www.ijmer.com Vol. 3, Issue. 3, May.-June. 2013 pp-1763-1767 ISSN: 2249-6645

DISCLOSURE PREVENTION IN PRIVACY-PRESERVING DATA PUBLISHING

V.Kavitha[1] , M.Poornima[2]

**(Department of Information Technology, Paavai College of Engineering, Anna University, India)*
*** (Department of Computer Sciences and Engineering ,Anna University, India)*

ABSTRACT: *The advancement of information technologies has enabled various organizations (e.g., census agencies, hospitals) to collect large volumes of sensitive personal data (e.g., census data, medical records. Data in its original form, however, typically contains sensitive information about individuals, and publishing such data will violate individual privacy. The current practice in data publishing relies mainly on policies and guidelines as to what types of data can be published and on agreements on the use of published data. In order to protect sensitive information, the simplest solution is not to disclose the information. However, this would be overkill since it will hinder the process of data analysis over the data from which we can find interesting patterns. Moreover, in some applications, the data must be disclosed under the government regulations. Alternatively, the data owner can first modify the data such that the modified data can guarantee privacy and, at the same time, the modified data retains sufficient utility and can be released to other parties safely. This process is usually called as privacy-preserving data publishing. This thesis identifies a collection of privacy threats in real life data publishing, and presents a unified solution to address these threats.*

KEYWORDS: *Privacy, Sensitive data, Data Publishing, Information security.*

I. INTRODUCTION

In recent years, advances in hardware technology have lead to an increase in the capability to store and record personal data about consumers and individuals. This has lead to concerns that the personal data may be misused for a variety of purposes. In order to alleviate these concerns, a number of techniques have recently been proposed in order to perform the data publishing tasks in a privacy-preserving way. A task of the utmost importance is to develop methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved. This undertaking is called *privacy-preserving data publishing* (PPDP).In the past few years, research communities have responded to this challenge and proposed many approaches.

This paper is organized as follows, section 2 deals with the analysis of data .Section 3 discusses about various protection methods.

Section 4 deals with the various limitations of the privacy models, while section 5 deals with enhancing the anonymization methods. Section 6 concludes the featured description of various protection methods.

II. PRIVATE DATA ANALYSIS

Data analysis is the process of extracting hidden predictive information from large amount of datasets. This analysis can be performed by the data owner or the data owner can outsource the data analysis to other parties. In any case, the privacy concerns of the involved individuals should be addressed and considered at all times. According to Michal Sramka [2010], private data analysis is achievable in the following ways:

II.1 PRIVATE DATA ANALYSIS OVER ORIGINAL DATA

In this scenario, computations are performed over the original private or even confidential data.

- Data analysis is performed by the data owner. No other party will learn the data, and the results of the analysis will stay "in house".
- Data mining is performed over the original data and then the obtained knowledge is published. The published knowledge is protected against privacy leaks in a way that it does not reveal sensitive information about the underlying data. This is achievable by sanitizing the learned knowledge and referred to as the privacy-preserving knowledge publishing as stated by Atzori et al [2008].
- One or several parties own confidential data and another party perform a computation over them. According to Lindell [2002] Secure multiparty computation over distributed data sets are fields that study cryptographic tools that allow to compute a function over confidential data without learning anything else than what can be learned from the output of the function.

II.2 DATA ANALYSIS OVER SANITIZED DATA

In this scenario, data is sanitized and then shared or published for analysis. This is referred to as privacy preserving data publishing (PPDP). Sanitization is usually achieved as a transformation of the data that

www.ijmer.com Vol. 3, Issue. 3, May.-June. 2013 pp-1763-1767 ISSN: 2249-6645

provides pseudonymity, anonymity, or privacy risk reduction by generalizing, masking, randomizing, or even suppressing some data.

III. CLASSIFICATION OF MICRO DATA PROTECTION METHODS

A microdata set can be viewed as a file with *n* records, where each record contains *m* attributes on an individual respondent. The attributes can be classified in four categories which are not necessarily disjoint:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are the passport number, social security number, name, surname, etc.
- *Quasi-identifiers or key attributes*. These are attributes which identify the respondent with some degree of ambiguity. (Nonetheless, a combination of quasi-identifiers may provide unambiguous identification.) Examples are address, gender, age, telephone number, etc.
- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes*. Those attribute which do not fall in any of the categories above.

