Can the Utility of Anonymized Data be used for Privacy Breaches?

Raymond Chi-Wing Wong[1], Ada Wai-Chee Fu[2], Ke Wang[3], Philip S. Yu[4], Jian Pei[3]
[1] *The Hong Kong University of Science and Technology*
[2] *The Chinese University of Hong Kong*
[3] *Simon Fraser University*
[4] *University of Illinois at Chicago*
raywong@cse.ust.hk, adafu@cse.cuhk.edu.hk,
{wangk,jpei}@cs.sfu.ca, psyu@cs.uic.edu

Group based anonymization is the most widely studied approach for privacy preserving data publishing. Privacy models/definitions using group based anonymization includes k-anonymity, l-diversity, and t-closeness, to name a few. The goal of this paper is to raise a fundamental issue on the privacy exposure of the approaches using group based anonymization. This has been overlooked in the past. The group based anonymization approach by bucketization basically hides each individual record behind a group to preserve data privacy. If not properly anonymized, patterns can actually be derived from the published data and be used by the adversary to breach individual privacy. For example, from the medical records released, if patterns such as people from certain countries rarely suffer from some disease can be derived, then the information can be used to imply linkage of other people in an anonymized group with this disease with higher likelihood. We call the derived patterns from the published data the foreground knowledge. This is in contrast to the background knowledge that the adversary may obtain from other channels as studied in some previous work. Finally, our experimental results show that the attack is realistic in the privacy benchmark dataset under the traditional group based anonymization approach.

Categories and Subject Descriptors: H.2.0.a [**Information Technology and Systems**]: Database Management—*General*

General Terms: Algorithms, Experimentation, Security

Additional Key Words and Phrases: privacy preservation, data publishing, l-diversity, k-anonymity

1. INTRODUCTION

A major technique used in privacy preserving data publishing is *group based anonymization*, whereby records in the given relation are partitioned into groups and each group must ensure some property such as diversity so as to satisfy the privacy requirement while maintaining sufficient data utility. There are many privacy models associated with group based anonymization such as k-anonymity [Sweeney 2002; Nergiz and Clifton 2007], l-diversity [Machanavajjhala et al. 2006], t-closeness [Li and Li 2007], (k, e)-anonymity [Zhang et al. 2007], Injector [Li and Li 2008] and m-confidentiality [Wong et al. 2007]. It *seems* that this technique is sound for privacy preserving data publishing. However, when examined more carefully, they suffer from one fundamental privacy violation problem, which is overlooked in the past. The main cause of this problem is that the *utility* that is maintained in the anonymized table can help the adversary to breach individual privacy.

In the literature, background knowledge [Machanavajjhala et al. 2006; Kifer and Gehrke 2006; Martin et al. 2007; Wong et al. 2007; Li and Li 2008] such as the rarity of a disease among a certain ethnic group or the pattern of age or gender for a disease can often be used by the adversary. In this paper, we show that such knowledge can be *mined* from the *published data* or the *anonymized data*. In fact, one of the purposes of data publishing is for data mining which is mainly about the discovery of patterns from the published data.

Let us illustrate the problem with an example. Suppose a table T is to be anonymized for publication. Table T has two kinds of attributes, the quasi-identifier (QI) attributes and the sensitive attribute.

The QI attributes can be used as an identifier in the table. [Sweeney 2002] points out that in the United States, most individuals can be uniquely identified by QI attributes, namely birthdate, zipcode and gender. Information about the QI attributes can often be obtained from some external tables such as a voter registration list. An example of a voter registration list is shown in Table II. In this table, there are three attributes, namely Name, Nationality and Zipcode. Assume that the table contains information about a set of individuals, and the information about each such individual is contained in exactly one tuple. We also say that the individual *owns* the tuple. Assume attributes Nationality and Zipcode are the two QI attributes. Then the adversary can uniquely identify the tuple for Alex with the attributes Nationality and Zipcode.

The sensitive attribute contains some sensitive values. For example, if T is Table I, the sensitive attribute is "Disease", which contains sensitive values such as Heart Disease and HIV. Assume that each tuple in the table is owned by an individual and each individual owns at most one tuple.

After the anonymization, we publish the anonymized dataset T^*. T^* consists of a set of *QI-groups*, where each QI-group is a set of tuples linked with a multi-set of sensitive values. Depending on the anonymization mechanism, each QI-group may correspond to either a set of quasi-identifer (QI) values or a single generalized QI value. An attribute GID is added for the ID of the QI-group. We shall refer to a QI-group by its GID. Such *group-based anonymization* is commonly adopted in the literature of data publishing [Aggarwal et al. 2005; LeFevre et al. 2005; Xiao and Tao 2006; Wong et al. 2007; Li and Li 2008; 2007] (including k-anonymity, l-diversity, t-closeness and a vast number of other privacy models). The linkage between individual records and the sensitive attribute in each QI-group must be broken. One way to break the linkage is *bucketization*, forming two tables, called 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^*.

For illustration, a simplified setting of the l-diversity model [Machanavajjhala et al. 2006] is used as a privacy requirement for published data T^*. This simplified setting has been adopted in a lot of followup papers [Xiao and Tao 2006; Wong et al. 2007; Xiao and Tao 2007; Li et al. 2009] due to its lucid illustration of the concept behind l-diversity. A QI-group is said to be *l-diverse* or satisfy *l-diversity* if in the QI-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 QI-groups in it are l-diverse. Table I is anonymized to Table III by *bucketization*. The first group containing the first two tuples is given a GID QI_1. The GID for each group is shown in Table III. The intention is that each individual cannot be linked to a disease with a probability of more than 0.5. However, *does this table protect individual*

Name	Nationality	Zipcode	Disease
Alex	Malaysian	45501	Heart Disease
Bob	Japanese	45502	Flu
	Japanese	55503	Flu
	Japanese	55504	Stomach Virus
	Chinese	66601	HIV
	Japanese	66601	Diabetes
	Indian	77701	Flu
	Singaporean	77701	Diabetes
	...	...	...

Table I. An example

Name	Nationality	Zipcode
Alex	Malaysian	45501
Bob	Japanese	45502
Chris	Japanese	55503
David	Japanese	55504
Emily	Chinese	66601
Fred	Japanese	66601
Gary	Indian	77701
Henry	Singaporean	77701
...	...	...

Table II. Voter registration list

Nationality	Zipcode	GID
Malaysian	45501	QI_1
Japanese	45502	QI_1
Japanese	55503	QI_2
Japanese	55504	QI_2
Chinese	66601	QI_3
Japanese	66601	QI_3
Indian	77701	QI_4
Singaporean	77701	QI_4
...	...	...

(a) QI Table

GID	Disease
QI_1	Heart Disease
QI_1	Flu
QI_2	Flu
QI_2	Stomach Virus
QI_3	HIV
QI_3	Diabetes
QI_4	Flu
QI_4	Diabetes
...	...

(b) Sensitive table

Table III. A 2-diverse dataset anonymized from Table I by bucketization

privacy sufficiently?

