Data Privacy in Data Engineering, the Privacy Preserving Models and Techniques in Data Mining and Data Publishing: Contemporary Affirmation of the Recent Literature

Fuad Ali Mohammed Al-Yarimi
Ph.D student at School of Computer & Systems Science, Jawaharlal Nehru University
New Delhi 110067 – India

Sonajharia Minz
Professor at School of Computer & Systems Science Jawaharlal Nehru University
New Delhi 110067 – India

ABSTRACT

Privacy preserving for data engineering methods like mining and publishing etc., with the advancement of the rapid development of technologies like Internet and distributed computing has turned out to be one of the most important research areas of interest and has also triggered a serious issue of concern in accordance with the personal data usage in the recent times. Effective analysis result and gathering accurate data is desired by data users in specific, in contrast to the data owners who are concerned as their data contains personal information like the ones in government departments, Health insurance organizations and hospitals and data mining and warehouse utilities, where privacy is an issue to be taken rather seriously. Hence various proposals have been designated in data engineering methods publishing and mining for the purpose of preserving privacy. This paper briefs about the classification of the various privacy preserving approaches in data engineering, scans the current state of the art in lieu of preserving privacy of data, as also reviewing of the pros and cons of these specified approaches.

General Terms

Security, Algorithms , privacy preserving.

Keywords

Data Mining, Data publishing, privacy preserving, anonymity, data engineering, k-anonymity, t-closeness, l-diversity

1. INTRODUCTION

Where individual sensitive information exists, privacy is an issue of concern, when in recent times, data collection is an easy task and data mining methodologies are turning out to be more and more efficient. Various fields like computer science, statistics, economics, and social science have contributed in researching data privacy of individuals and also the confidentiality of data. This paper explains research in the field of preserving privacy in data mining and publishing. Data custodians such as hospitals, government agencies, insurance companies, and other businesses are preferred who possess data that can be granted to researchers, analysts etc. Evaluation of economic models, identification of social trends, search of opportunities in various fields etc are few intentions for which data usage is preferred as such data consists of personal information such as medical records, salaries etc, which does not at any cost facilitate release of such sensitive information and so on, so that a straightforward release of data is not appropriate. This issue can be tackled with data users signing a non-disclosure agreement, which further requires legal resources and enforcement mechanisms and might prove to be a hindrance to wide distribution of the concerned data.

A more intensely researched area is preserving privacy in data mining [Aggarwal and Yu 2008c]. The main intent of privacy-preserving data mining (PPDM), materialized in 2000 [Agrawal and Srikant 2000] is to mask confidential information in the modified data with the conventional data mining techniques. Modification of data and recovery of result of data mining from the modified dara is the main issue of concern. The various data mining algorithms involve the solutions, however PPDP may not be concerned with some specific data and during data publishing, data mining task will be in wraps.

Data truthfulness is also laid emphasis on by few PPDP solutions at the record level, but they fail to preserve a property. The study of non interactive query model statistical disclosure control [Adam and Wortman 1989; Brand 2002] is another area of study which deals data recipients submitting queries to the systems in existence. This may fail to address the data needs of the mentioned data recipients as construction of a query in a single go may sometimes prove to be difficult. Studies on the interactive query model include by the area where attackers submit a queries sequence depending upon the received query results, by [Blum et al. 2005; Dwork 2008; Dinur and Nissim 2003]. The only constraint in any privacy-preserving query system is that only sub linear quries can be answered else a data recipient, rather an attacker, will reconstruct all except $1 - o(1)$ fraction of the original data [Blum et al. 2008], which actually is a very strong violation of privacy. To avoid privacy leak, the system must be closed when the maximum number of queries is reached. The degree of privacy assigned by an interactive model cannot be achieved by a non-interactive model.

This survey reviews the taxonomy and current state of the art in anonymous approaches in preserving privacy for data engineering such as Data mining and Data publishing.

The paper is organized as follows: Section 2 talking about privacy-preserving data publishing (PPDP) vs. privacy-preserving data mining (PPDM) and also summarized the Data Mining and privacy preserving, and Data Publishing and privacy preserving Section 3 talking about and summarizing the nomenclature of privacy preserving techniques, where the Section 4 present the contemporary affirmation of recent literature in privacy preserving approaches, conclusion present at Section 5.

