A General Framework for Privacy Preserving Data Publishing

A.H.M. Sarowar Sattar[a,*], Jiuyong Li[a,*], Xiaofeng Ding[a], Jixue Liu[a], Millist Vincent[a]

[a] *School of Information Technology and Mathematical Science, University of South Australia, Mawson Lakes, SA-5095, Australia*

Abstract

Data publishing is an easy and economic means for data sharing, but the privacy risk is a major concern in data publishing. Privacy preservation is a major task in data sharing for organizations like bureau of statistics, hospitals, etc. While a large number of data publishing models and methods have been proposed, their utility is of concern when a high privacy requirement is imposed. In this paper, we propose a new framework for privacy preserving data publishing. We cap the belief of an adversary inferring a sensitive value in a published data set to as high as that of an inference based on public knowledge. The semantic meaning is that when an adversary sees a record in a published data set, s/he will have a lower confidence that the record belongs to a victim than not. We design a method integrating sampling and generalization to implement the model. We compare the method with some state-of-the-art methods on privacy-preserving data publish-

[*] Corresponding author

Email addresses: `sarowar@gmail.com` (A.H.M. Sarowar Sattar), `Jiuyong.Li@unisa.edu.au` (Jiuyong Li)

[1] School of Information Technology and Mathematical Science, Mawson Lakes, SA-5095. Mob. +61420863356

[2] D3-07, School of Information Technology and Mathematical Science, Mawson Lakes, SA-5095.

ing experimentally, our proposed method provides sound semantic protection of individuals in data and, provides higher data utility.

Keywords: Privacy protection, Data publishing, Anonymization

1. Introduction

Recently, the privacy preservation in data publishing has received considerable attention from researchers. Compared with data publishing through the format of aggregated results or statistical ones, the release of microdata offers an advantage in terms of information availability, which makes it particularly suitable for scientific analysis in a variety of domains such as public health, demographic studies, etc. However, the release of microdata causes privacy concerns of disclosing sensitive information of individuals. Simply removing explicit identifiers like names or IDs has been shown to be vulnerable to privacy breach, since other personal identifying attributes, such as age, gender and zip code, called quasi-identifier (QID) which usually remain in the published data for data analysis, allow individuals' sensitive information to be revealed when they are linked with publicly available information. For example, by combining a public voter registration list with a released information of health insurance, Sweeney was able to identify the medical record about a former governor of Massachusetts [1]. Many techniques have been proposed to address the problem.

There are generally two types of definitions for privacy.

One type of definitions is microdata based. k-anonymity [1] and l-diversity [2] are two typical examples. k-anonymity requires that a published data set should have at least k rows (called a group) sharing the same QID value. So the probability for identifying an individual in a published data set is $1/k$. The k-anonymity model protects an individual from being identified in a data set with a high confidence. The l-diversity model requires that the number of the sensitive values in a QID group is at least l. So an adversary could not tell which sensitive value belongs to an individual in a group. There are many improved models (definitions) along this line [3, 4, 5, 6]. They all associate with one or a few user specified thresholds, like k and l in the above works, and it is difficult for users to set the right thresholds.

Another type of definitions is probabilistic. Differential privacy [7] is a typical example. It assumes that even if an adversary knows all other sensitive values but the victim's, the adversary could not infer victim's sensitive value when knowing the randomized aggregated result with a certain confidence. This requirement is strong and causes a big utility loss.

Most privacy protection principles are to bind the leakage of sensitive information. In general form, the leakage is the difference between the posterior probability and the prior probability. The posterior possibility is easy to be quantified. However, the prior probability is difficult to estimate, and different estimations lead to different privacy protection models. For example, the l-diversity model assumes the uniform distribution of sensitive values. ϵ-differential privacy does not distinguish between sensitive and non-sensitive attributes. One major disadvantage of such models is that the requirement of a small leakage will cause published data set to have little utility due to, for example, too much generalization or too much noise. We will need to search for an alternative approach for sound privacy protection and better data utility.

In this paper, we explore an alternative model that is semantically sound and gives a published data set more utility. When an adversary accesses a published data set, s/he may infer that a record belongs to a victim (adversary knows that the victim's record must be in the published data set). However, if this record is what everyone expects to see in a data set, for example, a 40-50 old male with flu in a medical data set, does this breach the privacy of the victim? We say no, even if the adversary gets the sensitive value of the victim right (note that we do not mean that flu is not sensitive, and we will elaborate this example more later.).

We argue that the damage of a privacy breach is not directly associated with whether the adversary obtains the sensitive value right, but is associated with the confidence level of the inference. For example, if an adversary claims that a victim suffers from prostate cancer with a convincing inference in a published data set, but the claim is wrong since the victim actually suffers from bowel cancer. Even though the inference is wrong, the damage has been made to the victim by the claim. Since the claim is convincing, most people believe in it, and this brings damage to the victim. If an adversary alleges a victim suffering from HIV with a weak inference (as strong as a random guess), the victim will not have to do any defence regardless if the allegation is true or not. Few people will believe in the allegation.

Consequently the importance of privacy protection is not to give an adversary strong belief to build an allegation. If the belief of an allegation in a published data set is the same as the confidence of a random guess, this will be a sufficient protection for the privacy of an individual in data since the believability of an allegation is low. The question is how to model a random guess in a published data set. In this paper, we will discuss a model towards such a protection.

Our idea is that the belief of an adversary obtained from a published data set should be at most the same as the belief obtained from the public knowledge. In other words, when an adversary sees a record in a published data set, the adversary should expect to see the same record in a randomly generated data set following the public knowledge. The occurrence of a record in a published data set does not relate to whether the victim's record is in the published data set or not. In the previous example, the 45 year old male patient does not care the claim that he suffers from flu because the adversary sees a record "40-50, male, flu" in the published data set of a hospital where the patient visited because the adversary is expected to see the same record even if the 45 year old male patient's record is not in the published data set (note that in our model, only a sample of records are published). Therefore, the privacy of the patient is protected.

In this paper, we propose a new framework for privacy preserving data publishing based on the above motivations, and propose an effective hybrid method of sampling and generalization for privacy preserving data publishing. Contributions of the work are listed as the following.

- This new model is semantically sound and offers good data utility. Semantically, it provides a strong protection for the privacy of individuals since it does not give an adversary a stronger belief from an inference in a published data set than the belief from an inference on public knowledge. Practically, it allows many records to be published with a light generalization and a large sample rate. The method integrates generalization with sampling. Sampling is essential in our method. We note that good sampling does not reduce the quality of data. The sampling techniques have been used for many rigorous studies for a long time. Furthermore, a major goal for data publishing is to support the shared data analysis in a large community. In data analysis the aggregated results are often derived. When data sets are randomly sampled, the bias in the aggregated results will be low.

- This model controls privacy risk of individuals at the record level. This supports local generalization of each record irrespective with other records. This provides an easy and effective criterion to judge whether a record is publishable. The method only restricts a few records with values of very low frequencies, such as 95 year old male and Huntington's disease, from being published. It provides good data utility for those publishable records. We note that data publishing is not a right means for data sharing with rare values (for example, some rare diseases). If we try to accommodate those rare cases, the overall quality of published data will suffer badly.

- This model links privacy risk to data set size, which is crucial in privacy risk analysis. The data size has not been utilized in previous data publishing models. For example, consider data sets with 100 records and 100,000 records respectively. Intuitively, an individual in the data set of 100 records has higher privacy risk than an individual in the data set of 100,000 records.

The rest of this paper is organized as follows. Section 2 introduces preliminaries and principle of the new privacy framework. Section 3 and Section 4 formally define the way of estimating the adversary's expected confidence and observed confidence respectively, followed by a hybrid method to published data sets after satisfying the new privacy criterion in Section 5. Section 6 shows the experimental results, followed by some related works in Section 7. Finally, Section 8 concludes this paper with future direction.

