A Survey on Privacy Measurement and Enhancement Models

Arifa M[1] Meera K[2]
[1,2]Cochin College of Engineering, Valanchery, India

Abstract— Census data and medical data are referred as micro data. Data publish schemes are used to provide private data for analysis. Privacy preservation is used to protect private data values. Anonymity is considered in the privacy preservation process. Individual identity and private data values are protected with privacy preservation models. Sensitive attribute values are hided in the privacy preservation process. Statistical and Anonymization techniques are used for privacy preservation. Generalization and suppression protocols are used for the privacy. The privacy preservation process is applied on data and information in various types of applications. The survey is conducted on five different categories. They are Privacy for Personalized Mobile Applications, Privacy using Post Randomization via Information Theory, Utility-Privacy Tradeoff in Databases, Distance-Aware Privacy-Preserving Record Linkage and Cross-Bucket Generalization for Information Privacy. Privacy measurement models are applied to guarantee the privacy for user required levels. Data publishing methods are build with privacy measurement and enhancement models.

Key words: Privacy Preserved Data Mining, Privacy Measures, Utility Measures and Privacy Preserved Data Publish

I. INTRODUCTION

Data mining and knowledge discovery in databases are two new research areas that investigate the automatic extraction of previously unknown patterns from large amounts of data. Recent advances in data collection, data dissemination and related technologies have inaugurated a new era of research where existing data mining algorithms should be reconsidered from a different point of view, this of privacy preservation. It is well documented that this new without limits explosion of new information through the Internet and other media, has reached to a point where threats against the privacy are very common on a daily basis and they deserve serious thinking.

Privacy preserving data mining, is a novel research direction in data mining and statistical databases, where data mining algorithms are analyzed for the side-effects they incur in data privacy. The main consideration in privacy preserving data mining is two fold. First, sensitive raw data like identifiers, names, addresses and the like should be modified or trimmed out from the original database, in order for the recipient of the data not to be able to compromise another person's privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithms should also be excluded, because such knowledge can equally well compromise data privacy. The main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process. The problem that arises when confidential information can be derived from released data by unauthorized users is also commonly called the "database inference" problem.

II. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

There are many approaches which have been adopted for privacy preserving data mining. They are classified on the basis of the following dimensions: Data distribution, data modification, data mining algorithm, data or rule hiding and privacy preservation. The first dimension refers to the distribution of data. Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places. The second dimension refers to the data modification scheme. In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

- perturbation, which is accomplished by the alteration of an attribute value by a new value,
- blocking, which is the replacement of an existing attribute value with a "?",
- aggregation or merging which is the combination of several values into a coarser category,
- swapping that refers to interchanging values of individual records, and
- sampling, which refers to releasing data for only a sample of a population.

The third dimension refers to the data mining algorithm, for which the data modification is taking place. This is actually something that is not known before and but it facilitates the analysis and design of the data hiding algorithm. The problem of hiding data as included for a combination of data mining algorithms. For the time being, various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

The fourth dimension refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion". The last dimension which is the most important, refers to the privacy preservation technique used for the selective modification of the data. Selective

modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been applied for this reason are:

- heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values
- cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and
- Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data.

It is important to realize that data modification results in degradation of the database performance. Two metrics are used to quantify the degradation of the data. The first one, measures the confidential data protection, while the second measures the loss of functionality.

III. PRIVACY PRESERVATION MODELS

The privacy preservation process is applied on data and information in various types of applications. The survey is conducted on five different categories. They are Privacy for Personalized Mobile Applications, Privacy using Post Randomization via Information Theory, Utility-Privacy Tradeoff in Databases, Distance-Aware Privacy-Preserving Record Linkage and Cross-Bucket Generalization for Information Privacy. Privacy measurement models are applied to guarantee the privacy for user required levels. Data publishing methods are build with privacy measurement and enhancement models.

A. Privacy for Personalized Mobile Applications

Mobile devices today are increasingly equipped with a range of sensors such as GPS, microphone, accelerometer, light and proximity sensors. These sensors can be effectively used to infer a user's context including his location from GPS, transportation mode from accelerometer, social state from the microphone and other activities from a combination of sensors. Consequently, a large and increasing number of applications in popular smart phone platforms such as the iPhone, the Android and the Windows Phone utilize user contexts in order to offer personalized services. Examples of such applications include GeoReminder that notifies the user when he is at a particular location, JogBuddy that monitors how much he jogs in a day, PhoneWise that automatically mutes the phone during meetings, Social Group on that delivers coupons or recommendations when he is in a group of friends, etc.