2. DATA ENGINEERING AND PRIVACY PRESERVING

The two essential aspects of engineering are amount of data that need to be processed which extract some useful information and also publishing of data with interests laid on analysis and retrieval of information. Hence, to achieve optimum result in accordance with time and data utility, various data mining and publishing techniques are adopted. There has been a considerable increase in the amount of personal data that is being collected and analyzed for the purpose of data mining and publishing as also the usage of supporting data engineering tools to deduce trends and patterns. To avoid hampering of individual privacy, there should be a restriction on access of data containing personal information. The solution can be obtained by releasing a part of the entire database and answering adequate queries by taking care not to reveal any sensitive information. The queries are supposed to be formulated by the researchers without accessing any data. Data can be put on anonymous mode by using the sanitization approach which is helpful in hiding the data's accurate values. Data values can sometimes be suppressed taking care to release data values exactly, which may probably hamper the utility. Research work like k-anonymity has garnered attention from scientists which includes the concept that every piece of undisclosed data is exactly equal to at least k-1 other pieces of disclosed data commencing over a set of privacy sensitive attributes.

2.1 PPDM versus PPDP

In the data collection phase, we collect data from record owners (Databases). In the second phase, the data collected in first case will releases to the data miner called the data recipient, who will then conduct data mining on the collected data. (In this case our concern will be PPDM) in the other case we will releases the collected data to the public (In this case our concern will be PPDP) So Data publisher may want to publish some data in real life applications, but fails to show interest in data mining results and its algorithms. This can be accomplished by privacy-preserving data publishing (PPDP), which is unique in its own way as compared to PPDM. Published records should be meaningful as PPDP focuses on data and noton the data mining results which therefore implies that encryption and randomization are inapplicable, as individuals identity is hidden by PPDP but not the insensitive data. The conventional data mining methods analyzes the anonymized data which leaves no room for new data mining techniques.

2.2 Data Mining and privacy preserving

Violation of privacy may take place if there lacks sufficient protection and abusing of private data, as data misuse is the main cause. As in case of individuals and organizations, data mining can be hazardous if data contains private characteristics. The development of the algorithms can be done by protecting existed private data and knowledge in PPDM and also assist in sharing the critical and private data for analytical aims. There exist two scenarios in the concept of, Privacy Preserving Data Mining: the Multi-party collaborations scenario and Data publishing scenario. Data is distributed between one or more sites, out of which private data is owned as also a data mining algorithm is computed when their sites and union of databases collaborate wherein only the results of data mining are revealed. This scenario mainly works on Secure Multi-party Computation, wherein to acquire results, owners share or publish their data to which privacy preservation techniques are applied. Data modification and Data sanitization are the two classified categories which work in accordance with the specifications of privacy preservation. Data sanitization approaches aim to hide the critical rules and patterns existed in dataset. However, the Data modification approaches are hiding critical data and aiming to acquire valid results of data mining while private data cannot be reached directly and precisely. In these techniques, major concerns are to maximize the quality of the released data, data mining results accuracy and protecting the data privacy as well.

Data Publishingand privacy preserving sufficiently private views are published by various government and corporate institutions to facilitate data analysis, it should be ensured that sensitive information of individuals is not disclosed by views and enough data is available for the data analysis process to take place. Privacy of sensitive information is guaranteed by few formal definitions of privacy and techniques to formulate data publishing.

Access control has been implemented in the earlier work on privacy in databases, Data is encoded cryptographically in the mentioned technique [2,3]. Authentication of users via credentials, as in the TrustBuilder project [4] includes other techniques. E-tables are used to finitely represent large set of possible worlds and projects $\prod_2^p$ -complete data complexity for checking that the sets of possible worlds represented by two c-tables are the same. It is a compact formalism which was introduced by Abiteboul et al [5]. C-tables are not sufficiently expressive to model the set of possible worlds given by a view instance. Database templates were introduced by Gosta Grahne et al [6] which shows how to compute them using the chase, but does not address the comparison of the sets of possible worlds.

The proposal of Evfimievski et al [7] solves the problem of limiting privacy breaches in a scenario in which the aggregation of a set of private client data items is computed at the server. A privacy breach is essentially defined as a significant difference between the a posteriori and the a priori probability distributions. Evfimievski et al [7] provides not only a diagnostic tool, it also scrambles the data to improve privacy. The model assumes independence among the private values at the clients. Thus, the techniques do not apply directly to our scenario, where the secret tuples are not independent of each other (indeed they are correlated via the possible worlds in which they appear). On the other hand, we do not handle aggregation, which is at the center of the model in [7]. Michal Bielecki et al [8] takes aggregation into account and shows that exposing the result of counting queries allows the retrieval of an isomorphic copy of the structure of the database. Shariq Rizvi [9] takes a dual approach to ours. While we use queries to specify that what cannot be disclosed, [9] uses conjunctive query views to specify what may be seen by outsiders. In this setting, conjunctive client queries asked against the proprietary database are answered only if they have a rewriting using the allowable views.