2. Preliminaries and the principle

A data owner has a data set D_1, where each record t contains information about an individual, like 'id', 'age', 'sex', 'zip code', along with the sensitive information, such as a disease or the salary, of that individual. For simplicity, we consider that there is only one sensitive value in each row (multiple sensitive values can be considered as a set of sensitive values.). The attributes that uniquely identify an individual are called unique identifiers (IDs), such as social security number and name. The attributes that potentially conjunctively identify an individual are called quasi-identifiers (QID), such as 'age', 'sex' and 'zip code'. Consider that D_1^* is a published data set of D_1, where the attribute ID has been removed, QID and sensitive attributes are kept in D_1^*. Some of the QID attribute's value may be generalized [3] due to legislation [8].

Now we consider an adversary whose goal is to infer whether a victim individual v has a sensitive value s. We assume that an adversary has the following background knowledge.

Definition 1 (The background knowledge of an adversary). *We assume that a victim is an individual v in D_1. The adversary knows*

1. D_1^**, the published version of* D_1.
2. *the QID values of* v.
3. *global statistics of the population from which* D_1 *has been generated.*

[3] Generalization of an attribute means its current value is replaced by the value of higher level node from its taxonomy. For example, in Figure 1(a), if the attribute is 'age' and its value is 20, the generalized value can be 13-25.

4. *v is in D_1 and v is in D_1^* with a probability because of sampling used in generating D_1^*.*

We note that the adversary uses QID values of v to identify a group in D_1^* containing v to narrow down the possible sensitive values of the victim.

Let us assume that the victim v's record is not in the published data set D_1^*. An adversary is still expected to see a record with the same generalized QID values as v's in D_1^* just purely by chance. We call the confidence of seeing such a record expected confidence.

Definition 2 (Expected confidence). *Let q_i be the quasi identifier value of a victim v, q_i' be the generalized value of q_i and s be a sensitive value that may or may not be v's. The expected confidence of record (q_i', s) is the probability of (q_i', s) to be included in D_1^* regardless whether v is in D_1 or not.*

We note that the expected confidence relates to the QID values of v, but not whether v is in D_1 or not.

After observing the published data set D_1^*, the adversary may have a different confidence about the victim's record. This confidence is called the observed confidence of an adversary and is formally defined as follows.

Definition 3 (Observed confidence). *Let q_i' be the generalized value of q_i of a victim v. The observed confidence of an adversary about v is the probability of (q_i', s) belonging to v.*

Thus, a robust privacy preserving criterion should place an upper bound on the observed confidence of an adversary. Instead of setting an arbitrary user threshold to cap the observed confidence we will argue to use the expected confidence as the as the upper bound on the observed confidence. Therefore, we have the following principle.

Definition 4 (Privacy principle). *The observed confidence of an adversary inferring that a record belongs to a victim should be no more than the expected confidence of the adversary seeing the same record in the data set without assuming that the victim's record is in.*

To achieve this, we employ the following process. Firstly, the data set D_1 is sampled. Secondly, the sampled data set is generalized. The data owner publishes the sampled and generalized data set D_1^*. We will discuss how to estimate the expected confidence and observed confidence in the following sections.

3. Estimating expected confidence

In this section, we show a method for estimating expected confidence of an adversary from the public knowledge.

We firstly explain the public knowledge that an adversary has. Consider a population of individuals, which is described by a set of attribute values. The distributions of attribute values of this population are known to public. For example, in the population of patients in a country, age, gender, and disease distributions are known to public. Each published data set contains a sample of individuals of the population.

We now model the expected confidence. Let τ be the record space and t be a record with attribute values (QID, s). t is associated with a probability $Pr(t)$ (Definition 5). Each data set is a sample of τ with the probability $Pr(t)$ with replacement. Duplicated records are allowed in a data set. We call these randomly

Table 1: Notations

Notation	**Description**
τ	the record space
D_0	the hypothesized data set, $\lvert D_0\rvert = \lvert D_1^*\rvert$
D_1	the original data set
D_1^*	the publishing data set of D_1
ID	the identifier attribute
QID	the quasi-identifier attribute
q_i	i'th QID attribute
s	the sensitive attribute
$Pr(E)$	the probability of event E happens
$S(t)$	the sensitive value of record t
$ID(t)$	the identifier attribute of record t

sampled data sets as hypothesized data sets, which are available to an adversary to estimate the expected confidence.

For a published data set D_1^*, a sampled and generalized version of D_1, D_0 of the size of D_1^* is a hypothesized data set of D_1^*. D_0 and D_1^* have the same attribute domains. D_0 is a random sample of τ with probability $Pr(t)$. Some common notations used in our paper are shown in Table 1. We assign the probability to each t as follows.

Definition 5 (Record probability). *We assume that attribute values and the sensitive value in a record are independent. $Pr(q_i)$ and $Pr(s)$ are the probabilities of selecting value q_i and sensitive value s respectively in the population*

Table 2: Data before anonymization

Name	**Age**	**Sex**	**Zip Code**	**Diseases**
Alice	22	F	5095	Cancer
Bob	24	M	5085	HIV
Paul	20	F	5001	Anemia
Clark	23	M	5005	Flu

by a random draw, which are known to the public. The probability of a record $t = \{q_1, \ldots, q_d, s\}$, denoted as $Pr(t)$, is assigned as the following.

$$\begin{aligned} Pr(t) &= Pr(q_1) \times \cdots \times Pr(q_d) \times Pr(s) \\ &= (\prod_{i=1}^{d} Pr(q_i)) \times Pr(s) \end{aligned}$$

For example, let us assume that $Pr(age = [20{-}30]) = 0.15$, $Pr($gender = [female]$) = 0.5$, and $Pr($disease = [diabetes]$) = 0.05$ are obtained from the patient population. Let $t = \{20 - 30, \text{female}, \text{diabetes}\}$. $Pr(t) = 0.00375$. Note that the public knowledge may include that a 40-60 female has higher probability of diabetes, say 0.02. Such public knowledge can be modeled also. The independency assumption is used when we do not have other public knowledge.

In the above estimation, the independency between QID and sensitive value are assumed. When they are not, their relationship can be modeled by other data mining models. For example, the confidence of an association rule $(40 - 60, M)$ $\rightarrow Prostate\ Cancer$, can be used to model the probability of a group of people to a disease.

An adversary wants to determine whether a victim v is in the published data set. Note that, t is a record that could be anyone including that of the victim v.

Since the victim v is an individual, the inclusion or exclusion of v in a data set does not affect the overall distributions of a data set in general. We consider that the probability of t occurring in D_0 as the expected confidence of an adversary.

Definition 6 (Expected confidence 2). *The expected confidence of an adversary on a record t with the sensitive value s_i in a published data set D_1^* is the probability of t occurring in D_0. It is represented by $Pr(S(t) = s_i|D_0)$. This probability is formulated by the binomial distribution with a success probability (sampled) of $Pr(t)$ and $n = |D_0|$ trials.*

$$\begin{aligned} Pr(S(t) = s_i|D_0) &= 1 - f(0; n, Pr(t)) \\ &= 1 - (1 - Pr(t))^n \end{aligned} \tag{1}$$

$Pr(t)$ is the probability that record a t is picked in a random draw from τ to D_0, and $Pr(S(t) = s_i|D_0)$ is the probability of t with sensitive value s_i occurring in D_0 (could be more than once). For example, let $t = \{20 - 30, \text{female}, \text{diabetes}\}$, $Pr(t) = 0.00375$, and $n = 100$. Based on this knowledge, the probability of t in D_0 will be 0.313.