Let us examine the QI-group QI_1 as shown in Table III. In QI_1, Heart Disease and Flu are values of the sensitive attribute Disease. It *seems* that each of the two individuals, Alex and Bob, in this group has a 50% chance of linking to Heart Disease (Flu). The reason why the chance is interpreted as 50% is that the analysis is based on this group *locally* without any additional information.

However, from the *entire published table* containing *multiple* groups, the adversary may discover some interesting patterns *globally*. For example, suppose the published table consists of many QI-groups like QI_2 with all Japanese with no occurrence of Heart Disease. At the same time, there are many QI-groups like QI_3 containing some Japanese without Heart Disease. The pattern that Japanese rarely suffer from Heart Disease can be uncovered. Note that it is very likely that such an anonymized data is published by conventional anonymization methods, given the fact that Heart Disease occurs rarely among Japanese. With the pattern uncovered, the adversary can say that Bob, being a Japanese, has less chance of having Heart Disease. S/he can deduce that Alex, being a Malaysian, has a higher chance of having Heart Disease. The intended 50% threshold is thus violated.

1.1 Foreground Knowledge Attack

The anonymized data can be seen as an *imprecise* or *uncertain data* [Burdick et al. 2005; Burdick et al. 2007], and an adversary can uncover interesting patterns since the published data must maintain high data utility [Xiao and Tao 2006; Zhang et al. 2007; Wong et al. 2007]. We call the uncovered patterns the *foreground knowledge* (which is *implicitly* inside the table) in contrast to the *background knowledge*, studied by existing works [Machanavajjhala et al. 2006; Li and Li 2007; Zhang et al. 2007; Wong et al. 2007], which requires much adversary *effort* to obtain from somewhere outside the table. Since it is easy to obtain the foreground knowledge from the anonymized dataset, most existing works suffer from privacy breaches.

In Table III, there are only two *local possible worlds* for assigning the disease values to the two individuals in QI_1: (1) w_1 : Alex is linked to Heart Disease and Bob is linked to Flu and (2) w_2 : Alex is linked to Flu and Bob is linked to Heart Disease. To construct a probability distribution over the domain of the real world, a simplest definition is based on the assumption that *all the possible worlds are equally likely*, or *each world has the same probability.*

If we publish a group QI_1 alone, the random world assumption is a good principle in the absence of other information. However, when several groups are published together as typically the case, the groups with Japanese contribute to a statement that their members are not likely linked to Heart Disease. This statement means that the *probability* (or *weight*) of the possible world w_1 is much greater than that of w_2.

Most previous privacy works such as l-diversity [Machanavajjhala et al. 2006], t-closeness [Li and Li 2007], (k, e)-anonymity [Zhang et al. 2007] and m-confidentiality [Wong et al. 2007] adopt the random world assumption *locally.* In this paper, the source of attack of the adversary is to apply the more complete model of the *weighted possible worlds.* We call this kind of attack *foreground knowledge attack.*

1.2 Generalization-Based Anonymization

The above example shows that the foreground knowledge attack appears in a table generated by bucketization. Although the example is based on bucketization anonymization, the same issue arises with a *generalization* based method [Machanavajjhala et al. 2006; Wong et al. 2006; Li and Li 2007]. 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 a QI-group. For example, if the QI values of QI_4 in Table III are generalized to { Asian, 777** }, and Gary and Henry are the only Asians with a Zipcode of 777** in the external table, Table II, then the adversary can determine that they are the owners of the two tuples in QI_4. Hence, the exact QI values of {Indian, 77701} and {Singaporean, 77701} will be disclosed for QI_4, and the data for QI_4 becomes as detailed as that

Nationality	Zipcode	Disease
Asian	455**	Heart Disease
Asian	455**	Flu
Asian	555**	Flu
Asian	555**	Stomach Virus
Asian	666**	HIV
Asian	666**	Diabetes
Asian	777**	Flu
Asian	777**	Diabetes
...	...	...

Table IV. A 2-diverse dataset anonymized from Table I by global recoding

Nationality	Zipcode	Disease
{Malaysian, Japanese}	45501-45502	Heart Disease
{Malaysian, Japanese}	45501-45502	Flu
Japanese	55503-55504	Flu
Japanese	55503-55504	Stomach Virus
{Chinese, Japanese}	66601	HIV
{Chinese, Japanese}	66601	Diabetes
{Indian, Singaporean}	77701	Flu
{Indian, Singaporean}	77701	Diabetes
...	...	...

Table V. A 2-diverse dataset anonymized from Table I by local recoding

from bucketization. [Sweeney 2002] points out that in the United States, 87% of individuals can be uniquely identified by QI attributes, namely birthdate, zipcode and gender, and therefore such information disclosure is not uncommon. Once such details are determined, the adversary can determine the revised probabilities. It is worth mentioning that it is more difficult for the adversary to perform the foreground knowledge attack on the table generated by generalization compared with the table generated by bucketization. This is because in some cases, it is more likely that a given individual such as Alex can be mapped to *multiple* QI-groups in the table generated by generalization (instead of a single QI-group in the table generated by bucketization).

In the literature, there are two kinds of generalization techniques, namely *global recoding* and *local recoding*. We want to emphasize that the foreground knowledge attack occurs in the table generated by either global recoding or local recoding. Under global recoding, all occurrences of a *single* attribute value are recoded to the same value. Table IV shows a 2-diverse dataset anonymized from Table I by global recoding. With the external table (Table II), the adversary can figure out the original values of each tuple in each group in Table IV. For example, the QI values for the first two records must be {Malaysian, 45501}, and {Japanese, 45502}. Thus, foreground knowledge attack is valid. Under local recoding, occurrences of the same value of an attribute may be recoded to different values. Table V shows a 2-diverse dataset anonymized from Table I by local recoding. Note that notation "{Malaysian, Japanese}" means that this value is either Malaysian or Japanese.

Similar to global recoding, once the adversary can figure out the original QI value of each tuple, the same principle for foreground knowledge attack can be applied. Note that since global recoding often incurs a higher *information loss* than local recoding, it is more resistent to foreground knowledge attack. The major focus of this paper is to study a new form of attack. For ease of illustration, we will show how the attack can be successful for the case of bucketization.

1.3 Contributions

Our contributions can be summarized as follows. Firstly, we define and study data anonymization issues in data publication with the consideration of foreground knowledge attack. The concept of the foreground knowledge attack was derived independently of [Kifer 2009], and the paper by [Kifer 2009] and this paper were written concurrently. Secondly, we show how an adversary can breach privacy by computing the probability that an individual is linked to a sensitive value by using foreground knowledge.

Finally, we have conducted experiments to show how the adversary can succeed in foreground knowledge in different anonymization schemes, including *Anatomy* [Xiao and Tao 2006], *MASK* [Wong et al. 2007], and *Injector* [Li and Li 2008], and also under different privacy requirements such as t*-closeness* [Li and Li 2007].

