A Study Of Novel Anonymization Techniques For Secure Data Publishing

G. Vijay Kanth[1], B. Santha Kumar[2]
[1]*M.Tech(CSE), Sri Kottam Tulasi Reddy Memorial College of Engineering, Kondair,Andhra Pradesh, India.*
[2]*Asst.Professor, Sri Kottam Tulasi Reddy Memorial College of Engineering, Kondair, Andhra Pradesh, India.*

Abstract

The information available around the world is sharing on the internet has greatly improved the productivity of our society but also increased the risk of privacy violations. Privacy preserving data publishing renders approaches and methods for sharing useful information in the form of publication while preserving data privacy.

This study provides a research motivated by real world problems with certain challenging issues to be addressed and helps us to identify challenges, focus on research efforts and highlight the future directions.

In this paper, we present brief yet systematic review of several anonymization techniques such as generalization and bucketization, have been designed for privacy preserving micro data publishing.

Keywords: Privacy, Preservation, Publication, Anonymization, Generalization, Bucketization

1. Introduction

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

The explosive growth of interest has increased the dependence of both organizations and individuals on sharing information universally. This has led to an ever-increasing demand on protect information from unintended use and to guarantee the privacy.

Privacy Preserving data publishing, abbreviated as PPDP, emerged to address the privacy issues.

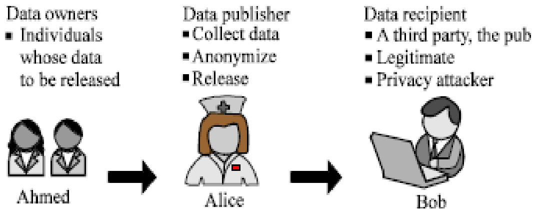


Figure 1: Process Model of PPDP

Privacy Principles:

Generally speaking, privacy is the claim of individuals to control when, how and to what extend information about them is communicated to others. A Privacy protection principles enables users to specify the level of privacy protection against a certain type of privacy risk. In PPDP, k-anonymity and l-diversity are well known principles.

Data Anonymization:

Data anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Detailed data (also called as microdata) contains information about a person, a household or an organization. Most popular anonymization techniques are *Generalization and Bucketization.* There are number of attributes in each record which can be categorized as

1) *Identifiers* such as *Name or Social Security Number* are the attributes that can be uniquely identify the individuals.
2) Some attributes may be Sensitive Attributes(SAs) such as *disease* and *salary* and
3) Some may be Quasi-Identifiers (QI) such as *zipcode*, *age*, and *sex* whose values, when taken together, can potentially identify an individual.

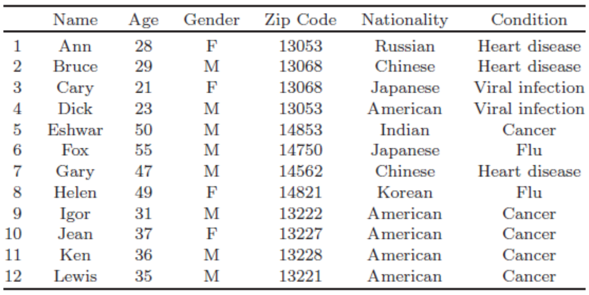
2. K-Anonymity and L-Diversity

This section clearly studies how k-anonymity and l-diversity are the well known principles for privacy preservation.

K-anonymity: Sweeney in demonstrated that releasing a data table by simply removing identifiers (e.g., names and social security numbers) can seriously breach the privacy of individuals whose data are in the table. By combining a public voter registration list and a released medical database of health insurance information, she was able to identify the medical record of the governor of Massachusetts. In fact, according to her study of the 1990 census data, 87% of the population of the United States can be uniquely identified on the basis of their five-digit zip code, gender, and date of birth.