3. PRIVACY PRESERVING TECHNIQUES

3.1 Anonymization Techniques

De-identification is the first solution in publishing raw-critical data with the intent of privacy preserving wherein after removal of key identities of the records, raw-critical dataset is spread. A "Quasi Identifier" is used to identify few

personal details outside the external database. Anonymization approach is hence used which helps in modifying the QI values the summarized of data modification based techniques of PPDM presented in figure1.

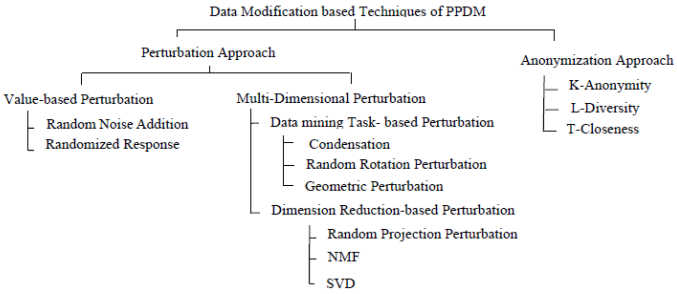

Figure1: Data Modification-based framework for classification the PPDM Techniques

K-anonymity: Every record is indistinguishable with other k-1 record within the anonymity table in accordance with a set of QI attributes in a dataset. If the values of QI attributes are identical when compared with the other records, a table is said to be K-anonymous. Generalization or suppression can be made use of [10, 11] to achieve the K-anonymity requirement. There exist few limitations which can be described as follows. It is difficult to check whether the attributes are available or nt in the external tables for a database owner. Linkage attack is considered by the model which fails to sufficiently preserve sensitive attributes against Homogeneity attack (similarity of the sensitive attributes values in an anonymized group) and Background Knowledge attack (awareness about the relationship between sensitive and QI attributes).

L-diversity: The Homogeneity attack of K-anonymity technique can be solved using this technique [12]. It saves various sensitive attributes in every group and also minimum size of K group. Atleast 1 well-represented value should be held by every anonymized group. The disadvantage is it fails to prevent attribute disclosure like in similarity attack.

T-Closeness: Regardless of the distribution of the data, L-diversity technique treats all the values of the attribute in a similar manner, which may not happen for real datasets where sensitive attributes values may not hold the same sensitivity level and hence, using Background Knowledge Attack, the exact values of sensitive attributes may be inferred. The distance between the distributions of a sensitive attribute in an anonymized group to that of the whole table should be less than 1 threshold. The distance criterion should be maintained so as to reflect the semantic gap in between the quantities.

Anonymization techniques are therefore very simple and hence scalable in case of privacy preservation, however, they fail to efficiently prevent the records' critical values deduction against attacks. Optimal anonymization is an "NP-Hard" problem [14] and is not proved to be effective as it may reveal the identity of the underlying record owners [15] when data is combined with public or background information.

3.2 Value-based decomposition Techniques

Random noise can be added to data values using this approach which is based on the fact that few data mining results use just the distribution of records and don't necessarily require the individual records. Data mining goals can be achieved by reconstruction of required aggregate distributions, wherein every data dimension is distributed independently. However there are possibilities of missing the implicit information which is available multidimensional records and on the other hand it is required to develop new distribution-based data mining algorithms.

3.3 Random Noise Addition Technique

In order to prevent a linkage attack, noise is added to the data in an occur-ring. By accounting for the extra variability from the added noise, the perturbed data can be correctly analyzed. Noise addition is discussed in Fuller in continuous data (1993) and Post Randomization Method (PRAM) by Gouweleeuw et al. (1998) can be applied for discrete data. To allow an accurate estimation of main data to take place, it is assumed that noise variance is large enough.

3.4 Randomized Responses Technique

Data is scrambled in this technique [16] with better probabilities than the defined threshold as to whether the data that has been sent back by the respondent is correct or not. Related–Question and Unrelated- Question models are the two described models in this technique wherein Related-Question model involves an interviewer asking a couple of questions related together to every respondent. The respondent will answer randomly and with θ probability to the first question and with 1- θ to the second question. Although the interviewer finds out the answers (yes or no), he does not know which question has been answered by the respondent; hence, the respondent's privacy preserving will be saved.

The aggregate information can be estimated with decent accuracy if the number of users is significantly large even though information from each individual user is scrambled. For providing information with response model and for processing categorical data, randomized response technique is used. This technique can be applied ti various dimensions [17].

