Volume 3, Issue 6, June 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Privacy Preserving for High-dimensional Data using Anonymization Technique

Neha V. Mogre
ABHA Gaikwad-Patil
College of Engineering, Nagpur
India

Prof. Girish Agarwal
ABHA Gaikwad-Patil
College of Engineering, Nagpur
India

Prof. Pragati Patil
ABHA Gaikwad-Patil
College of Engineering, Nagpur
India

Abstract— In recent years, privacy-preserving data publishing has seen rapid advances that have lead to an increase in the capability to store and record personal data about consumers and individuals. Maintain the privacy for the high dimensional database has become important aspect. 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 mining tasks in a privacy-preserving way. These techniques for performing privacy-preserving data mining are drawn from a wide array of related topics such as data mining, cryptography and information hiding. In this paper, we provide a state-of-art methods for privacy for the high dimensional databases.

Keywords— Data anonymization, Privacy preservation, Data publishing, Data security, Privacy Threats

I. INTRODUCTION

In recent years, due to increase in ability to store personal data about users and the increasing sophistication of data mining algorithms to leverage this information the problem of privacy-preserving data mining has become more important. A number of anonymization techniques have been researched in order to perform privacy-preserving data mining. Most of existing work is formulated in the following context: Several organizations, such as hospitals, publish detailed data (also called *microdata*) about individuals (e.g. medical records) for research or statistical purposes. However, sensitive personal information may be disclosed in this process, due to the existence in the data of quasi-identifying attributes, or simply quasi-identifiers (QID), such as age, zip code, etc. An attacker can join the QID with external information, such as voting registration lists, to reidentify individual records. Existing privacy-preserving techniques focus on anonymizing personal data, which have a fixed schema with a small number of dimensions. Through *generalization* or *suppression*, existing methods prevent attackers from re- identifying individual records. For example, should businesses trust their employees with the critical role of protecting sensitive corporate information? Industry analysts would probably say “never” and with good reason. According to one recent Forrester study, 80 percent of data security breaches involve insiders, employees or those with internal access to an organization, putting information at risk. The big challenge for companies today – particularly as email and the Internet make sharing and distributing corporate information easier than ever - is to strike the right balance between providing workers with appropriate access and protecting sensitive information as much as possible. For example, database users traditionally are assigned a database administrator (DBA) role or granted multiple system privileges. As companies continue to consolidate databases and streamline operations to maximize efficiency and the protection of data from external threats, this user- and role-based security model no longer complies with “need-to-know” security best-practices. Today, to help ensure the safety, integrity and privacy of corporate information, more companies are pursuing a comprehensive, multi-factored security approach. For example, in a database which is having large datasets with a high dimension data such as Customer personal data such as Customer ID, Address, Phone No., Account details, Purchase details etc., such database table should be secure. Such data when shown to outer world should be secure enough.The rest of the paper is organized as follows: Section II describes about Related Work for privacy preservation. Section III describes about Attack Models and Privacy Models. Section IV outline about Anonymization Techniques for Privacy Preservation. Privacy Threats. Section V discuss the Problem Statement. Section VI describes the Proposed Work for the privacy preservation and finally Section VII concludes this paper.

II. RELATED WORK

A. Privacy-Preserving Data Mining

Generally when people talk of privacy, they say .keep information about me from being available to others. It is this intrusion, or use of personal data in a way that negatively impacts someone's life, that causes concern. As long as data is not misused, most people do not feel their privacy has been violated. The problem is that once information is released, it may be impossible to prevent misuse. Utilizing this distinction, ensuring that a data mining would not enable *misuse* of personal information opens opportunities that complete privacy would prevent. To do this, technical and social solutions that ensure data will not be released. The same basic concerns also apply to collections of data. Given a collection of data, it is possible to learn things that are not revealed by any individual data item. An individual may not be care about someone knowing their birth date, mother's maiden name, or social security number; but knowing all of them enables identity theft. This type of privacy problem arises with large, multi-individual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such corporate information is generally the goal of data mining, but some results may still lead to concerns (often termed a secrecy, rather than privacy, issue.) The difference between such corporate privacy issues and individual privacy is not that significant. If we view disclosure of knowledge about an entity (information about an individual) as a potential individual privacy violation.

