Probabilistic Inference Protection on Anonymized Data

Raymond Chi-Wing Wong[1], Ada Wai-Chee Fu[2], Ke Wang[3], Yabo Xu[4], Jian Pei[3], Philip S. Yu[5]

[1]The Hong Kong University of Science and Technology
[2]The Chinese University of Hong Kong, [3]Simon Fraser University
[4]Sun Yat-sen University, [5]University of Illinois at Chicago

raywong@cse.ust.hk, adafu@cse.cuhk.edu.hk,
{wangk,jpei}@cs.sfu.ca, xuyabo@mail.sysu.edu.cn, psyu@cs.uic.edu

***Abstract*—Background knowledge is an important factor in privacy preserving data publishing. Probabilistic distribution-based background knowledge is a powerful kind of background knowledge which is easily accessible to adversaries. However, to the best of our knowledge, there is no existing work that can provide a privacy guarantee under adversary attack with such background knowledge. The difficulty of the problem lies in the high complexity of the probability computation and the non-monotone nature of the privacy condition. The only solution known to us relies on approximate algorithms with no known error bound. In this paper, we propose an algorithm that overcomes the difficulties of the problem by introducing a bounding condition on probability deviations in the anonymized data groups, which is much easier to compute and which is a monotone function on the grouping sizes. This bounding condition is also in harmony with the utility preservation objective. Our empirical studies show that our method preserves data utilities at a higher or comparable level when compared with some state-of-the-art algorithms that provide less protection.**

I. Introduction

With the increasing collections of data containing information of vast populations, which are potentially useful for different kinds of analysis, the issue of privacy preserving data publishing has become an important topic for database communities.

In our problem, a table such as Table I is to be anonymized for publication. We assume that each tuple in the table is owned by an individual and each individual owns at most one tuple. The table has two kinds of attributes, (1) the quasi-identifier (QI) attributes and (2) the sensitive attribute. The QI attributes can be an individual identifier in the table. In our example, the QI attributes are Gender and Age. [10] points out that in a real dataset, with the help of a publicly available *external table* such as a voter registration list, about 87% of individuals can be uniquely identified by only three QI attributes, namely sex, date of birth and 5-digit zip code. An example of a voter registration list is shown in Table II. The sensitive attribute contains some sensitive values that should be protected. In our example, the sensitive attribute is "Disease" with sensitive values such as Lung Cancer and HIV. Note that the attribute "Name" is an obvious identifier and will be removed before publication.

Our target is to anonymize T and publish an anonymized dataset T^* to satisfy some privacy requirements. A common technique is to horizontally partition T into multiple tuple groups, also called anonymized groups or *A-groups*. Let L be a resulting group. Each group is given a unique ID called GID. The linkage between individual records and the sensitive attribute in each A-group is broken. One way to achieve this is *bucketization*, forming two tables, the *QI table* (Table III(a)) for the QI attributes and the *sensitive table* (Table III(b)) for the sensitive attribute. These two tables form the anonymized dataset T^*.

Let us consider a simplified setting of the l-diversity model [8] as a privacy requirement for published data T^*. An A-group is said to be *l-diverse* or satisfy *l-diversity* if in the A-group the number of occurrences of any sensitive value is at most $1/l$ of the group size. A table satisfies l-diversity (or it is l-diverse) if all A-groups in it are l-diverse. Suppose that Table I is anonymized to Table III. It is easy to see that Table III satisfies 2-diversity. The target of 2-diversity is that each individual cannot be linked to a disease with a probability of more than 0.5. However, we show below that this table does not meet this target if we consider distribution based background knowledge. In the following, we simply refer to the A-group with GID equal to L_i by L_i.

Example 1: Consider L_1 in Table III. In L_1, Lung Cancer and Hypertension are values of the sensitive attribute Disease. If we are given the voter registration list as shown in Table II, one can determine that the two tuples in L_1 correspond to Alan and Betty. Without additional information one concludes that each of Alan and Betty has a 50% chance of linking to Lung Cancer (Hypertension). However, suppose we are given Table IV which discloses that the probability of a male patient being linked to Lung Cancer is 0.1 and that of a female patient is 0.003. With this distribution, the adversary can deduce that Betty, being a female patient, has less chance of having Lung Cancer while Alan, being a male patient, has a higher chance. The intended protection guarantee of 50% threshold is thus violated. □

The above example shows that background knowledge has important impact on privacy preserving data publishing.