In recent years, numerous algorithms have been proposed for implementing *k*-anonymity via generalization and suppression. Samarati [2001] presents an algorithm that exploits a binary search on the domain generalization hierarchy to ¯find minimal *k*-anonymous table. Sun *et al.* [2008] recently improve Samarati's algorithm by integrating the hash-based technique. Bayardo and Agrawal [2005] present an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal *k*-anonymous table, exploiting ad hoc pruning techniques. LeFevre *et al*. [2005] describe an algorithm that uses a bottom-up technique and a priori computation. Fung *et al*. present a top-down heuristic to make a table to be released *k*-anonymous. As to theoretical results, Meyerson and Williams [2004] and Aggarwal *et al*
prove that optimal *k*-anonymity is NP-hard (based on the number of cells and number of attributes that are generalized and suppressed) and describe approximation algorithms for optimal *k*-anonymity. Sun *et al*. [2008] prove that *k*-anonymity problem is also NP-hard even in the restricted cases. While focusing on identity disclosure, *k*-anonymity model fails to protect attribute disclosure. Several

Gender	Age	Zip Code	Diseases
Male	25	4370	Hypertension
Male	25	4370	Hypertension
Male	22	4352	Depression
Female	28	4373	Chest Pain
Female	28	4373	Obesity
Female	34	4350	Flu

Table 1: The microdata

Gender	Age	Zip Code	Diseases
Male	22-25	43**	Hypertension
Male	22-25	43**	Hypertension
Male	22-25	43**	Depression
Female	28-34	43**	Chest Pain
Female	28-34	43**	Obesity
Female	28-34	43**	Flu

Table 2: A 3-anonymous table

models such as *p*-sensitive *k*- anonymity, *l*-diversity, (α *k*)-anonymity and *t*-closeness are proposed in the literature in order to deal with the problem of *k*-anonymity. Although these models can achieve privacy properties to some extent, they are not enough for privacy protection.

A key difficulty of data anonymization comes from the fact that data utility (i.e., data quality) and data privacy are conflicting goals. Intuitively, data privacy can be enhanced by hiding more data values, but it decreases data utility; on the other hand, revealing more data values increases data utility, but it may decrease data privacy. Thus, it is necessary to devise solutions that best address both the utility and the privacy of data.

Publishing high dimensional data is part of daily operations in commercial activities and public services. A classic example of high dimensional data is transaction databases. Examples of transactions are web queries, click streams, emails, market baskets, and medical notes. Such data often contain rich information and are excellent sources for data mining. Narayanan and Shmatikov showed that an attacker only needs a little bit information of an individual to identify the anonymized movie rating transaction of the individual in the data set. Such breach occurs when an attacker only needs a little bit information of an individual to re-identify the anonymized rating transaction of the individual in the data set. Existing research on privacy-preserving data publishing focuses on relational data and the objective is to enforce privacy-preserving paradigms (e.g., *k*-anonymity, *l*-diversity, etc) while minimizing the information loss incurred in the anonymizing process. However, methods developed on low dimensional relational data are very inefficient on high dimensional and sparse transactional data.

www.ijmer.com Vol. 3, Issue. 3, May.-June. 2013 pp-1763-1767 ISSN: 2249-6645

III.1 K-ANONYMITY: An anonymization algorithm finds a release candidate that is both useful and safe (according to privacy criterion) from this space of search. K-anonymity is defined as:

Each release of the data must be such that every combination of values of quasi-identifiers can be indistinguishably matched to at least k respondents.

The approach uses domain generalization hierarchies of the quasi-identifiers in order to build k-anonymous tables. The concept of k-minimal generalization has been proposed by Samarati .P [2001] in order to limit the level of generalization for maintaining as much data precision as possible for a given level of anonymity. Subsequently, the topic of k-anonymity has been widely researched.

- K-anonymity having several techniques are P-Sensitive k-anonymity, p+-sensitive k-anonymity, (p,α) sensitive k-anonymity.
- P-sensitive k-anonymity: The modified micro data table T' satisfies p-sensitive k-anonymity property if it satisfies k-anonymity, and for each QI-group in T' , the number of distinct values for each sensitive attribute is at least p within the same QI group.
- P+-sensitive k-anonymity: The modified Micro data table T' satisfies p+-sensitive k-anonymity property if it satisfies k-anonymity, and for each QI-group in T' , the number of distinct categories for each sensitive attribute is at least p within the same QI-group.
- (P,α)-sensitive k-anonymity:The modified microdata table T' satisfies (P,α)-sensitive k-anonymity property if it satisfies k-anonymity, and each QI-group has at least p distinct sensitive attribute values with its total weight at least α.