A person may not wish for their medical records to be revealed to others. This may be because they have concern that it might affect their insurance coverage's or employment. Or it may be because they would not wish for others to know about medical or psychological conditions or treatments which would be embarrassing. Revealing medical data could also reveal other details about one's personal life. Privacy Breach Physicians and psychiatrists in many cultures and countries have standards for doctor-patient relationships which include maintaining confidentiality. In some cases the physician-patient privilege is legally protected. These practices are in place to protect the dignity of patients, and to ensure that patients will feel free to reveal complete and accurate information required for them to receive the correct treatment.[5] The United States has laws governing privacy of private health information, see HIPAA and the HITECH Act.

B. Privacy-Preserving Data Publishing

These techniques tend to study different transformation methods associated with privacy. These techniques include methods such as randomization [1], k-anonymity [16, 7], and l-diversity [11]. Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining [15]. Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

[1]In the most basic form of privacy-preserving data publishing (PPDP), the data holder has a table of the form D(Explicit Identifier, Quasi Identifier, Sensitive Attributes, Non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners; Quasi Identifier is a set of attributes that could potentially identify record owners; Sensitive Attributes consist 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 [40]. Most works assume that the four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner. Anonymization [52, 56] refers to the Privacy Preserving Data Publishing approach that seeks to hide the identity and/or the sensitive data of record owners, assuming that sensitive data must be retained for data analysis.

III. ATTACK MODELS AND PRIVACY MODELS

Two main Privacy preserving paradigms have been established: *k*-anonymity [3], which prevents identification of individual records in the data, and *l*-diversity [8], which prevents the association of an individual record with a sensitive attribute value.

A. k-anonymity

The database is said to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least k-1 other rows. K-Anonymity thus prevents definite database linkages. K-Anonymity guarantees that the data released is accurate. K-anonymity proposal focuses on two techniques in particular: generalization and suppression.[2] To protect respondents' identity when releasing microdata, data holders often remove or encrypt explicit identifiers, such as names and social security numbers. De-identifying data, however, provide no guarantee of anonymity. Released information often contains other data, such as birth date, sex, and ZIP code that can be linked to publicly available information to re-identify respondents and to infer information that was not intended for release. One of the emerging concept in microdata protection is *k-anonymity*, which has been recently proposed as a property that captures the protection of a microdata table with respect to possible re-identification of the respondents to which the data refer. *k*-anonymity demands that every tuple in the microdata table released be indistinguishably related to no fewer than *k* respondents. One of the interesting aspect of *k*-anonymity is its association with protection techniques that preserve the truthfulness of the data. The first approach toward privacy protection in data mining was to perturb the input (the data) before it is mined. The drawback of the perturbation approach is that it lacks a formal framework for proving how much privacy is guaranteed. At the same time, a second branch of privacy preserving data mining was developed, using cryptographic techniques. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining. One definition of privacy which has come a long way in the public arena and is accepted today by both legislators and corporations is that of k-anonymity [3]. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals. Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data.

*[4]**Limitations of k-anonymity are:*** (1) it does not hide whether a given individual is in the database, (2) it reveals individuals' sensitive attributes , (3) it does not protect against attacks based on background knowledge , (4) mere

knowledge of the k-anonymization algorithm can violate privacy, (5) it cannot be applied to high-dimensional data without complete loss of utility , and (6) special methods are required if a dataset is anonymized and published more than once.

B. l- diversity

The next concept is "l-diversity". Say you have a group of k different records that all share a particular quasi-identifier. That's good, in that an attacker cannot identify the individual based on the quasi-identifier. But what if the value they're interested in, (e.g. the individual's medical diagnosis) is the same for every value in the group. The distribution of target values within a group is referred to as "*l*-diversity". [8] Currently, there exist two broad categories of *l*-diversity techniques: *generalization* and *permutation*-based. An existing generalization method would partition the data into disjoint groups of transactions, such that each group contains sufficient records with *l*-distinct, well represented sensitive items.

