# Privacy Characterization and Quantification in Data Publishing

C.Suguna Devi, Assistant Professor
S.M.Annu Karishma, Student, C.Navitha, Student, J.Akhila, Student,
V.Chandra Lekha, Student, S.Mohammed Burhan, Student
Computer Science And Engineering
Annamacharya Institute of Technology And Sciences,Rajampet,India.

**Abstract—**The increasing interest in collecting and publishing large amounts of individuals' data as public for purposes such as medical research, market analysis, and economical measures has created major privacy concerns about individual's sensitive information. To deal with these concerns, many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed in literature. However, they lack a proper privacy characterization and measurement. In this paper, we first present a novel multi-variable privacy characterization and quantification model. Based on this model, we are able to analyze the prior and posterior adversarial belief about attribute values of individuals. Based on our framework and the proposed metrics, we can determine that all the existing PPDP schemes have limitations in privacy characterization. Our proposed privacy characterization and measurement framework contributes to better understanding and evaluation of these techniques. Thus, this paper provides a foundation for design and analysis of PPDP schemes.

**Index Terms—**Data privacy, data security, data publishing, big data, data mining, privacy quantification, privacy leakage

## 1 Introduction

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 whose data are in the table. To avoid possible identification of individuals from records in published data, uniquely identifying information such as names and social security numbers are generally removed from the table. 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 than would be possible without releasing the table. 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. A table satisfies k-anonymity if each quasi-identifier attribute in the table is indistinguishable from at least k ! 1 other quasi-identifier attributes; such a table is called a k-anonymous table. While k-anonymity protects identity disclosure of individuals by linking attacks, it is insufficient to prevent attribute dis-closure 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.

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. We show that 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.

## DATA PUBLISHING AND ATTACKS ON DATASETS

Privacy-Preserving Data Publishing. Datasets publishing naturally consists of two phases. Different parties first collect data from record owners in a phase known as the data col-lection phase. It is then managed by the data publisher and is released in a phase known as the data publishing phase. This data is published to a certain data recipient for the purpose of data mining or to the public for the purpose of pro-viding useful societal information that could be utilized in different areas including research.

Data is commonly published in two models, untrusted and trusted model. In the untrusted model, the data pub-lisher attempts to extract or manipulate sensitive information about record owners. To avoid such attempts, record owners apply cryptographic operations on the published data to prevent the publisher from accessing sensitive information. In the trusted model, the data publisher is assumed to be honest. In this model, record owners are not concerned about uploading their record to the publisher. However, when data is released to the public, the publisher guarantees that sensitive information or identity of the record owner is not revealed to any possible adversary.

Utility-Privacy Trade off. Data utility is in a natural conflict with data privacy. It is trivial that, from the perspective of data utility, it is best to publish a dataset as is, while from the perspective of data privacy, it is best to publish a mostly generalized dataset or even an empty one. Although this is easy to understand, as far as we know, including the information theoretic approaches proposed in, there is not yet a tight closed form relationship that fully model the utility-privacy trade off. We believe that the first step on the track of finding such a relationship is to better characterize and quantify both sides of the trade off. We note that the importance of studying data utility is undeniable and of great value as it definitely contributes to resolving the trade off modeling. In this paper, we focus on the data privacy side.

Data Disclosure Model. Data is usually released in the form of tables, where the rows are the records of individuals and columns are their corresponding attributes. Some of the attributes are for information only and not sensitive, while others are sensitive. For the information that is not being viewed as sensitive, when multiple records or maybe side information are combined, the individual maybe potentially identified. These attributes are generally referred to as quasi-identifiers QID, which may include information such as Zip-Code, Age, and Gender. The sensitive information may include attributes that can uniquely identify the individuals such as the social security or the driving license numbers. These attributes are called explicit-identifiers. Another type of information being considered sensitive may include information such as disease and salary. When datasets are published, all explicit-identifiers are removed. Sensitive attribute disclosure occurs when the adversary learns information about an individual's sensitive attribute. This form of privacy breach is different and incomparable to learning whether an individual is included in the data-base, which is the focus of differential privacy.

Generalization and Anonymization. As the original dataset contains abundant information that could help an adversary link records to certain individuals, datasets are not published before being modified. Modifications could be accomplished in many ways. Basically, all modifications are listed under the anonymization operations. These operations might be in the form of generalization, suppression, anatomization, permutation, or perturbation. In generalization and suppression values of quasi-identifiers are somehow relaxed in case of generalization, or suppressed in case of suppression, to increase the range of individuals that carry the same quasi-identifier values and therefore increase the uncertainty of a possible adversary about certain individual's record. On the other hand, anatomization and permutation operations achieve anonymization by dissociation of quasi-identifiers and sensitive attributes. Perturbation mainly adds some noise to the whole dataset based on the statistical properties of the original data.