We emphasize that, similar to l-diversity, most group-based anonymization algorithms [Xiao and Tao 2006; Wong et al. 2007; Li and Li 2008; 2007] by bucketization also suffer from possible *privacy breaches* due to the *utility* of the published table. We believe that this work is significant in pointing out this overlooked issue, and that follow-up works would need to deter foreground knowledge attack.

The rest of the paper is organized as follows. Section 2 formulates the problem. Section 3 describes how the adversary can breach individual privacy with the foreground knowledge obtained from the anonymized data. Section 4 shows how the adversary can obtain the foreground knowledge from the anonymized data. An empirical study is reported in Section 5. Section 6 reviews the related work. The paper is concluded in Section 7.

2. 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. 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, namely $A_1, A_2, ..., A_q$, 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.

Let P be a partition of table T. We give a unique ID called GID to this partition P and append an additional attribute called GID to this partition where each tuple in P has the same GID value. Existing group-based anonymization defines a function β on P to form a QI-group such that the linkage between the QI attributes and the sensitive attribute in the QI-group is lost. There are two ways in the literature for this task. One is *generalization* by generalizing all QI values to the same value. The other is *bucketization* by forming two tables, called the *QI table* and the *sensitive table*, where P is projected on all QI attributes and attribute GID to form the QI table, and on the sensitive attribute and attribute GID to form the sensitive table.

$p()$	Heart Disease	Not Heart Disease
Malaysian	0.1	0.9
Japanese	0.003	0.997
Chinese	0.05	0.95

Table VI. A global distribution of attribute "Nationality" for our motivating example

A table T is *anonymized* to a dataset T^* if T^* is formed by first partitioning T into a number of partitions, then forming a QI-group from each partition by β and finally inserting each QI-group into T^*. For example, Table I is anonymized to Table III by bucketization.

We assume that there is a mapping which maps each tuple in T to a QI-group in T^*. For example, the first tuple t_1 in Table I is mapped to QI-group QI_1.

In the following, we focus on discussing the anonymized table generated by bucketization.

In the literature [Xiao and Tao 2006; Wong et al. 2007; Li and Li 2008; 2007], it is assumed that the knowledge of the adversary includes (1) the published dataset T^*, (2) the QI value of a target individual, and (3) an external table T^e such as the voter registration list that helps to map QIs to individuals [Sweeney 2002; LeFevre et al. 2005]. We also follow these assumptions in our analysis.

The aim of privacy preserving data publishing is to deter any attack from the adversary on linking an individual to a certain sensitive value. Specifically, the data publisher would try to limit the probability that such a linkage can be established. Let us consider an arbitrary sensitive value x for the analysis. We denote any value in X which is not x by $\overline{x}$.

In this paper, we consider that an adversary can obtain additional information from the published dataset T^* in the form of *global distribution*, which can lead to individual privacy breach. In the example in Section 1, we can mine from the published table that the chance of Japanese suffering from Heart Disease is low compared with Malaysian. This pattern is from the *global distribution* for the attribute set {"Nationality"}.

Each possible value in attribute "Nationality" is called a *signature*. There are three possible signatures in our example: "Japanese", "Malaysian" and "Chinese". In general, there are other attribute sets, such as {"Gender", "Nationality"}, with their correspondence global distributions.

Definition 1 Signature. *Let T^* be the published dataset. Given a QI attribute set $\mathcal{A}$ with r attributes $A_1, ..., A_r$. A* signature *s of $\mathcal{A}$ is a set of attribute-value pairs $(A_1, v_1), ..., (A_r, v_r)$ 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 $i = 1, 2, ..., r$.*

For example, a signature s can be {("Nationality", "Malaysian"), ("Gender", "Male")} if the attribute set $\mathcal{A}$ is {"Nationality", "Gender"}. For convenience, we often drop the attribute names in a signature, and thus we refer to {"Malaysian", "Male"} instead of {("Nationality", "Malaysian"), ("Gender", "Male")}. The first tuple t_1 in Table III(a) matches {"Malaysian"} but the second tuple does not.

Definition 2 Sample Space $\Omega(s, x)$. *The* sample space *$\Omega(s, x)$ for a signa-*

ture s and sensitive value x is a set of two elements: (1) s is linked to x and (2) s is not linked to x (or s is linked to $\overline{x}$).

Consider an arbitrary sensitive value "Heart Disease". Assume that Table VI shows the *global distribution* of attribute set {"Nationality"}, which consists of the probabilities that a Japanese, a Malaysian or a Chinese is linked to Heart Disease. Each such probability in the table is called a *global probability.* The global probabilities are based on sample spaces for different signatures.

Definition 3 Global Distribution. *Given an attribute set $\mathcal{A}$, the* global 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 probability that a tuple matching signature s is linked to x given the published dataset T^*.*

For example, if G contains ("Japanese":"Heart Disease", 0.003) and ("Malaysian":"Heart Disease", 0.1), then the probability that a Japanese patient is linked to Heart Disease is equal to 0.003 while that of a Malaysian patient is 0.1.

The global distribution G derived from the published dataset T^* is called the *foreground knowledge*. We will describe how the adversary derives G from the published table.

Problem 1 Foreground Knowledge. *Given any arbitrary attribute set $\mathcal{A}$, we want to find the global distribution G of $\mathcal{A}$ from the published dataset T^*.*

From Section 1, we show that with the global distribution G of attribute set {"Nationality"}, we can deduce that the chance of Alex, a Malaysian, suffering from Heart Disease is high. Let t be Alex and x be Heart Disease. The chance can be formulated by $p(t : x)$, the probability that t is linked to x given G.

Problem 2 Privacy Breach. *Given a published dataset T^*, for any individual t, and any sensitive value x, we want to determine whether the probability that t is linked to x denoted by $p(t : x)$ is greater than $1/r$. Individual t is said to suffer from privacy breaches if the probability is greater than $1/r$.*

We should point out here that knowing that a Japanese has a 0.003 probability of Heart Disease does not lead to a conclusion that a Japanese t in a QI-group QI in T^* also has a 0.003 chance of Heart Disease. A very simple counter example is when the QI-group does not contain a record with Heart Disease. In this example, the probability is obviously 0. However, how to derive such a probability in general is non-trivial and will be the main focus in the next two sections. Another point to clarify is that Problems 1 and 2 need to be solved as one problem. The global distribution G is dependent on the probabilities $p(t : x)$ and vice versa. This is because T^* consists of QI-groups in which the probabilities $p(t : x)$ from each group contribute to G, and conversely, G in turn affects $p(t : x)$.

In this paper, we study Problems 1 and 2. In Section 3, we will first describe how we solve Problem 2 assuming that we are given the foreground knowledge. Then, in Section 4, we will describe how we can mine the foreground knowledge from the published dataset T^* for Problem 1. We shall show that the two problems are intertwined and they are solved as one problem.