These context-aware mobile applications raise serious privacy concerns. Today, people already believe that risks of sharing location information outweigh the benefits in many location based services [6]. One reason why risks are high is that many mobile applications today aggressively collect much more personal context data than what is needed to provide their functionalities. Moreover, applications rarely provide privacy policies that clearly state how users' sensitive information will be used and with what third-parties it will be shared. To avoid the risks, a user can decide not to install these application or not to release any context information to them; but then the user might not be able to enjoy the utility provided by these applications. In order to explore a better tradeoff between privacy and utility, we can let the user control at a fine granularity when and what context data is shared with which application. For example, a user might be okay to release when he is at lunch but he might be hesitant to release when he is at a hospital. With such fine-grained decisions, a user can choose a point in the privacy-utility tradeoff for an application and can still enjoy its full functionality when he chooses to release his context information or when his context information is not actually needed.

MASKIT is a system that addresses the above question with two novel privacy checks deciding in an online fashion whether to release or suppress the current state of the user. The probabilistic check flips for each context a coin to decide whether to release or suppress it. The bias of the coin is chosen suitably to guarantee privacy. The simulatable check makes the decision only based on the released contexts so far and completely ignores the current context. That way, the decision does not leak additional information to the adversary. Both checks provably provide privacy, but interestingly their relative benefit varies across users —there are situations where the probabilistic check provides higher utility than the simulatable check and vice versa. We explain how to select the better check among the two for a given user.

B. Privacy using Post Randomization via Information Theory

The classification of attributes as key or confidential need not be disjoint or objectively unique. Ultimately, it relies on the specific application the microdata set is intended for. There are several privacy models to anonymize microdata sets. k-Anonymity [4] is probably the best known. However, it presents several shortcomings which have motivated the appearance of enhanced privacy models reviewed below. t-Closeness is one of those recent proposals. Despite its conceptual appeal, t-closeness lacks computational procedures which allow reaching it with minimum data utility loss.

Here, we define a privacy measure similar to the idea of t-closeness and provide an information-theoretic formulation of the privacy-distortion trade-off problem in microdata anonymization. This is done in such a way that the knowledge body of information theory can be used to find a solution to it. The resulting solution turns out to be the post randomization (PRAM) masking method in the discrete case and a form of noise addition in the general case.

C. Utility-Privacy Tradeoff in Databases

Just as information technology and electronic communications have been rapidly applied to almost every sphere of human activity, including commerce, medicine and social networking, the risk of accidental or intentional disclosure of sensitive private information has increased. The concomitant creation of large centralized searchable data repositories and deployment of applications that use them has made "leakage" of private information such as medical data, credit card information, power consumption data, etc. highly probable and thus an important and urgent societal problem. Unlike the secrecy problem, in the *privacy* problem,

disclosing data provides informational utility while enabling possible loss of privacy at the same time [5]. Thus, in the course of a legitimate transaction, a user learns some public information, which is allowed and needs to be supported for the transaction to be meaningful and at the same time he can also learn/infer private information, which needs to be prevented. Thus, every user is also an adversary.

The utility of a data source lies in its ability to disclose data and privacy considerations have the potential to hurt utility. Indeed, utility and privacy are competing goals in this context. For example, one could sanitize all or most of the entries in the gender attribute to 'M' to obtain more privacy but that could reduce the usefulness of the published data significantly. Any approach that considers only the privacy aspect of information disclosure while ignoring the resultant reduction in utility is not likely to be practically viable. To make a reasoned tradeoff, we need to know the maximum utility achievable for a given level of privacy and vice versa, i.e. an analytical characterization of the set of all achievable U-P tradeoff points. We show that this can be done using an elegant tool from information theory called rate distortion theory: utility can be quantified via fidelity which, in turn, is related to *distortion*. Rate distortion has to be augmented with privacy constraints quantified via *equivocation*, which is related to entropy.

D. Distance-Aware Privacy-Preserving Record Linkage

Record linkage is a two-step process. The goal of the first step, known as blocking, is to formulate as many as possible matching pairs and, simultaneously, maintain the number of non-matching pairs as small as possible. In the second step, termed as matching, the distances between the pairs formed during the blocking step are calculated. Approximate matching lies at the core of record linkage, since values contained in records that are owned by different data custodians, but refer to the same real-world entity, usually exhibit variations, errors, misspellings and typos. Therefore, applying exact matching on record pairs would typically generate results of low quality.

