ijournals International Journal of Software & Hardware Research in Engineering
ISSN No: 2347-4890 Volume 2 Issue 1/ January 2014

Study of Functional and Matching Dependencies for Privacy Preservation in Micro Data

Deepika.R[1]; Karthik.J[2]
Avinasilingam University; Anna University

ABSTRACT
Privacy-preserving data mining is an area of data mining which is used to safeguard the sensitive information from third party [1]. It considers the problem of running data mining algorithm on confidential information like medical database, survey or census database etc. that is not supposed to be privacy breeched while performing data mining functions. In this paper aims to study the various dependencies like functional and matching dependencies present in micro data that may lead to privacy breech. Also, privacy preservation techniques that can be applied on these dependent data so that their privacy can be preserved are studied.

Keywords
Privacy Preservation Data Mining, Micro data, Functional Dependencies, Matching dependencies.

1. INTRODUCTION
Nowadays, worldwide networked society places great demand on the broadcasting and sharing of information, which is possibly becoming the most important and demanded storage but in the past released information was mostly in tabular and statistical form which is called as **macro data [6]**. But today many situations call for the release of specific data which only belongs to the particular domain called as the **micro data** [6]. Micro data, contains the specific data in its original form, it said to be as contrast to macro data which reporting pre-computed statistics data, provide the convenience of allowing the final recipient to perform on them analysis as needed. Micro data domain such as public health and population studies there are many possibilities to violate individual privacy. This leads to concerns that the personal data may be misused for a variety of purposes. A field of research namely privacy preserving data mining works on techniques to alleviate these concerns. These techniques of privacy-preservation are drawn from a wide array of related topics such as data mining, cryptography and information hiding [1].Privacy preservation consists of two types:

1. Individual privacy preservation
2. Collective privacy preservation

The goal of **Individual privacy preservation** privacy preservation is to protection the personal identification information. **Collective privacy preservation** not only protecting the personal identification information and also some patterns and trends that are not supposed to be reveal.

1.1 Privacy-Preserving Data Publishing
These techniques tend to study different transformation methods associated with privacy. These approaches include methods such as randomization, k-anonymity [7] and l-diversity [5]. A related issue is how the perturbed data can be used along with classical data mining methods such as association rule mining [9]. Other related problems include that of determining privacy preserving methods to keep the underlying data useful or the problem of studying the various privacy definitions, and how they compare in terms of effectiveness in different states. The various techniques of privacy preserving data mining [4] are shown in Fig.1.

1.2 Cryptographic techniques for Distributed Privacy
In many cases, the data may be distributed across many sites, and the owners of the data across these different sites may wish to compute a common function. In those cases, a variety of cryptographic protocols [10] may be used to communicate among various sites, so that secure function computation is possible without revealing the sensitive information.

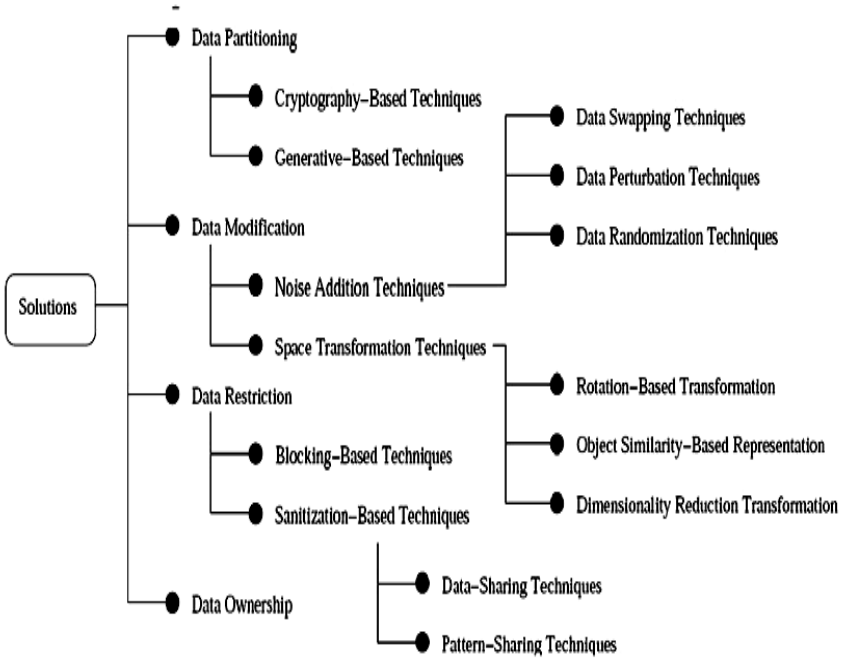

Fig 1: Technique of PPDM: [8]

2. MICRO DATA ATTRIBUTES

Micro data are analyses for privacy preservation shows that there are various dependencies in data like functional dependencies (FDs), conditional functional dependencies (CFDs), matching dependencies (MDs) which can be used as part of adversary knowledge to violate privacy. The aim of this paper is to study the various techniques in finding and preservation privacy breaches in these data. . Classification of Attributes in Micro data: [5]
The attributes in the micro data can be classifies as follows based on the information contained in them:

1. Key attributes/primary data/identifier
2. Quasi identifier
3. Sensitive attribute

Examples of **Key attribute:** Name, address, phone number - uniquely identifying.
Examples of **Quasi-identifiers:** ZIP code, gender, birth date uniquely and which can be used for linking anonymized dataset with other datasets
Examples of **Sensitive attributes:** Medical records, salaries, etc. Sensitive attributes is need for the researchers, so they mostly released the micro data with these attributes.

