Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080321

A Survey of Privacy Preserving Data Publishing using Generalization and Suppression

Yang Xu[1], *Tinghuai Ma*[2,*], *Meili Tang*[3] *and Wei Tian*[1]

[1] School of Computer and Software, Nanjing University of Information Science & Technology University, Nanjing, 210044, China
[2] Jiangsu Engineering Centre of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[3] School of Public Administration, Nanjing University of Information Science & Technology University, Nanjing, 210044, China

Received: 23 May. 2013, Revised: 27 Sep. 2013, Accepted: 28 Sep. 2013
Published online: 1 May. 2014

Abstract: Nowadays, information sharing as an indispensable part appears in our vision, bringing about a mass of discussions about methods and techniques of privacy preserving data publishing which are regarded as strong guarantee to avoid information disclosure and protect individuals' privacy. Recent work focuses on proposing different anonymity algorithms for varying data publishing scenarios to satisfy privacy requirements, and keep data utility at the same time. *K*-anonymity has been proposed for privacy preserving data publishing, which can prevent linkage attacks by the means of anonymity operation, such as generalization and suppression. Numerous anonymity algorithms have been utilized for achieving *k*-anonymity. This paper provides an overview of the development of privacy preserving data publishing, which is restricted to the scope of anonymity algorithms using generalization and suppression. The privacy preserving models for attack is introduced at first. An overview of several anonymity operations follow behind. The most important part is the coverage of anonymity algorithms and information metric which is essential ingredient of algorithms. The conclusion and perspective are proposed finally.

Keywords: Data Publishing, Privacy Preserving, Anonymity Algorithms, Information Metric, Generalization, Suppression

1 INTRODUCTION

Due to the rapid growth of information, the demands for data collection and sharing increase sharply. A great quantity of data is used for analysis, statistics and computation to find out general pattern or principle which is beneficial to social development and human progress. Meanwhile, threats appear when tremendous data available for the public. For example, people can dig privacy information by getting together safe-seeming data, consequently, there is a great possibility exposing individuals privacy. According to the study, approximately 87 % of the population of the United States can be uniquely identified by given dataset published for the public. To avoid this situation getting worse, measures are taken by security department of many countries, for example, promulgating privacy regulation (e.g. privacy regulation as part of Health Insurance Portability and Accountability Act in the USA [1]). The requirement for data publisher is that data to be published must fit for the predefined conditions. Identifying attribute needs to be omitted from published dataset to guarantee that individuals privacy cannot be inferred from dataset directly. Removing identifier attribute is just the preparation work of data processing, several sanitization operations need to be done further. However, after data processing, it may decrease data utility dramatically, while, data privacy did not get fully preserved.

In face of the challenging risk, some researches have been proposed as a remedy of this awkward situation, which target at accomplishing the balance of data utility and information privacy when publishing dataset. The ongoing research is called Privacy Preserving Data Publishing (PPDP). In the past few years, experts have taken up the challenge and undertaken a lot of researches. Many feasible approaches are proposed for different privacy preserving scenario, which solve the issues in PPDP effectively. New methods and theory come out

* Corresponding author e-mail: thma@nuist.edu.cn

continuously in experts' effort to complete privacy preserving.

1.1 Privacy Preserving Data Publishing

Generally, the process of Privacy Preserving Data Publishing has two phases, data collection and data publish phase. It refers to three kinds of roles in the process who are data owner, data publisher and data recipient. The relationship of two phases and three roles involved in PPDP is shown in figure 1. In the data collection phase, data publisher collects dataset from data owner. Then, in the data publishing phase, data publisher sends the processed dataset to data recipient. It is necessary to mention that raw dataset from data owner cannot be directly sent to data recipient. The dataset should be processed by data publisher before being sent to data recipient.

Fig. 1: the relationship of phases and roles in PPDP

In [2], data publisher can be divided into two categories. In the untrusted model, data publisher is tricky who is more likely to gain privacy from dataset. In the trusted model, data publisher is reliable and any data in their hands is safe and without any risk. Owing to the difference of data publishing scenarios affected by varying assumptions and requirements to data publisher, data recipients purposes and other factors, it gives four scenarios for further detailed discussion that maybe appear in real privacy preserving data publishing in [3]. The first scenario is *the non-expert data publisher*. In this scenario, data publisher does not need to have specific knowledge about research fields. What they need to do is make data be published satisfying the requirements of data utility and information privacy. The second one is *the data recipient could be an attacker*. This scenario is more commonly-accepted and many proposed solutions make it as the requisite hypothesis. The third one is *the publish data is not the data mining result*. It indicates that dataset provided by data publisher in this scenario is not merely for data mining. That is to say, published dataset is not data mining result. The last one is *truthfulness at record level*. Data publisher should guarantee the authenticity of data to be published whatever processing methods will be used. Thus, randomization and perturbation cannot meet the requirements in this scenario.

1.2 K-*Anonymity*

When referring to data anonymization, the most common data is two-dimensional table in relational database. For privacy preserving, the attributes of table are divided into four categories which are *identifier*, *quasi-identifiers*, *non-quasi attributes* and *sensitive attribute*. *Identifier* can uniquely represent an individual. Obviously, it should be removed before data processing. *Quasi-identifiers* are a specific sequence of attributes in the table that malicious attackers can take advantage of these attributes linking released dataset with other dataset that has been already acquired, then breaking privacy, eventually gaining sensitive information. Data sanitization operated by data publisher mainly targets on quasi-identifiers. Due to uncertainty of the number of quasi-identifiers, each approach of PPDP assumes the quasi-identifiers sequence in advance. Only in this way can the following processing carry out. *Non-quasi attributes* have less effect on data processing. For this reason, sometimes, these attributes does not turn up in the progress of data processing which tremendously decrease memory usage and improve the performance of the proposed algorithm. *Sensitive attribute* contains sensitive information, such as disease, salary. From table 1(2), this is a two-dimensional table to be published. According to above introduction, we can get the conclusion that *ID* is identifier. If table 1(1) is a known table which attacker will use as background knowledge, then we know *Birthday*, *Sex* and *ZipCode* are quasi-identifiers, *Work* is non-quasi attribute and *Disease* is sensitive attribute.

From the example above, we know why data processing steps mainly work on quasi-identifiers. Only in this way can we reduce the correlation of dataset to be published and other dataset. In PPDP, the progress of data processing is called data anonymization. *K*-anonymity is one of anonymization approaches proposed by Samarati and Sweeney[4] that each record in dataset cannot be distinguished with at least another (*k-1*) records under the projection of quasi-identifiers of dataset after a series of anonymity operations (e.g. replace specific value with general value). *K*-anonymity assures that the probability of uniquely representing an individual in released dataset will not great than *1/k*. For example in table 1, we learn about Miss Yoga has diabetes by linking census data table with patient data table by *Birthday*, *Sex* and *ZipCode* attributes even removing identifier. What if it cannot uniquely determine a record? Thus attacker has no ability to identify sensitive information with full confidence. How to make patient table in Table 1 meet 2-anonymity? One of practical ways is that replacing data with year for

Table 1: Illustrate anonymization and *k*-anonymity

(1) Census Data

Name	Birthday	Sex	ZipCode
Myron	1990/10/01	Male	210044
Yoga	1980/05/22	Female	210022
James	1782/06/23	Male	210001
Sophie	1992/03/12	Female	210012

(2) Patient Data

ID	Work	Birthday	Sex	ZipCode	Diease
231001	Student	1990/10/01	Male	210044	Cardipathy
231002	Clerk	1980/05/22	Female	210022	Diabetes
231003	Official	1990/08/12	Male	210021	Flu
231004	HR	1980/02/25	Female	210012	Caner

Birthday attribute and using * replace the last two character of *ZipCode* attribute. *K*-anonymity has been extensively studied in recent years [5,6,7,8]. After 2-anonymity, it cannot infer that Miss Yoga has diabetes, or maybe she has cancer. Because in patient data table, there are two records that can be linked to one record in census data table about Miss Yoga. We can see that *k*-anonymity has an effective impact on this scenario.