$p()$	x	$\overline{x}$
s_1	f_1	$\overline{f_1}$
s_2	f_2	$\overline{f_2}$
$\vdots$	$\vdots$	$\vdots$

Table VII. Global distribution

3. FINDING PRIVACY BREACHES

In this section, we derive a formula to calculate the probability that an individual t is linked to a sensitive value x. The major idea is similar to what we discussed in Section 1. Consider an individual t (e.g., Alex) in a QI-group QI_k of the published table T^*. Suppose QI_k contains N individuals and N values in the sensitive attribute. We can enumerate all possible assignments (or possible worlds) between a set of N individuals and a set of N values for QI_k. In Section 1, since N is equal to 2, there are two possible worlds for QI_k. Assume that G is the global distribution of a certain attribute set $\mathcal{A}$ (e.g., {Nationality}). According to G, different possible worlds have different probabilities. Finally, according to the probabilities of these possible worlds, we calculate the probability that t is linked to a sensitive value x.

Formally, we derive the formula as follows. We assume that the attack is based on the linkage of an attribute set $\mathcal{A}$ to a sensitive value x. We denote by $\bar{x}$ any value not equal to x. In this section, we assume that the global distributions G for $\mathcal{A}$ and x have been determined and we show how an adversary can use G to find privacy breaches. How the global distributions can be derived is explained in Section 4.

Consider the motivating example in Section 1. In Table I, attribute “Nationality” contains “Malaysian”, “Japanese” and “Chinese”. Each value in “Nationality” is called a *signature* of attribute “Nationality”. Thus, there are three possible signatures of a *single* attribute, namely “Nationality”. In general, there are signatures of a *attribute set* containing multiple attributes. An example of an attribute set can be {Nationality, Zipcode}. (“Malaysian”, 5501) is a signature of {Nationality, Zipcode}.

Formally, suppose there are m possible signatures for attribute set $\mathcal{A}$, namely $s_1, s_2, ..., s_m$. The global distribution G of $\mathcal{A}$ is shown in Table VII. To simplify our presentation, the probability that s_i is linked to x ($\overline{x}$), $p(s_i : x)$ ($p(s_i : \overline{x})$), is denoted by f_i ($\overline{f}_i$).

Given G, the formula for $p(t : x)$, the probability that a tuple t is linked to sensitive value x, is derived here. Suppose t belongs to QI-group QI_k. For the ease of reference, let us summarize the notations that we use in Table VIII.

Definition 4 Sample Space $\Omega(QI_k)$. *Given a QI-group QI_k in the form of a set of tuples and a multi-set of sensitive values V, we define a sample space $\Omega(QI_k)$ for QI_k to be the set of all possible one-to-one assignments of the sensitive values V to the tuples in QI_k.*

Definition 5 Possible World. *Consider a QI-group QI_k with N tuples, namely $t_1, t_2, ..., t_N$, with sensitive values $\gamma_1, \gamma_2, ...\gamma_N$, where γ_i is either x or $\overline{x}$ for $i = 1, 2, ..., N$. A possible world w for QI_k 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 QI_k.*

QI_k	a QI-group in the anonymized dataset
$\mathcal{A}$	set of attributes e.g. {"Nationality", "Gender"}
$t_1, ..., t_N$	tuples in an A-group
$s_1, ..., s_m$	signatures for $\mathcal{A}$, e.g.{"Malaysian", "Male"} multiple tuples t_j's can map to the same s_i
x	a sensitive value
$\bar{x}$	any value not equal to x
$p(t_j : x)$	probability that tuple t_j is linked to value x
$p(s_i : x)$	probability that signature s_i is linked to x
f_i	a simplified notation for $p(s_i : x)$
$\bar{f}_i$	$1 - f_i$
w	a possible world: an assignment of the tuples in QI-group QI_k to the sensitive values x and $\bar{x}$
$\mathcal{W}_k$	set of all possible worlds w for QI_k
$\mathcal{W}_k^{(t_j:x)}$	set of all possible worlds w in $\mathcal{W}_k$ in which t_j is assigned value x.
$p(w)$	probability that w occurs given the anonymized dataset and based on $\mathcal{A}$
$p(w\|QI_k)$	conditional probability that w occurs given QI-group QI_k
$p_{j,w}$	the probability that t_j is linked to a value in the sensitive attribute as specified in w
$\mathcal{QI}_{s_i}$	set of QI-groups containing tuples matching s_i
$QI_k(s_i)$	the set of tuples in QI_k matching s_i.
$c_k(s_i : x)$	the expected number of tuples which match s_i and are linked to x in the QI-group QI_k

Table VIII. Notations

Thus, each element in $\Omega(QI_k)$ is a possible world.

DEFINITION 6 PRIMITIVE EVENTS, PROJECTED EVENTS. *A mapping $t : \gamma$ from an individual or tuple t to a sensitive value γ (x or $\bar{x}$) is called a* primitive event. *Suppose t matches signature s. Let us call an event for the corresponding signature, "$s : \gamma$", a* projected event *for t.*

Hence, a primitive event $(t : x)$ is an event defined by a subset of $\Omega(Q)$ consisting of the possible worlds where t is assigned to x. The probability of this event, $p(t : x)$, is a probability of interest for the adversary. A projected event is a corresponding event $(s : x)$ where $p(s : x)$ appears in the global distribution G.

Suppose that we are given a QI-group QI_k with a set of tuples and a multi-set of sensitive values. For each possible world w for QI_k, according to the global distribution G based on attribute set $\mathcal{A}$, we compute the probability $p(w)$ that w occurs. The sample space for $p(w)$ consists of all the possible assignments of x or $\overline{x}$ to a set of N tuples with the same signatures as those in QI_k.

EXAMPLE 1. In our motivating example, consider the first QI-group QI_1 and x is "Heart Disease". There are two possible world for QI_1: (1) w_1: Alex is linked to x and Bob is linked to $\overline{x}$ and (2) w_2: Alex is linked to $\overline{x}$ and Bob is linked to x. According to the global distribution of attribute "Nationality" as shown in

Table VI, we would like to compute the probability that w_1 occurs, denoted by $p(w_1)$, and the probability that w_2 occurs. Consider w_1. From Table VI, we know that the probability that a Malaysian is linked to x is 0.1 and the probability that a Japanese is linked to $\overline{x}$ is 0.997. Since Alex is Malaysian, the probability that he is linked to x is 0.1. Similarly, since Bob is a Japanese, the probability that he is linked to $\overline{x}$ is 0.997. Similar to [Machanavajjhala et al. 2006; Xiao and Tao 2006; Wong et al. 2007], we assume that the linkage of a sensitive value to an individual is independent of the linkage of a sensitive value to another individual. For example, whether a Malaysian suffers from Heart Disease is independent of whether a Japanese suffers from Heart Disease. We conclude that $p(w_1)$ is equal to $0.1 \times 0.997 = 0.0997$. Similarly, $p(w_2)$ is equal to $0.003 \times 0.9 = 0.0027$. □

We have just illustrated the major idea of computing $p(w)$ for a possible world w. In the following, we will give a formal derivation for computing $p(w)$.

Formally, suppose that in a possible world w for QI_k, tuple t_j is linked to γ, where γ is either x or $\overline{x}$. Let $p_{j,w}$ be the probability that t_j is linked to γ.