Name	Gender	Age	Disease
Alan	Male	41	Lung Cancer
Betty	Female	42	Hypertension
Catherine	Female	63	Flu
Diana	Female	64	HIV
	...	...	...

Table I
GIVEN DATASET T

Name	Gender	Age
Alan	Male	41
Betty	Female	42
Catherine	Female	63
Diana	Female	64
...	...	...

Table II
VOTER REGISTRATION LIST

Gender	Age	GID
Male	41	L_1
Female	42	L_1
Female	63	L_2
Female	64	L_2
...	...	...

(a) QI Table

GID	Disease
L_1	Lung Cancer
L_1	Hypertension
L_2	Flu
L_2	HIV
...	...

(b) Sensitive table

Table III
2-DIVERSE T^*

$p()$	Lung Cancer	Not Lung Cancer
Male	0.1	0.9
Female	0.003	0.997

Table IV
A QI BASED PROBABILITY DISTRIBUTION FOR "GENDER"

Although in the example, the anonymization is based on bucketization, the same issue arises with a generalization based method. The reason is that the adversary has at his/her disposal the external table with which he/she may be able to look up the details of individuals who are mapped to an A-group. For example, if the QI values of L_2 in Table III are generalized to { Female, 6* }, and if Catherine and Diana are the only female patients with a Age of 6* in the external table, Table II, then the adversary can determine that they are the owners of the two tuples in L_2 and all their exact QI values can be determined. Once the details are determined, the adversary can estimate the revised probabilities.

In this paper, we consider background knowledge in the form of QI based distribution, which is the distribution of the values in the sensitive attribute restricted to individuals with the same values on some QI attributes. For example, the distribution of the sensitive attribute values according to female patients may be encoded as {(Female:"Lung Cancer", 0.003), (Female:"Hypertension", 0.21), ...} where (Female:x, p) denotes that the probability that a female patient is linked to a value x is p. This is called the *apriori distribution* since it is assumed to be known by the adversary before T^* is published. It can be seen that such background knowledge is not difficult to come by given a lot of statistics available from the government or other agencies (e.g., statistical reports from the US Department of Health and Human Services and other statistical data sources given in [7], [9]).

Given that the linkage probability can be affected by an apriori probabilistic distribution that is known by the adversary, one obvious approach is to incorporate the revised probability into an existing anonymization method so that the new probability is measured against the threshold in each validation step. This is in fact the strategy in [7]. The authors have adopted the algorithm of Mondrian [3] and incorporated bucketization for the result generation. However, this straight-forward approach has a major obstacle. The complexity of computing the above linkage probability, also called *posterior belief* in [7] (as opposed to the apriori distribution), is very high. As pointed out in [7], this problem is # P-complete. Also, with Mondrian or other anonymization methods, the computation is carried out many times during the state space search. Therefore, they have resorted to an approximation algorithm for computing the probabilities. To our knowledge, [7] is the only previous work that has dealt with probabilistic adversary knowledge of QI based distribution. However, the use of approximated probability computation implies that there is no solid guarantee on the privacy protection. Such a solution may not be desirable since it compromises the one issue that users are most concerned about. In this paper, we propose the first algorithm that solves this problem with a solid guarantee.

The key to our solution is that one does not need to compute the exact linkage probabilities in order to provide a guarantee. Instead, we prove that once the anonymization satisfies certain conditions that are easy to compute, the privacy is guaranteed. The essence of our method is the principle of *similar linkage*. Specifically, we observe that privacy is breached whenever an individual in an A-group has a *much higher* chance of linking to a sensitive value compared with another individual in the A-group according to the QI based distribution. Based on this observation, we propose the principle of similar linkage and develop a solution which generates a dataset such that all individuals in each A-group have "similar" chances of linking to any sensitive value in the group, according to the distribution. Since they have "similar" chances, it is not possible for the adversary to pinpoint any linkage of an individual to a sensitive value with a higher chance. This observation has motivated us to translate this idea into a concrete formulation of some computable properties of the A-groups. We define a measurement called Greatest Probability Deviation, $\triangle_{max}$, to model the fluctuation of probabilities in an A-group. A bound is derived for this measurement to ensure privacy. Instead of enforcing the required probabilities, we enforce this required condition on $\triangle_{max}$ which is much easier to compute.