The various attributes representation in the micro data are shown in Table.1. It shows three types of attributes as shown in the above list.

Table 1: Micro data table with attribute representation

Key attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

3. DEPENDENCIES IN MICRODATA

Micro data contain specific data in its original form. There may be number of dependencies within micro data, which may lead to privacy breach.
The types of dependencies in micro data are:

1. Full Functional dependencies (FFD)
2. Matching dependencies (MDs)

3.1 Fully Functional Dependencies

It is a type of integrity constraint. Given two attribute X and Y, R (*A*1, *A*2... *An*) : a relation schema , *X* and Y are attributes in R. *r*(*R*): a specific relation of type *R*, which satisfies the *functional dependency* (fd) X → Y. if *each* specific relation (relational VALUE) r(R) satisfies X→Y. A relation value r satisfies X → Y. if each X value in r is associated with a unique Y value in r. In other words, a relation value r satisfies X → Y if for any two tuples *t*1 and *t*2 in r, *t*1[X] = *t*2[X] → *t*1[Y] = *t*2[Y].

The full functional dependency (FFD) [3], can be used as part of adversary knowledge. FFD expose the cross-attribute correlations among the data. For example FFD :Phone→Zipcode [3] states the fact that any two same phone numbers must correspond to the same zip code and imagine that the attacker having the knowledge of F in a micro data can bring potential vulnerability to privacy

3.2 Matching Dependencies

The concept of matching dependencies (MDs) [2], has recently been proposed for specifying matching rules for object identification. Similar to the functional dependencies, matching dependencies can also be applied to various data quality applications such as detecting the violations of integrity constraints. Consider a relation with schema R (*A*1, *A*2... *An*). Following similar syntax of FDs, we define MDs as following. A matching dependency (MD) φ has the form (X → Y, λ), where X and Y are two sets of attributes in relation R, and λ is a threshold pattern of similarity thresholds on attributes in X ∪ Y,

e.g., λ[A] denotes the similarity threshold on attribute A ∈ X ∪ Y.

The MDs can be regarded as a generalization of FDs, which are based on the equality of values (i.e., having matching similarity equal to 1.0 exactly) [2]. Thus, FDs can be represented by the syntax of MDs as well.

It specifies the dependency between two set of attributes according to their matching quality measured by some similarity matching operators, such as Euclidean distance and cosine similarity [2].we may have an MD as ([Street] → [City]) [2] which states that for any two tuples from Contacts, if they agree on attribute Street then the corresponding City attribute should match as well [2].The high value of MDs will help to attacker to threat the privacy of the individual.

4. CONCLUSION

This paper presented a brief study about the role of dependencies in data like full functional dependencies and Matching Dependencies that can lead to breach of privacy among the micro data.

5. REFERENCES

[1] Xinjing Ge and Jianming Zhu," Privacy Preserving Data Mining" School of Information, Central University of Finance and Economics Beijing, China

[2] Shaoxu Song et al "Efficient discovery of similarity constraints for matching dependencies", Data and Knowledge Engineering ,2013, pp- 146–166

[3] Hui Wang and Ruilin Lui,"Privacy Preserving Publishing Microdata with Full Functional Dependencies",Data and Knowledge Engineering 2011, pp 249–268

[4] C.Aggarwal and Philip S. Yu, "Privacy-Preserving Data Mining: Models and Algorithms". University of Illinois at Chicago,Springer, 2008.

[5] Ashwin Machanavajjhala, ohannes Gehrke, Daniel Kifer, Muthuramakrishnan Venkitasubramaniam , "L-Diversity: Privacy Beyond K-Anonymity"

[6] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity" Universitμa degli Studi di Milano, 26013 Crema

[7] Meyerson A., Williams R. "On the complexity of optimal k-anonymity", 2004

[8] Osmar R. Za¨ane et al., "Privacy-Preserving Data Mining on the Web Foundations and Techniques" Department of Computing Science University of Alberta Edmonton, AB, Canada

[9] J.Vaidya and C.Clifton, "Privacy Preserving Association Rule Miningin Vertically Partitioned Data," Proc Eighth ACMSIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD'02), pp.23-26,July 2002.

[10] M.Shaneck and Y.Kim, "Efficient Cryptographic Primitives for Private Data Mining," Proc.43rd Hawaii Int'l Conf. System Sciences (HICSS), pp.1-9, 2010