Expected confidence increases when the number of draws increases. When $n = 1000$, the probability will be 0.977. This captures the intuition that a record in a large data set has low privacy risk.

4. Estimating observed confidence

In this section, we estimate adversary's observed confidence. Let us assume that the data publishers publish a sample of the of original data set. If an adversary looks for a victim's record in the published data set, the sampling rate quantifies the confidence of the adversary regarding the presence (due to sampling victim's record may not be chosen) of the victim in that published data set. This confidence is called probability of publishing a victim's record. We formally define the confidence as follows,

Definition 7 (Probability of publishing). *Let t be a record of a victim v in data set D_1. D_1^* is the published data set which is sampled from D_1 with sampling rate β. Therefore, the probability of publishing*

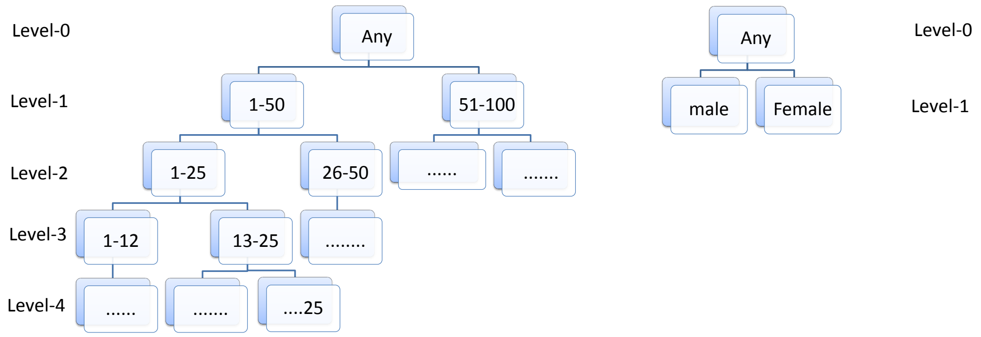
$$Pr(t \in D_1^* | ID(t) = v) \quad = \quad \beta$$

We note that the probability of publishing is a public knowledge.

Data set will not be published with all its original attributes' values due to legislation [8]. For example, consider that according to the legislation, the 'age' attribute of a record should not have specific value in a published data set. Intuitively, in this paper, we use a rule that the value of 'age' attribute must be published no less than 5 years interval. This will be considered as a minimum requirement for publishing 'age' attribute. For example, to satisfy the minimum requirement, the generalized value of age 22 should be 20-25 in the published data set. Generally speaking, if we follow such minimum requirement for data publishing, we have different equivalence groups (Definition 8) in a published data set.

Definition 8 (Equivalence group). *An equivalence group of a data set with respect to an QID attribute set is the set of all records in the data set containing identical values for the QID attribute set.*

For example, records 1 and 2, and records 3 and 4 in Table 3, a published data set of the original data set in Table 2, form two equivalence groups with respect to attribute set $\{$age, sex, postcode$\}$. Their corresponding values are identical.



(a) Age

(b) Gender

Figure 1: Taxonomy of age and gender attributes

In a published data set, records in each equivalence group can have same or different sensitive values. The diversity of sensitive values in an equivalence group reduces the confidence of an adversary inferring the victim's sensitive value. For example, an adversary knows with 100% confidence (when $\beta = 1$) that the victim's record in an equivalence group which has sensitive values $\{A, A, A, B, B, C\}$. The adversary has 50% , 33.33% and 16.67% confidence to infer the sensitive values of the victim as A, B, and C respectively. We call this confidence as group confidence. Formal definition of group confidence is given by Definition 9.

Definition 9 (Group confidence). *Let t be the record of a victim v in an equivalence group E of a published data set D_1^*. The group confidence of the sensitive value s_i belonging to victim v is measured as follows.*

$$Pr(S(t) = s_i|E) = \frac{\#s_i}{|E|}$$

where $\#s_i$ represents the number of a specific sensitive value s_i and $|E|$ is the size of the equivalence group E.

Table 3: Anonymous data

Age	**Sex**	**Zip Code**	**Diseases**
20-25	M	5005	Flu
20-25	M	5085	HIV
20-25	F	5001	Anemia
20-25	F	5095	Cancer

After observing the published data set, an adversary's observed confidence is calculated as follows.

$$\begin{aligned} Pr(S(t) = s_i | D_1^* \wedge ID(t) = v) &= Pr(t \in D_1^* | ID(t) = v) \times Pr(S(t) = s_i | E) \\ &= \beta \times \frac{\# s_i}{|E|} \end{aligned} \qquad (2)$$

Adversary's observed confidence is calculated based on the sampling rate and the number of sensitive values in each equivalence group. We control this confidence by adjusting the sampling rate. However, for better data utility we always try to use a higher sampling rate like 90%.

Based on the estimations of the expected and the observed confidences, the privacy preserving criterion in Definition 4 is given by Equation 3.

$$Pr(S(t) = s_i | D_1^* \wedge ID(t) = v) \leq Pr(S(t) = s_i | D_0) \qquad (3)$$

In the next Section, we study how to achieve our desired goal and an algorithm for the implementation.

5. An algorithm

Data publishers can control both expected and observed confidences of an adversary. The following two sections explain how the data publishers control the confidences. In order to preserve individual privacy, the data publishers will publish only those records that satisfy the privacy principle in Equation 3.

5.1. Manipulating adversary's confidences

In this section, we elaborate a method of controlling an adversary's confidences. By controlling these confidences we can protect the sensitive information of a victim. The expected confidence of an adversary depends on the record probability (Definition 5), and the record probability depends on the generalization level of QID values. For example, if age 25 is generalized to 20-40, then more records would be in this group and the expected confidence increases.

Adversary's observed confidence depends on both the sampling rate and the sizes of equivalence groups in the published data set. There would be different equivalence groups for different generalized values of the QID of a record. When the sampling rate is fixed, record generalization can change the observed confidence of an adversary.

Both expected confidence and observed confidence relate to the levels of generalization. We can control both confidences by applying local recoding method (a form of generalization) [9], where the generalization is applied independently without considering other records. The local recoding will be applied to each record until the record satisfies Equation 3.

Moreover, when the process (local recording) generalizes all QID attributes of a record to its top generalized value, say the record 1 in Table 2 is generalized to {age:any, sex:any, zip code:any} (consider 'any' is the root value of taxonomies of all QID attributes), the record can give us a desired confidences but at the same time this makes the record useless in its published form. Therefore, it is required to control levels of the local recoding. We do it by using a distortion metric (Definition 10). The generalization will continue until the distortion of a record is acceptable. Note that, the acceptable value of distortion is usually given by a data publisher.

Definition 10 (Distortion metric). *Let t be a record in the data set D_1 and t' represents its generalization in D_1^*. The distortion of t is represented by δ and is measured by,*

$$\delta = \| t - t' \| = \frac{1}{d} \sum_{i=1}^{d} (1 - \frac{\textit{Current level of } (q_i')}{\textit{Maximal level in } (q_i)}) \tag{4}$$

where, q_i represents the i'th QID attribute of t and q_i' represents its generalized version.

For example, consider a record t = {age:25, sex:male, disease:cancer} and its generalization t' = {age:20-40, sex: any , disease: cancer }. In the taxonomies 'age' and 'sex' in Figure 1, the age of '25' is at Level 4 and 20-40 is at Level 3. In the case of 'sex' attribute the original value and the generalized values are at Level1 and Level 0 respectively. So, based on Definition 10, the distortion of this generalization is 0.625. Generally speaking, distortion '0' represents the original form of the record, whereas distortion '1' means all attributes take their top generalized value.

By using this distortion metric, the local recoding method helps us to find out the generalized record that is publishable with acceptable distortion. Algorithm 1 shows the procedure.