Note that we assume that the linkage of a sensitive value to an individual is independent of the linkage of a sensitive value to another individual. For a possible world w for QI_k, the probability that w occurs is the product of the probabilities of the corresponding projected events for the tuples $t_1, ...t_N$ in QI_k.

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

Suppose t_j matches signature s_i. If t_j is linked to x in w, then $p_{j,w} = f_i$. Otherwise, $p_{j,w} = \overline{f_i}$.

$p(w)$ corresponds to the *weight* of w, which we mentioned in the introduction.

We have just given a formal derivation of computing $p(w)$. Note that $p(w)$ considers the likelihood that w occurs in the *entire* anonymized table but it does not consider any *particular* QI-group. In order to consider a particular QI-group QI_k, in the following, we first illustrate how we derive a formula of computing the probability that w occurs when we consider QI_k only, denoted by $p(w|QI_k)$. After that, we give a formal derivation for this formula.

Example 2. From Example 1, we know that $p(w_1) = 0.0997$ and $p(w_2) = 0.0027$. Consider QI_1. We know that there are only two possible worlds, namely w_1 and w_2, for QI_1. The probability that QI_1 occurs given the anonymized table T^* is equal to $0.0997 + 0.0027 = 0.1024$. Thus, the probability that w_1 occurs given QI_1, denoted by $p(w_1|QI_1)$, is

$$\frac{0.0997}{0.1024} = 0.9736.$$

Similarly, $p(w_2|QI_1)$ is equal to

$$\frac{0.0027}{0.1024} = 0.0264.$$

It is easy to verify that $p(w_1|QI_1) + p(w_2|QI_1) = 1$. □

Formally, the probability of QI_k given T^* is the sum of the probabilities of all the possible worlds consistent with T^* for QI_k. Let the set of these worlds be $\mathcal{W}_k$.

For $w \in \mathcal{W}_k$, we have

$$p(w|QI_k) = \frac{p(w)}{\sum_{w' \in \mathcal{W}_k} p(w')} \tag{2}$$

It is easy to verify that $\sum_{w \in \mathcal{W}_k} p(w|QI_k) = 1$.

Our objective is to find the probability that an individual t_j in QI_k is linked to a sensitive value x, denoted by $p(t_j : x)$.

Example 3. Consider that we are interested in knowing the probability that Alex in QI_1 is linked to Heart Disease (i.e., x). There are two possible worlds for QI_1 and there is only one possible world that Alex is linked to x (i.e., w_1). Let t_j be Alex. The probability that Alex is linked to x, denoted by $p(t_j : x)$, is equal to $p(w_1|QI_1) = 0.9736$. □

Formally, $p(t_j : x)$ is given by the sum of the conditional probabilities $p(w|QI_k)$ of all the possible worlds w where t_j is linked to x.

$$p(t_j : x) = \sum_{w \in \mathcal{W}_k^{(t_j:x)}} p(w|QI_k) \tag{3}$$

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

One can verify that $p(t_j : x) + p(t_j : \overline{x}) = 1$.

Example 4. Consider a QI-group QI_k in a published table T^*. Suppose there are four tuples, t_1, t_2, t_3 and t_4, and four sensitive values, $x, x, \overline{x}$ and $\overline{x}$ in QI_k. Suppose the published table T^* satisfies 2-diversity.

Consider the global distribution G based on a certain QI attribute set $\mathcal{A}$ which contains two possible signatures s_1 and s_2 as shown in Table IX(a).

Suppose t_1, t_2, t_3 and t_4 match signatures s_1, s_1, s_2 and s_2, respectively. There are six possible worlds w as shown in Table IX(b). For example, the first row is the possible world w_1 with mapping $\{t_1 : x,\ t_2 : x,\ t_3 : \overline{x},\ t_4 : \overline{x}\}$. The table also shows the probability $p(w)$ of the possible worlds. Take the first possible world w_1 for illustration. From the global distribution in Table IX(a), $p(s_1 : x) = 0.5$ and $p(s_2 : \overline{x}) = 0.8$. Hence, $p(w_1) = 0.5 \times 0.5 \times 0.8 \times 0.8 = 0.16$. The sum of probabilities $p(w)$ of all possible worlds from Table IX(b) is equal to 0.16 + 0.04 + 0.04 + 0.04 + 0.04 + 0.01 = 0.33. Consider w_1 again. Since $p(w_1) = 0.16$, $p(w_1|QI_k) = 0.16/0.33 = 0.48$.

Suppose the adversary is interested in the probability that t_1 is linked to x. We obtain $p(t_1 : x)$ as follows. w_1, w_2 and w_3, as shown in Table IX(b), contain "$t_1 : x$". Thus, $p(t_1 : x)$ is equal to the sum of the probabilities $p(w_1|QI_k), p(w_2|QI_k)$ and $p(w_3|QI_k)$. $p(t_1 : x) = 0.48 + 0.12 + 0.12 = 0.72$ which is greater than 0.5, the intended upper bound for 2-diversity that an individual is linked to a sensitive value. □

Let $|QI_k|$ be the size of the QI-group containing t_j and $|\mathcal{W}_k|$ be the number of possible worlds in a QI-group QI_k. We will generate $|\mathcal{W}_k|$ possible worlds. For each possible world, we calculate $p(w)$ and $p(w|QI_k)$ in $O(|QI_k|)$ time. Thus, the time complexity is $O(|QI_k| \cdot |\mathcal{W}_k|)$.

$p()$	x	$\overline{x}$
s_1	0.5	0.5
s_2	0.2	0.8

(a) Global distribution

w	t_1	t_2	t_3	t_4	$p(w)$	$p(w\|QI_k)$
	(s_1)	(s_1)	(s_2)	(s_2)		
w_1	x	x	$\overline{x}$	$\overline{x}$	$0.5 \times 0.5 \times 0.8 \times 0.8 = 0.16$	$0.16/0.33 = 0.48$
w_2	x	$\overline{x}$	x	$\overline{x}$	$0.5 \times 0.5 \times 0.2 \times 0.8 = 0.04$	$0.04/0.33 = 0.12$
w_3	x	$\overline{x}$	$\overline{x}$	x	$0.5 \times 0.5 \times 0.8 \times 0.2 = 0.04$	$0.04/0.33 = 0.12$
w_4	$\overline{x}$	x	x	$\overline{x}$	$0.5 \times 0.5 \times 0.2 \times 0.8 = 0.04$	$0.04/0.33 = 0.12$
w_5	$\overline{x}$	x	$\overline{x}$	x	$0.5 \times 0.5 \times 0.8 \times 0.2 = 0.04$	$0.04/0.33 = 0.12$
w_6	$\overline{x}$	$\overline{x}$	x	x	$0.5 \times 0.5 \times 0.2 \times 0.2 = 0.01$	$0.01/0.33 = 0.03$