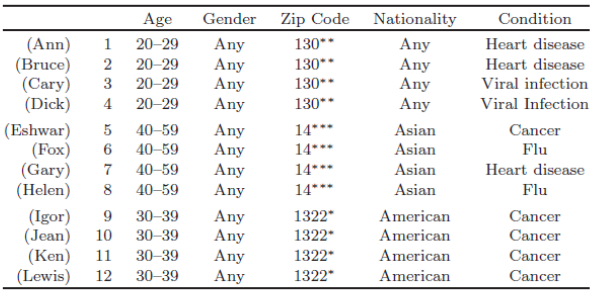
This kind of attack is called *linking attack* Take table 2 for example. Suppose that we remove the Name attribute and release the resulting table. It is common that the adversary has access to several public databases. For instance, he can easily obtain a public voter registration list as shown in Table 3. Assume the area of zip code 13068 is a small town and Ann is the only 28 year old female living in that town. When the adversary looks at Table 2 with names removed, he can almost be sure that the first record with *Age* = 28, *Gender* = F, and *Zip code* = 13068 is Ann's record by matching that record with Ann's record in the voter registration list. The goal of a linking attack is to find the identity of an individual in a released data set that contains no identifying attributes by linking the records in the data set to a public data set that contains identifying attributes. This linkage is performed with a set of *quasi-identifier* (QI) attributes that are in both data sets. In the above example, *Age*, *Gender*, and *Zip code* are QI attributes.

Table 1: Sample Record Table



	Name	Age	Gender	Zip Code	Nationality	Condition
1	Ann	28	F	13053	Russian	Heart disease
2	Bruce	29	M	13068	Chinese	Heart disease
3	Cary	21	F	13068	Japanese	Viral infection
4	Dick	23	M	13053	American	Viral infection
5	Eshwar	50	M	14853	Indian	Cancer
6	Fox	55	M	14750	Japanese	Flu
7	Gary	47	M	14562	Chinese	Heart disease
8	Helen	49	F	14821	Korean	Flu
9	Igor	31	M	13222	American	Cancer
10	Jean	37	F	13227	American	Cancer
11	Ken	36	M	13228	American	Cancer
12	Lewis	35	M	13221	American	Cancer

Table 2: Generalized Record Table



		Age	Gender	Zip Code	Nationality	Condition
(Ann)	1	20–29	Any	130**	Any	Heart disease
(Bruce)	2	20–29	Any	130**	Any	Heart disease
(Cary)	3	20–29	Any	130**	Any	Viral infection
(Dick)	4	20–29	Any	130**	Any	Viral Infection
(Eshwar)	5	40–59	Any	14***	Asian	Cancer
(Fox)	6	40–59	Any	14***	Asian	Flu
(Gary)	7	40–59	Any	14***	Asian	Heart disease
(Helen)	8	40–59	Any	14***	Asian	Flu
(Igor)	9	30–39	Any	1322*	American	Cancer
(Jean)	10	30–39	Any	1322*	American	Cancer
(Ken)	11	30–39	Any	1322*	American	Cancer
(Lewis)	12	30–39	Any	1322*	American	Cancer

To protect data from linking attacks, Samarati and Sweeney proposed *k*- anonymity. Let *D* (e.g., Table 2) denote the original data table and *D*∗ (e.g., Table 1.2) denote a release candidate of *D* produced by the generalization mechanism.

Definition (*k*-Anonymity). Given a set of QI attributes *Q*1,...,*Qd*, release candidate *D*∗ is said to be *k*-anonymous with respect to *Q*1,...,*Qd* if each unique tuple in the projection of *D*∗ on *Q*1,...,*Qd* occurs at least *k* times.

Table 2 is 4-anonymous. Now, no matter what public databases the adversary has access to, he can only be sure that Ann's record is one of the first four. While k-anonymity successfully protects data from linking attacks, an individual's private information can still leak out. For example, the last four individuals of Table 2 have cancer. Although the adversary is not able to know which record belongs to Jean, he is sure that Jean has cancer if he knows Jean's age, gender, and zip code from a public database. This motivated Machanavajjhala et al., who propose the principle of -diversity, which is presented in the next section.

In practice, multiple criteria should be enforced at the same time in order to protect data from different kinds of attacks. We note that, for a given data publication scenario, the issues of setting the parameter k and deciding which attributes to include in the set of QI attributes have not been well-addressed in the literature. For the second question, a simple approach that has often been taken is to conservatively include

all of the non-sensitive attributes in the set of QI attributes. However, further research is still needed to develop principles to help determine the right k value for a given scenario.