Note also that the above linkage probability or posterior belief is not monotone in that an A-group violating privacy can be split into two groups that preserve privacy. For example, an A-group with 2 female patients and 2 male patients may violate privacy, but when it is split into a group of 2 female patients and another group of 2 male patients, the privacy can be preserved. However, most existing anonymization methods depend on the monotone property of

the privacy guarantee. The implication is that a supposedly exhaustive algorithm like Incognito [4] is neither exhaustive nor optimal.

Our contributions can be summarized as follows. To the best of our knowledge, our algorithm is the first solution to handle probabilistic adversary knowledge of QI based distribution without compromising the privacy guarantee. We derive an interesting and useful theoretical property for the anonymization and based on this property, our algorithm generates a dataset protecting individual privacy in the presence of the probabilistic adversary knowledge. Our empirical study confirms that our algorithm can be executed in reasonable runtime and provides comparable or lower information loss in terms of a widely accepted utility measure when compared to a number of state-of-the-art anonymization algorithms.

II. Problem Definition

Let T be a table. We assume that one of the attributes is a sensitive attribute X where some values of this attribute should not be linkable to any individual. These values are called sensitive values. The value of the sensitive attribute of a tuple t is denoted by $t.X$. A *quasi-identifier* (QI) is a set of attributes of T, $A_1, A_2, ..., A_q$, that may serve as identifiers for some individuals. Each tuple in the table T is related to one individual and no two tuples are related to the same individual. With publicly available voter registration lists (like Table II), the QI values can often be used to identify a unique individual.

The first step of privacy preserving data publication is to determine the target of protection. In our problem setting, the target of protection is to limit the probability of a linkage from an individual to some sensitive value based on the knowledge of an adversary. In the literature [14], [12], [6], [5], it is assumed that the knowledge of an adversary includes (1) the published dataset T^*, (2) a publicly available external table T^e such as a voter registration list that maps QIs to individuals [10], [12] and (3) some background knowledge. We also follow these assumptions in our analysis. We focus on the QI based distribution as background knowledge.

The QI is made up of a set of attributes. Each possible value for an attribute set such as "Gender" in our example is called a *signature*. In general, there can be different attribute sets, such as {"Gender", "Age"}, for which a signature s can be {("Gender", "Male"), ("Age", "41")}. For convenience, we often drop the attribute names, and thus we have {"Male", "41"} for the above signature. The first tuple in Table III(a) matches {"Male"} but the second does not.

Definition 1 (Signature $t.s$): Given a QI attribute set $\mathcal{A}$ with q attributes $A_1, ..., A_q$. A *signature* s of $\mathcal{A}$ is a set of attribute-value pairs $(A_1, v_1), ..., (A_q, v_q)$ which appear in the published dataset T^*, where A_i is a QI attribute and v_i is a value. A tuple t in T^* is said to match s if $t.A_i = v_i$ for all $i = 1, 2, ..., q$. We also say that $t.s = s$. □

The QI based apriori distribution for the attribute set {"Gender"} is described in Table IV. Each probability in the table is called an *aprori probability*. The **sample space** for each such discrete probability distribution consists of the possible assignments of the sensitive values such as x to an individual with the particular gender. For signature s, the sample space is denoted by Ω_s.

Definition 2 (Apriori Distribution G): Given an attribute set $\mathcal{A}$, the *QI-based distribution* G of $\mathcal{A}$ contains a set of entries $(s : x, p)$ for each possible signature s of $\mathcal{A}$, where p is equal to $p(s : x)$ which denotes the apriori probability that a tuple matching signature s is linked to x. □

For example, G may contain ("Female":"Lung Cancer", 0.003) and ("Male":"Lung Cancer", 0.1). This involves two sample spaces Ω_{Female} and Ω_{Male}.

Definition 3 (r-robustness): Assume that an adversary has the background knowledge of the QI-based distribution. A dataset T^* is said to satisfy r-robustness (or T^* is r-robust) if, for any individual t and any sensitive value x, the probability that t is linked to x, $p(t : x)$, does not exceed $1/r$. □

In this paper, we study the following problem.