The distortion measure we use specializes the measure in [29] by removing the user intervention. After this specialization, the semantic of this distortion metric becomes clearer, and easy to analyze, and does not rely on user's input to work.

The distortion increases when the generalization level closes to the root of the taxonomy.

5.2. Generalization

In this section, we elucidate the local recording method that we use for generalization. Generalization can be applied either by global recording [24] or local recording [9]. Here we use the technique of local recording, where each record is generalized independently with respect to some defined parameters.

We use the bottom-up local recording approach. The idea of this technique is to firstly keep the record's QID and sensitive values in their original form, then to calculate adversary's expected and observed confidences for this record. If the observed confidence is higher than expected confidence, then generalization starts with the QID attribute that has the largest left domain size (Definition 11).

Definition 11 (Left domain size). *Given a value u of an attribute q_i, the left domain size of q_i is the number of nodes at the current level of u and above in the taxonomy of q_i.*

When a record does not meet the requirement of privacy, a choice of which attribute to be generalized needs to make. Our method chooses to use the attribute having the largest left domain size. For example, suppose 'age' and 'sex' attribute have current values '26-50' and 'male' respectively. From the taxonomies (Figure 1), the left domain sizes of the QID attributes 'age' and 'sex' are 7 and 3 respectively. The attribute to be generalized will be 'age'.

5.3. Algorithm overview

Here we present an overview of our algorithm with an elaboration of the key steps. We analyze the complexity of the algorithm at the end of this section.

Algorithm 1

Input: Data set D_1 with d attributes and n' records, sampling rate β, acceptable distortion δ', attribute generalization taxonomies $T_1, T_2, \ldots, T_d$ for all attributes. Probabilities of all values in the taxonomies in D_0.

Output: D_1^*

```
Take a random sample in $D_1^*$ from $D_1$ with the rate $\beta$
Calculate the number of sampled records $n = n' \times \beta$
Generalize those attributes of all records that require minimum generalization  ▷ See Section 4
while there is a record in $D_1^*$ that does not satisfy the privacy criterion do
    for each record $t$ in $D_1^*$ and let $t' = t$ do
        Calculate the expected confidence from Equation 1
        Calculate the observed confidence from Equation 2
        Calculate $\delta$ = distortion($t'$) from Equation 4
        if Expected confidence $<$ Observed confidence then
            $t' = generalize(t')$
        else
            if $\delta > \delta'$ then
                $t'$ is not published and remove $t$ from $D_1^*$
            else
                Replace $t$ with $t'$ in $D_1^*$
            end if
        end if
    end for
end while
Output $D_1^*$.
```

Sampling and initialization are done in Steps 1-3: Algorithm-1 randomly samples the original data set D_1 with a given sample rate β and initializes D_1^* with those records. Step 2 calculates the possible number of records that is going to be published, denoted by n. It is possible that the number of published records will be less than n. The unpublished records may pose high privacy risk according to the privacy criterion. Step 3 is generalizes those attributes that are required a minimum generalization, as described in Section 4.

Checking of the generalized data set is done in Step 4: Step 4 checks whether there is any record in D_1^* that does not satisfy the privacy criterion. If so, the control goes at Step 5, otherwise at Step 20.

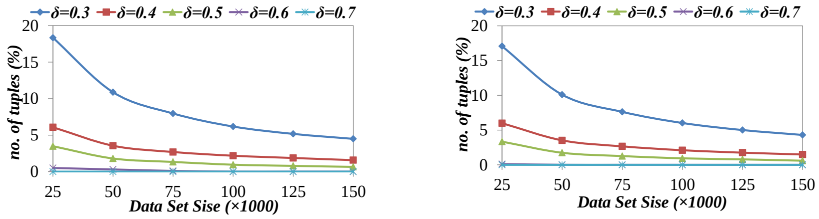
Expected and observed confidences, and distortion of a record are estimated in Steps 5-8: These steps are used to estimate the adversary's expected and observed confidences, and the distortion of a generalized record. We need to repeat the calculation of both confidences, because after each step of generalization, the generalized records move into larger equivalence groups in the generalized data set.

Generalizing and checking the satisfaction of the criterion are done in Steps 9-15: Step 9 is used to check whether the record publishing criterion is satisfied or not (Equation 3). If not, generalization is applied to the record and the control goes back to Step 6. A local recoding generalization is applied here. Step 12 is used to check the distortion of the record. If the distortion is acceptable and the observed confidence is less than the expected confidence, then the record t in D_1^* is replaced with the record t'.

Output tables is done in Steps 20: Step 20 outputs the generalized data set D_1^*.

Table 4: Distinct values of different attributes of data sets

Attribute's name Data set's name	Age	Education	Marital Status	Gender	Race	Birth Place	Sensitive Attribute	
							Salary	Occupation
OCCUPATION	77	14	6	2	6	41	50	-
SALARY	77	14	6	2	6	41	-	50
ADULTS	74	16	7	2	5	41	2	-



(a) Salary

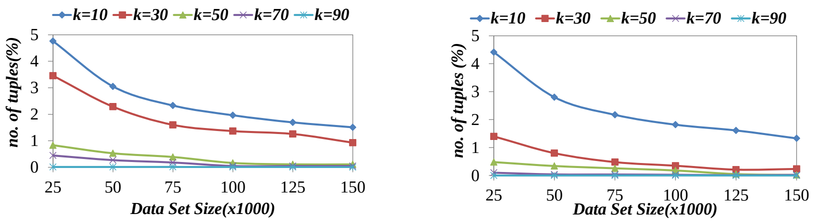
(b) Occupation

Figure 2: The percentages of suppressed records for different distortion thresholds that do not satisfy the privacy criterion in data sets (a) Salary (b) Occupation

The complexity of the algorithm lies in Step 4 to Step 19. The 'while' loop at Step 4 iterates $O(n)$ times. With each iteration, the 'for' loop at Step 5 iterates $O(n)$ times. There are $O(c)$ instructions to be executed in each time inside the 'for' loop. Thus the overall complexity of the algorithm is $O(n^2)$.

6. Experiments

In this section, our objectives are to study the impact of our privacy criterion on the data utility and to evaluate the scalability of the proposed algorithm for handling large data sets. We quantify the data utility of a published data set in terms of aggregated query accuracy and classification accuracy. For the query accuracy, we compare the difference between the answer from the anonymized data and the answer from the original data. For the classification accuracy, we compare



(a) Salary

(b) Occupation

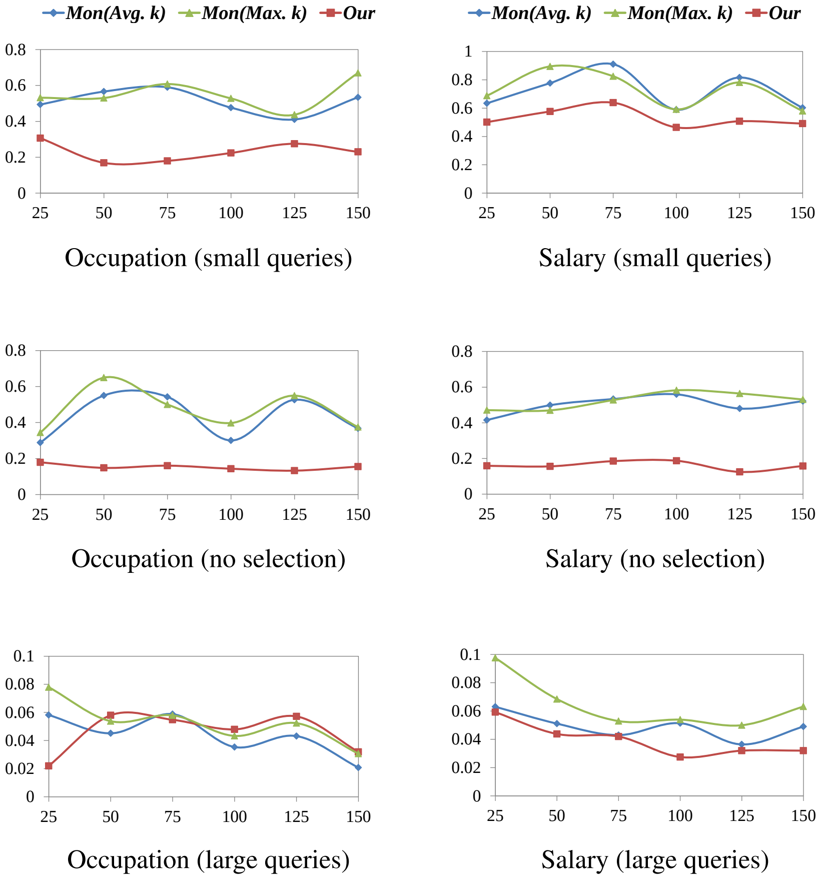
Figure 3: The percentages of unsatisfied records for different k values that do not satisfy the privacy criterion in data sets (a) Salary (b) Occupation