L-Diversity: k-Anonymity ensures that individuals cannot be uniquely re-identified in a data set and thus guards against linking attacks. However, Machanavajjhala et al. showed that adversaries with more background knowledge, also called adversarial knowledge, can infer sensitive information about individuals even without re-identifying them. The following two attacks—*homogeneity attack* and *background knowledge attack illu*strate such adversaries.

In order to guarantee privacy against such adversaries, Machanavajjhala et al. first propose a formal but impractical definition of privacy called Bayes-Optimal privacy. The attributes in the input table are considered to be partitioned into non-sensitive QI attributes (called *Q*) and sensitive attributes (called *S*). The adversary is assumed to know the complete joint distribution *f* of *Q* and *S*. Publishing a generalized table breaches privacy according to Bayes-Optimal privacy if the adversary's prior belief in an individual's sensitive attribute is very different from the adversary's posterior belief after seeing the published generalized table. More formally, adversary Alice's prior belief, *α*(*q*,*s*), that Bob's sensitive attribute is *s* given that his non-sensitive attribute is *q*, is her background knowledge:

$$\alpha_{(q,s)} = P_f(t[S] = s \mid t[Q] = q) = \frac{f(s,q)}{\sum_{s' \in S} f(s,q)},$$

where *t*[*S*] and *t*[*Q*] denote the sensitive value and the vector of QI attribute values of individual *t*, respectively; *Pf* denotes the probability computed based on distribution *f*. On observing the published table *T* which is generalized from *T*, and in which Bob's quasi-identifier *q* has been generalized to *q*∗, her posterior belief about Bob's sensitive attribute is denoted by *β*(*q*,*s*,*T*) and is equal to:

$$\beta_{(q,s,T^\star)} = P_f(t[S] = s \mid t[Q] = q \text{ and } T^\star \text{ and } t \in T)$$

Given the joint distribution *f* and the output table *T*, Machanavajjhala et al. derived a formula for *β*(*q*,*s*,*T*).

Definition (Recursive (*c*,*l*)-Diversity): In a given *q*-block, let *ri* denote the number of times the *i*-th most frequent sensitive value appears in that *q*_-block. Given a constant *c*, the *q*-block satisfies *recursive* (*c*,*l*)*-diversity* if *r*1 < *c*(*r*_ + *r*_+1 + · · · + *rm*). A table *T* satisfies recursive (*c*, *l*) diversity if every *q*-block satisfies recursive _-diversity. We say that l-diversity is always satisfied.

The recursive (*c*,*l*)-diversity, thus, can be interpreted in terms of adversarial background knowledge. It guards against all adversaries who possess atmost 2 statements of the form "Bob does not have heart disease". We call such statements as *negation* statements.

3. Anonymization Techniques

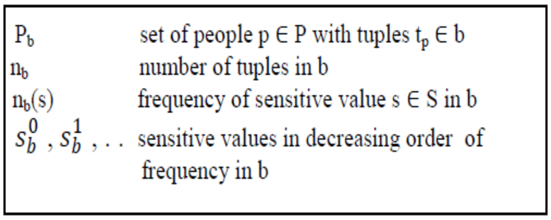
Anonymization refers to the PPDP approaches that aim to hide the identify and sensitive information of individuals by transforming data to observe a particular privacy principles. Anonymization techniques can be broadly categorized as generalization and suppression and perturbation.

Generalization and Suppression Techniques: Generalization involves replacing specific values with a more general one. Suppression can be deemed as generalizing a value to unknown value. LeFevre et.al(2005) categorizes generalization models into two classes. The first class is hierarcy-based models which use fixed value generalization hierarchies and are more suitable for categorical data. The second class is partition-based models which require the domain of an attribute to be a totally ordered set, define generalizations by partitioning the set into disjoint ranges and are most suitable for numerical data. Generalization models can be classified as follows.