1.3 Paper Overview

This paper mainly refers to four topics that are privacy model, anonymity operation, information metric and anonymization algorithm. Due to different kinds of attacks to steal privacy, it forms different privacy preserving models for these attacks accordingly. Every privacy preserving model has its feature, so that researchers propose some theory and method for each type of attack. Algorithm implementation is based on specific theory and methodology. So each anonymity algorithm belongs to the specific privacy preserving model. As to anonymity operation and information metric, they are the details of algorithms. Anonymity operation is the core of algorithm, an algorithm often keep one or two operations in mind, and finally make the processed dataset to meet privacy requirement. The information metric is incorporated into the algorithm to guide its anonymity process or execution, and finally get better result rather than just get a rare result. Therefore, these four topics are essential parts of privacy preserving data publishing.

There are several essential operations to implement data anonymization that are generalization, suppression, anatomization, permutation and perturbation. Generalization and suppression usually replace the specific value of quasi-identifiers with general value. Generally, there exists a taxonomy tree structure for each quasi-identifier that is used for replacement. Anatomization and permutation decouple the correlation of quasi-identifier and sensitive attribute by separating then in two datasets. Perturbation distorts dataset by the means of adding noise, exchanging value or generating synthetic data that must keep some statistical properties of original dataset.

This paper focuses on the anonymization algorithms using generalization and suppression which are the frequent used anonymity operations to implement *k*-anonymity. In chapter 2, it gives a representation of privacy preserving models of PPDP and puts more emphasis on privacy model for record linkage attack. In chapter 3, it introduces the category of anonymity operation of PPDP. The majority of this paper is to introduce different information metric criteria and relevant algorithms using generalization and suppression which are put in chapter 4. Finally, chapter 5 is a summarized conclusion of this paper.

2 PRIVACY PRESERVING MODEL FOR ATTACKS

The rigorous definition of privacy protection by Dalenius [9] is that addressing to the published dataset should not increase any possibility of adversary to gain extra information about individuals, even with the presence of background knowledge. However, it is impossible to quantize the scope of background knowledge. Therefore, a transparent hypothesis taken by many PPDP literatures is that adversary has limited background knowledge. According to adversaries' attack principle, attack model can be classified into two categories, which are linkage attack and probabilistic attack.

2.1 Privacy Model for Attacks

The linkage attack is that adversary steals sensitive information by the means of linking with released dataset. It has three types of linkage, *record linkage*, *attribute linkage* and *table linkage*. Quasi-identifiers are known by

adversary beforehand is the common characteristic of linkage attack. Furthermore, adversary also grasps the basic information of individuals and wants to know their sensitive information under the scenarios of record linkage and attribute linkage. While, table linkage attack puts more emphasizes on the point that whether known individual's information presents in released dataset. The privacy model of record linkage will be elaborately described in the next section, which is the important part of this paper.

For the attack of attribute linkage, the adversary could infer sensitive information from the released dataset based on the distribution of sensitive value in the group that the individual belongs to. A successful inference is possible working on the published dataset that satisfies the qualifications of *k*-anonymity. The common effective solution for the attribute linkage attack is to lessen the correlation of quasi-identifiers and sensitive attributes of original dataset. Certainly, others models also bloom recently for capturing this kind of attack, like ℓ-diversity[10] and recursive (c, ℓ)-diversity[11], (X, Y)-Anonymity[12], (a, *k*)-Anonymity[13], (*k*, e)-Anonymity[14], t-closeness by Li et al.[15], personalized privacy by Xiao and Tao[16] and so on.

Table linkage is different from both record linkage and attribute linkage. In the table linkage attack, the presence or absence of individual record in released table has already revealed the sensitive information of the specific individual. Nergiz et al. proposed the theory of δ-presence to prevent table linkage and further bound the probability inferring occurrence of individual record within a given range[17].

The probabilistic attack can be depicted in the scenarios that adversary will not immediately scratch sensitive information from released dataset, while, the released dataset can do some favor for adversary through increasing his/her background knowledge to some extent. This kind of attack is called probabilistic attack that it turns up a visible deviation for gaining sensitive information after accessing the released dataset. Probabilistic attack is not like linkage attack which precisely knows individual information, then gain sensitive information combined with existed background knowledge, but it focuses on changing adversary's probabilistic confidence of getting privacy information after acquiring published dataset. The privacy model to this attack needs to ensure that the change of probabilistic confidence is relatively small after obtaining the published dataset. Some insightful notions for probabilistic attack are (c, t)-isolation[18], ε-differential privacy[19], (d, γ)-privacy[20], distributional privacy[21] and so on. Different privacy preserving model has its unique features determined by details of the vicious attack, so related algorithms which belong to a specific privacy model are customized and targeted at settling particular attack situation.

2.2 Privacy Model for Record Linkage

For record linkage attack, we must learn about the definition of equivalence class at first. When the values under the projection of quasi-identifiers of dataset are same, the certain numbers of records form a group. Many groups make up the dataset. Those groups are called equivalence class. In the original dataset, the size of equivalence class varies dramatically. If attackers known record of released dataset matching a group with only one record at the worst situation, unfortunately, the privacy information of individual related to the only one record will be leaked. For example, the dataset in Table 2(1) needs to be released. If publishing it without carrying out any anonymity operations and assuming that adversary has the background knowledge of Table 2(2). We can readily find that Myron who is born in Nanjing, China on 1990 has regular headache by linking the two datasets in table 2 on *Birthday*, *Sex* and *ZipCode*. These three attributes are called quasi-identifiers of this attack from the definition introduced above. *K*-anonymity is a method to solve record linkage attack which guarantees that the size of each equivalence class is greater or equal than the given value *k* by the means of replacing specific value with general value. The probability of uniquely inferring the sensitive information of individual known by the adversary is less than *1/k*, thus, it can safeguard individuals' privacy to a large extent. Each quasi-identifier has a taxonomy tree structure of which generalization extent increases from leaf to root node. Empirically, every categorical quasi-identifier has a predetermined taxonomy tree, while, the taxonomy tree of numerical quasi-identifier will be dynamically generated in the execution of anonymity algorithm, and, in addition, a specific value of numeric attribute will be replaced by a well-partitioned range in generalization. The taxonomy tree structures of two quasi-identifiers are shown in figure 2. For example, in taxonomy tree structures of *Job* attribute, the root node *ANY* is more general than node *Student*. The parent node *Student* is more general than its child node *Graduate*.