4. Data Mining Task-based decomposition Techniques:

Original data is modified in this technique as also the various properties that are preserved in a perturbed dataset turn out to be task specific information data mining tasks. Hence, privacy can be preserved without loss of any specific information of data mining tasks so as to strike a suitable balance between privacy and data mining results accuracy. Also, direct applications are allowed in case of data mining algorithms without the need of development of new data mining algorithms on the perturbed dataset.

4.1 Condensation Technique

Original dataset is modified into anonymized datasets which helps preserve the covariance matrix for multiple columns. Groups with pre-defined size K is formed from the data available and for every group of records, a sequence of statistical information allied to the mean and associations across the dissimilar dimensions will be potted. Anonymized data is generated using the statistical records which possess similar statistical characteristics to the original dataset in the server. Simple classifier is created for the K Nearest Neighbor (KNN) [18], however in [19], it is weak in protecting the private data. The KNN-based data groups result in some serious conflicts between preserving covariance information and preserving privacy.

4.2 Random Rotation Perturbation Technique

The original dataset with fields and records will be decomposed by randomly rotating the part of the dataset, which is an orthogonal matrix. The Euclidean distance, inner product and geometric shape hyper in a multi-dimensional space are preserved in rotation transformation. When trained and tested with the rotation perturbed data, there exists similar model accuracy, kernel methods, Support Vector Machine (SVM) classifiers with certain kernels and hyper plane-based classifiers are invariant to rotation perturbation [19]. Privacy violations may be caused due to various attacks like Independent Component Analysis (ICA), attack to rotation center and distance-inference attack [20, 21] as a result of which random rotation perturbation may become involved, as shown by earlier researches.

4.3 Geometric Perturbation Technique

Rotation, Translation and Noise addition perturbation techniques are all combined in this technique. While preserving the data quality for classification modeling, the additional components ψ and Δ are used to address the weakness of rotation perturbation. The attack to rotation center is addressed by the random translation matrix and adds additional difficulty to ICA-based attacks and the noise addition addresses the distance-inference attack. This technique is fixed for Kernel, SVM and linear classifiers and is invariant against geometrical modification and also has high-great Privacy Preserving guarantees as compared to Rotation perturbation and condensation.

5. Dimension Reduction-based Perturbation Techniques

A compact representation with reduced-rank to the original dataset is obtained reserving dominant data patterns in this technique. The dimensionality and the exact value of every element present in the original data are kept confidential in this technique.

5.1 Random Projection Perturbation Technique

Projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace refers to the technique of Random projection [20]. The Johnson-Lindenstrauss Lemma [22] explained the possibility of maintaining distance-related statistical properties simultaneously with dimension reduction for a dataset and hence can be used for a variety of tasks such as including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc. It however, fails to preserve the distance and inner product that is acquired during the modification process when compared with the geometric and random rotation techniques.

5.2 Singular Value Decomposition (SVD) Technique

SVD [23] is a well known method that is used to reduce dimensions in data mining process. It has the property of capturing the maximum variation among the objects in the dimension due to the organization of singular descending values in the matrix Σ. The rest of the variations are hence captured in the subsequent dimensions. Hence, to represent the structure of the original matrix, a transformed matrix with a much lower dimension needs to be constructed.

6. CONTEMPORARY AFFIRMATION OF RECENT LITERATURE IN PRIVACY PRESERVING APPROACHES:

Privacy-preserving query-answering systems data publishing was explored and analyzed by **Ashwin Machanavajjhala et al[24].** Guaranteeing that the sensitive information will be kept secret, these systems answer queries. A query QS is used to express a secret in perfect privacy and another QV is answered if and only if no information is disclosed about QS. The problem of checking out if QV fails to maintain secrecy about QS is considered as $\Pi p2$-complete, if in case QS and QV are considered as arbitrary conjunctive queries. It is shown in this paper that perfect privacy for conjunctive queries in large subclasses is tractable. This connection is therefore used to relate the complexity so as to enforce perfect privacy to the complexity of query containment. Hence, Synthetic data generation was first proposed by Ashwin Machanavajjhala et al [25] which is used to analyze the data anonymization technique for publishing. There exists a mapping program which portrays the patterns of commuting of the United States population. The source for this application was gathered by the U.S. Census Bureau. Synthetic data is generated so as to statistically mimic the original data, maintaining privacy guarantees, which is hence used as a surrogate for the original data. Sparse data cannot be anonymized by the existing solutions as the data in it is sparse.