Full domain generalization proposed by Samarati and Sweeney(1998) requires that all generalized values of an attribute in the anonymized data must be on the same level of the taxonomy tree of the attribute's domain.

Full subtree generalization proposed by lyengar (2002) requires that the child values sharing a common parent value are either all or none generalized and each generalization is applied to all records. This techniques is more flexible than the full domain generalization.

The first, which we term *bucketization*, is to partition the tuples in T into *buckets*, and then to separate the sensitive attribute from the non-sensitive ones by randomly permuting the sensitive attribute values within each bucket. The sanitized data then consists of the buckets with permuted sensitive values. In this paper we use bucketization as the method of constructing the published data from the original table T, although all our results hold for full-domain generalization as well. We now specify our notion of bucketization more formally. Partitition the tuples into buckets (i.e., horizontally partition the table T according to some scheme), and within each bucket, we apply an independent random permutation to the column containing S-values. The resulting set of buckets, denoted by B, is then published. For example, if the underlying table T, then the publisher might publish bucketization B .Of course, for added privacy, the publisher can completely mask the identifying attribute (Name) and may partially mask some of the other non-sensitive attributes for a bucket b € B the following notation.



P_b	set of people p ∈ P with tuples t_p ∈ b
n_b	number of tuples in b
$n_b(s)$	frequency of sensitive value s ∈ S in b
s_b^0 , s_b^1 , . . .	sensitive values in decreasing order of frequency in b

While bucketization has better data utility than generalization, it has several limitations. First, bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. As shown in, 87 percent of the individuals in the United States can be uniquely identified using only three attributes (Birthdate, Sex, and Zipcode). A microdata (e.g., census data) usually contains many other attributes besides those three attributes. This means that the membership information of most individuals can be inferred from the bucketized table.

Second, bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. Third, by separating the sensitive attribute from the QI attributes, bucketization breaks the attribute correlations between the QIs and the SAs.

Bucketization first partitions tuples in the table into buckets and then separates the quasi identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. The anonymized data consist of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high-dimensional data. However, their approach assumes a clear separation between QIs and SAs. In addition, because the exact values of all QIs are released, membership information is disclosed.

4. Challenges

PPDP is quite challenging as the legitimate data recipient is a privacy adversary which results in the hardness to reach the optimal balance between privacy and data utility. Therefore, several questions remain open: can be optimal solution with a flexible anonymization model gain utility significantly can be customized privacy principle improve utility, is it practical to find an optimal solution efficiently for real world data?

While emerging applications of privacy provide their own set of challenges, there are some challenges that are application independent.

a. The curse of dimensionality
b. Sequential Releases and Compensability
c. Obtaining Privacy Preferences and Setting Parameters.

5. Conclusion

In an increasingly data-driven society, personal information is often collected and distributed with ease. In this paper, we have presented an overview of recent technological advances in defining and protecting individual privacy and confidentiality in data publishing. In particular, we have focused on organizations, such as hospitals and government agencies, that compile large data sets, and must balance the privacy of individual participants with the greater good for which the aggregate data can be used. While technology plays a critical role in privacy protection for personal data, it does not solve the problem in its entirety. In the future, technological advances must dovetail with public policy, government regulations, and developing social norms.

The research community has made great strides in recent years developing new semantic definitions of privacy, given various realistic characterizations of adversarial knowledge and reasoning. However, many challenges remain, and we believe that this will be an active and important research area for many years to come.

References

[1]. Neha V. Mogre Prof. Girish Agarwal Prof. Pragati Patil – “ A Review On Data Anonymization Technique For Data Publishing”, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[2]. Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy “Slicing: A New Approach for Privacy Preserving Data Publishing” Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 3, MARCH 2012.

[3] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati On K-Anonymity. In Springer US, Advances in Information Security (2007).

[4]. Junquang Liu – “ Privacy Preserving Data Publication: Current status and New Directions”, Information Technology Journal 11, ISSN 1812-5638, 2012.

[5]. Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala – “Privacy Preserving Data Publishing”, Foundations and Trends in Databases Vol. 2, Nos. 1–2 (2009) 1–167