C. t-closeness

t-closeness that formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*). This effectively limits the amount of individual-specific information an observer can learn. Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of approach is that separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

Information Disclosure Threats:

When publishing microdata, there are three types of information disclosure threats [23,24,25].

A. Membership Disclosure Protection:

The first type is *membership disclosure*, when the data to be published is selected from a larger population and the selection criteria are sensitive (e.g., when publishing datasets about diabetes patients for research purposes), it is important to prevent an adversary from learning whether an individual's record is in the data or not.

B. Identity Disclosure Protection:

The second type is *identity disclosure*, which occurs when an individual is linked to a particular record in the released table. In some situations, one wants to protect against identity disclosure when the adversary is uncertain of membership. In this case, protection against membership disclosure helps protect against identity disclosure. In other situations, some adversary may already know that an individual's record is in the published dataset, in which case, membership disclosure protection either does not apply or is insufficient.

C. Attribute Disclosure Protection:

The third type is *attribute disclosure*, which occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the attributes of an individual more accurately than it would be possible before the release. Similar to the case of identity disclosure, we need to consider adversaries who already know the membership information. Identity disclosure leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive value is revealed. Attribute disclosure can occur with or without identity disclosure, e.g., when the sensitive values of all matching tuples are the same.

IV. ANONYMIZATION TECHNIQUES FOR PRIVACY-PRESERVATION :

Two widely studied data anonymization technique are generalization and bucketization. The main difference between the two anonymization techniques lies in that bucketization does not generalize the QI attributes.

A. Generalization

Generalization is one of the commonly anonymized approach, which replaces quasi-identifier values with values that are less-specific but semantically consistent. Then, all quasi-identifier values in a group would be generalized to the entire group extent in the QID space. [12] If at least two transactions in a group have distinct values in a certain column (i.e. one contains an item and the other does not), then all information about that item in the current group is lost. The QID used in this process includes all possible items in the log. Due to the high-dimensionality of the quasi-identifier, with the number of possible items in the order of thousands, it is likely that any generalization method would incur extremely high information loss, rendering the data useless [8]. In order for generalization to be effective, records in the same bucket must be close to each other so that generalizing the records would not lose too much information. However, in high-dimensional data, most data points have similar distances with each other poosible. This is an inherent problem of generalization that prevents effective analysis of attribute correlations.

***Limitation of Generalization*:-**

Our intention is not to eliminate generalization; there is no doubt that generalization is an important technique, partly proved by the fact that it has received much attention in the literature. Instead, our goal is to present an alternative option for privacy preservation, which has its own advantages, since it can retain a larger amount of data characteristics [14]. The main problems with generalization are: 1) it fails on high-dimensional data due to the curse of dimensionality [1] and

2) It causes too much information loss due to the uniform-distribution assumption.

B. *Bucketization*

Bucketization, is to partition the tuples in T into *buckets*, and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. The sanitized data then consists of the buckets with permuted sensitive values.

Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high-dimensional data. However, their approach assumes a clear separation between QIs and SAs. In addition, because the exact values of all QIs are released, membership information is disclosed.

Limitations of Bucketization:

While bucketization [1] has better data utility than generalization, it has several limitations. 1) Bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. In the United States, 87 percent of the individuals can be uniquely identified using only three attributes (Birthdate, Sex, and Zipcode). A microdata (e.g., census data) usually contains many other attributes besides those three attributes. This means that the membership information of most individuals can be inferred from the bucketized table.

2) Bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. 3) By separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