Disadvantages caused in measuring utility of current heuristic approaches was proposed by **Daniel Kiferet al[26]** who designated a formal approach for publishing data anonymity in measuring utility. K-anonymous and l-diverse tables are injected with additional information which guarantees and maintains k-anonymity and l-diversity frameworks.

A technique called I-diversity was proposed by **Ashwin Machanavajjhala et al[27]** which explained that a k-anonymized dataset has some subtle, but severe privacy problems wherein the values of sensitive attributes can be discovered by an attacker, who often possesses background knowledge, revealed that k-anonymity fails to provide privacy guarantee against suck attackers. A unique and powerful privacy definition called ℓ-diversity is proposed in [27] which gives information about the detailed analysis of the mentioned two attacks. It was indicated that ℓ-diversity can be implemented efficiently and is practical in an experimental evaluation.

A unique mechanism for data privacy in publishing was proposed by **Alin Deutsch et al[28]** where the proposed model strongly depends on data owner's awareness in predicting the attribute sensitivity in given data and provides privacy guarantee to data owners. The owner is supposed to use a secret query to identify the sensitive proprietary data against the proprietary database. The potential attackers may not learn about the secret information as it is considered as modification of the attacker's a-priori probability distribution on the set of possible secrets. This can be assumed under the pretext when secret and views are expressed as unions of conjunctive queries with non-equalities, under integrity constraints which help to solve the problem by using arbitrary a-priori distributions. The key insight on privacy diagnostics is based on the fact that the modeling of the attacker's

knowledge should start from possible worlds or at least plausible secrets. The individual tuples in the secret are correlated by appearing together in possible worlds.

Ruilin Liu et al[29] proposed a unique anonymization approach for data publishing, which is fully functional dependencies centric which provides maximum level utility with low information loss. This model considers an even set of attributes and the dependent relation between them. Consider that all attributes of set X are related to that of set Y, then a fully functional dependent relation can be defined between X and Y. Determinants are the attributes of set X while dependents are attributes of set Y. The relation between X and Y is defined as Conditional Functional dependency if in case the relation between X and Y are limited to only a few attributes. I-diversity [27] can be used to apply the attribute dependency relation model. The following steps explain the process of anonymizing the quasi identifiers.

- Sort the quasi identifiers based on their frequency
- Group the quasi identifiers such that each group justifies the l-diversity and d-closeness.
- Intersect each two sequence groups

The impact of Full Functional Dependencies (FFD) to avoid information loss was not explored as no empirical study was made. Anonymization would be vulnerable to avoid information loss if the degree of FFD is high.

The concept of anonymized tables being susceptible to corruption of attacks was proposed by **Tao et al. [30].** To prevent corruption attacks, Perturbed Generalization (PG) was proposed. A percentage p of SA-values is retained initially after which QI attributes are generalized to create k-sized anonymity groups. One perturbed record from every group is sampled. Distortions are introduced in the data based on the perturbation, generalization, and sampling. Distribution on SA renders things difficult for reconstruction for record sampling in the same way as information is lost in generalization of significant queries.

The concept of utility of perturbed records was focused by **Rhonda Chaytor et al[32]** who explained and consented that the anonymity group which is used to hide the identity of an individual is the root of all corruption attacks. All the group members are at a higher risk when the SA value of a group member is corrupted. By perturbing the group member's SA value, all QI attributes are published indignantly. No two individuals possess information about each other's QA values, hence preventing corruption attacks.

Corruption attacks exist in privacy preserving data publishing in the conventional anonymity-group approach. Perturbation, was hence deduced as the main methodology for data publishing, which was earlier used for privacy preserving data mining. Usage of fine-grain perturbation, a new perturbation operator is considered which is helpful in minimizing loss of information. Better utility is retained as deduced from the experimental results for ad hoc tasks, all the while conducting experiments for utility optimization for highly skewed datasets in aggregate data mining tasks.

The concept of preserving utility without compromising anonymity was proposed by **Ling Guo et al [33],** who preferred a more randomized approach**.** Taking into cue the techniques [34]-[38], [39] deduced earlier, a much general randomized framework was proposed and linking attacks causing attribute disclosure was investigated upon. This gave rise to optimal randomization parameters to provide efficient solutions in case of both QI and sensitive attributes. The following steps explain the randomization process proposed earlier.

- Find discloser probability of s under given QI.
- According to the determined probability threshold apply one of the fallowing
    - Randomize s only
    - Randomize QI only
    - Randomize both s and QI

The proposed concept's performance is considered to be efficient and effective as compared with the existing concepts [34]-[38], [39]. There exists a proportionality relationship between performance of randomization and quantity of sensitive attributes and quasi identifiers QI.