III.2 *l* -DIVERSITY METHOD

Clearly, while k-anonymity is effective in preventing identification of a record, it may not always be effective in preventing inference of the sensitive values of the attributes of that record. Therefore, the technique of l-diversity was proposed which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes. Therefore, the l-diversity model for privacy is defined as follows: *Let a $q*$-block be a set of tuples such that its non-sensitive values generalize to $q*$. A $q*$-block is l-diverse if it contains l "well represented" values for the sensitive attribute S. A table is l-diverse, if every $q*$-block in it is l-diverse.*

A number of different instantiations for the l-diversity definition is available. When there are multiple sensitive attributes, then the l-diversity problem becomes especially challenging because of the curse of dimensionality, methods have been proposed in for constructing l-diverse tables from the data set, though the technique remains susceptible to the curse of dimensionality. Other methods for creating l-diverse tables are discussed in, in which a simple and efficient method for constructing the l-diverse representation is proposed.

III.3 t-CLOSENESS MODEL

The t-closeness model is a further enhancement on the concept of l-diversity. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. This is rarely the case for real data sets, since the attribute values may be very skewed. This may make it more difficult to create feasible l-diverse representations.

Often, an adversary may use background knowledge of the global distribution in order to make inferences about sensitive values in the data. Furthermore, not all values of an attribute are equally sensitive. For example, an attribute corresponding to a disease may be more sensitive when the value is positive, rather than when it is negative. According to Venkatasubramanian.S [2007], a t-closeness model was proposed which uses the property that the distance between the distribution of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold t. The Earth Mover distance metric is used in order to quantify the distance between the two distributions. Furthermore, the t-closeness approach tends to be more effective than many other privacy-preserving data mining methods for the case of numeric attributes.

IV. LIMITATIONS OF CURRENT PRIVACY PRINCIPLES

Initially anonymization was the first technique to prevent disclosure. Since k-anonymity model is not enough to protect sensitive information, several models such as *p*-sensitive *k*-anonymity, *l*-diversity, (*α k*)-anonymity and *t*-closeness have been proposed.

IV.1 LIMITATION OF P-SENSITIVE K-ANONYMITY:

The purpose of *p*-sensitive *k*-anonymity is to protect against attribute disclosure by requiring that there be at least *p* different values for each sensitive attribute within the records sharing a combination of quasi-identifier. This approach has the limitation of implicitly assuming that each sensitive attribute takes values

www.ijmer.com Vol. 3, Issue. 3, May.-June. 2013 pp-1763-1767 ISSN: 2249-6645

uniformly over its domain; that is, that the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving the required level of privacy may cause a huge data utility loss.

IV.2 LIMITATION OF L-DIVERSITY

The *l*-diversity model protects against sensitive attribute disclosure by considering the distribution of the attributes. The approach requires *l* "well-represented" values in each combination of quasi-identifiers. This may be difficult to achieve and, like *p*-sensitive *k*-anonymity, may result in a large data utility loss. Further, as previously identified, *l*-diversity is insufficient to prevent similarity attack.

IV.3 LIMITATION OF t-CLOSENESS:

The *t*-closeness model protects against sensitive attributes disclosure by defining semantic distance among sensitive attributes. The approach requires the distance between the distribution of the sensitive attribute in the group and the distribution of the attribute in the whole data set to be no more than a threshold *t*. Whereas Li et al. [2007] elaborate on several ways to check *t*-closeness, no computational procedure to enforce this property is given. If such a procedure was available, it would
greatly damage the utility of data because enforcing *t*-closeness destroys the correlations between quasi-identifier attributes and sensitive attributes.

V. DEVELOPING *K*-ANONYMITY ALGORITHMS

***Hash-based Technique*:** *k*-anonymity is a technique that prevents "linking" attacks by generalizing and/or suppressing portions of the released microdata so that no individual can be uniquely distinguished from a group of size *k*. A practical model of *k* anonymity, called full-domain generalization describes, a Hash-based technique previously used in mining associate rules and present an efficient hash-based algorithm to find the minimal *k*-anonymous table, which improves the previous binary search algorithm first proposed by Samarati.

Restricted K-Anonymity:

There are two new variants of the k-anonymity problem, namely, the Restricted k-anonymity problem and Restricted k- anonymity problem on attribute and discuss the connection between the Restricted k-anonymity and the general k-anonymity problems which stresses the significance of investigating this new class of anonymity problem. The theoretical results for restricted k-anonymity problem also provide an alternative NP-hardness proof of general k- anonymity problem.