(b) $p(w)$ and $p(w|QI_k)$

Table IX. An example illustrating the computation of $p(t_j : x)$

The time complexity depends on two factors. One is $|QI_k|$ and another is $|\mathcal{W}_k|$. (1) $|QI_k|$ is bounded by the greatest size of the QI-group which depends on the anonymization techniques. For example, $|QI_k|$ is equal to l or $l+1$ for algorithm Anatomy [Xiao and Tao 2006] which restricts that each QI-group contains either l or $l+1$ tuples. In our experiment, $|QI_k|$ is at most 23 for algorithm MASK [Wong et al. 2007] where $l = 2$. (2) $|\mathcal{W}_k|$ is equal to C^N_n where n is the number of tuples with x in this QI-group of size N and C^N_n denotes the total number of possible ways of choosing n objects from N objects. Note that $|\mathcal{W}_k|$ is typically small because n is usually equal to a small number. For algorithm Anatomy [Xiao and Tao 2006], as we discussed, N (which corresponds to the size of the QI-group) is either l or $l+1$. In this algorithm, since x appears in the QI-group at most once, n (which corresponds to the number of tuples with x in this QI-group) is at most 1. Thus, for each possible x, $|\mathcal{W}_k|$ is at most $l+1$. For Algorithm MASK [Wong et al. 2007], in our experiment with $l = 2$, the greatest frequency of x in a QI-group is 8. The size of this QI-group is 23. $|\mathcal{W}_k|$ is equal to $C^{23}_8 = 490,314$. When $l = 10$, the greatest possible value of $|\mathcal{W}_k|$ is 140,364,532. These values are small compared with the excessive number of possible worlds studied in uncertain data [Imielinski and Jr. 1984; Burdick et al. 2005; Burdick et al. 2007; Antova et al. 2007; Cheng et al. 2008] (e.g., 10^{10^6} in [Antova et al. 2007])). In the experimental setups in existing works [Machanavajjhala et al. 2006; Xiao and Tao 2006; Li and Li 2007; Wong et al. 2007; Li and Li 2008], $l \leq 10$. In other words, $\mathcal{W}_k$ can be generated within a reasonable time.

4. MINING FOREGROUND KNOWLEDGE

We first describe how we find the global distribution G of a certain attribute set $\mathcal{A}$ from the anonymized data in Section 4.1. Next, we introduce a pruning strategy to prune our search space of attribute sets in Section 4.2. Finally, we describe the algorithm for finding the global distribution of multiple attribute sets and discuss its complexity in Section 4.3.

4.1 Foreground Knowledge

In the previous section, we assume that the values of f_i are given. Here we consider how to derive f_i from the published table T^*. We will develop m equations involving the m variables f_i, $1 \leq i \leq m$.

Let the set of QI-groups in T^* be $QI_1, ..., QI_u$. Let $QI_k(s_i)$ be the set of tuples in QI_k matching signature s_i. For example, in Table III, let $s_i =${"Malaysian"}. Then, $QI_1(s_i)$ contains only the first tuple.

Let $\mathcal{QI}_{s_i}$ be a set of QI-groups containing tuples which match s_i. That is, $\mathcal{QI}_{s_i} = \{QI_k | QI_k(s_i) \neq \emptyset\}$.

f_i is equal to the expected number of tuples which match s_i and are linked to x in T^* divided by the number of tuples which match s_i in T^*. Let $c_k(s_i : x)$ be the expected number of tuples which match s_i and are linked to x in the QI-group QI_k. Then, we can express f_i as follows.

$$f_i = \frac{\sum_{QI_k \in \mathcal{QI}_{s_i}} c_k(s_i : x)}{\sum_{QI_k \in \mathcal{QI}_{s_i}} |QI_k(s_i)|} \tag{4}$$

The denominator is simply equal to the number of occurrences of s_i in T^* and which can be easily found from the dataset. Let us consider the term $c_k(s_i : x)$ in the numerator.

Without additional knowledge to govern otherwise, we assume that the event that a tuple matching s_i in QI_k is linked to x is independent of the event that another tuple also matching s_i in QI_k is linked to x. Then we have the following.

$$c_k(s_i : x) = |QI_k(s_i)| \times p(t_j : x) \tag{5}$$

where t_j is any tuple in QI_k matching s_i. Note that any t_j in QI_k matching s_i can be used here since all such $p(t_j : x)$ values are equal. Substitute Equations (3) and (2) into the above equation, we get

$$c_k(s_i : x) = |QI_k(s_i)| \times \sum_{w \in \mathcal{W}_k^{(t_j : x)}} \frac{p(w)}{\sum_{w' \in \mathcal{W}_k} p(w')} \tag{6}$$

Hence, $c_k(s_i : x)$ is expressed in terms of probabilities $p(w)$ which in turn are expressed in the m variables f_i (see Equation (1) where $p_{j,w}$ is equal to f_i or $\overline{f}_i$). Here note that $\overline{f_i} = 1 - f_i$.

There are m equations of the form of Equation (4) for the expression of f_i, $1 \leq i \leq m$. These equations involve m variables, f_i. This is a classical problem of a system of simultaneous non-linear equations, which occurs in many applications. It can be solved by conventional methods such as Newton's method and Bairstow's iteration. Since Newton's method [Chapra and Canale 2002] has been known to be effective and feasible, we choose this method for our study in this paper.

Example 5. *Given a table T containing six tuples, $t_1, t_2, ..., t_6$, as shown in Table X. If the objective of the privacy requirement is 2-diversity, T does not satisfy 2-diversity. Thus, an anonymized dataset T^* in Table XI with three QI-groups, QI_1, QI_2 and QI_3, is published (for each sensitive value x and each QI-group, the fraction of tuples with x is at most 0.5). Note that QI_3 satisfies 2-diversity. This is because since $\overline{x}$ corresponds to a value not equal to x, in QI_3, the first $\overline{x}$ corresponds to a value y and the second $\overline{x}$ corresponds to another value z.*

$\mathcal{A}$	...	X
s_1	...	x
s_1	...	x
s_1	...	$\overline{x}$
s_2	...	$\overline{x}$
s_2	...	$\overline{x}$
s_2	...	$\overline{x}$

Table X. A raw table

t	$\mathcal{A}$	...	GID
t_1	s_1	...	QI_1
t_2	s_2	...	QI_1
t_3	s_1	...	QI_2
t_4	s_1	...	QI_2
t_5	s_2	...	QI_3
t_6	s_2	...	QI_3

(a) QI Table

GID	X
QI_1	x
QI_1	$\overline{x}$
QI_2	x
QI_2	$\overline{x}$
QI_3	$\overline{x}$
QI_3	$\overline{x}$

(b) Sensitive Table

Table XI. An example illustrating the computation of the global distribution

Consider the global distribution of attribute set $\mathcal{A}$. There are two possible signatures based on $\mathcal{A}$, namely s_1 and s_2. Thus, we have two equations with two variables, namely f_1 and f_2, the probabilities in the global distribution G of $\mathcal{A}$ as shown in Table VII.

