Privacy and Utility in Data Publishing with Full Functional Dependencies

P.V.N. Prasoona[#1], M. Vasumathi Devi[*2] and K.V. Narasimha Reddy[#3]

[#1]*PG Scholar, Department of Computer Science,*
Vignan's Nirula Institute of Technology & Science for Women, Guntur, AP, India.

***Abstract*— A number of methods have recently been proposed for privacy preserving data mining of multidimensional data records. One of the methods for privacy preserving data mining is that of anonymization, in which a record is released only if it is indistinguishable from k other entities in the data. Data publishing has generated much concern on individual privacy. Recent work has shown that different background knowledge can bring various threats to the privacy of published data. We distinguish the safe FFDs that will not jeopardize privacy from the unsafe ones. We design robust algorithms that can efficiently anonymize the microdata with low information loss when the unsafe FFDs are present. Our results clarify several common misconceptions about data utility and provide data publishers useful guidelines on choosing the right tradeoff between privacy and utility.**

***Keywords*— Privacy-preserving, data publishing, functional dependency, utility, data reconstruction.**

I. INTRODUCTION

Data Mining which is sometimes also called as Knowledge Discovery Data (KDD) is the process of analyzing data from different perspectives and summarizing it into useful information. Today, data mining is used by many companies with a strong consumer focus such as retail, financial, communication, and marketing organizations. Extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering. It has been said that knowledge is power, and this is exactly what data mining is about. It is the acquisition of relevant knowledge that can allow to make strategic decisions. which will further allow for the successful business or organization.

Data anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Detailed data (also called as microdata) contains information about a person, a household or an organization. Most popular anonymization techniques are Generalization and Bucketization. [1]There are number of attributes in each record which can be categorized as 1) Identifiers such as Name or Social Security Number are the attributes that can be uniquely identify the individuals. 2) some attributes may be Sensitive Attributes(SAs) such as disease and salary and 3) some may be Quasi-Identifiers (QI) such as zipcode, age, and sex whose values, when taken together, can potentially identify an individual.

Data from which the patient cannot be identified by the recipient of the information. The name, address, and full post code must be removed together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient. Unique numbers may be included only if recipients of the data do not have access to the key‟ to trace the identity of the patient. Technology that converts clear text data into a nonhuman readable and irreversible form, including but not limited to preimage resistant hashes (e.g., one-way hashes) and encryption techniques in which the decryption key has been discarded. Data is considered anonymized even when conjoined with pointer or pedigree values that direct the user to the originating system, record, and value (e.g., supporting selective revelation) and when anonymized records can be associated, matched, and/or conjoined with other anonymized records. Data anonymization enables the transfer of information across a boundary, such as between two departments within an agency or between two agencies, while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization.

II. BACKGROUND WORK

Generalization and Bucketization:

One popular anonymization method is generalization Generalization is applied on the quasi-identifiers and replaces a QI value with a "less-specific but semantically consistent value". As a

result, more records will have the same set of quasi-identifier values. We define an equivalence class of a generalized table to be a set of records that have the same values for the quasi- Two main Privacy preserving paradigms have been established: k-anonymity [7], which prevents identification of individual records in the data, and l-diversity [1], which prevents the association of an individual record with a sensitive attribute value.

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.

III. METHODS

Anonymization algorithm

In this section, we explain the details of the algorithmthat constructs the QI-groups for anonymization. Given an unsafe FFD A→B (recall that unsafe FFDsmust include at least one sensitive attribute in its determinant attribute), a naive anonymization method is to apply DG, CG and IG grouping strategies directly on all distinct values of the sensitive attributes in A. This may incur tremendous information loss by tuple suppression, especially for the dataset whose sensitive values are of skewed frequency distribution. Therefore, first, to reduce the information loss by tuple suppression, we split the sensitive values into smaller disjoint segments, and apply one of DG, CG, and IG groupings on these segments, depending on which returns the smallest number of removed tuples. We call this the phase-1 partition. The second phase of anonymization is QI-group construction, by which we reduce the information loss by data generalization while construct QI-groups from the phase-1 partitions. In the next section, we explain the details of these two phases.