Definition 4 (problem): Given a dataset T, generate an anonymized dataset T^* from T which satisfies r-robustness and at the same time minimizing the information loss. □

There have been different definitions for information loss in the literature. In our experiments, we shall adopt the measurement of accuracy in query results from T^* versus that from T. We assume that the published data is meant for data analysis and most data analysis can be modeled by certain aggregate queries on the dataset. It has been found in previous studies [14], [12], [6] that the accuracy in certain types of queries can give an indication of the utility of the published dataset.

The only previous work that deals with QI-based distribution is [7]. Their problem definition, however, is based on a relative bound on a distance measure between the aprior linkage probability before the published table is given and the posterior probability after the data is published. We believe that both an absolute bound such as $1/r$ in our definition and a relative bound have their merits. An absolute bound may be too rigid in case the prior probability already exceeds the given bound, though this problem can be handled by setting r appropriately. On the other hand, a good relative bound may be too complex to be understandable for naive users, for example, the distance measure used in [7] involves kernel smoothing and JS divergence. In this paper, we focus on an absolute bound.

III. Probability Formulation

There are two common approaches for anonymization, which generates T^* from T: generalization and bucketiza-

tion. In both approaches, the tuple set of T is partitioned into multiple anonymized groups or A-groups. Bucketization is more challenging since the adversary is saved the effort to uncover the QI values for individuals in an A-group. We focus on bucketization but our results apply readily to generalization. With anonymization, there is a mapping which maps each tuple in T to an A-group in T^*. For example, the first tuple t_1 in Table I is mapped to A-group L_1.

Suppose there are m possible signatures for attribute set $\mathcal{A}$, namely $s_1, s_2, ..., s_m$. Let G be the background knowledge consisting of the set of all QI based distributions. In G, the probability that s_i is linked to a sensitive value x is given by $p(s_i : x)$. Given G, the formula for $p(t : x)$, the probability that a tuple t is linked to sensitive value x, is derived below.

Definition 5 (Possible World): Consider an A-group L with N tuples, namely $t_1, t_2, ..., t_N$, with corresponding values in sensitive attribute X of $\gamma_1, \gamma_2, ...\gamma_N$. A possible world w for L is a possible assignment mapping the tuples in set $\{t_1, t_2, ..., t_N\}$ to values in multi-set $\{\gamma_1, \gamma_2, ...\gamma_N\}$ in L. □

Given an A-group L with a set of tuples and a multi-set of the values in X. Considering all possible worlds, we form a sample space. More precisely, the **sample space** Ω_L consists of all the possible assignments of the sensitive values in L to the N tuples in L. We call each possible assignment a *possible world*. For each such possible world w, according to the QI based distribution G based on attribute set $\mathcal{A}$, we can determine the probability $p(w|L)$ that w occurs given L.

Definition 6 (Primitive Events, Projected Events): A mapping $t : x$ from an individual or tuple t to a value x in the set of sensitive attributes is called a *sensitive event*. Such an event corresponds to the set of possible worlds in Ω_L where t is assigned x. Denote the signature of tuple t by $t.s$. Let us call an event for the corresponding signature, "$t.s : x$", a *projected event* for t. □

The probability of a sensitive event, $p(t : x)$, is the probability of interest for the adversary. The projected event, $(s : x)$, is an event of sample space Ω_s which consists of possible worlds of assigning different values to s. The probability $p(s : x)$ is assumed to be known since we assume the knowledge of the set of QI based distributions G. Note that $p(s : x)$ is independent of L.

The probability that w occurs given L is proportional to the product of the probabilities of the corresponding projected events for the tuples $t_1, ...t_N$ in L, we shall denote this product as $p(w)$:

$$p(w) = p_{1,w} \times p_{2,w} \times ... \times p_{N,w} \quad (1)$$

where $p_{j,w}$ is the probability that t_j is linked to a value in the sensitive attribute specified in w. Suppose t_j matches signature s_i. If t_j is linked to x in w, then $p_{j,w} = p(s_i : x)$.

Let the set of all the possible worlds for L be $\mathcal{W}$. The sum of probabilities of all the possible worlds given L must be 1, since they form the sample space Ω_L. Therefore, we want to make sure that $\sum_{w \in \mathcal{W}} p(w|L) = 1$.

Hence, the probability of w given L is given by:

$$p(w|L) = \frac{p(w)}{\sum_{w' \in \mathcal{W}} p(w')} \quad (2)$$

With the above equation, it is easy to verify that $\sum_{w \in \mathcal{W}} p(w|L) = 1$.