Consider f_1. Since only QI-groups QI_1 and QI_2 contain the tuples matching s_1, $\mathcal{QI}_{s_1} = \{QI_1, QI_2\}$.

$$f_1 = [\textstyle\sum_{QI_k \in \mathcal{QI}_{s_1}} c_k(s_1 : x)]/[\textstyle\sum_{QI_k \in \mathcal{QI}_{s_1}} |QI_k(s_1)|]$$

QI_1 contains one tuple t_1 matching s_1 and QI_2 contains two tuples t_3, t_4 matching s_1, $|QI_1(s_1)| = 1$ and $|QI_2(s_1)| = 2$. Thus,

$$f_1 = [1 \times p(t_1 : x) + 2 \times p(t_3 : x)]/(1 + 2) \tag{7}$$

Consider QI_1. There are only two possible worlds, $w_1 = \{t_1 : x, t_2 : \overline{x}\}$ and $w_2 = \{t_1 : \overline{x}, t_2 : x\}$. Note that t_1 and t_2 match signatures s_1 and s_2, respectively. $p_{1,w_1} = f_1, p_{2,w_1} = \overline{f_2}, p_{1,w_2} = \overline{f_1}$ and $p_{2,w_2} = f_2$. Thus, $p(w_1) = p_{1,w_1} \times p_{2,w_1} = f_1 \times \overline{f_2}$ and $p(w_2) = p_{1,w_2} \times p_{2,w_2} = \overline{f_1} \times f_2$. We derive that

$$p(t_1 : x) = p(w_1|QI_1) = f_1\overline{f_2}/(f_1\overline{f_2} + \overline{f_1}f_2)$$

Similarly, consider QI_2. There are two possible worlds, $w_3 = \{t_3 : x, t_4 : \overline{x}\}$ and $w_4 = \{t_3 : \overline{x}, t_4 : x\}$. Similarly, $p(w_3) = f_1 \times \overline{f_1}$ and $p(w_4) = \overline{f_1} \times f_1$. We have

$$p(t_3 : x) = p(w_4|QI_2) = f_1\overline{f_1}/(f_1\overline{f_1} + \overline{f_1}f_1) = 1/2$$

From (7), we obtain

$$\begin{aligned} f_1 &= [f_1\overline{f_2}/(f_1\overline{f_2} + \overline{f_1}f_2) + 1]/3 \\ &= [f_1(1 - f_2)/(f_1(1 - f_2) + (1 - f_1)f_2) + 1]/3 \end{aligned}$$

Similarly, since QI_1 contains one tuple t_2 matching s_2 and QI_3 contains two tuples t_5, t_6 matching s_2,

$$\begin{aligned} f_2 &= [1 \times p(t_2 : x) + 2 \times p(t_5 : x)]/(1 + 2) \\ &= [\overline{f_1}f_2/(f_1\overline{f_2} + \overline{f_1}f_2) + 0]/3 \\ &= [(1 - f_1)f_2/(f_1(1 - f_2) + (1 - f_1)f_2)]/3 \end{aligned}$$

With the above two equations involving two variables, we adopt Newton's method to solve for these variables.

Finally, we obtain $f_1 = 0.666667$ *and* $f_2 = 0.000000$*. Thus, we derive* $\overline{f_1} = 0.333333$ *and* $\overline{f_2} = 1.000000$*.* □

4.2 Pruning Attribute Sets

The adversary may choose to attack with as many attribute sets as possible. Although there are many attribute sets in the anonymized data, it is not always true that the global distribution of each attribute set is *reliable* because if the global distribution derived is based on a small sample or a small set of tuples matching the same signature, the distribution is not accurate. For example, consider attribute set $\mathcal{A}$="Nationality" and the signature {"Malaysian"}. Suppose there are only a few Malaysians, says 10 Malaysians, in the published table T^*. Intuitively, 10 Malaysians cannot represent a meaningful global distribution. We will make use of the sample size studied in the literature of statistics to determine whether the distribution is reliable or not. The adversary can launch an attack only based on reliable distributions.

Based on studies in statistics [Toivonen 1996], we use the following theorem to determine the acceptable sample size (i.e., the size of the set which contains the tuples matching the same signature s). Let S be a random sample of tuples for a signature s, and p be the expected fraction of tuples in S with the sensitive value x. Let $\widetilde{p}$ be the observed fraction of tuples with the sensitive value x in the sample S. Then the following theorem applies.

THEOREM 1 SAMPLE SIZE [TOIVONEN 1996]. *Given an error parameter* $\epsilon \geq 0$ *and a confidence parameter* $\sigma \geq 0$*, if random sample* S *has size* $|S| \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\sigma}$*, the probability that* $|\widetilde{p} - p| > \epsilon$ *is at most* σ*.* □

In case the sample size is not large enough to satisfy the error bound, then uniform distribution will be assumed. The sample size satisfies the monotonicity property. Formally, without loss of generality, assume that there are u attributes, namely $A_1, ..., A_u$. Let $v_1 \in A_1, ..., v_u \in A_u$. Let $y(v_1, ..., v_i)$ be the number of tuples with attributes $(A_1, ..., A_i)$ equal to $(v_1, ..., v_i)$. Given a positive integer J, if $y(v_1, ..., v_i) < J$, then $y(v_1, ..., v_i, v_{i+1}) < J$. With the above monotonicity property, whenever we find that the sample size of $y(v_1, ..., v_i)$ is not large enough, we do not need to count the number of the tuples with values $v_1, ..., v_{i+1}$ because $y(v_1, ..., v_i, v_{i+1})$ is also not large enough. Thus, this can help to prune the search space.

4.3 Algorithm

In this section, we will describe how to compute the set $\mathcal{G}$ of all global distributions of multiple attribute sets with the use of the sample size just described. The steps are shown in Algorithm 1.

In the algorithm, Step 1 is to find all signatures with sufficient sample size for each attribute set $\mathcal{A}$. Similar to frequent pattern mining, this step is typically computed within a reasonable time. Let α be the time for this step. After we have determined the sample sizes, $\mathcal{G}$ is used to store the global distributions of all attribute sets each of which contains signatures with sufficient sample size.

Step 2 is to calculate the global distribution of $\mathcal{A}$ according to non-empty $\mathcal{S}_{\mathcal{A}}$ for each attribute set $\mathcal{A}$. In other words, it finds each global distribution in $\mathcal{G}$.

Algorithm 1 Computation of the global distributions

—*Step 1:* For each attribute set $\mathcal{A}$, we first identify the set $\mathcal{S}_{\mathcal{A}}$ of signatures s_i with respect to $\mathcal{A}$ where each s_i is matched by some tuples in T^* and has sufficient sample size. For example, for $\mathcal{A} = \{$"Nationality", "Gender" $\}$, a signature equal to $\{$ "Malaysian", "Male"$\}$ is matched by the first tuple in Table III(a). If it has sufficient sample size, it is stored in $\mathcal{S}_{\mathcal{A}}$.

—*Step 2:* For each attribute set $\mathcal{A}$, if $\mathcal{S}_{\mathcal{A}}$ is non-empty, we calculate the global distribution of $\mathcal{A}$ according to $\mathcal{S}_{\mathcal{A}}$ for each sensitive value x.