The limitations of l-diversity is revealed by **Ninghui Li et al[40],** who proposed the concept of "closeness", stating that the release of useful information is limited by t-closeness. The whole table is considered as the large population which explains an altogether different theory that is as follows. "An equivalence class E1 is said to have (n, t)-closeness if there exists a set E2 of records that is a natural superset of E1 such that E2 contains at least n records, and the distance between the two distributions of the sensitive attribute in E1 and E2 is no more than a threshold t. A table is said to have (n, t)-closeness if all equivalence classes have (n, t)-closeness". This is known as the (n, t)-closeness principle¸ wherein the breadth of the observer's background knowledge is denoted by n and the amount of sensitive information available is bounded by t. Smaller the value of n, smaller is the amount of sensitive information known to the observer concerning a smaller group of records and also implies a stronger privacy requirement. As part of (n, t)-closeness, two individual approaches are considered, mainly generalizing and hiding sensitive attributes, which may be taken into keen consideration for future researches.

Tiancheng Li[41] explained that **s**licing preserves data utility in a more better way than generalization which partitions data in a horizontal and vertical manner and is used for membership disclosure protection. Generalized and bucketing models that handle high dimensional data follow the slicing technique. Attributes are first partitioned into columns which contain a subset of attributes, partitioning the table vertically. The tuples are also partitioned into buckets, which contain a subset of tuples, partitioning the table horizontally. The linking between different columns is broken up by randomly per mutating the values in each column. This model is considered to be advantageous over the traditional k-anonymity, t-closeness and l-diversity models, which fails in proving its efficiency in the random grouping process in order to achieve anonymity in the context of data utilization.

Original data should be reconstructed jointly at different trust levels for preventing data miners to combine copies. . **Yaping Li et al[42]** suggested a theory stating that noise across copies can be correlated and additive perturbation based PPDM to multi-level trust (MLT)'s scope can be expanded at different trust levels, hence allowing generation of differently perturbed data copies at various trust levels by leaving free an implicit assumption of single-level trust in exiting work. The design of noise covariance matrix helps in prevention of diversity gain for joining and reconstructing the original data so as to possess corner-wave property as stated in the

acclaimed empirical study. However there is no detailed information on expansion of scope of other approaches in the proposed MLT-PPDM, which, for example are random rotation based data perturbation, k-anonymity, and retention replacement, to multi-level trust, fall under the aegis of partial information hiding. Only linear attacks are considered in the MLT-PPDM model which is obviously a disadvantage. However, original data can be derived and more information can be recovered by applying nonlinear techniques.

Under reasonable security assumptions, **Murat Kantarcıoglu et al[43]** debated that efficient distributed association rule mining can be done in a procedural way on data partitioned in a horizontal manner based on mine distributed association rules, which is again considered on multi party computations in a cryptographic model.

The "padding" set F is defined to be infinite, so that the probability of collision among these items is 0.

Collisions among the padded itemsets seem secure and size of set F is specified in order to advance the collision probability in real itemsets. There, hence exists, a constant relationship between the collision probabilities in F and real itemsets, where under secure multi-party communication definitions, the protocol is considered to be less secure. Prediction of fully encrypted real item sets from fake itemsets is enabled once the collision probability among items chosen from F is known, thus supporting a number of itemsets at every site allowing a probabilistic upper bound estimate on each one of them. A probabilistic estimate is allowed based on the number of itemsets supported in common by subsets of the sites that is tighter than the number of collisions found in the RuleSet. The privacy of protocols can be proved so as to meet strict secure multi-party computation definitions which are practically superior.

A tool for privacy preserving data mining explored by **Kun Liu et al[44]** mentioned the usage of random projection matrices which makes use of distance-related statistical properties without sensitizing the dimensionality and the exact data values after perturbation. **Johnson-Lindenstrauss Lemma [22]** introduced the key notion of random projection which projects a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace. This lemma shows that any set of s points in m-dimensional Euclidean space can be embedded into an O(log s/e2)-dimensional space such that an arbitrarily small factor is the only point of distance between the pair-wise distance of any two points.

Different kinds of data mining tasks, including inner product/Euclidean distance estimation, correlation matrix computation, clustering, outlier detection, linear classification, etc demonstrate in the experimental results that this technique can be successfully applied. When applied with some other geometric transformation techniques like scaling, translation, and rotation, the random projection-based technique may prove to be even more powerful.