However, unlike statistical databases publishing individuals' data, also known as micro-data, requires that data remains intact after being released. Therefore not all the previously mentioned techniques are good candidates for anonymization of microdata. To keep data intact, and as much useful as possible, it is obvious that only generalization and suppression operations could be applied in privacy-preserving micro-data publishing techniques.

Attacks on Datasets. Generally, there are two types of attacks on datasets, record linkage and attribute linkage. The record linkage occurs when some values of quasi-identifier attributes can lead to the identification of a smaller number of records in the published dataset. In this case, an individual having these attribute values is vulnerable to being linked to a limited number of records. On the other hand, attribute linkage occurs if some sensitive values are predominate in a group, where an attacker has no difficulty to infer such sensitive values for the record owner belonging to this group.

Attribute linkage mainly consists of two types, homogeneity and background knowledge attacks. In homogeneity attacks, protection model may create groups that leak information due to lack of diversity in the sensitive attribute. In fact, some protection process is based on generalizing the quasi-identifiers but does not address the sensitive attributes that can reveal information to an attacker. In background knowledge attacks, an attacker can have prior knowledge that enables him to guess sensitive data with high confidence. These kinds of attacks depend on other information available to an attacker. Using this background knowledge, an adversary can disclose information in two ways, positive and negative disclosure. In positive disclosure, an adversary can correctly identify the value of a sensitive attribute with high probability. On the other hand, in

negative disclosure, the adversary can correctly eliminate some possible values of sensitive attribute with high probability. We also note that a background knowledge attack is difficult to prevent as compared to homogeneity attack.

# 3 ANALYSIS OF THE EXISTING PPDP SCHEMES

In this section, some representative PPDP schemes will be analyzed.

## 3.1 k-Anonymity

A table satisfies k-anonymity if every record in the table is indistinguishable from at least k ! 1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table. To satisfy this condition, before groups that share values of QIDs. Each group, named as an equivalence class ½C &, shares the same combination of quasi-identifiers and has at least k records. The idea of k-anonymity was proposed to combat record linkage attacks. Authors show that k-anonymity does not provide sufficient protection against attribute linkage.

To address the limitations of k-anonymity, introduced l-diversity as a stronger notion of privacy.

## 3.2 l-Diversity

An equivalence class is said to have l-diversity if there are at least l well-represented values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity. l-diversity represents an important step beyond k-anonymity in protecting against attribute linkage. However, it is susceptible to attacks such as skew-ness and similarity attacks. When the over-all distribution is skewed, satisfying the l-diversity does not prevent attribute linkage.

This leakage of sensitive information occurs because l-diversity does not take into account the semantical closeness of attribute values.

## 3.3 t-Closeness

An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.

# 4 OUR PROPOSED PUBLISHING MODEL AND PRIVACY CHARACTERIZATION

All previous approaches to characterize and quantify privacy have only investigated the privacy risk of publishing a sensitive attribute by focusing only on the change of belief of an adversary about the probability distribution of this attribute. However, we believe that any attribute by itself is not sensitive. The sensitivity of an attribute comes from combining it with other attributes. For example, cancer in a medical records dataset, high or low salaries in an employees dataset, are not sensitive unless they are linked to a certain geographical area, age-range or race. To obtain a meaningful definition of data privacy, it is necessary to characterize and quantify the knowledge about sensitive attributes that the adversary gains from observing the published dataset taking into consideration the combinational relation of different attributes. In our approach to characterize privacy, we employ a multi-dimensional scheme of privacy risk analysis attached with combining different attributes. Thus, we introduce the following combinational characterization of privacy.

Similarly we can find the privacy loss of individuals having other attribute values (other diseases) within the same class. One of our goals is to quantify this loss. In the next section, we propose two privacy metrics that are able to measure privacy leakage from two different perspectives.

# 5 OUR PROPOSED PRIVACY QUANTIFICATION

There is an immense amount of existing privacy loss quantification metrics in literature. The state-of-the-art approaches to measure privacy can be mainly sub-categorized into uncertainty, information gain or loss, similarity and diversity, and indistinguishability metrics. Uncertainty metrics measure the uncertainty in the adversarial estimate. The more uncertain the adversary is, the higher the achieved privacy in the published dataset. Information gain or loss met-rics quantify the amount of information gained by the adversary, or the amount of information lost by users after data publishing. High adversarial gain and high user's loss of information corresponds to low privacy. Similarity and diver-sity metrics measure the similarity or diversity between the original and the published dataset. High similarity or low diversity between the two datasets corresponds to low privacy. Indistinguishability measures the ability of an adversary to distinguish between two outcomes of a privacy preserving data publishing technique. Privacy is high if it is hard for an adversary to distinguish between any pair of outcomes.