Privacy-Preserving Record Linkage (PPRL) investigates how to perform the steps described above in a secure manner, by respecting the privacy of the individuals who are represented in the data. For this reason, input records undergo an anonymization process that embeds them into a space, where the underlying data is kept private. The anonymization of numerical values is of paramount importance in a PPRL process operating under the three-party model [1], where a Trusted Third Party (TTP), which is not a simple broker but a trusted entity, receives the anonymized records from the data custodians and performs their linkage. Consider, for example, two hospitals that submit their anonymized patients' medical records to a trusted public agency, such as a Ministry of Health, whose duty is to identify records that refer to the same patient. The fields of those records could be the 'Name' and 'Address' of patients, whose values are strings and the 'Year of Birth', 'Height', or 'Cholesterol level', whose values are numerical. These numerical fields may be invaluable to identify common real-world entities, but their values might exhibit numerical variations.

The Euclidean distances are calculated using homomorphic computations. Homomorphic techniques generate large ciphers, which increase significantly both the communication and computation requirements. Especially in the presence of either large data sets or high dimensional data structures, homomorphism falls short to achieve a scalable solution. Recently, Vatsalan and Christen proposed a scheme relies on Bloom filters to anonymize numerical values. This scheme suffers from serious shortcomings, among which is the lack of accuracy guarantees between the distances of the numerical values and the corresponding distances generated in the employed Bloom filter space.

E. Cross-Bucket Generalization for Information Privacy

In the past several years, microdata release has posed threats to individual privacy and organizational confidentiality. According to Sweeney, 87 percent of the population in the United States had reported characteristics that likely made them unique based on particular attributes. Access to these data need to be safeguarded for the safety and security of the people. An adverse effect could be the unwarranted use of microdata.

Generalization, or transforming the QI values into more general forms, divides the tuples into equivalence groups in which the values of each QI attribute are the same [2]. Thus, the records in the same equivalence group are indistinguishable. The size of each equivalence group is at least k. Bucketization partitions the tuples into buckets in which the relation between the sensitive attributes and QI attributes is broken, such that each record in the bucketized corresponds to multiple sensitive values. The bucketized table always complies with l-diversity principle, i.e., for each tuple, the probability that the sensitive value is exposed is at most 1/l. For example, provides the bucketized satisfies the 4-diversity condition. Any tuple associated with a sensitive value inside its bucket has an equal probability that is no more than 1/4.

The requirement of attribute protection is higher than that of identity protection and proposes cross-bucket generalization as a solution. This anonymization technique prevents both identity and attributes disclosure while preserving significant information utility requires stringent measures for the protection of sensitive values. It separates the protection for identity and sensitive values. Cross-bucket generalization partitions the tuples into equivalence groups that satisfies the requirement of identity protection, then divides the generalized tuples into buckets to break their linkages between QI values and sensitive values.

The advantages of cross-bucket generalization are as follows. First, it provides separate protection for identity and sensitive values by establishing different sets of requirements for identity protection and attribute protection. The level of protection can be flexibly adjusted based on actual demands. For instance, the cross-bucket generalized complies with 2-anonymity and 4-diversity, which is more flexible. Cross-bucket generalization reduces the sizes of equivalence groups and buckets as far as possible by satisfying the protection requirements only. The size of each bucket is likewise 2, but the power of the protection for sensitive values is the same as the bucketized.

IV. PRIVACY-PRESERVED DATA PUBLISHING

Nowadays, datasets are considered a valuable source of information for the medical research, market analysis and economical measures. These datasets can include information about individuals that contain social, medical, statistical and customer data. Many organizations, companies and institutions publish privacy related datasets. While the shared dataset gives useful societal information to researchers, it also creates security risks and privacy concerns to the individuals. To avoid possible identification of individuals from records in published data, uniquely identifying information such as names and social security numbers are generally removed. While the obvious personal identifiers are removed, the quasi-identifiers such as zip-code, age and gender may still be used to uniquely identify a significant portion of the population since the released data makes it possible to infer or limit the available options of individuals. Correlating this data with the publicly available side information, such as information from voter registration list for Cambridge Massachusetts, medical visits about many individuals could be easily identified [3]. This study estimated that 87% of the population of the United States could be uniquely identified using quasi-identifiers through side information based attacks, including the medical records of the governor of Massachusetts in the medical data.

The spate of privacy related incidents has spurred a long line of research in privacy notions for data publishing and analysis, such as k-anonymity, l-diversity and t-closeness. K-anonymity if each quasi-identifier attribute in the table is indistinguishable from at least k - 1 other quasi-identifier attributes is called a k-anonymous table. While k-anonymity protects identity disclosure of individuals by linking attacks, it is insufficient to prevent attribute disclosure with side information. By combining the released data with side information, it makes it possible to infer the possible sensitive attributes corresponding to an individual. Once the correspondence between the identifier and the sensitive attributes is revealed for an individual, it may harm the individual and the distribution of the entire table. To deal with this issue, `l-diversity requires that the sensitive attributes contain at least ` well represented values in each equivalence class. `l-diversity has two major problems. One is that it limits the adversarial knowledge, while it is possible to acquire knowledge of a sensitive attribute from generally available global distribution of the attribute. Another problem is that all attributes are assumed to be categorical, which assumes that the adversary either gets all the information or gets nothing for a sensitive attribute.