V. PROBLEM STATEMENT

***Database Privacy*:** When people speak about database privacy, they usually are referring to the protection of information contained within digital databases and of the databases themselves. It can include security issues surrounding the database and the classification of its information. Database privacy is a concept that is important to organizations and private citizens alike. However, organizations have the responsibility to protect clients' information, because their clients entrust them to do so. There are a number of steps that organizations can take to help safeguard databases and the data they hold. Some of these steps include making sure that servers are configured correctly, assigning proper authentication levels to database workers, providing unique authentication credentials for each application, preventing the theft of authentication credentials and protecting the database against software designed to compromise it or the information it contains. Privacy professionals also can secure storage systems against theft involving servers, hard drives, desktops and laptops. Organizations should ensure that storage management interfaces and all database backups, whether on-site or off-site, maintain their integrity. If attacks on a database occur, it is an organization's responsibility to take defensive measures. This might first entail the immediate classification of data according to importance. Then, encryption methods might be employed to help protect applications and data based on their sensitivity levels. Of course, the best method of protecting a database's privacy is prevention. One method of database privacy protection might include assessing a database regularly for exploits and signs that it has been compromised. If an organization can detect exploits or indications of database compromising before the threat becomes real and unmanageable, the database might be able to be rectified with little and reversible damage.

VI. PROPOSED WORK

As Privacy-preservation for the high dimensional database has become important in many ways. Database of any organization, company's, medical is a confidential database. Such database must be preserved, so that no confidential information gets open into the real world. In this paper, we introduce a data anonymization technique called slicing to improve the current state of the art.[6] Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Note that when the data set contains QIs and one SA, bucketization has to break their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute. The key intuition that slicing provides privacy protection is that the slicing process ensures that for any tuple, there are generally multiple matching buckets. Slicing first partitions attributes into columns. Each column contains a subset of attributes. Slicing also partition tuples into buckets. Each bucket contains a subset of tuples. This horizontally partitions the table. Within each bucket, values in each column are randomly permutated to break the linking between different columns.

VII. CONCLUSION

An important research problem is for handling high-dimensional data. As per the above, Privacy Preservation for high dimensional database is important. There are two popular data anonymization technique Generalization and Bucketization. These techniques are designed for privacy preserving microdata publishing. Our Proposed work include a slicing technique which is better than generalization and bucketization for the high dimension data sets. Slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data.

ACKNOWLEDGMENT

We would like to thanks all the authors whom we referred for giving their suggestions and making the material available for us to refer.

REFERENCES

[1] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu, "Privacy Preserving Data Publishing Concepts and Techniques" ,*Data mining and knowledge discovery series (2010).*

[2] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).

[3] Latanya Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557–570, 2002.

[4] J. Brickell and V. Shmatikov, "The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008*

[5] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" *Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.*

[6] Neha V. Mogre, Girish Agarwal, Pragati Patil:"A Review On Data Anonymization Technique For Data Publishing" *Proc. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181*

[7] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and '-Diversity," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.*

[8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. "*l*-diversity: Privacy beyond k-anonymity". In ICDE, 2006.

[9] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. "Worst-case background knowledge for privacy preserving data publishing". In ICDE, 2007.

[10] G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," *Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.*

[11] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k- Anonymization," in *Proc. of ICDE*, 2005, pp. 217–228.

[12] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-domain k-Anonymity," in *Proc. of ACM SIGMOD*, 2005, pp. 49– 60.

[13] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," in *Proc. of ICDE*, 2006.

[14] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in *Proc. Of IEEE Transactions on Knowledge and Data Engineering* February 2011 (vol. 23 no. 2) pp. 161-174.

[15] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy- Preserving Data Publishing," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.*

[16] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation*," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.*

[17] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," *Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009.*

[18] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Data Sets," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 217-228, 2006.*

[19] T. Li and N. Li, "On the Tradeoff between Privacy and Utility in Data Publishing," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 517-526, 2009.*

[20] Y. Xu, K. Wang, A.W.-C. Fu, and P.S. Yu, "Anonymizing Transaction Databases for Publication," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 767-775, 2008.*

[21] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, "Utility- Based Anonymization Using Local Recoding," *Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 785-790, 2006.*

[22] C. Dwork, "Differential Privacy: A Survey of Results," *Proc. Fifth Int'l Conf. Theory and Applications of Models of Computation (TAMC), pp. 1-19, 2008.*

[23] G. T. Duncan and D. Lambert, "Disclosure-limited data dissemination," *Journal of The American Statistical Association*, pp. 10–28, 1986.

[24] D. Lambert, "Measures of disclosure risk and harm", Journal *of Official Statistics*, vol. 9, pp. 313–331, 1993.

[25] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pp. 665–676, 2007.