We aim to find the probability that an individual t_j in L is linked to a sensitive value x. This is given by the sum of the probabilities $p(w|L)$ of all the possible worlds w where t_j is linked to x.

$$p(t_j : x) = \sum_{w \in \mathcal{W}^{(t_j:x)}} p(w|L) \quad (3)$$

where $\mathcal{W}^{(t_j:x)}$ is the set of all possible worlds w in $\mathcal{W}$ in which t_j is assigned value x.

Note that our probabilistic formulation is basically the same as that in [7] despite different terminologies. As pointed out in [7], we can compute $p(t : x)$ by enumerating all the possible worlds in Ω_L. However, the total number of possible words is exponential in the size of L. If the sensitive values in L are $a_1, ..., a_m$, and the frequency of a_i in L is f_i, then the number of possible worlds is $\frac{N!}{\prod_{i=1}^{m} f_i!}$, where $\sum_i f_i = |L|(= N)$.

IV. Theoretical Properties

Given the problem definition our next task is to find an anonymization algorithm. Most known algorithms belong to one of two main categories: top-down and bottom-up. With top-down approach [2], we start with the whole table as a single A-group and recursively split the current groups until the privacy condition is violated and report the smallest qualified A-groups. Smaller A-groups are more favorable since they tend to incur less information loss. Note that the privacy condition is checked at each splitting. The bottom-up approach [11] goes in the reverse direction: starting with single tuple A-groups, we merge A-groups until the privacy condition is met. Both approaches depends on a monotone property of the privacy condition: if an A-group violates privacy, then splitting the group into smaller groups will also violate privacy.

A naive approach for r-robustness is to adopt some known anonymization algorithm A and replace the probability measure in A by $p(t : x)$. However, the complexity of computing $p(t : x)$ is very high. Moreover, r-robustness is not monotone so that an A-group that violates r-robustness may be split into small groups that are r-robust, while top-down or bottom-up algorithms are based on monotone privacy conditions.

In this section, we presents an important theoretical property for this problem which can help us to overcome the above difficulties. Section V describes our proposed

algorithm, ART, which is based on this theoretical property. Our algorithm will be a bottom-up algorithm, starting with singleton A-groups and merging them to form bigger groups when necessary. Our theoretical property allows us to set up a new privacy condition that does not require the computation of $p(t : x)$.

In Section I, we observe that privacy is breached easily whenever an individual in an A-group has a much higher chance of linking to a sensitive value compared with another individual in the A-group. For example, consider the A-group L_1 in Table III. From the QI-based distribution (Table IV), it is more likely that a male patient is linked to Lung Cancer compared with a female patient. Note that the apriori probability of a male patient linking to Lung Cancer, f_1, is 0.1 and that of a female patient, denoted by f_2, is 0.003. The difference in the apriori probabilities is $0.1 - 0.003 = 0.097$. This difference is the culprit that aids privacy breach.

Consider a tuple t_v in an A-group L and a sensitive value x. We want to show that if L satisfies a certain condition, the privacy of t_v can be guaranteed (i.e., $p(t_v : x) \leq 1/r$). The condition essentially limits the deviations in the apriori probabilities in terms of the group size.

In the following, we require that for *each* sensitive value x in X, each A-group L contains at most one occurrence of x. This requirement (called m-*uniqueness* in [15]) helps to increase the number of possible sensitive values in L and weaken the linkage to any such value. It also allows us to determine any privacy breach in $O(1)$ time. It can be easily satisfied if the frequency of each sensitive value is not high. Note that similar requirements are found in other models, including m-invariance [15] and Anatomy [14], which requires that each sensitive value appears at most once in each group for ℓ-diversity. Conceptually, if the frequency of a value is high, then it is a common phenomenon, and common phenomena are typically not sensitive.

In the following, we consider the QI based apriori distribution G on a certain attribute set $\mathcal{A}$. The algorithm to be described later will consider multiple attribute sets.

Definition 7 (Probability Deviation, $\triangle_v$): Let L be an A-group in T^* with tuples $t_1, t_2, ...t_N$ and $N \geq r$. Let x be a sensitive value that appears exactly once in L. Let the signature of t_v be $t_v.s$, $v \in [1, N]$. Suppose that for tuple t_v, the apriori probability $p(t_v.s : x) = f_v$.