There are many exquisite methods to solve the problem of data anonymization, which obey qualification of *k*-anonymity or its extension. The detailed description will be introduced in subsequent chapter. With regard to *k*-anonymity, most of recent works assume that there exists only one quasi-identifier sequence containing all possible attributes. With the number of quasi-identifier increasing, not only does it take more effort to carry out one anonymity operation, but also level of data distortion increases respectively. So, some researchers propose a distinct standpoint taking multi quasi-identifier sequences into account which is more flexible than one quasi-identifier sequence. However, whatever way is chosen, the determination of attributes in quasi-identifier needs many attempts. No method or theory can deal with all issues in the specific research area. Afterwards, extensions of *k*-anonymity are proposed. Such as (X,

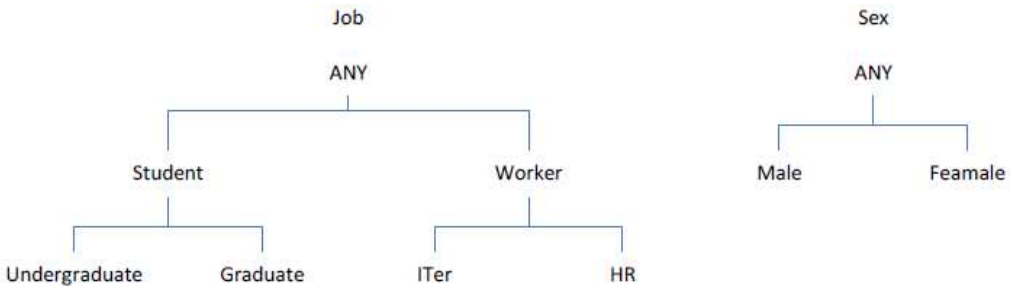

Fig. 2: taxonomy tree structure of quasi-identifier

Table 2: Illustrate record linkage

(1) Patient Data

Work	Birthday	Sex	ZipCode	Diease
Student	1990/10/01	Male	210044	Headache
Clerk	1980/05/22	Female	220022	Diabetes
Official	1990/08/12	Male	210021	Flu
HR	1980/02/25	Female	220012	Caner

(2) Background Knowledge

Name	Birthday	Sex	ZipCode
Myron	1990/10/01	Male	210044
Yoga	1980/05/22	Female	210022
James	1782/06/23	Male	210001
Sophie	1992/03/12	Female	210012

(3) 2-anonymous patient data

Work	Birthday	Sex	ZipCode	Diease
Student	1990	Male	2100**	Headache
Clerk	1980	Female	2200**	Diabetes
Official	1990	Male	2100**	Flu
HR	1980	Female	2200**	Caner

Y)-anonymity and MultiRelational *k*-anonymity[22]. (X, Y)-anonymity has more strict constraints than *k*-anonymity by appending additional requirements, and it is for the scenario that an individual is mapped to more than one record in released dataset, which means that the distinct number of Y attribute must greater or equal than the given *k* on the projection of X. MultiRelational *k*-anonymity expands the boundary of *k*-anonymity, which is for anonymizing multiple datasets instead of only one dataset. Basically, *k*-anonymity, (X, Y)-anonymity and MultiRelational *k*-anonymity constitutes the theory basis for privacy model for record linkage.

3 ANONYMITY OPERATIONS

A series of anonymity operations works on original dataset to make it fulfill the privacy requirement during data anonymization. The frequently used anonymity operations are generalization, suppression, anatomization, permutation and perturbation. Diverse algorithms toward privacy preserving data publishing differ in the choice of anonymity operations. Or, to put it in another way, the idea of algorithm is based on some specific anonymity operations.

3.1 Anonymity Operations

Generalization and suppression are the most common anonymity operations used to implement *k*-anonymity and its extension which are further depicted in the next session. Using one sentence to explain generalization is that replacing specific value of quasi-identifiers with more general value. Suppression is the ultimate state of generalization operation which uses special symbolic character to replace its authentic value (e.g. *, &, #), and makes the value meaningless. Unlike generalization and suppression, anatomization and permutation does not make any modification of original dataset, while decrease the correlation of quasi-identifiers and sensitive attribute. Generally, quasi-identifiers and sensitive attribute are published separately. Quite a few researches make use of this two anonymity operations[23,24,25]. When just referring to the purpose of information statistic, perturbation operation has merits of simplicity and efficiency. The main idea of perturbation is to substitute original value for synthetic data, and, ensures the statistical characteristic of original dataset. After perturbation operation, the dataset is completely not the presentation of original dataset which is its remarkable trait. Adding noise, swapping data and generating synthetic data are the three common means of perturbation [26,27,28,29,30].

3.2 Generalization and Suppression

Achieving *k*-anonymity by generalization and suppression will lead to not precise, but consistent representation of original dataset. Comprehensive consideration needs to be

taken about three vital aspects referred to PPDP, which are privacy requirement, data utility and algorithm complexity.

There are roughly four types of generalization with difference in scope and principle which are full-domain generalization, subtree generalization, cell generalization and multidimensional generalization. By the way, specification is the reverse anonymity operation of generalization.

1)**Full-domain generalization**[31] is proposed in early research of PPDP, it has the smallest search space in four types of generalization, while it leads to large data distortion. The key of full-domain generalization is that the value of quasi-identifier must be generalized to the same level in given taxonomy tree structure. We will use the taxonomy tree structure in figure 2 to explain. Before any anonymity operation, all values stay at the bottom of taxonomy. If node *Undergraduate* is generalized to its parent node *Student*, then node *Graduate* must be generalized to node *Student*, at the same time, nodes *ITer* and *HR* need to be generalized to node *Worker*.

2) **Subtree generalization**[32,33], its boundary is smaller than full-domain generalization. When a node in taxonomy tree structure generalizes to its parent node, all child nodes of the parent node need to be generalized to the parent node. For example, in figure 2, if node *Undergraduate* is generalized to its parent node *Student*, it needs to generalize node *Graduate* to its parent node *Student* to meet the requirement of subtree generalization. Unrestricted subtree generalization[34] is similar to the subtree generalization, except that siblings of the generalized node could remain unchanged. For example, in figure 2, if node *Undergraduate* is generalized to its parent node *Student*, node *Graduate* is unnecessary to generalize to its parent node *Student*.

3) **Cell generalization**[35] is slightly different from generalization ways above. Cell generalization is for single record, while, full-domain generalization is for all of records in the dataset. The search space with this generalization is significantly larger compared to other generalization, but the data distortion is relatively small. For example, in figure 2, when node *Undergraduate* generalizes to its parent node *Student*, it can maintain the record with *Undergraduate* value in the dataset. When the anonymous dataset is used for classification of data mining, it suffers from data exploration problem. For example, classifier may not know how to distinguish *Undergraduate* and *Student*. Those problems are the common traits of local recoding scheme.

4) **Multidimensional generalization**[34,35,36] emphasizes different generalization for different combination of values of quasi-identifiers. For example, in figure 2, [*Undergraduate*, *Female*] can be generalized to [*Student*, *ANY*], while [*Graduate*, *Male*] generalizes to [*ANY*, *Male*]. This scheme has less data distortion compared to full domain generalization. It generalizes records by combination of quasi-identifiers with different value.

In most situations of generalization schemes, it mixes suppression operations in its process of data anonymization. It is without any doubts that there exists some theory or techniques to go on data anonymization only using suppression operation [37,38,39]. Like the category of generalization, there are five kinds of suppression, attribute suppression[34], record suppression[40], value suppression[41], cell suppression[42] and multidimensional suppression[43]. Attribute suppression suppresses the whole values of the attribute. Record suppression means suppressing the records. Value suppression refers to suppressing the given value in the dataset. While cell suppression compared to cell generalization, works on small scope and suppresses some records with the given value in dataset.

4 ALGORITHMS USING GENERALIZATION AND SUPPRESSION

Before introducing anonymity algorithms using generalization and suppression, it is necessary to introduce the definition of minimal anonymity and optimal anonymity. Minimal anonymity means that original dataset through a series of anonymity operations satisfying the predefined requirements, and the sequence of used anonymity operations cannot be cut down. Compared with minimal anonymity, the requirements of optimal anonymity algorithms seem more rigorous. It demands that dataset after data anonymity operations satisfies privacy requirements, and, most importantly, the anonymous dataset needs to contain maximal quantity of information by the chosen information metric which will be described below. Hence, we can infer that the anonymous dataset with maximal information is definitely chosen from a collection of anonymous datasets which all meet the given privacy requirement. However, some previous works have proven that accomplishing optimal anonymity is NP-hard. It is generally accepted that generating anonymous dataset satisfying minimal anonymity requirement makes sense and can be achieved efficiently. The typical anonymity algorithms merely using generalization and suppression that will be narrated below and all algorithms belong to minimal anonymity or optimal anonymity algorithm. First, lets see the explanation of information metric criteria first.