Another interesting direction can be explored when the Random projection model is combined with SMC-based techniques.

7. CONCLUSION

Routine activity of many individuals, companies, organizations, and government agencies involves information sharing to which Privacy-preserving data publishing is a promising approach taking care to protect sensitive information and to preserve individual privacy. The taxonomy of anonymization techniques and current state of the art in privacy preserving techniques for data publishing and mining is reviewed in this manuscript whose main objective is to provide an anonymous form to the original data so as to preserve the utility of owners' sensitive information.

A list of desirable properties of a privacy-preserving data engineering methods is given and difference between privacy-preserving data publishing and privacy-preserving data mining is presented. The traditional and current state of the art methods in view of privacy models, anonymization operations, information metrics, and anonymization algorithms is reviewed and a single release from a single publisher is assumed, which ensures the first release or the first recipient protects the data up to its level. Several challenging works have been considered on purpose of mining, including multiple releases for publishing and mining, sequential release for publishing and mining, streamlining for publishing and mining, and collaborative data mining and publishing have also been reviewed. Policy-making, technology, psychology, and politics comprises of social complex issues which is privacy protection. Technical solutions to the problem can be gained by research of privacy protection in the field of computer science. Policy makers in governments and decision makers in companies and organizations, with their due cooperation can assist in successful application of privacy preserving technology. Radio Frequency Identification (RFID) and social networks, with their deployment limit the implementation of privacy-preserving technology in real-life applications. There may be an increase in the number of incidents and the scope of privacy breach with the increasing gap. However, to adopt privacy-preserving technology, general public, decision makers, and systems engineers should be facilitated to garner few potential research directions in privacy preservation along with some desirable properties.

Privacy preserving in Data Engineering is still at a developing stage. As the privacy problem is more complex, this technology needs to be researched further. By analyzing the existing work, three research directions of privacy preserving approaches can be concluded in data publishing.

1) Issue of the research of personalized privacy preservation.

2) Improving implementation efficiency and ensuring result availability so as to meet various requirements.

3) How to combine the advantage of above approaches.

Few research aspects are striking a better balance between privacy and accuracy, improving the efficiency of the algorithms, different types of privacy preserving generality, different Data mining tasks etc can be considered as privacy preserving methods for data publishing and mining for further research in the future.

8. REFERENCES

[1] Stanley R. M. Oliveira, and Osmar R. Zaïane1, "Towards Standardization in Privacy-Preserving Data Mining", In ACM SIGKDD 3rd Workshop on Data Mining Standards, 2004, pp. 7–17.

[2] G. Miklau and D. Suciu; Cryptographically enforced conditional access for xml; In WebDB, 2002

[3] Gerome Miklau and Dan Suciu; Controlling access to published data using cryptography. In VLDB, 2003

[4] Winslett et. al. The TrustBuilder Project; Publications Available from http://drl.cs.uiuc.edu/security/pubs.html

[5] S. Abiteboul, P. Kanellakis, and G. Grahne; on the representation and querying of sets of possible worlds; Theoretical Computer Science, 78:159{187, 1991

[6] Gosta Grahne and Alberto O. Mendelzon. Tableau techniques for querying information sources through global schemas; In ICDT, 1999

[7] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In PODS, 2003

[8] Michal Bielecki and Jan Van den Bussche. Database interrogation using conjunctive queries; In ICDT, pages 259, 269, 2003

[9] Shariq Rizvi, Alberto O. Mendelzon, S. Sudarshan, and Prasan Roy; Extending query rewriting techniques for fine-grained access control. In SIGMOD Conference, 2004

[10] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", In Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[11] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledgebased Systems, 2002, pp. 557-570.

[12] A.Machanavajjhala, J.Gehrke, and D.Kifer, "ℓ-diversity: Privacy beyond k-anonymity", In Proc. of ICDE, Apr.2006.

[13] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-anonymity and ℓ-Diversity", In Proc. of ICDE, 2007, pp. 106-115

[14] A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity", In Proceedings of PODS'04, pages 223–228, New York, NY, USA, 2004. ACM

[15] C. Aggarwal. "On k-anonymity and the curse of dimensionality", In Proceedings of VLDB'05, pages 901–909. VLDB Endowment, 2005

[16] L. Warner. "Randomized response: A survey technique for eliminating evasive answer bias," The American Statistical Association, 60(309):63–69, March 1965

[17] W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 505{510, Washington, DC, USA, August 2003

[18] AGGARWAL, C. C., AND YU, P. S. " A condensation approach to privacy preserving data mining." Proc. of Intl. Conf. on Extending Database Technology (EDBT) (2004)

[19] Chen, K., and Liu, L. "Privacy Preserving Data Classification with Rotation Pertubation", Proc. ICDM, 2005, pp.589-592

[20] K. Liu, H. Kargupta, and J. Ryan, "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," IEEE Transactions on Knowledge and Data Engineering, January 2006, pp. 92–106

[21] Keke Chen, Gordon Sun, and Ling Liu. Towards attack-resilient geometric data perturbation.In Proceedings of the 2007 SIAM International Conference on Data Mining.,April 2007.