Let $f_{max} = \max_{v \in [1,N]} f_v$. The probability deviation of t_v given f_{max} is:

$$\triangle_v = f_{max} - f_v$$

□

Now we are ready to introduce a property whereby an A-group can guarantee r-robustness.

Theorem 1 ($\triangle$ Bounding Condition): Let r be the privacy parameter in r-robustness where $r > 1$. Following the symbols in Definition 7, if for all $v \in [1, N]$,

$$\triangle_v \leq \frac{(N-r)f_{max}}{f_{max}(r-1)/(1-f_{max}) + (N-1)} \tag{4}$$

then for all $v \in [1, N]$, $p(t_v : x) \leq 1/r$ □

A proof of this theorem can be found in [13].

Definition 8 ($\triangle_{ceil}$, $\triangle_{max}$): $\triangle_{ceil}$ is defined to be the R.H.S. of Inequality (4). That is,

$$\triangle_{ceil} = \frac{(N-r)f_{max}}{f_{max}(r-1)/(1-f_{max}) + (N-1)}$$

Define the Greatest Probability Deviation as

$$\triangle_{max} = \max_{v \in [1,N]} \{\triangle_v\}$$

$\triangle_{max}$ is the greatest difference in the apriori probabilities linking to x in an A-group. Note that $\triangle_{ceil} \geq \triangle_{max} \geq 0$. Rewriting Theorem 1, we have:

Corollary 1 ($\triangle$ bounding condition): Let r be the privacy parameter in r-robustness where $r > 1$.

$$\text{If } \triangle_{max} \leq \triangle_{ceil} \tag{5}$$

then for all $v \in [1, N]$, $p(t_v : x) \leq 1/r$ □

The inequality in Theorem 1 (or Corollary 1) corresponds to the principle of *similar linkage* we mentioned in Section I. Intuitively, the greatest difference in the aprioir probabilities linking to x in an A-group should be bounded. In other words, the apriori probabilities should be "similar".

Note that the computation of $\triangle_{ceil}$ takes $O(1)$ time. After we obtain the value of $\triangle_{ceil}$, we can determine whether $p(t_v : x) \leq 1/r$ by Inequality (4) in O(1) time, which is quite efficient.

Let us consider the effects of the values of f_{max} and N to understand the physical meaning of Theorem 1. If $f_{max} = 1$ or $f_{max} = 0$, then $\triangle_{max} = 0$. Hence, the QI based distributions of all tuples in L should be the same to guarantee privacy.

Table V shows the values of $\triangle_{ceil}$ with some chosen values of N, r and f_{max}. It can be seen that $\triangle_{ceil}$ is small when f_{max} is near the extreme values of 0 or 1, since the apriori probability of a tuple is more pronounced.

Consider Inequality (4) again. If $N \to \infty$, then $\triangle_{max} \leq f_{max}$. Since f_{max} is the greatest possible apriori probability

N	r	f_{max}	$\triangle_{ceil}$
3	2	0.1	0.0474
3	2	0.3	0.1235
3	2	0.5	0.1667
3	2	0.9	0.0818
4	2	0.3	0.1750
6	2	0.3	0.2211
6	3	0.3	0.1537
6	4	0.3	0.0955

Table V
VALUES OF $\triangle_{ceil}$ WITH SOME CHOSEN VALUES OF N, r AND f_{max}

in L, it means that $\triangle_{max}$ can be any feasible value (i.e., $0 \leq \triangle_{max} \leq f_{max}$). Therefore, when the A-group is extremely large, under Theorem 1, there will be no privacy breach. When $N = r$, $\triangle_{max} \leq 0$. That is, the apriori probabilities of all tuples in L should be equal. Otherwise, there may be a privacy breach. Furthermore, N has the following relation with $\triangle_{ceil}$.

Theorem 2 (Monotonicity): $\triangle_{ceil}$ is a monotonically increasing function on N.

A proof of the above theorem can be found in [13]. From the above, in order to guarantee $p(t_v : x) \leq 1/r$, we can increase the size N of the A-group L. With a greater value of N, the upper bound $\triangle_{ceil}$ increases, and the constraint as dictated by Inequality (4) is relaxed, making it easier to reach the guarantee.