4.1 Information Metric Criteria

During the process of data anonymization, the anonymity algorithms need maximally ensuring data utility, and satisfying privacy requirement at the same time. As a result, at each step of anonymity algorithm, it needs some qualified metrics criteria to guide the proceeding of algorithms. As the vital ingredient part of algorithms, it is necessary to take some effort to describe common

information metric frequently referred by privacy preserving algorithms, thus, it gives an overall and detailed scene of information metric next. Metric for Privacy Preserving Data Publishing (PPDP) can be roughly divided into two categories, which are data metric and search metric. Data metric is used to compare the discrepancy of anonymous and original dataset, which mainly refers to the aspect of data utility. While search metrics are usually incorporated into anonymity algorithms used for guiding execution of algorithm to get superior result. The availability of search metric directly affects efficiency of algorithm and effectiveness of anonymous result. According to the purpose of metric, general metric and trade-off metric naturally appear. From the name of general purpose, it is known that anonymous dataset restricted by it can apply for all kinds of application area. Referring to trade-off metric, it adds some additional measurement for specific purpose so that anonymous dataset confined by it has narrow scope of application area, but good performance. For uncertainty of application area consuming anonymous dataset, it is the best choice that uses metric with general purpose on the condition of satisfying relevant requirements. However, sometime, for the sake of gaining better analysis result from anonymous dataset in specific area, it is generally acknowledged that using trade-off metric instead of general purpose metric.

4.1.1 General Metric

General metric is used for the scenario that data publisher has no knowledge about application area of published dataset. For this reason, general metric leverages all kinds of factors, make it fit for different situation as much as possible by measuring discrepancy of anonymous and original dataset. In early works[42,43,44,45], the definition of data metric is not totally formed, and it just has the concept of minimal distortion that corresponding algorithm must comply with. For example, one measure of information loss was the number of generalized entries in the anonymous dataset and some algorithm judges the data quality by the ratio of the size of anonymous and original dataset.

Loss Metric (LM) for records is computed by summing up a normalized information loss of each quasi-identifier. Information loss (IL_{attr}) of each attribute is compute by

$$IL_{attr} = (N_g - 1)/(N - 1) \tag{1}$$

In the formula, N_g is represented for child count of the parent node which current value generalizes to and N is the child count of this quasi-identifier. For example, if *Undergraduate* generalizes to *Student* in figure 2, from the taxonomy tree structure of *Job* attribute, we can get that IL_{attr} is 1/3. If IL_{attr} equals 0, it means that the value is not generalized. This formula only fits for categorical attribute. While, calculation of information loss (IL_{attr}) for numeric attribute is slightly different from categorical attributes, as shown below. Usually, the value of numeric attribute with domain [U, L] is generalized to a specific range, like [U_g, L_g]. So the information loss of this generalization can be calculated by the formula 2. For example, if value 45 of *Age* attribute in figure 3 generalizes to the specific rage (0-45], and total range of age attribute is (0-99], so information loss (IL_{attr}) for this generalization is 45/99.

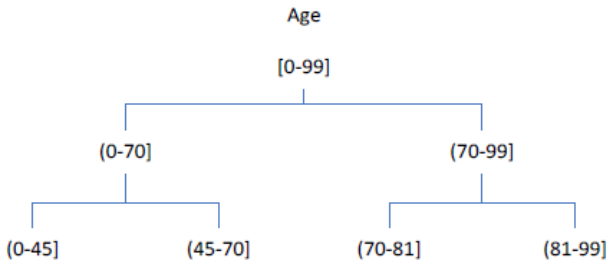

Fig. 3: taxonomy tree structure of numeric quasi-identifier

$$IL_{attr} = (N_g - L_g)/(U - L) \tag{2}$$

Loss metric (IL_{record}) of a record is calculated by the sum of information loss of quasi-identifiers under the assumption that each quasi-identifier has equal weight.

$$IL_{record} = \sum_{attr \in record} IL_{attr} \tag{3}$$

Therefore, we can infer that Loss metric for the whole dataset (IL_{table}) is based on information loss of generalized record like the formula shown below. This series of information metric also adopted by[46,47].

$$IL_{table} = \sum_{record \in table} IL_{record} \tag{4}$$

Quality measure is based on the size of the equivalence class E in dataset D. the discernibility metric (C_{DM}) assigns each record t in dataset D a penalty determined by the size of the equivalence class containing t. If a record belongs to an equivalent class of size s, the penalty for the record is s. If a tuple is suppressed, then it is assigned a penalty of $|D|$. This penalty reflects the fact that a suppressed tuple cannot be distinguished from any other tuple in the dataset. In the formula 5, the first sum computes penalties for each non-suppressed tuple, and the second for suppressed tuples.

$$C_{CM}(g,k) = \sum_{|E| \geq k} |E|^2 + \sum_{|E| < k} |D||E| \tag{5}$$

Another metric is proposed by Iyengar [32] which is used in privacy preserving data mining. This classification metric assigns no penalty to an unsuppressed tuple if it

belongs to the majority class within its induced equivalence class. All other tuples are penalized a value of 1. The minority function in formula 6 accepts a set of class-labeled tuples and returns the subset of tuples belonging to any minority class with respect to that set.

$$C_{CM}(g,k) = \sum_{|E|\geq k} (|minority(E)|) + \sum_{|E|<k} |E| \quad (6)$$

Normalized average equivalence class size metric (C_{AVG}) is proposed as an alternative. The intention of the metric is to measure how well the partitioning reaches the optimal case where each record is generalized in the equivalent class of k indistinguishable records. This sort of information metric also has many supporters[48].

$$C_{AVG} = (\frac{total\_records}{total\_equiv\_classes})/(k) \quad (7)$$

Precision metric is based on taxonomy tree structure. If the original dataset $D(A_1,\ldots,A_{N_a})$ and the anonymous dataset $D(A_1,\ldots,A_{N_a})$. Then

$$prec(D') = 1 - \sum_{j=1}^{N_a}\sum_{i=1}^{|D|} h/|VGH_{A_j}|/|D|\cdot|N_a| \quad (8)$$

N_a is the count of quasi-identifiers and h represents the height of taxonomy tree structure of A_j after generalization. VGH_{A_j} is total height of taxonomy tree structure of A_j. For any quasi-identifier A_j, the smaller of $prec(D')$, the worse of information loss.

4.1.2 Trade-off Metric

Trade-off Metric takes both of maximal data information and minimal privacy disclosure into consideration, and makes a superior balance of the two requirements at each step of algorithm execution. The difference between trade-off metric and general metric is that trade-off metric puts more emphasis on application scope of anonymous dataset provided by data publisher, not let two requirements alone at the same times. And, general metric considers more about the extent of data distortion between anonymous and original dataset.