Table 5: Different k values for anonymizing different sizes of data sets by the Mondrian

Data Set Size	**Max.k**	**Avg.k**
25,000	90	8
50,000	90	10
75,000	90	11
100,000	70	12
125,000	70	13
150,000	70	14

the classification accuracies of various classification models built on the different anonymized data sets. With these measurements, we compare our method with a benchmark utility-aware anonymization algorithm, InfoGain Mondrian [10] and differential privacy [7]. Both aggregated query and classification are frequently used in data mining tasks. For example, the classification algorithm learns a classification model (i.e., decision trees) from the training data sets for the future use of classifying unseen data.

We perform experiments with real world data sets from U.S. Census Bureau (http://ipums.org). We split the data set into two independent data sets 1) Occu-



Occupation (small queries)

Salary (small queries)

Occupation (no selection)

Salary (no selection)

Occupation (large queries)

Salary (large queries)

Figure 4: The average query errors in comparison with a generalization based method with increasing data set size (× 1000)

pation and 2) Salary. Each data set consists of 600k records. The Occupation data set includes six quasi-identifer attributes age, sex, education, marital status, race, birth place and one sensitive attribute occupation. The Salary data set contains the same QID attributes, and its sensitive attribute is salary. All QID attributes are discrete except ‘age’ and ‘education’. The sizes of their domains are given in Table 4.

We create six groups of data sets in the following ways. We firstly define the

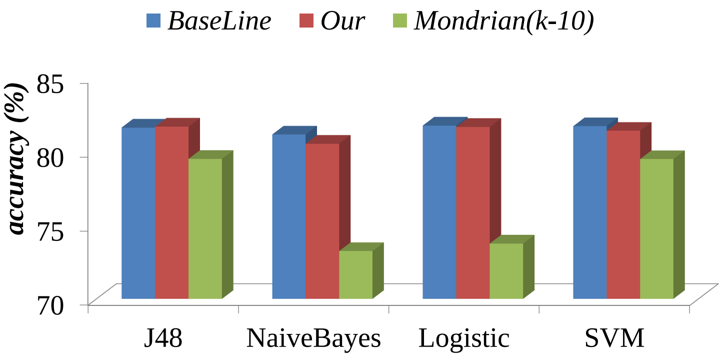


Figure 5: Classification accuracy comparison with the Mondrian

sizes of different groups of data sets as 25k, 50k, 75k, 100k, 125k and 150k. This partition is to study the scalability of the algorithm with different data set sizes. We then make six disjoint data groups from each of the data sets (Salary and Occupation) with 25k, 50k, 75k, 100k, 125k and 150k randomly drawn records respectively.

We also employ a publicly available Adult data set from UCIrvine Machine Learning Repository [11], which has been used for testing many anonymization algorithms [2, 12, 13, 14, 15, 16] for classification accuracy. The data set has 45,222 records. We make use of 6 attributes as quasi-identifiers and the salary attribute as the sensitive information. We discretise salary as $< 50k$ and $\geq 50k$ as the class attribute.

In these experiments, we assume that data distributions in the data sets are the same as in the population. On the basis of this assumption, we use the local distribution to estimate the record probability (Definition 5). For initial generalization, we intuitively apply the rule that “the value of ‘age’ attribute should have at least five years interval”. All experiments were conducted on an Intel Core $i5$ 3.30GHz

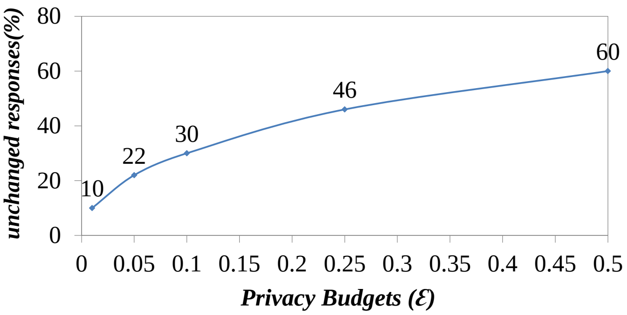


Figure 6: Unchanged responses through differentially private mechanism

PC with 4GB RAM.

When there are no parameters specified for our method in the following experiments, sampling $\beta = 90\%$ and acceptable distortion threshold $\delta' = 0.6$.

6.1. Comparison with a generalization method

We use a benchmark workload-aware anonymization algorithm Mondrian [13] to compare with our algorithm.

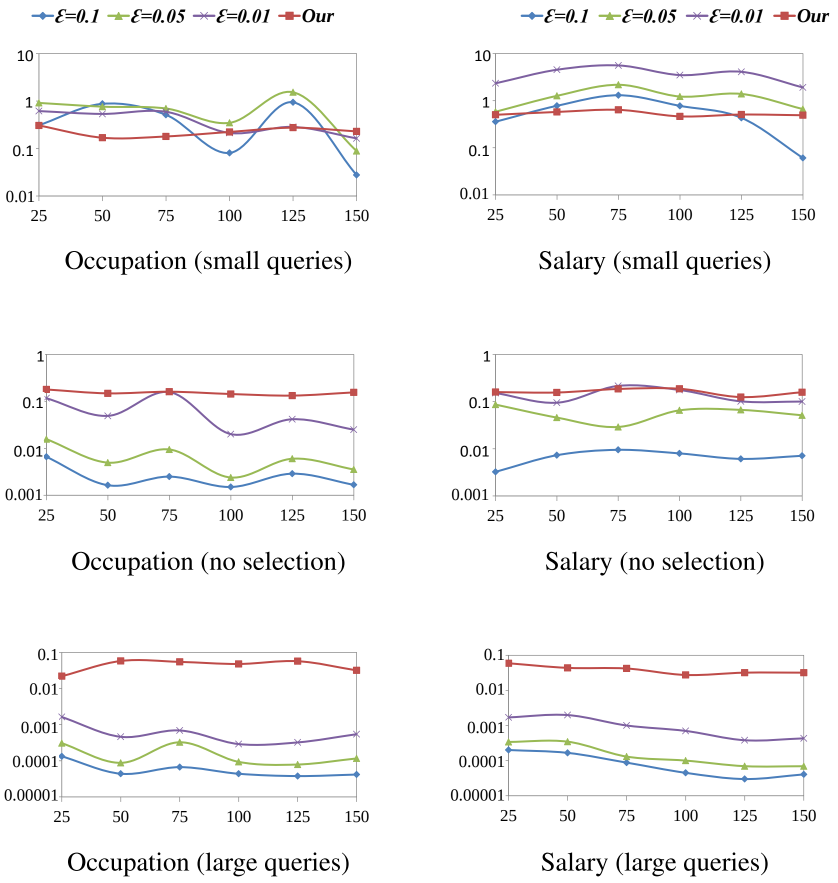
6.1.1. Query accuracy

We assess the utility of published data sets by the accuracy of answering range queries. We randomly generate 1000 queries using the following template.