Our approach to quantify privacy mainly depends on understanding when information leakage happens and how this leakage could be measured. To have a better under-standing of when leakage occurs, we revisit the two states of knowledge of an adversary before and after a table T is published. At the first state of knowledge, based on public information of sensitive

attribute's distribution, an adversary has some prior belief about the attribute value of an individual. This prior belief is in the form of probability distributions of attributes and joint distributions of their combinations.

After publishing the table, an adversary moves to the second state of knowledge to gain some more information about the individual. This amount of information is the leakage that we need to capture where it enables us to mea-sure the extent to which this data-publishing model minimizes privacy leakage. We now analyze this leakage and find a set of appropriate metrics that contribute to a better quantification of privacy represented in the amount of uncertainty an adversary has about an individual's sensitive attribute value after a table is published.

# 6 EMPIRICAL ANALYSIS AND SIMULATION RESULTS

This section is divided into two parts. In the first part, based on our findings, we introduce a wide set of empirical examples for different case scenarios that support our findings. The provided examples aim to help understand the implications of the proposed metrics and show how these metrics contribute to analyzing, comparing and evaluating the previously mentioned existing privacy-preserving data-publishing techniques. In the second part of this section, aided with our simulation results, we focus on instances where different PPDP techniques assume to achieve an intended privacy level. However, based on our proposed metrics, they fail to express, and therefore fail to avoid, a considerable amount of privacy leakage. Throughout this section, we assume that an adversary has no other side information about dataset statistics or the user of interest other than the determined quasi-identifier values.

## 6.1 Empirical Analysis

We begin by giving examples to show how the distribution and the entropy leakages are two different measures of privacy leakage.

## 6.2 Simulation Results

In our simulations, we investigate the effectiveness of differ-ent PPDP techniques based on our privacy metrics. Simulation results give us a more insightful understanding of privacy leakage. Specifically, our analysis gives a spotlight on several instances where published tables are believed to achieve privacy based on the PPDP techniques utilized, while based on our metrics, they do leak private valuable information about users in the datasets. We also show how our proposed metrics enable a data publisher to have more control over the privacy of a specific group of users having certain sensitive attribute values.

# 7 CONCLUSION

In this paper, we introduced comprehensive characterization and novel quantification methods of privacy to deal with the problem of privacy quantification in privacy-pre-serving data publishing. In order to consider the privacy loss of combined attributes, we presented data publishing as a multi-relational model. We re-defined 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. Our experiments demonstrate how we could gain a better judgment of existing techniques and help analyze their effectiveness in reaching privacy.

Our work opens doors to a wide range of research problems and questions including whether two metrics are sufficient to evaluate privacy or there exist other independent metrics that could help achieve better privacy quantification. Another open problem is the optimization of the original data generalization as to achieve maximum privacy based on our proposed metrics. Typically, we believe that equivalence classes should be designed in such a way that keeps both the entropy leakage and the distribution leakage below a certain pre-determined level. This motivates us to think of a typical publishing scenario. We also leave as an open problem for further research, optimization of the chosen set of quasi-identifiers with an objective of minimizing distribution and entropy leakages within the published table or specific classes of higher privacy concerns.

# REFERENCES

[1] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowl.-Based Syst., vol. 10, no. 5, pp. 557– 570, 2002.

[2] L. Sweeney, "Uniqueness of simple demographics in the U.S. pop-ulation," LIDAPWP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, 2000.

[3] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proc. Secur. Privacy, 2008, pp. 111–125.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubrama-niam, "'-diversity: Privacy beyond k-anonymity," ACM Trans. Knowl. Discovery Data, vol. 1, Mar. 2007, Art. no. 3.

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in Proc. IEEE 23rd Int. Conf. Data Eng., 2007, pp. 106–115.

[6] N. Li, W. Qardaji, D. S. Purdue, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy defi-nitions," in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2013, pp. 889–900.

[7] I. Wagner and D. Eckhoff, "Technical privacy metrics: A system-atic survey," eprint arXiv:1512.00327, 2015.

[8] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "Privbayes: Private data release via bayesian networks," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2014, pp. 1423– 1434.

[9] M. Gotz,€ S. Nath, and J. Gehrke, "Maskit: Privately releasing user context streams for personalized mobile applications," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2012, pp. 289–300.

[10] Y. Rubner, C. Tomasi, and , L. J. Guibas, "The earth mover's dis-tance as a metric for image retrieval," Int. J. Comput. Vis., vol. 40, no. 2, pp. 99–121, 2000.