Wang et al.[49] adapt the information loss metric based on information entropy. Their application scenario is that anonymous dataset can be used for classification of data mining, and it requires that classification model generated by anonymous dataset has rough equivalent effectiveness of the classification model by original dataset. Hence, the information metric needs considering the factor that how anonymous dataset is used for the construction of classification model. That is to say, it needs to know approach and mechanism of generating classification model. Generating classification model by the means of decision tree is one of common approaches. Meanwhile, using information gain based on entropy is one of viable measures to select splitting attribute, and further, generating the structure of decision tree step by step. It is reasonable that using entropy-based information metric to restrict data anonymization of original dataset at each step. It builds some underlying connection with classification model at the whole process of data anonymization by using entropy-based information metric. The formula for information loss is shown as formula 10. A generalization ($\{c \rightarrow p\}$) is denoted by G. R_c denotes the set of records in the dataset containing c and R_p denotes the set of records containing p after applying generalization G. We know that $|Rp|=\sum_c |Rc|$, where symbol $|x|$ represents the count of records in set x. The effect of a generalization G ($IP(G)$) is evaluated by both information metrics of information loss and anonymity gain.

$$IP(G) = I(G)/P(G) \quad (9)$$

The formula of information loss(I(G)) shows below.

$$I(G) = Info(R_p) - \sum_c \frac{|R_c|}{|R_p|} \star Info(R_c) \quad (10)$$

where the formula of Info(Rc) is shown below, which stands for the entropy of R_c. Selecting quasi-identifier with high entropy means that it uses less information to classify records in relevant class label and reflects minimal randomness or purity of the involved records, that is to say, records in the equivalent class have consistent class label which is an attribute of original dataset. When generating classification model, the deviation of records partition is small accordingly. $freq(R_x, cls)$ represents the count of records in R_x with class label cls

$$Info(R_c) = -\sum_{cls} \frac{freq(R_X,cls)}{|R_X|} \times \log_2 \frac{freq(R_X,cls)}{|R_X|} \quad (11)$$

The anonymity gain is calculated by

$$P(G) = A_G(VID) - A(VID) \quad (12)$$

VID denotes the sequence of quasi-identifier ($VID=\{D_1,\ldots,D_k\}$). $A(VID)$ and $A_G(VID)$ denote the anonymity before and after applying generalization G. Without doubt, $A_G(VID)\geq A(VID)$ and $A(VID)\geq k$. $A_G(VID)$-$A(VID)$ is the redundancy of the anonymity extent. It is unnecessary to have redundancy of anonymity under the requirement of k-anonymity, and the redundancy of k-anonymity means more information loss even though good for privacy protection. Comprehensively considering both two factors of information gain and anonymity gain, it comes to the formula 9. Information-Privacy metric has the tendency to select the quasi-identifier with minimal information loss for each extra increment of anonymity gain. If $P(G)$ equals 0, $IP(G)$ is ∞. On this occasion, using information loss $I(G)$ as metric selects quasi-identifier to be

generalized. Another formula for Information-Privacy metric is

$$IP(G) = I(G) - P(G) \tag{13}$$

However, *I(G)/P(G)* is superior to *I(G)-P(G)* by handling different quantification relationship. Because *IP(G)* considers anonymity requirement in algorithm execution, it helps focus the search process on the privacy goal, and has certain look-ahead effect.

In [50], on the basis of entropy, it proposes the monotone entropy metric and non-uniform entropy measure. They are simple variant of the entropy metric that respects monotonicity, and their definitions show below.

Definition of monotonicity: Let D be a dataset, let $g(D)$ and $g'(D)$ be two generalizations of D, and let $\prod$ be any measure of loss of information. Then, $\prod$ is called monotone if $\prod(D, g(D)) \leq \prod(D, g'(D))$ whenever $g(D) \in g'(D)$.

Definition of monotone entropy metric: Let $D=\{R_1,\ldots,R_n\}$ be a dataset having quasi-identifiers A_j, $1 \leq j \leq r$, and $g(D)=\{R_1,\ldots, R_n\}$ be a generalization of D. Then,

$$\prod_{me}(D, g(D)) = \sum_{i=1}^{n}\sum_{j=1}^{r} Pr(\overline{R}_i(j)) \cdot H(X_j|\overline{R}_i(j)) \tag{14}$$

is the monotone entropy measure of information loss caused by generalizing D into g(D). The monotone entropy metric coincides with the entropy metric when considering generalization by suppression only, while it penalizes generalizations more than the entropy measure does.

Definition of non-uniform entropy metric: Let $D=\{R_1,\ldots,R_n\}$ be a dataset having quasi-identifiers A_j, $1 \leq j \leq r$, and $g(D)=\{R_1,\ldots,R_n\}$ be a generalization of D. Then,

$$\prod_{me}(D, g(D)) = \sum_{i=1}^{n}\sum_{j=1}^{r} -\log Pr(R_i(j)|\overline{R}_i(j)) \tag{15}$$

is the non-uniform entropy metric of the loss of information caused by generalizing D into $g(D)$.

Both the entropy and the monotone entropy metric are uniform for all records. For example, the domain of a quasi-identifier is {1, 2}, while there are 99 % records with the quasi-identifier equals to 1 and the left 1% records are that the value of the quasi-identifier is 2. However, when applying entropy or monotone entropy metric, the information loss of generalizing 1 and 2 to its parent are same. Apparently, it is a bit unreasonable. So, the non-uniform entropy metric appears to tackle with this situation.

4.2 Optimal Anonymity Algorithms

Samarati [31] proposes an algorithm for finding minimal full-domain generalization to implement optimal anonymity using binary search. In [51], it proposes an algorithm for finding minimal generalization with minimal distortion that is called MinGen. MinGen try to examine all of potential full-domain generalization, so that find the optimal generalization according to relevant information metric. However, in the work, it points out that an exhaustive search of all possible generalization is impractical even for the modest sized dataset.

Incognito [34] is an efficient anonymity algorithm by producing minimal full-domain generalization to achieve k-anonymity, which takes advantage of two key features of dynamic programming, namely, bottom-up aggregation along dimensional hierarchies and a priori aggregate computation. Moreover, it uses generalization lattice and three properties to facilitate finding minimal full-domain generalization. There is a brief explanation for figure 4. The leftmost two structures are taxonomy tree structure of *Zipcode* and *Sex* quasi-identifier. And the rightmost structure is two-attribute lattice assembling with the tree structure of *Zipcode* and *Sex*. The related three properties are generalization property, rollup property and subset property which are showed below in detail. Incognito algorithm generates all possible k-anonymous full-domain generalization of original dataset based on subset property. It checks whether single-attribute reaches the requirement of k-anonymity at the start of algorithm execution, then iterates by increasing the size of quasi-identifiers forming larger multi-attribute generalization lattice till reach the scale of given quasi-identifiers.

Generalization Property: Let T be a dataset, and let P and Q be sets of attributes in T such that $D_P <_D D_Q$. If T is k-anonymous with respect to P, then T is also k-anonymous with respect to Q.

Rollup Property: Let T be a dataset, and let P and Q be sets of attributes in T such that $D_P <_D D_Q$. If we have f_1, the frequency set of T with respect to P, then, we can generate the count of f_2, the frequency set of T with respect to Q, by summing the set of counts in f_1 associated by γ with each value set of f_2.

Subset Property: Let T be a dataset, and let Q be a set of attributes in T. If T is k-anonymous with respect to Q, then T is k-anonymous with respect to any set of attributes P such that $P \subseteq Q$.

K-Optimize [40] algorithm is an algorithm of implementing optimal anonymity by pruning unsatisfied released table using a set enumeration tree. The difference between K-Optimize and Incognito is that K-Optimize uses subtree generalization and record suppression, while, Incognito incorporates full-domain generalization into scheme.

4.3 Minimal Anonymity Algorithms

The μ-argus algorithm [52] is the earliest work achieving privacy preserving by generalization and suppression. It can handle a class of disclosure control rules based on

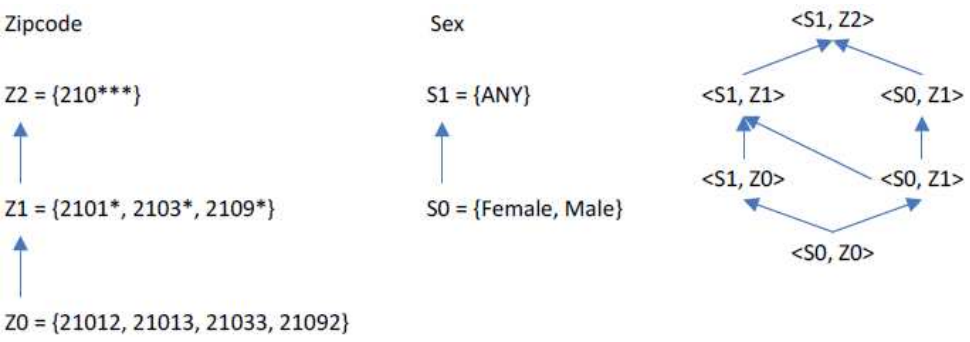

Fig. 4: multi-attribute generalization lattice

checking frequency of certain attribute combinations (quasi-identifiers). It makes decision in favor of bin sizes to generalize and suppress values of the quasi-identifier. The bin size represents the number of records matching the characteristics. User provides an overall bin size and mark sensitive attribute by assigning a value to each attribute which ranges from 0 to 3. Then μ-argus algorithm identifies and handles combination (equivalent class) with small size. The handling covers two disclosure control measures, namely subtree generalization and cell suppression. They are called global recoding and local suppression in the work respectively which are applied simultaneously to dataset. Eventually, it implements *k*-anonymity by computing the frequency of all 3-value combinations of domain values and greedily using anonymity operation. μ-argus algorithm has a serious limit that it is only suitable for combination of 3 attributes. Therefore the anonymous dataset may not meet the requirement of *k*-anonymity when the size of quasi-identifier reaches more than three.