Select Count (*) from D_1^* where ($t[A_1] = x_1$ AND $t[A_2] = x_2$ AND $\ldots$ AND $t[A_m] = x_m$ AND $t[S] = s$)

where x_1, x_2, $\ldots$, x_m and s are some random ranges and values which are not aligned with generalized values.

For a query, we obtain its true result R_{act} from the original data set, and compute an estimated answer R_{est} from its anonymized data set. The relative error of a query is defined as $\frac{|R_{act} - R_{est}|}{R_{act}}$. We measure the workload error as the average



Occupation (small queries)

Salary (small queries)

Occupation (no selection)

Salary (no selection)

Occupation (large queries)

Salary (large queries)

Figure 7: The average query errors in comparison with differential privacy

relative error of all the queries of all data sets. As in our method (due to 90% sampling) the anonymized data set size is less than that of Mondrian. So, after getting the R_{est} from our anonymized data set we increase the query result count by 10% in R_{est} to make it equivalent to that of Mondrian.

To make comparison with the Mondrian we employ two different approaches. In the first approach, we apply our algorithm to all disjoint data sets. We use different distortion thresholds to anonymize data sets. With different distortion thresholds we have different number of suppressed records. Figures 2(a) and 2(b) show the percentage of suppressed records for different distortion thresholds with increasing data set size for the data sets Salary and Occupation respectively. With the distortion threshold, $\delta = 0.6$, all the records satisfy the privacy criterion. We apply the same method for the Mondrian to find out, for which k (equivalence group size) all the records satisfy the privacy criterion. Figure 3(a) and 3(b) show the percentage of unsatisfied records with increasing data set size. We name the k for which all records satisfy the privacy criterion as $Max.k$. Then, we use those data sets that are anonymized by the Mondrian ($k = Max.k$) to compare the query accuracies with the data sets that are anonymized by our method ($\delta = 0.6$).

In the second approach, firstly we anonymize the data sets by our method ($\delta = 0.6$). Later on, we find out average equivalence group size (we name as $Avg.k$) of those anonymous data sets and use the average group size to anonymize the corresponding data sets by the Mondrian ($k = Avg.k$). Table 5 shows the $Agv.k$ and the $Max.k$ that are used to anonymize the data sets by the Mondrian.

Figure 4 lists the average query errors on anonymized data sets by our method and the Mondrian. Each average value is obtained from a set of 1000 random queries. We list errors of small queries (returns less than 1% count of the original data set), of large queries (returns more than 10% count of the original data set) and of unbounded quires (any count will be considered). The results show that the query errors of our method are always less than those of Mondrian.

6.1.2. Classification accuracy

To evaluate the classification accuracy, we divide the Adult data (45222 records) set into training and testing sets. Each training and testing data sets contain 40700 records and 4522 records respectively. The accuracy is obtained from 10 cross-validation based on stratified sampling. A test data set is independent from its corresponding training data set. Generalization levels are determined by the training data sets solely and then applied to the test data sets. For classification models, we use four classifiers $J48$ (an implementation of well-known $C4.5$ classifier [17] in weka [18]), $Naive\ Bayes$, $Logistic\ Regression$ and SVM. For better visualization we provide an additional measure *Baseline Accuracy (BA)*, which is the classification accuracy of the raw data without anonymization.

Figure 5 shows that in all cases we have better accuracy than the Mondrian. The accuracy of our method is very close to the base line accuracy.

6.2. Comparison with differential privacy

In the differential privacy it is crucial to choose a right ϵ (privacy budget). We ran 100,000 random queries on our data sets, and count the number of unchanged responses after adding Laplacian noise to the original count value (we replaced the fraction number with the nearest integer). The ratio of unchanged responses with different privacy budgets are shown in Figure 6. When $\epsilon > 0.1$, more than 30% of the query results have not been changed. The higher ratio of unchanged responses, the lower the privacy. We set the upper bound of ϵ as 0.1.

In these experiments, we use both the interactive[4] and the non-interactive[5] settings to make the comparison. To compare with the interactive setting and the non-interactive setting we use the measures query accuracy and classification accuracy respectively.

[4] In an interactive framework, a data miner/recepient can pose aggregate queries through an anonymization technique, and a data set owner answers these queries in response. [16]

[5] In a non-interactive framework, the data set owner first anonymizes the raw data and then releases the anonymized version. [16]

6.2.1. Query accuracy

Firstly, we use the same set of queries in the previous Subsection for the assessment of the utility of the differential privacy mechanism. For implementing differential privacy, Laplacian noises are added to the true results and errors are calculated by using the relative error of a query that is defined in Section 6.1.1. We report the averages of a set of 1000 queries in Figure 7. The query errors of differential privacy are significantly higher in small query. Yet again, to show that our method preserve better utility than the differential privacy we do the following experiment.

In this experiment, we randomly define some equivalence classes and create histograms of sensitive values (all sensitive values in its domain) for those classes in the original data sets, in the anonymous data sets by our method and in the noisy responses through the differentially private mechanism. A histogram of size m is defined as a set $H = \{\#s_i\}_{i=1}^{m}$, where i is the i'th sensitive value and m is the total number of sensitive values or the domain size of sensitive attribute. We use a probabilistic distance measure *Kullback Leibler* [19] distance and a vector distance measure *City-Block* distance to measure the distances of the histograms (histogram in the anonymous data sets by our method and histogram in the noisy responses through differentially private mechanism) from the histogram in the original data set. The smaller a distance, the better the preservation of the distribution of the original data set.

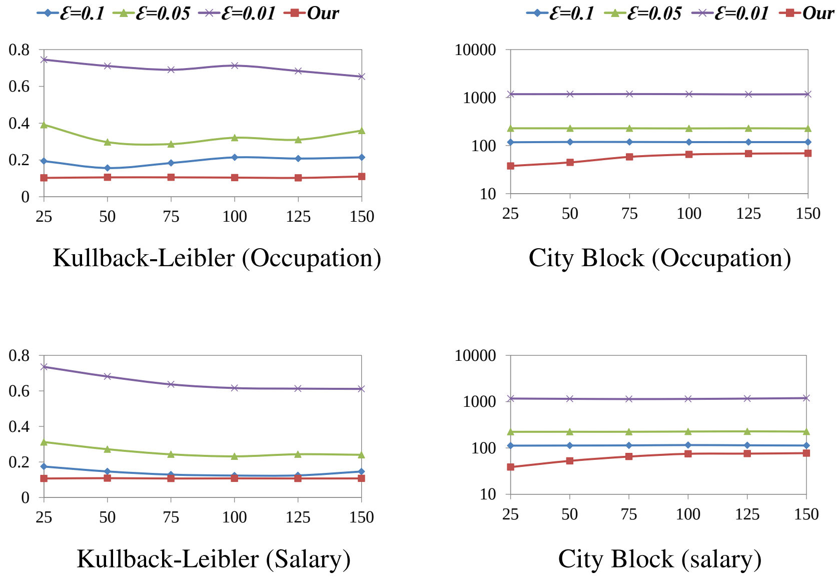
Figure 8 shows that both *Kullback-Leibler* and *City-Block* distances between the differentially private responses and the original data sets are significantly higher than the distances between our anonymized data sets and the original data sets. The reported distances are aggregated distances of 1000 random equivalence classes. The results indicate that the utility of our anonymized sets are significantly better than the utility of the differentially private responses.