[22] W.B. Johnson and J. Lindenstrauss, "Extensions of Lipshitz Mapping into Hilbert Space," Contemporary Math., vol. 26,pp. 189-206, 1984

[23] Golub GH, van Loan CF (1996) Matrix computations, 3rd edn. John Hopkins University, Columbia, MD

[24] Ashwin Machanavajjhala and Johannes Gehrke; On the Efficiency of Checking Perfect Privacy; In Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2006)

[25] Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: From Theory to Practice on the Map. In Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008), Cancun, Mexico, April 2008

[26] Daniel Kifer and J. E. Gehrke. Injecting Utility into Anonymized Datasets . In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data (SIGMOD 2006)

[27] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthu Venkitasubramaniam. l-Diversity: Privacy Beyond k-Anonymity. In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006), Atlanta, Georgia, April 2006

[28] Alin Deutsch, Yannis Papakonstantinou: Privacy in Database Publishing. ICDT 2005: 230-245

[29]Ruilin Liu; Hui Wang; , "Privacy-preserving data publishing," Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on , vol., no., pp.305-308, 1-6 March 2010 doi: 10.1109/ICDEW.2010.5452722

[30] Y. Tao, X. Xiao, J. Li, and D. Zhang, "On Anti-Corruption Privacy Preserving Publication," Proc. ICDE 2008

[31] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," Proc. VLDB 2006

[32] Rhonda Chaytor, Ke Wang, Patricia L. Brantingham: Fine-Grain Perturbation for Privacy Preserving Data Publishing. ICDM 2009: 740-745

[33] Ling Guo; Xiaowei Ying; Xintao Wu; , "On Attribute Disclosure in Randomization Based Privacy Preserving Data Publishing," Data Mining Workshops (ICDMW), 2010 IEEE International Conference on , vol., no., pp.466-473, 13-13 Dec. 2010; doi: 10.1109/ICDMW.2010.76

[34]W. Du and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining," KDD, 2003

[35] S. Rizvi and J. Haritsa, "Maintaining data privacy in association rule mining," in VLDB, 2002

[36] L. Guo, S. Guo, and X.Wu, "Privacy preserving market basket data analysis," in PKDD, 2007

[37] L. Guo, and X. Wu, "Privacy preserving categorical data analysis with unknown distortion parameters," in Transaction on Data Privacy, 2009

[38] Z. Teng and W. Du, "Comparisons of k-anonymization and randomization schemes under linking attacks," in ICDM, 2006

[39] Z. Huang and W. Du, "Optrr: Optimizing randomized response schemes for privacy-preserving data mining," in ICDE, 2008, pp. 705–714

[40] Ninghui Li; Tiancheng Li; Venkatasubramanian, S.; , "Closeness: A New Privacy Measure for Data Publishing," Knowledge and Data Engineering, IEEE Transactions on , vol.22, no.7, pp.943-956, July 2010; doi: 10.1109/TKDE.2009.139

[41] Tiancheng Li; Ninghui Li; Jian Zhang; Molloy, I.; , "Slicing: A New Approach for Privacy Preserving Data Publishing," Knowledge and Data Engineering, IEEE Transactions on , vol.24, no.3, pp.561-574, March 2012; doi: 10.1109/TKDE.2010.236

[42] Yaping Li, Minghua Chen, Qiwei Li and Wei Zhang;"Enabling Multi-level Trust in Privacy Preserving Data Mining"; IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2011

[43] Kantarcioglu, M.; Clifton, C.; , "Privacy-preserving distributed mining of association rules on horizontally partitioned data," Knowledge and Data Engineering, IEEE Transactions on , vol.16, no.9, pp. 1026- 1037, Sept. 2004; doi: 10.1109/TKDE.2004.45

[44] Kun Liu; Kargupta, H.; Ryan, J.; , "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining," Knowledge and Data Engineering, IEEE Transactions on , vol.18, no.1, pp. 92- 106, Jan. 2006; doi: 10.1109/TKDE.2006.1