Our evaluation methodology has a number of advantages when compared with existing work. First, one can use this methodology to compare datasets anonymized using different requirements. E.g., both diversity and t-closeness are motivated by protecting against attribute disclosure, by choosing one privacy loss measure, one can compare datasets anonymized with diversity for different values and those anonymized with t-closeness for different t values. Second, we measure utility loss against the original data rather than utility gain. Utility gain is not well-defined in data publishing. In order to measure utility gain, a baseline dataset must be defined. Because only correct information contributes to utility, the baseline dataset must contain correct information about large populations. In [5], Brickell and Shmatikov used the trivially-anonymized data as the baseline, in which every distribution is estimated to be the overall distribution and therefore causes incorrect information. Third, we measure utility for aggregate statistics, rather than for classification. This is because, as several studies have shown, the utility of the anonymized data in classification is limited when privacy requirements are enforced. Finally, we measure privacy loss in the worst-case and measure the accumulated utility loss. Our methodology thus evaluates the privacy loss for every individual and the utility loss for all pieces of useful knowledge.

IV. SYSTEM MODEL

Name	Age	Gender	Zipcode	Disease
Ann	20	F	12345	AIDS
Bob	24	M	12342	Flu
Cary	23	F	12344	Flu
Dick	27	M	12344	AIDS
Ed	35	M	12412	Flu
Frank	34	M	12433	Cancer
Gary	31	M	12453	Flu
Tom	38	M	12455	AIDS

TABLE1 – ORIGINAL TABLE

Age	Gender	Zipcode	Disease
[20-38]	F	12***	AIDS
[20-38]	M	12***	Flu
[20-38]	F	12***	Flu
[20-38]	M	12***	AIDS
[20-38]	M	12***	Flu
[20-38]	M	12***	Cancer

[20-38]	M	12***	Flu
[20-38]	M	12***	AIDS

TABLE 2: GENERALIZATION

Age	Gender	Zipcode	Disease
[20-27]	*	1234*	AIDS
[20-27]	*	1234*	Flu
[20-27]	*	1234*	Flu
[20-27]	*	1234*	AIDS
[35-38]	*	124**	Flu
[35-38]	*	124**	Cancer
[35-38]	*	124**	Flu
[35-38]	*	124**	AIDS

TABLE 3: BUCKETIZATION

Age	Gender	Zipcode	Disease
20:1,24:1,23:1,27:1	M:2,F:2	12344:2,12342:1, 12345:1	AIDS
20:1,24:1,23:1,27:1	M:2,F:2	12344:2,12342:1, 12345:1	Flu
20:1,24:1,23:1,27:1	M:2,F:2	12344:2,12342:1, 12345:1	Flu
20:1,24:1,23:1,27:1	M:2,F:2	12344:2,12342:1, 12345:1	AIDS
35:1,34:1,31:1,38:1	M:4,F:0	12412:1,12433:1, 12453:1, 12455:1	Flu
35:1,34:1,31:1,38:1	M:4,F:0	12412:1,12433:1, 12453:1, 12455:1	Cancer
35:1,34:1,31:1,38:1	M:4,F:0	12412:1,12433:1, 12453:1, 12455:1	Flu
35:1,34:1,31:1,38:1	M:4,F:0	12412:1,12433:1, 12453:1, 12455:1	AIDS

TABLE 4: MULTI-SET BASED GENERALIZATION

(Age, Gender, Disease)	(Zip-code, Disease)
20,F,Flu	12345,Flu
24,M,AIDS	12342,AIDS
23,F,AIDS	12344,AIDS
27,M,Flu	12344,Flu
35,M,Flu	12412,Flu
34,M,AIDS	12433,AIDS
31,M,Flu	12453,Flu
38,M,Cancer	12455,Cancer