Datafly [53] system proposed by Sweeney in 1998 guarantees anonymity using medical data by automatically generalizing, substituting, inserting and removing specific information without losing substantial information of original dataset. In the work, it summarizes three major difficulties in providing anonymous dataset. The first problem is that anonymity is in the eye of the beholder. The second is that concerns of unique and unusual information appearing within the dataset when producing anonymous dataset. The last one is that how to measure the degree of anonymity in released data. Datafly system is constructed in the process of considering three difficulties above. Datafly is an interactive model that users provide it with an overall anonymity level that determines the minimum bin size allowable for each field. Like bin size in μ-argus algorithm, the minimal bin size reflects the smallest number of records matching the characteristics. As the minimal bin size increases, the more anonymous the dataset is. The Datafly system calculates the size of each equivalent class emerged by quasi-identifiers stored in an array and generalizes those equivalent classes whose size is less than *k* based on a heuristic search metric which select the quasi-identifier with the largest number of distinct values. By the way, Datafly system uses full-domain generalization and record suppression to implements data anonymization.

It is the first time that genetic algorithm is introduced to achieve privacy preserving which is proposed by Iyengar[32]. The core of this work is that how to model the problem of privacy preserving data publishing and makes the model fit for genetic algorithm. With the help of taxonomy tree structure, it encodes each state of generalization as a chromosome. At each time of iterative procedure, new solutions (that is new chromosome) are generated to achieve better result by the means of crossover and mutation. A better solution is based on the information metric. Applying a solution mainly includes two parts, namely conduct generalization operation exactly subtree generalization and carry out suppression as needed. Because execution of genetic algorithm needs setting some parameters, like size of the population, the probability of mutation and the times of iteration, in this work, the size of chromosome is 5000, 0.5 million times of iteration, and probability of mutation was set to 0.002. It takes approximately 18 hours to transform the training set of 30k record. From the result of experiment and the consumed time, we can infer that using genetic algorithm to achieve privacy preserving is time consuming, impractical and inefficient for very large dataset.

Wang et al. [49] propose bottom-up generalization which is a data mining solution to privacy preserving data publishing. The brief description of this algorithm is that it computes the *IP(G)* of each generalization *G*, find the best generalization G_{best} according to forgoing *IP(G)* which is the information metric of this algorithm, and applies the best generalization G_{best} for dataset until the anonymous dataset satisfy the requirement of *k*-anonymity. It is well known that calculating all *IP(G)* for each generalization *G* is impossible. In this situation, it gives the definition of Critical Generalization that is generalization *G* is critical if $A_G(VID) > A(VID)$. *A(VID)* and $A_G(VID)$ are the anonymity before and after applying generalization *G*. So, using compute $A_G(VID)$ for each

critical generalization replaces the operation of computing the *IP(G)* of each generalization *G*. The purpose of modifying algorithm steps is to reduce unnecessary calculation by the means of pruning redundant generalization judged by proved theory. It uses TEA (Taxonomy Encoded Anonymity) index, two observations and one theorem prune unnecessary generalization to improve the efficiency of this algorithm. TEA index for the sequence of quasi-identifiers is a tree of *k* level, each level of TEA index represents current generalization state of a quasi-identifier. Each leaf node stores the size of equivalent class that is determined by the path from this leaf node to the root. With the theorem shows below, it is sufficient to find the critical generalization.

Theorem 1 *G* is critical only if every anonymity *vid* is generalized by some size-*k* segment of *G*, where $k>1$. At most $|VID|$ generalizations satisfy this "only if" condition, where $|VID|$ denotes the number of attributes in *VID* (*VID* is the set of quasi-identifiers).

Top-down specialization [54,55] for privacy preserving can be learned about by comparing with bottom-up generalization. Specialization is a reverse operation of generalization. TDS (Top-Down Specialization) starts at the most general state of dataset according to taxonomy tree structure of quasi-identifiers, and choose best specialization attributes measured by trade-off metric to specialize dataset till further specialization would breach anonymity requirement. Compared with bottom-up generalization, TDS algorithm is more flexible that it can stop at any specialized state of dataset according to users' requirement. Another merit of TDS is that it takes multi-sequence of quasi-identifiers into account. Mondrian multidimensional algorithm proposed by LeFevre [36] is a kind of anonymity algorithm to achieve minimal anonymity by the means of top-down specialization. The difference between TDS scheme and Modrian Multidimensional algorithm can specializate one equivalent class not all equivalent class containing the specific value of selected quasi-identifier. Fung et al. also proposes *k*-anonymity algorithm to precede privacy preserving for cluster analysis in [56].

Using multidimensional suppression for *k*-anonymity is proposed by Slava et al. [57]. The algorithm (KACTUS: *K*-Anonymity Classification Tree Using Suppression) can generate effective anonymous dataset applied for the scenario of classification model. Value suppression is applied only on certain tuples depending on other quasi-identifier values. Because of only using suppression, it is unnecessary to build taxonomy tree structure for each quasi-identifier which is a relaxation compared with priori works equipped with predetermined taxonomy tree for categorical attributes and runtime-generated taxonomy tree structure for numeric attributes. The goal of this work is to generate anonymous dataset to reach the target that the classifier performance trained on anonymous dataset is approximately similar to the performance of classifier trained on the original dataset. KACTUS is composed of two main phases. In the first phase, it generates a classification tree trained on the original dataset. Some diverse top-down classification inducers are prepared for generating classification tree (e.g. C4.5 algorithm). Need to mention, the decision tree is trained on the collection of the given quasi-identifiers. In the second phase, the classification tree generated in the first phases is used by algorithm to emerge the anonymous dataset. The process of data anonymization on classification tree proceeds on bottom-up manner. Every leaf node of classification tree take notes of the number of records has the same equivalent class which is represented by the path from root node to leaf node. If all of leaf nodes in classification tree meet the requirement of *k*-anonymity, there is no need to suppress any quasi-identifiers existed in the classification tree currently. While, if leaves do not meet the requirement of *k*-anonymity, it can obtain a new leaf which may comply with *k*-anonymity by adequately pruning them.

The anonymity operations of generalization and suppression are still widely used in latest research. Generalization is employed in [58] for improve DNALA. An appropriately robust privacy model is proposed for data anonymity which is called β-likeness in [59], one of the anonymization schemes is based on generalization. In [60], it focuses on the generalization hierarchy of numeric attributes, extends previous method and validates their proposal with the help of existed information metric. Information-based algorithm is proposed by Li et al. [61] for classification utility using generalization and suppression. Sun et al. [62] improves *p*-Sensitive *k*-anonymity model and proposes two private requirements, namely p^+-sensitive *k*-anonymity and (p, α)-sensitive *k*-anonymity properties. It generates anonymous dataset with a top-down specialization manner by specializing values step by step. Wong et al. find that anonymization error can be reduced by employing a non-homogeneous generalization[63]. Generalization hierarchy is incorporated to implement user-oriented anonymization for public data in [64]. Adeel et al.[65] also use generalization operation coping with the problem of sequential release under arbitrary update. Mahesh et al.[66] propose a new method to preserve individuals sensitive data from record and attribute linkage attacks by setting range values and record elimination.

