background knowledge, Posterior background knowledge and revised sensitive values background knowledge. Adversary's background knowledge is revised each time when data is released. The main goal of the method is to maximize the similarity of probability distribution of sensitive value. For that Jenson-Shannon divergence is used.

III. RECOMMENDER SYSTEM

This system mainly studies the privacy preserving data publishing for recommender systems (e.g. Netflix Prize dataset). To help consumers make intelligent buying decisions, lots of websites provide so called recommender systems. Recommender systems form or work from a specific type of information filtering system technique that attempts to recommend information items (films, television, video on demand, music etc.) that are likely to be of interest to the user. Typically, recommendations are usually based on user ratings and logs. These data may contain sensitive user information such as buying history and movie-rating records. One of the most commonly used recommendation technique is called "collaborative filtering", by which predicts a subscriber's future choices from his past behavior using the knowledge of what similar consumers did. i.e. try to make prediction of values of some items using a combination of other attributes. Before data publishing privacy called to set security code. Each and every person need to register and get security code. Another one is the public semantic searching and getting result for public person. This public person is not considered anonymous. Clearly, the released data containing such information about individuals should not be considered anonymous. Sometimes getting information via searching in particular/filter particular name wise can't visible full information for the public person. Incase public person search for a particular person information, the result is each and every splitting data's then blocking or set substring of asterisk (*) using l-diversion and closeness. Here public person or unauthorized person is considered anonymous. We can analyze the percentage of possible privacy loss.

IV. IMPLEMENTATION

In this paper an interactive user interface program is developed to extract data from neutral format using JSP and this proposed system consists of four modules.

Publishing privacy: Doesn't need to set security for your publishing data's but yours is safe. Administrator only can see full details.

The third party can't fully details. Third searching for people in this records database can view splitting/blocking records using l-diversion and closeness and actual data is shown in Table 1.

Table 1 Patient data

NAME	AGE	COUNT	PINCODE	DIESES
kanmani	23	1	534201	cold
raju	23	1	534203	cold
hema	24	2	534202	fever
sekar	28	1	534209	flu
chandran	27	1	534207	tumor
guhan	28	2	534206	cold
sheela	29	1	534208	fever

L-diversion and closeness: L-diversion and closeness is derived formula can using secured data publishing. Here using the formula of l-diversion **Error! Reference source not found.**. Logically we will process this formula getting data's recursively then splitting row wise data's. For example *Count* that indicates the number of individuals.

Closeness: To be calculate the distance closeness and checking for this privacy process.

Identity of indiscernible: An adversary has no information gain if her belief does not change.

Non-negativity: When the released data is available, the adversary has a non-negative information gain.

Probability scaling: The belief change from probability α to $\alpha+\gamma$ is more significant than that from β to $\beta+\gamma$ when $\alpha < \beta$ and α is small.

Zero-probability definability: D[P,Q] should be well-defined when there are zero probability values in P and Q.

These are all data for searching with different things as shown in Table 2

Table 2 closeness measure

	l-diversion and closeness	Non-negativity	Zero-probability definability
Search with Name			
raju	0.11111111	7	6
Hema	0.22222222	7	6
chandran	0.11111111	7	6
Search with Dieses			
cold	0.44444445	7	4
fever	0.22222222	7	6
Search with Pincode			
534201	0.11111111	7	6
534202	0.5555556	7	3
Search with age			
23	0.22222222	7	5
28	0.33333334	7	5

Data Processing: This is one of property for designing the distance Measure. Searching for in particular is set on substring of asterisk (*). We can identify easy to see closeness ratio. L-diversion and closeness is very low the security mode very high. Incase l-diversion and closeness is very high the security mode is very low.

V.CONCLUSION

Privacy preservation in data publishing is one of the tedious tasks in data publishing. This survey describes several existing data publishing methods such as k-anonymity, l-diversity, t-closeness, JS-Reduce. Among these methods JS reduce is the only method which models sequential background knowledge attack. The proposed study provides better model which consider sequential background knowledge attack as well as anonymize data which provides better privacy protection to individual.

REFERENCES

1. Hafner K. And if you liked the movie, a Netflix contest may reward you handsomely [Report] , New York Times, 2006.
2. Fung, Benjamin, Ke Wang, Rui Chen, and Philip S. Yu. , "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys (CSUR) 42, 2010.
3. Samarati, Pierangela. , "Protecting respondents identities in microdata release", IEEE Transactions on Knowledge and Data Engineering, pp.1010-1027,2001.
4. Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets", IEEE Symposium on Security and Privacy,pp. 111-125,2008. IEEE, 2008.
5. Verykios, Vassilios S., Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. , "State-of-the-art in privacy preserving data mining", ACM Sigmod Record 33, pp. 50-57, 2004.
6. Aggarwal, Charu C., and S. Yu Philip. A general survey of privacy-preserving data mining models and algorithms. Springer US, 2008.
7. Sweeney, Latanya. , "K-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, pp. 557-570, 2002.
8. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "K-Anonymity", Springer US, Advances in Information Security, 2007.
9. Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), 2007.
10. Ninghui Li, Tiancheng Li; Venkatasubramanian, S, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", IEEE 23rd International Conference on Data Engineering , pp.106-115, 2007.
11. Riboni, D, Pareschi, L,Bettini, C, "JS-Reduce: Defending Your Data from Sequential Background Knowledge Attacks", IEEE Transactions on dependable and Secure Computing, pp.387-400, 2012.