TABLE 5: OVERLAPPING SLICING

V. PROPOSED WORK

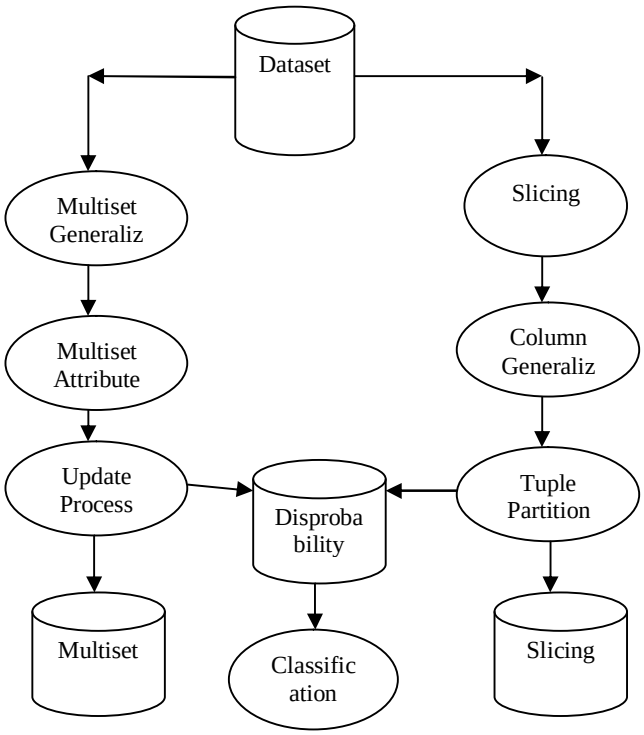

Figure 1: Overall process diagram

VI. CONCLUSION

In this paper, we studied the problem of privacy-preserving publishing of data that contains full functional dependencies. we present our methodology for evaluating privacy utility tradeoff. Our results give data publishers useful guidelines on choosing the right tradeoff between privacy and utility. For future work, first, we plan to further improve the heuristic approaches in the phase-1 partition step. One possibility is to make the construction of the initial partitions for the bottom-up approach driven by the frequency distribution. A similar idea applies to the choice of the split point for the top-down approach also. Inference cannot effectively defend against the privacy attack by conditional functional dependencies (CFDs), we will move to the study of privacy-preserving publishing microdata that contains CFDs.

REFERENCES

[1] [1] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, V. Verykios, Disclosure limitation of sensitive rules, Workshop on Knowledge and Data Engineering Exchange (KDEX), 1999, pp. 45–52.

[2] [2] M. Atzori, F. Bonchi, F. Giannotti, D. Pedresch, K-anonymous patterns, Proceedings of the Ninth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2005, pp. 10–21.

[3] [3] R.J. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization, Proceedings of the International Conference on Data Engineering (ICDE), 2005, pp. 217–228.
[4] [4] S. Chawla, C. Dwork, F. McSherry, A. Smith, H. Wee, Toward privacy in public databases, Second Theory of Cryptography Conference (TCC), 2005, pp. 363–385.
[5] [5] B.-C. Chen, K. LeFevre, R. Ramakrishnan, Privacy skyline: privacy with multidimensional adversarial knowledge, Proceedings of the International Conference on Very Large Data Bases (VLDB), 2007, pp. 770–781.
[6] [6] T. Dalenius, S.P. Reiss, Data swapping: a technique for disclosure control, Journal of Statistical Planning and Inference, 1982.
[7] [7] W. Du, Z. Teng, Z. Zhu, Privacy-maxent: integrating background knowledge in privacy quantification, Proceedings of ACM International Conference on Special Interest Group on Management of Data (SIGMOD), 2008, pp. 459–472.
[8] [8] A. Evfimievski, J. Gehrke, R. Srikant, Limiting privacy breaches in privacy preserving data mining, Proceedings of ACM Symposium on Principles of Database Systems (PODS), 2003, pp. 211–222.
[9] [9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey on recent developments. *ACM Computing Survey*, 2009.
[10] [10] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.
[11] [11] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *SIGMOD*, pages 1–12, 2000.
[12] [12] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *KDD*, pages 279–288, 2002.
[13] [13] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, pages 217–228, 2006.

P.V.N.Prasoona pursing M.Tech(C.S.E) at Vignan's Nirula Institute of technology and science for women. Interested area is Data Mining.

M. Vasumathi Devi possessed her B.Tech(C.S.E) in 2003 from JNTU and M.Tech(C.S) in 2010 from JNTUK. She worked as Assitant professor in SSN engineering college at ongle, VIGNAN college at Guntur, and now she is working in VIGNAN'S NIRULA INSTITUTE OF TECHNOLOGY AND SCIENCE FOR WOMEN at GUNTUR. She having nine years experience . she attend for workshop on Unified Modeling Language conducted in the year 2005 at Kottam Tulasi Reddy engineering college at Mahabubnagar. She guided many areas majorly in Computer Networks, Data Mining, and Image processing. She had attend for National Conference. She is supposed to do her paper and also research work in the area of Complier Design.

K.V.Narasimha Reddy received the B.Tech(CSE) from JNTUH, M.Tech(C.S.E) from JNTUK he is currently working as an Assistant Professor & Head of the Department of Computer Science & Engineering at Vignan's Nirula Institute Of Technology & Science for Women, Guntur. He guided many projects in the area of image processing for CSE & IT Departments. His research interests are in the areas of Datamining and Image Processing.