5 CONCLUSION

Information sharing is becoming indispensable part of individuals and organizations, privacy preserving data publishing comes to receive increasing attentions from all over the world, which is regarded as an essential guarantee for information sharing. To put it simply, the role of privacy preserving data publishing is to transform the original dataset from one state to the other state so as to avoid privacy disclosure and withstand diverse attacks.

In this paper, first, we discuss privacy model of PPDP in details, mainly introduce privacy preserving model for record linkage and anonymity operations. Information metric with different purposes are collected which is an important part of anonymity algorithms. Subsequently, more emphasis is put on the anonymization algorithms with specific anonymity operations, exactly, generalization and suppression operation. This paper may be used for researcher to scratch the profile of anonymity algorithms for PPDP by the means of generalization and suppression. Our further research shows below.

A) Hybrid *k*-anonymity algorithm. Algorithm implementation of *k*-anonymity is simple and can adapt to different scenario. So, it will be an effective scheme to mix *k*-anonymity with other anonymity techniques.

B) Background knowledge attack simulation to make information safe. It is difficult to accurately simulate the background knowledge of attackers. While different background knowledge will cause privacy breach in varying degree. It will be a part of research to find out the way of simulating background knowledge of attackers, so that provide all-round protection for privacy.

C) Information metric. It has given an overall summary of information metrics in this paper. We can see that different metric fits for different scenario of PPDP. Studying new information metric or improving existed metric will be a part of further research.

D) Multi sensitive attributes anonymity constraint. Existing study focuses on anonymization of a single sensitive attribute, which cannot simply shift to solve the multi sensitive attribute problem. Therefore, we need to study effective anonymity algorithms with multidimensional constraint. In addition, the difficulties of implementing personalize anonymity efficiently and choosing quasi-identifiers exactly are all worthy of further thought and study.

ACKNOWLEDGEMENT

This work was supported in part by China Postdoctoral Science Foundation (No. 2012M511303), National Science Foundation of China (No. 61173143), and Special Public Sector Research Program of China (No. GYHY201206030) and was also supported by PAPD. The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] M. S. Wolf, C. L. Bennett, Local perspective of the impact of the HIPAA privacy rule on research, Cancer-Philadelphia Then Hoboken, **106**, 474-479 (2006).

[2] Johannes Gehrke, Models and methods for privacy-preserving data publishing and analysis, In Proceedings of the 22nd International Conference on Data Engineering (ICDE), **105**, (2006).

[3] B. Fung, K. Wang, R. Chen, P. Yu, Privacy-preserving data publishing: A survey of recent developments, ACM Computing Surveys, **42**, 1-53 (2010).

[4] L. Sweeney, K-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, **10**, 557-570 (2002).

[5] Dewri R., k-anonymization in the presence of publisher preferences, Knowledge and Data Engineering, IEEE Transactions, **23**, 1678-1690 (2011).

[6] Nergiz M. E., Multirelational k-anonymity, Knowledge and Data Engineering, IEEE Transactions, **21**, 1104-1117 (2009).

[7] Jiuyong Li, Wong, R. C. W. Wai-chee Fu, A., Jian Pei, Transaction anonymization by local recoding in data with attribute hierarchical taxonomies, Knowledge and Data Engineering, IEEE Transactions, **20**, 1181-1194 (2008).

[8] Tamir Tassa, Arnon Mazza, k-Concealment: An Alternative Model of k-Type Anonymity, Transactions on Data Privacy, 189-222 (2013).

[9] Dalenius T., Towards a methodology for statistical disclosure control, Statistik Tidskrift, **15**, 429-444 (1977).

[10] Ahmed Abdalaal, Mehmet Ercan Nergiz, Yucel Saygin, Privacy-preserving publishing of opinion polls, Computers & Security, 143-154 (2013).

[11] Machanavajjhala A., Gethrke J., Kifer D., Venkitasubramaniam M., l-diversity: Privacy beyond k-anonymity, In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), **24**, (2006).

[12] Ke Wang, Benjamin C. M. Fung, Anonymizing sequential releases, In Proceedings of the 12th ACM SIGKDD Conference, 414-423 (2006).

[13] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, Ke Wang, (a, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing, In Proceedings of the 12th ACM SIGKDD, 754-759 (2006).

[14] Qing Zhang, Koudas N., Srivastava D., Ting Yu, Aggregate query answering on anonymized tables, In Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE), 116-125 (2007).

[15] Ninghui Li, Tiancheng Li, Venkatasubramanian S., t-closeness: privacy beyond k-anonymity and l-diversity, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 106-115 (2007).

[16] Xiaokui Xiao, Yufei Tao, Personalized privacy preservation, In Proceedings of the ACM SIGMOD Conference, 229-240 (2006).

[17] Mehmet Ercan Nergiz, Maurizio Atzori, Chris Clifton,Hiding the presence of individuals from shared databases, In Proceedings of ACM SIGMOD Conference, 665-676 (2007).

[18] Chawla S., Dwork C., Mcshrry F., Smith A., Wee H., Toward privacy in public databases, In Proceedings of the Theory of Cryptography Conference (TCC), 363-385 (2005).

[19] Dwork C., Differential privacy, In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), 1-12 (2006).

[20] Rastogi V., Suciu D., Hong S., The boundary between privacy and utility in data publishing, In Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB), 531-542 (2007).

[21] Blum A., Ligett K., Roth A., A learning theory approach to non-interactive database privacy, In Proceedings of the 40th