These results are quite consistent with the previous results. Differential privacy is not good for small queries. The magnitude of Laplacian noise is same for small and large value, and the noises reduce the accuracy of small query results significantly.

Count query in an interactive setting is a strength of differential privacy. However, our method is non-interactive and publishes anonymous data sets. More fair comparison should be with differential privacy of non-interactive setting. The following experiment is to compare a non-interactive setting of differential privacy with our method.

6.2.2. Classification accuracy

In this experiment, we use a non-interactive [16] setting of differential privacy to compare the classification accuracy with our method. We employ the well known classifier J48 to compare the performances. Figure 9 compares the accuracy of models built on our anonymized data set with the accuracy of models built on the differentially private data set in the non-interactive setting [16]. We use both Max and $InforGain$ algorithms in [16]. The level of specialization (for details see [16]) is given 10, as with this level the both algorithms have better accuracy than other levels. The privacy budget is set to $\epsilon = 0.1$ as explained before. The classification accuracies of classification models built on differentially private data sets are significantly lower than accuracies of classification models built on anonymous data set by our method. Since many small queries are used for building classification models.



Kullback-Leibler (Occupation)

City Block (Occupation)

Kullback-Leibler (Salary)

City Block (salary)

Figure 8: Distances between histograms of sensitive values in anonymized data and original data in comparison with differential privacy

6.3. Efficiency

In these experiments, we show the efficiency of our algorithm. The Figure 10(a) shows the execution time of our algorithm with different data set sizes. We further study the scalability of our algorithm over large data sets. By randomly adding records to the data set of 100,000 records, we generate different sizes of data sets. For each record we create θ variations of the record by replacing some of the attribute values randomly from its domain. Here θ is the blowup scale and thus the total number of record is $\theta\times$ 100,000 after adding random records. Figure 10(a) shows the execution time from 200,000 to 1 million records. Our method scales well with data set sizes. However, Figure 10(b) shows that the execution time of our algorithm is slightly higher than the Mondrian. This is due to the fact that our algorithm processes each record independently. Moreover, when there is

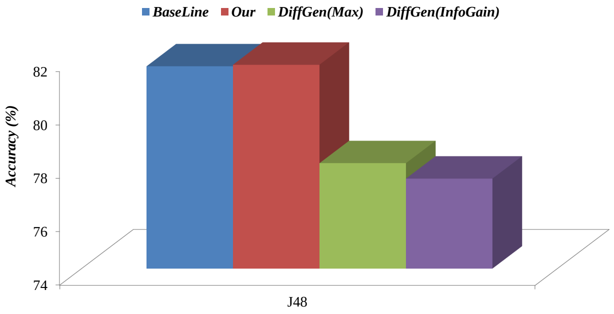
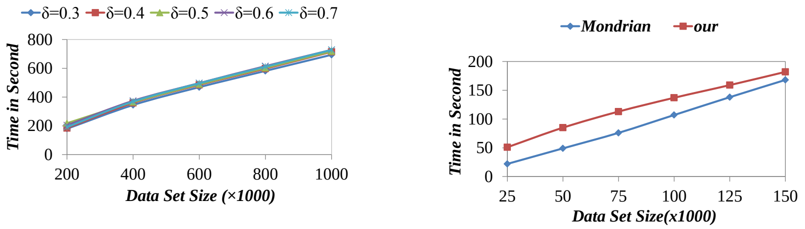


Figure 9: Classification accuracy comparison with differential privacy



(a) Execution time for different distortion thresholds

(b) Comparison with the Mondrian

Figure 10: Scalability of the designed algorithm

a generalization, our method checks all the records including those records that have already satisfied the privacy criterion at the previous iteration.

7. Related work

The privacy preserving data publishing works can be broadly classified into two categories. The first category consists of generalization-based approaches, where data values of some attributes such as age, sex and address are generalized to form small groups, so that an individual can not be identified and his/her sensitive value(s) can not be inferred with a high confidence. Among all the techniques in this category, *k*-anonymization [1] and its variations [2, 3, 4] are well studied in the literature. For details readers are referred to [20, 21]. The second category comprises of perturbation technique, where original values are noised, and hence it is difficult to pinpoint an individual in a published information. Among different techniques of perturbation methods, differential privacy [7] attracts a huge attention in the last few years.

7.1. Generalization based models and methods

Anonymization methods have been well studied in the literature [1, 13, 22, 23, 24, 25, 26, 27]. These methods are generally divided into two groups: task-specific and nonspecific. In former, the released tables are undergoing some specific data mining processes, such as building decision tree models. The purpose of the anonymization is to keep sufficient protection of sensitive information while maintaining the accuracy for a data mining task, such as classification. There have been a number of proposed methods in this group [23, 28]. On the other hand, when the data owners do not know the ultimate use of the released data, a general anonymization goal should not be associated with a data mining task but should minimize the distortions in the anonymized data sets. These methods are called nonspecific k-anonymization methods [12, 24, 29].

LeFevre et al. [13] have presented an interesting taxonomy to categorize alternative methods based on their "encoding schemas", which impose different constraints in generalizing QID values. They have divided multidimensional recoding methods into global recording and local recoding. In this paper, we use the local recoding method [9]. Global recoding methods generalize a table at the domain level. Many works of k-anonymization [12, 23, 24, 28] are based on the global recoding model. A typical global recoding generalization model is Incognito [24]. All these methods adopt k-anonymity [1] and/or its extension [2] as the underlying privacy principle. All these works are vulnerable to the recently discovered privacy attacks [30, 31, 32, 33]. A detailed discussion on generalization based categories can be found in a survey paper [21].

7.2. Randomized models and methods

An emerging line of work conforming to differential privacy [7] has received considerable attention recently. Differential privacy preserves the privacy by guaranteeing that the adversary should not be able to distinguish between two possibilities, i.e. an individual's record is in or not in the published data set. Numerous techniques have been proposed for ensuring ϵ-differential privacy [34, 35, 36, 37, 38, 39, 40, 41, 42, 43]. However, most of the research focuses on differential privacy on the interactive settings. Recently some researchers present several works [16, 34, 44, 45] to publish differentially private data sets.

Blum et al. [34] develop a technique for accurately answering range-count queries in a differentially private fashion. Another mechanism has shown to be optimal for a single counting query [47]. Xiao et al. [45] propose Privlet, a wavelet-transformation based approach that lowers the magnitude of added noise. In line to this, Hay et al. [42] also present a method to publish differential private histograms for a one-dimensional data set. Although Privelet and Hay et al.'s approaches achieve differential privacy, the latter one is applicable only to a one-dimensional data set.

In summary, existing differential privacy methods provide strong privacy guarantee and have good utility with count queries. However, the interactive settings cannot replace the microdata publishing. For details readers are referred to [48]. The non-interactive settings of differential privacy can be used as microdata publishing techniques. However, queries in a non-interactive setting has more deviations from the original data set than queries in an interactive setting. Our experiment is an evidence for this claim.

8. Conclusions

This paper presents a new privacy framework to prevent an adversary from gaining more information about an individual than an adversary can get from the public domain. We have proposed a new criterion for privacy preserving data publishing. The new privacy criterion allows a data publisher to assess the privacy risk of each record independently. We also design an effective method to implement the proposed model by integrating sampling and generalization. The empirical results show that the designed method anonymizes the data that supports better data analysis than the data anonymized by a benchmark utility-aware anonymization algorithm and the data releases by the differentially private mechanism. This work assumes that QID attributes and the sensitive attribute in a data set are independent. Our following works will consider the correlation of QID attributes and the sensitive attribute.