V.1 ENHANCING K-ANONYMITY MODEL

k-anonymity alone is not enough to protect privacy in data. There are more stronger
algorithms than the k-anonymity model and that protect both sensitive facts and private knowledge in data. The (p+, α)-sensitive k-anonymity model requires that in each combination of quasi-identifiers, there are at least p different sensitive values and the total weight in each combination of quasi-identifiers is at least α. The motivation for this model is the fact that although k-anonymity is effective in protecting identity disclosure, to some extent, it fails to protect sensitive attribute disclosure. (p+, α)-sensitive k-anonymity model, provides an ordinal distance system to evaluate the degree that the sensitive attribute contributes to the database.

VI. CONCLUSION

In an increasingly data-driven society, personal information is often collected and distributed with ease. In this survey, we have presented an overview of recent technological advances in defining and protecting individual privacy and confidentiality in data publishing. In particular, such as hospitals and government agencies, that compiles large data sets, and must balance the privacy of individual participants with the greater good for which the aggregate data can be used.

While technology plays a critical role in privacy protection for personal data, it does not solve the problem in its entirety. In the future, technological advances must combine with public policy, government regulations, and developing social norms.

Due to the wide use of the Internet and the trends of enterprise integration,, simultaneous cooperation and competition, and outsourcing in both public and private sectors, data publishing has become a daily and routine activity of individuals, companies, organizations, government agencies. Privacy-preserving data publishing is a promising approach for data publishing without compromising individual privacy or disclosing sensitive information.

In this thesis, we studied different types of linking attacks in the data publishing scenarios, analysis of data, sequential release, secure data integration and various limitations of the privacy models

www.ijmer.com Vol. 3, Issue. 3, May.-June. 2013 pp-1763-1767 ISSN: 2249-6645

REFERENCES

[1]. *X. Jin, N. Zhang, G. Das, Algorithm-safe privacy preserving data publishing, in: EDBT, 2010.*

[2]. ℓ-diversity: Privacy Beyond k-Anonymity, Ashwin Machanavajjhala Daniel Kifer Johannes Gehrke.

[3].]Protecting Respondents' Identities in Microdata Release, Pierangela Samarati.

[4]. X. Jin, M. Zhang, N. Zhang, G. Das, Versatile publishing for privacy preservation, in: KDD, 2010.

[5]. P.Samarati. Protecting respondent's identities in micro-data release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027. 2001

[6]. *Bayardo R. J., Agrawal R.: DataPrivacythrough Optimal k*-Anonymization. *Proceedings of the ICDE Conference*, pp. 217–228, 2005.

[7]. ASAP: Eliminating algorithm-based disclosure in privacy-preserving data publishing Xin Jin, NanZhang , GautamDas , 2011

[8]. P.Samarati, L.Sweeney, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,Technical Report, CMU, SRI, 1998.

[9]. K.LeFevre, D.J.DeWitt, R.Ramakrishnan, Mondrian multi-dimensional k-anonymity, in: ICDE, 2006, pp.25–35.

[10]. R.C. Wong, A.W. Fu, K. Wang, J. Pei, Minimality attack in privacy- preserving data publishing, in: VLDB, 2007,543–554.

[11]. N. Koudas, D. Srivastava, T. Yu, Q. Zhang, Distribution-based microdata anonymization, in: VLDB, 2009.

[12]. Machanavajjhala, J. Gehrke, M. Goetz, Data publishing against realistic adversaries, in: VLDB, 2009.

[13]. A.Meyerson, R.Williams, On the complexity of optimal k-anonymity, in: PODS, 2004, pp.223–228.

[14]. X. Jin, N. Zhang, G. Das, Algorithm-safe privacy preserving data publishing, in: EDBT, 2010.

[15]. N.Koudas, D.Srivastava, T. Yu, Q. Zhang, Distribution-based microdata anonymization, in: VLDB, 2009.

[16]. L. Sweeney. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness andKnowledge-based Systems, 10 (5), 2002; 557-570.

[17]. R. Bayardo and R. Agrawal. Data privacy through optimal *k*-anonymity. *In Proceedings of the 21st International Conference on Data Engineering (ICDE)*, 2005.

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

[19]. G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, A. Zhu, Achieving anonymity via clustering, in: PODS, 2006, pp. 153-162.

[20]. H. Park, K. Shim, Approximate algorithms for k-anonymity, in: SIGMOD, 2007, pp. 67-78.