Annual ACM Symposium on Theory of Computing (STOC), 609-618 (2008).
[22] Nergiz M. E., Clifton C., Nergiz A. E., Multirelational k-Anonymity, Knowledge and Data Engineering, IEEE Transactions on, **21**, 1104-1117 (2009).
[23] Xiangmin Ren, Research on privacy protection based on k-anonymity, Biomedical Engineering and Computer Science (ICBECS), 1-5 (2010).
[24] Vijayarani S., Tamilarasi A. Sampoorna M., Analysis of privacy preserving k-anonymity methods and techniques, Communication and Computational Intelligence (INCOCCI), 540-545 (2010).
[25] Wenbing Yu, Multi-attribute generalization method in privacy preserving data publishing, eBusiness and Information System Security (EBISS), 1-4 (2010).
[26] J. Domingo Ferrer, A survey of inference control methods for privacy preserving data mining, Advance in Database Systems, **34**, 53-80 (2008).
[27] N. Shlomo, T. De Waal, Protection of micro-data subject to edit constraints against statistical disclosure, Journal of Official Statistics, **24**, 229-253 (2008).
[28] I. Cano, V. Torra, Edit constraints on microaggregation and additive noise, In Proceedings of the International ECML/PKDD Conference on Privacy and Security Issues in Data Mining and Machine Learning (PSDML10), Springer, 1-14 (2010).
[29] Jing Zhang, Xiujun Gong, Zhipeng Han, Siling Feng, An improved algorithm for k-anonymity, Contemporary Research on E-business Technology and Strategy Communications in Computer and Information Science, 352-360 (2012).
[30] Xiaoling Zhu, Tinggui Chen, Research on privacy preserving based on k-anonymity, Computer, Informatics, Cybernetics and Applications Lecture Notes in Electrical Engineering, **107**, 915-923 2012.
[31] P. Samarati, Protecting respondents identities in microdata release, IEEE Trans. On Knowledge and Data Engineering, **13**, (2001).
[32] V. Iyengar, Transforming data to satisfy privacy constraints, In ACM SIGKDD, (2002).
[33] Xuyun Zhang, Chang Liu, Surya Nepal, Jinjun Chen, An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud, Journal of Computer and System Sciences, 542-555 (2013).
[34] Lefevre K., Dewitt D. J., Ramakrishnan R., Incognito: efficient full-domain k-annymity, In Proceedings of ACM SIGMOD, 49-60 (2005).
[35] Xu J., Wang W., Pei J., Wang X., Shi B., Fu A. W. C., Utility-based anonymization using local recoding, In Proceedings of the 12th ACM SIGKDD Conference, (2006).
[36] Lefevre K., Dewitt D. J., Ramakrishnan R, Mondrian multidimensional k-anonymity, In Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE), (2006).
[37] B.C.M. Fung, K. Wang, P.S. Yu, Anonymizing classification data for privacy preservation, IEEE Transaction Knowledge and Data Engneering, **19**, 711-725 (2007).
[38] Rui Chen, Benjamin C.M. Fung, Noman Mohammed, Bipin C. Desai , Ke Wang, Privacy-preserving trajectory data publishing by local suppression, Information Sciences, 83-97 (2013).
[39] Martin Serpell, Jim Smith, Alistair Clark, Andrea Staggemeier, A preprocessing optimization applied to the cell suppression problem in statistical disclosure control, Information Sciences, 22-32 (2013).
[40] Bayardo R. J., Agrawal R., Data privacy through optimal k-anonymization, In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE), 217-228 (2005).
[41] Wang K., Fung B. C. M., Yu P. S., Template-based privacy preservation in classification problems, In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM), 466-473 (2005).
[42] Meyerson A., Williams R., On the complexity of optimal k-anonymity, In Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART PODS, 223-228 (2004).
[43] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, Efficient multidimensional suppression for k-anonymity, IEEE TKDE, **22**, 334-347 (2010).
[44] Tripathy B. K., Kumaran K., Panda G.K., An improved l-diversity anonymization algorithm, Computer Networks and Intelligent Computing Communications in Computer and Information Science, **157**, 81-86 (2011).
[45] Asaf Shabtai, Yuval Elovici, Lior Rokach, Privacy, data anonymization, and secure data publishing, A Survey of Data Leakage Detection and Prevention Solutions SpringerBriefs in Computer Science, 47-68 (2012).
[46] Batya Kening, Tamir Tassa, A pratical approximation algorithm for optimal k-anonymity, Data Mining and Knowledge Discovery, **25**, 134-168 (2012).
[47] Florian Kohlmayer, Febian Prasser, Claudia Eckert, Alfons Kemper, Klaus A. Kuhn, Flash: efficent, stable and optimal K-anonymity, In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, 708-717 (2012).
[48] Y. He, J. F. Naughton, Anonymization of set-valued data via top-down, local generalization, In Proceedings of the Very Large DataBase (VLDB) Endowment, **2**, 934-945 (2009).
[49] Wang K., Yu P. S., Chakraborty S., Bottom-up generalization: A data mining solution to privacy protection, the fourth IEEE International Conference on Data Mining (ICDM2004), 249-256 (2004).
[50] Gionis A., Tassa T., k-Anonymization with Minimal Loss of Information, IEEE Transactions on Knowledge and Data Engineering, **21**, 206-219 (2009).
[51] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, **10**, 571-588 (2002).
[52] Hundepool A., Willenborg L., -and r-argus: Software for statistical disclosure control, In Proceeding of the 3rd Int. Seminar on Statistical Confidentiality, 208-217 (1996).
[53] Sweeney, L., Datafly: A system for providing anonymity in medical data, In Proceedings of the IFIP TC11 WG11.3 11th International Conference on Database Security XI: Status and Prospects, **113**, 356-381 (1998).
[54] B.C.M. Fung, K. Wang, P.S. Yu, Top-down specialization for information and privacy preservation, In Proceedings 21st IEEE International Conference Data Engneering (ICDE 05), 205-216 (2005).
[55] Mingquan Ye, Xindong Wua, Xuegang Hua, Donghui Hua, Anonymizing classification data using rough set theory, Knowledge-Based Systems, 82-94 (2013).

[56] Fung B. C. M., Wang K., Wang L., Hung P. C. K., Privacy-preserving data publishing for cluster analysis, Data Knowledge Engineering, **68**, 552-575 (2009).
[57] Slava Kisilevich, Lior Rokach, Yuval Elovici, Bracha Shapira, Efficient multidimensional suppression for k-anonymity, IEEE TKDE, **22**, 334-347 (2010).
[58] Guang Li, Yadong Wang, Xiaohong Su, Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices, Computer Method and Programs in Biomedicine, **108**, 1-9 (2012).
[59] Jianneng Cao, Panagiotis Karras, Publishing microdata with a robust privacy guarantee, In Proceeding of the VLDB Endowment, **5**, 1388-1399 (2012).
[60] Alina Campan, Nicholas Cooper, Traian Marius Truta, On-the-fly generalization hierarchies for numerical attributes revisited, Secure Data Management Lecture Notes in Computer Science, **6933**, 18-32 (2011).
[61] JiuYong Li, Jixue Liu, Muzammi Baig, Raymod Chi-Wing Wong, Information based data anonymization for classification utility, Data & Knowledge Engineering, **70**, 1030-1045 (2011).
[62] Xiaoxun Sun, Lili Sun, Hua Wang,Extended k-anonymity models against sensitive attribute disclosure, Computer Communications, **34**, 526-535 (2011).
[63] Wai Kit Wong, Nikos Mamoulis, David Wai Lok Cheung, Non-homogeneous generalization in privacy preserving data publishing, In Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, 747-758 (2010).
[64] Shinsaku Kiyomoto, Toshiaki Tanaka, A user-oriented anonymization mechanism for public data, In Proceedings of the 5th international Workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security, 22-35 (2010).
[65] Adeel Anjum, Guillaume Raschia, Anonymizing Sequential Releases under Arbitrary Updates, Proceedings of the Joint EDBT/ICDT 2013 Workshops, 145-154 (2013).
[66] R. Mahesh, T. Meyyappan, Anonymization Technique through Record Elimination to Preserve Privacy of Published Data, Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, 328-332 (2013).

Yang Xu received his Bachelor degree in Computer Science and Engineering from Nanjing University of Information Science & Technology, China in 2011. Currently, he is a candidate for the degree of Master of Computer Science and Engineering in Nanjing University of Information Science & Technology. His research interests include privacy preserving, data publishing etc.

Tinghuai Ma is a professor in Computer Sciences at Nanjing University of Information Science & Technology, China. He received his Bachelor (HUST, China, 1997), Master (HUST, China, 2000), PhD (Chinese Academy of Science, 2003) and was Post-doctoral associate (AJOU University, 2004). From Nov.2007 to Jul. 2008, he visited Chinese Meteorology Administration. From Feb.2009 to Aug. 2009, he was a visiting professor in Ubiquitous computing Lab, Kyung Hee University. His research interests are in the areas of Data Mining and Privacy Protected in Ubiquitous System, Grid Computing. He has published more than 80 journal/conference papers. he is principle investigator of several NSF projects. He is a member of IEEE.

Meili Tang is an associate professor in School of Public Administration at Nanjing University of Information Science & Technology, China. She received her master degree from Huazhong University of Science & Technology, China, 2000. Her main research interests are in the areas of e-government and data publishing.

Wei Tian is an assistant professor of Computer Sciences at Nanjing University of Information Science & Technology, China. He received his master degree from Nanjing University of Information Science & Technology, China, 2006. Now, he is a doctoral candidate of Applied Meteorology, Nanjing University of Information Science & Technology. His main research interests are in the areas of Cloud Copmputing and Meteorological Data Processing.