Authors propose a privacy notion called t-closeness. They first formalize the idea of global background knowledge and propose the base model t-closeness. This model requires the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table [7]. This distance was introduced to measure the information gain between the posterior belief and prior belief through the Earth Mover Distance (EMD) metric, which is represented as the information gain for a specific individual over the entire population. Moreover, as we show in this paper, the distance between two distributions cannot be easily quantified by a single measurement. t-closeness also has many limitations that will be described later. The state of the art PPDP techniques will be further analyzed in more details. Research on data privacy has purely been focused on privacy definitions, such as k-anonymity, l-diversity and t-closeness. While these models only consider minimizing the amount of privacy leakage without directly measuring what the adversary may learn, there is a motivation to find consistent measurements of how much information is leaked to an adversary by publishing a dataset.

In this paper, we begin by introducing our novel data publishing framework. The proposed framework consists of two steps. First, we model attributes in a dataset as a multi-variable model. Based on this model, we are able to re-define the prior and posterior adversarial belief about attribute values of individuals. Then we characterize privacy of these individuals based on the privacy risks attached with combining different attributes. This model is indeed a more precise model to describe privacy risk of publishing datasets.

For a given dataset, before it is released, we want to determine to what extent we can achieve privacy. Therefore, we introduce a new set of privacy quantification metrics to measure the gap between prior information belief and posterior information belief of an adversary, from both local and global perspectives. Specifically, we introduce two privacy leakage measurements: distribution leakage and entropy leakage. We discuss the rationale for these two measurements and illustrate their advantages through examples. We show how considering only one metric ignoring the effect of the other strongly contributes to the information leakage and in turn affects the privacy. An intuitive example for this problem is reviewing a blood work. The medical status of a patient cannot be determined based on only one measure even if this particular measure is the most sensitive one. Instead, a physician has to review the relation between combinations of all measures in the blood work. A minimized distribution leakage between sensitive attribute values distributions of the original and the published datasets does not essentially achieve the minimum entropy leakage that an adversary could gain. In fact, we show that distribution and entropy leakage are two different measures. We believe that for a published dataset to achieve better privacy, both metrics have to be taken into consideration.

V. CONCLUSION

Privacy characterization and quantification methods are applied to deal with the problem of privacy quantification in privacy preserving data publishing. In order to consider the privacy loss of combined attributes, we presented data publishing as a multi-relational model. We redefined the prior and posterior beliefs of the adversary. The proposed model and adversarial beliefs contribute to a more precise privacy characterization and quantification. Supported by insightful examples, we then showed that privacy could not be quantified based on a single metric. We proposed two different privacy leakage metrics. Based on these metrics, the privacy leakage of any given PPDP technique could be evaluated.

REFERENCES

[1] Dimitrios Karapiperis, Aris Gkoulalas-Divanis and Vassilios S. Verykios, "FEDERAL: A Framework for Distance-Aware Privacy-Preserving Record Linkage", Transactions on Knowledge and Data Engineering, February 2018.
[2] Boyu Li, Yanheng Liu, Xu Han and Jindong Zhang, "Cross-Bucket Generalization for Information and Privacy Preservation", IEEE Transactions On Knowledge And Data Engineering, Vol. 30, No. 3, March 2018.
[3] M.H. Afifi, Kai Zhou and Jian Ren, "Privacy Characterization and Quantification in Data Publishing" IEEE Transactions on Knowledge and Data Engineering, Volume: 30, Issue: 9, September 2018.
[4] D. Rebollo-Monedero, J. Forne and J. Domingo-Ferrer, "From tcloseness-like privacy to postrandomization via information theory," IEEE Trans. on Knowl. and Data Eng., vol. 22, pp. 1623–1636, Nov. 2010.
[5] L. Sankar, S. R. Rajagopalan and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," Trans. Info. For. Sec., vol. 8, pp. 838–852, June 2013.
[6] M. Gotz, S. Nath and J. Gehrke, "Maskit: Privately releasing user context streams for personalized mobile applications," in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12, pp. 289–300, ACM, 2012.
[7] D. Vatsalan and P. Christen, "Privacy-preserving matching of similar patients," JBI, vol. 59, pp. 285 – 298, 2016.