V. Algorithm

Based on Theorem 1, we propose a bottom-up Algorithm generating r-Robust Table called *ART*. If an A-group L satisfies the inequality in Theorem 1 with respect to attribute set $\mathcal{A}$ and, in L, each sensitive value occurs at most once, we say that L satisfies the $\triangle$ *bounding condition* with respect to $\mathcal{A}$. Otherwise, L violates the $\triangle$ bounding condition.

In the algorithm, initially, each individual forms an independent A-group. $\triangle_{max} = 0$ for each group. The algorithm repeatedly looks for any A-group such that there exists an attribute set $\mathcal{A}$ where it violates the $\triangle$ bounding condition with respect to $\mathcal{A}$. Such a group is merged with other existing groups so that the resulting group satisfies the condition. After merging, the number of tuples in L, N, is increased. Then, by Theorem 2, $\triangle_{ceil}$ is also increased. The constraint by Inequality (4) is relaxed and it is more likely to satisfy the $\triangle$ bounding condition. When a final solution is reached, each individual is linked to any sensitive value with probability at most $1/r$.

Specifically, algorithm ART involves two major steps.

- **Step 1** (Individual A-group Formation): For each tuple t in the table T, we form an A-group L containing t only.
- **Step 2** (Merging): For *each* sensitive value x, while there exists an A-group L and an attribute set $\mathcal{A}$ such that L violates the $\triangle$ bounding condition with respect to $\mathcal{A}$, we find a set $\mathcal{L}$ of A-groups such that, after merging all A-groups in $\mathcal{L}$ with L, the merged A-group satisfies the $\triangle$ bounding condition with respect to any attribute set $\mathcal{A}$.

The idea of Step 2 is to keep the $\triangle_{max}$ value in L with respect to $\mathcal{A}$ unchanged or only slightly increased after merging. At the same time, we also make sure that each merged A-group contains at most one x for any sensitive value x.

$\triangle_{max}$ is the greatest difference in the apriori probabilities among tuples in an A-group L. A set of more uniform QI values in L incurs less information loss, and also has a smaller difference in apriori probabilities. In this way, minimizing $\triangle_{max}$ can help to minimize information loss. Since we always start with $\triangle_{max} = 0$ in our bottom-up algorithm, and by the monotone property (discussed in Theorem 2), $\triangle_{ceil}$ will be increasing, we try to keep the $\triangle_{max}$ value unchanged or only slightly increased after merging. This goal can be achieved by merging an A-group L with another A-group L' which has a "similar" apriori probability to L. L' is said to be close to L.

Theorem 3 (Correctness): Any table T^* generated by Algorithm ART is r-robust.

VI. Conclusion

In this paper, we consider the background knowledge of QI based probabilistic distribution that may be possessed by the adversary in privacy-preserving data publishing. While the problem is difficult due to high complexity in the probability computation, the setting is realistic and powerful in that it covers some other known background knowledge such as positive associations and negative associations [1]. For future work, we may investigate how to anonymize the dataset with other kinds of probabilistic background knowledge, including the association among individuals such as members of the same family.

Acknowledgement: The research of Raymond Chi-Wing Wong is supported by HKRGC GRF 621309. The research of Ke Wang is supported by a Discovery Grant from NSERC. The research of Philip S. Yu is supported by US NSF through grants IIS-0914934, DBI-0960443, OISE-0968341 and OIA-0963278.

References

[1] B.-C. Chen, K. LeFevre, and R. Ramakrishnan. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB*, 2007.
[2] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, 2005.
[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, 2006.
[4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*, 2005.
[5] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.
[6] T. Li and N. Li. Injector: Mining background knowledge for data anonymization. In *ICDE*, 2008.
[7] T. Li, N. Li, and J. Zhang. Modeling and integrating background knowledge in data anonymization. In *ICDE*, 2009.
[8] A. Machanavajjhala, J. Gehrke, and D. Kifer. l-diversity: privacy beyond k-anonymity. In *ICDE*, 2006.
[9] D. J. Martin, D. Kifer, A. Machanavajjhala, and J. Gehrke. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*, 2007.
[10] L. Sweeney. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5), 2002.
[11] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, 2004.
[12] R. Wong, A. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, 2007.
[13] R. C.-W. Wong, A. W.-C. Fu, K. Wang, Y. Xu, J. Pei, and P. Yu. Probabilistic inference protection on anonymized data. In *http://www.cse.ust.hk/∼raywong/paper/probInferenceProtection-technical.pdf*, 2010.
[14] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.
[15] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.