References

[1] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5) (2002) 1–14.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, l-diversity: Privacy beyond k-anonymity, ACM Transaction on Knowledge Discovery from Data, 1 (1) (2007) 3–es.

[3] N. Li, T. Li, S. Venkatasubramanian, t-closeness : Privacy beyond k-anonymity and l-diversity, in: IEEE International Conference on Data Engineering, 2007, pp. 106–115.

[4] R. C. W. Wong, J. Li, A. W. C. Fu, K. Wang, (α,k)-anonymity : An enhanced k-anonymity model for privacy-preserving data publishing, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2006, pp. 754–759.

[5] J. Li, Y. Tao, X. Xiao, Preservation of proximity privacy in publishing numerical sensitive data, in: ACM SIGMOD International Conference on Management of Data, 2008, pp. 473–485.

[6] Q. Zhang, N. Koudas, D. Srivastava, T. Yu, Aggregate query answering on anonymized tables, in: IEEE International Conference on Data Engineering, 2007, pp. 116–125.

[7] C. Dwork, Differential privacy, in: International Colloquium on Automata, Languages and Programming, 33rd International Colloquium, 2006, pp. 1–12.

[8] S. Hawala, Microdata disclosure protection research and experiences at the US Census Bureau, Technical Repport 1 (2003).

[9] J. Li, R. C. W. Wong, A. W. C. Fu, J. Pei, Anonymization by local recoding in data with attribute hierarchical taxonomies, IEEE Transactions on Knowledge and Data Engineering, 20 (9) (2008) 1181–1194.

[10] K. LeFevre, D. J. DeWitt, R. Ramakrishnan, Workload-aware anonymization techniques for large-scale datasets, ACM Transactions on Database Systems 33 (3) (2008) 1–47.

[11] A. Asuncion, D. J. Newman, UCI machine learning repository (2007).

[12] R. Bayardo, R. Agrawal, Data privacy through optimal k-anonymization, in: IEEE International Conference on Data Engineering, 2005, pp. 217–228.

[13] K. LeFevre, D. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity, in: IEEE International Conference on Data Engineering, 2006, pp. 25–25.

[14] V. S. Iyengar, Transforming data to satisfy privacy constraints, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2002, pp. 279–288.

[15] K. Wang, B. C. M. Fung, P. S. Yu, Handicapping attacker's confidence: An alternative to k-anonymization, Knowledge and Information Systems 11 (3) (2006) 345–368.

[16] N. Mohammed, R. Chen, B. C. M. Fung, P. S. Yu, Differentially private data release for data mining, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2011, pp. 493–501.

[17] J. R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD Explorations 11 (1) (2009) 10–18.

[19] S. Kullback, R. A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 49–86.

[20] B. C. Chen, D. Kifer, K. LeFevre, A. Machanavajjhala, Privacy-preserving data publishing, Foundations and Trends in Databases 2 (12) (2009) 1–167.

[21] B. C. M. Fung, K. Wang, R. Chen, P. S. Yu, Privacy-preserving data publishing: A survey of recent developments, ACM Computing Surveys 42 (4) (2010) 1–53.

[22] R. Agrawal, R. Srikant, D. Thomas, Privacy preserving OLAP, in: ACM SIGMOD International Conference on Management of Data, 2005, pp. 251–262.

[23] B. Fung, K. Wang, P. Yu, Top-down specialization for information and privacy preservation, in: IEEE International Conference on Data Engineering, 2005, pp. 205–216.

[24] K. Lefevre, D. J. Dewitt, R. Ramakrishnan, Incognito : Efficient full-domain k-anonymity, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005, pp. 49–60.

[25] X. Xiao, Anatomy : Simple and effective privacy preservation, in: International Conference on Very Large Data Bases, 2006, pp. 139–150.

[26] G. Ghinita, Y. Tao, P. Kalnis, On the anonymization of sparse high-dimensional data, in: IEEE International Conference on Data Engineering, 2008, pp. 715–724.

[27] W. K. Wong, N. Mamoulis, D. W. Cheung, Non-homogeneous generalization in privacy preserving data publishing, in: ACM SIGMOD International Conference on Management of Data, 2010, pp. 747–758.

[28] K. Wang, P. Yu, S. Chakraborty, Bottom-up generalization: A data mining solution to privacy protection, in: IEEE International Conference on Data Mining, 2004, pp. 249–256.

[29] J. Li, R. C. W. Wong, A. W. C. Fu, J. Pei, Achieving k-anonymity by clustering in attribute hierarchical structures, in: Data Warehousing and Knowledge Discovery, 2006, pp. 405–416.

[30] R. C. W. Wong, A. W. C. Fu, Minimality attack in privacy preserving data publishing, in: International Conference on Very Large Data Bases, 2007, pp. 543–554.

[31] R. C. W. Wong, A. W. C. Fu, K. Wang, P. S. Yu, J. Pei, Can the utility of anonymized data be used for privacy breaches?, ACM Transaction on Knowledge Discovery from Data, 5 (3) (2011) 16:1–16:24.

[32] S. R. Ganta, S. Prasad, A. Smith, Composition attacks and auxiliary information in data, in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2008, pp. 265–273.

[33] D. Kifer, Attacks on privacy and deFinetti's theorem, in: ACM SIGMOD International Conference on Management of Data, 2009, pp. 127–138.

[34] A. Blum, K. Ligett, A. Roth, A learning theory approach to non-interactive database privacy, in: ACM Symposium on Theory of Computing, 2008, pp. 609–618.

[35] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, S. Vadhan, On the complexity of differentially private data release efficient algorithms and hardness results, in: ACM Symposium on Theory of Computing, 2009, pp. 381–390.

[36] M. Gupte, M. Sundararajan, Universally optimal privacy mechanisms for minimax agents categories and subject descriptors, in: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2010, pp. 135–145.

[37] M. Hardt, K. Talwar, On the geometry of differential privacy, in: ACM Symposium on Theory of Computing, 2010, pp. 705–714.

[38] D. Kifer, No free lunch in data privacy, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2011, pp. 193–204.

[39] D. Kifer, A. Machanavajjhala, A rigorous and customizable framework for privacy, in: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2012, pp. 77–88.

[40] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, L. Vilhuber, Privacy: Theory meets practice on the map, in: IEEE International Conference on Data Engineering, 2008, pp. 277–286.

[41] X. Xiao, G. Bender, M. Hay, J. Gehrke, iReduct: Differential privacy with reduced relative errors, Proceedings of the ACM SIGMOD International Conference on Management of Data, 2011, pp. 229–240.

[42] M. Hay, V. Rastogi, G. Miklau, D. Suciu, Boosting the accuracy of differentially private histograms through consistency, Proceedings of the VLDB Endowment 3 (1) (2010) 1021–1032.

[43] A. Roth, T. Roughgarden, Interactive privacy via the median mechanism, in: ACM Symposium on Theory of Computing, 2010, pp. 765–774.

[44] X. Xiao, G. Wang, J. Gehrke, T. Jefferson, Differential privacy via wavelet transforms, IEEE Transactions on Knowledge and Data Engineering, 23 (8) (2011) 1200–1214.

[45] Y. Xiao, L. Xiong, C. Yuan, Differentially private data release through multidimentional partitioning, in: International Conference on Very Large Data Bases, 2010, pp. 150–168.

[46] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, K. Talwar, Privacy, accuracy, and consistency too: A holistic solution to contingency table release, in: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2007, pp. 273–282.

[47] A. Ghosh, T. Roughgarden, M. Sundararajan, Universally utility-maximizing privacy mechanisms, in: ACM symposium on Theory of computing, 2009, pp. 351–360.

[48] N. Li, W. Qardaji, D. Su, Provably private data anonymization : Or , k-anonymity meets differential privacy, Technical report, Purdue University (2011).