As described in Section 4.1, for a particular global distribution, we formulate m equations with m variables where m is the total number of signatures for A. The average number of terms in each equation is $O(N \cdot |\mathcal{W}_k| \cdot |\mathcal{QI}_{s_i}|)$ where N is the average QI-group size, $|\mathcal{W}_k|$ is the average number of possible worlds in a QI-group QI_k and $|\mathcal{QI}_{s_i}|$ is the average number of QI-groups with tuples matching a signature s_i. If Newton's method takes β time to find a solution, the computation for a global distribution takes $O(m \cdot N \cdot |\mathcal{W}_k| \cdot |\mathcal{QI}_{s_i}| + \beta)$ time. Since there are $|\mathcal{G}|$ global distributions, Step 2 takes $O(|\mathcal{G}| \cdot (m \cdot N \cdot |\mathcal{W}_k| \cdot |\mathcal{QI}_{s_i}| + \beta))$ time.

Thus, the total running time is $O(\alpha + |\mathcal{G}| \cdot (m \cdot N \cdot |\mathcal{W}_k| \cdot |\mathcal{QI}_{s_i}| + \beta))$. Note that the values of m, N, $|\mathcal{W}_k|$ and $|\mathcal{QI}_{s_i}|$ are small and the complexity is dominated by $|\mathcal{G}|$ and β. But, as the attribute set size increases, the sample size quickly becomes insufficient, and so $|\mathcal{G}|$ is typically well-behaved.

From our experiments, in all of our cases, Step 2 with the system of m equations can be solved in a relatively short time. So, β is also a reasonable value. For the benchmark dataset, adult, foreground knowledge can be mined within 12 minutes in all our experiments.

The probabilistic analysis is similar in nature to that studied for uncertain databases [Burdick et al. 2005; Burdick et al. 2007; Antova et al. 2007] The computation complexity above is in fact much smaller than these previous works. In [Antova et al. 2007], all results are returned within 3 hours. The reason is that [Burdick et al. 2005; Burdick et al. 2007; Antova et al. 2007] analyze the possible worlds based on the entire uncertain table (which can be regarded as a single large QI-group) while we analyze the possible worlds based on a single small QI-group (which is typically smaller than the entire table).

4.4 Discussion

We have just discussed how to find the global distribution from the published table. One may argue that the global distribution $\widetilde{\mathcal{G}}$ found from the published table is just an *approximation* of the *true* global distribution $\mathcal{G}_o$ found from the original table. Thus, the privacy breaches found in Section 3 according to $\widetilde{\mathcal{G}}$ are invalid. However, we disagree with this argument with the following reasons.

Firstly, since the adversary does not have the true global distribution $\mathcal{G}_o$ (because s/he has not seen the original table), the best adversary's knowledge about the global distribution is $\widetilde{\mathcal{G}}$.

Secondly, an adversary with $\widetilde{\mathcal{G}}$ is more *powerful* and more *sophisticated* than another adversary without any knowledge about the global distribution. The former adversary is what we are studying in this paper and can breach individual privacy discussed in Section 3 while the latter adversary is the normal adversary studied in the privacy literature [Machanavajjhala et al. 2006; Xiao and Tao 2006; Wong et al. 2007] and cannot breach any individual privacy found by the former adversary.

Thirdly, previous work considers adversaries that are equipped with some *external* knowledge on the dataset (i.e., background knowledge). Here, we consider foreground knowledge attacks which can succeed without this external knowledge. Thus, our attacker is more sophisticated and powerful.

5. EMPIRICAL STUDY

A Pentium IV 2.2GHz PC with 1GB RAM was used to conduct our experiment. The algorithm was implemented in C/C++. We adopted the publicly available dataset, Adult Database, from the UCIrvine Machine Learning Repository [Blake and Merz 1998]. This dataset (5.5MB) was also adopted by [LeFevre et al. 2005; Machanavajjhala et al. 2006; Wang et al. 2004; Fung et al. 2005; Wong et al. 2007]. We used a configuration similar to [LeFevre et al. 2005; Machanavajjhala et al. 2006; Wong et al. 2007]. The records with unknown values were first eliminated resulting in a dataset with 45,222 tuples (5.4MB). Nine attributes were chosen in our experiment, namely Age, Work Class, Marital Status, Occupation, Race, Sex, Native Country, Salary Class and Education. By default, we chose the first five attributes and the last attribute as the quasi-identifer and the sensitive attribute, respectively. Similar to [Wong et al. 2007], in attribute "Education", all values representing the education levels before "secondary" (or "9th-10th") such as "1st-4th", "5th-6th" and "7th-8th" are regarded as a sensitive value set where an adversary checks whether each individual is linked to this set more than $1/r$, where r is a parameter.

There are 3.46% tuples with education levels before "secondary". We set $\epsilon = 0.01$ and $\sigma = 0.9$ for sampling. That is, the allowed relative error of sampling is 1/3.46 = 28.90%, which is considered large. A larger allowed error means less attribute sets can be pruned. Since there is a set $\mathcal{G}$ of multiple global distributions G, we can calculate $p(t : x)$ for different G's and different x's. We take the greatest such value to report as the probability that individual t is linked to some sensitive value since this corresponds to the worst case privacy breach.

5.1 Privacy Breach in l-diverse Tables

In this section, we will show that foreground knowledge attack is successful in the published data generated from the benchmark dataset, adult, by a well-known privacy algorithm, *Anatomy* [Xiao and Tao 2006]. We set $l = r$ where l is the parameter of l-diversity used in Anatomy. We implemented the formula in Section 3 to calculate the probability of a privacy breach and the formula in Section 4 to find the global distribution from the published data. If a tuple which appears in the published data is identified as a privacy breach by our algorithm, it is said to be a *problematic tuple*. The tuples linking to sensitive values in the original table are called *sensitive tuples*. In this case study, we evaluate privacy breaches with five measurements:

(1) *proportion of problematic tuples among sensitive tuples, (this is the recall in IR*

research).

(2) *proportion of non-sensitive tuples which are identified wrongly as problematic tuples by our algorithm,*

(3) *the average probability by which individual privacy is breached among all sensitive tuples*

(4) *the average absolute difference between $1/r$ and the probability by which individual privacy is breached among all sensitive tuples,*

(5) *the average square difference between $1/r$ and the probability by which individual privacy is breached among all sensitive tuples*

We have conducted experiments with the variation of r and the variation of the QI size. (1) Variation of r: When $r = 2$ with default settings, the *average probability that individual privacy breaches* among all sensitive tuples is $0.8917(> 1/2)$. When r is increased to 4, it becomes $0.4640(> 1/4)$. When r increases, there is a higher chance that a tuple forms a QI-group with other tuples. Thus, the average size of QI-groups is larger. Thus, the average probability of privacy breaches decreases.

We also studied the *proportion of problematic tuples* among all sensitive tuples and the *proportion of non-sensitive tuples identified wrongly as privacy breaches.* We found that, in most cases, more than 99% of sensitive tuples have privacy breaches and less than 6% of non-sensitive tuples are identified wrongly. This shows that the problem caused by the foreground knowledge is quite serious. The small percentage of false alarm confirms our concern that foreground knowledge can be used as a reliable information source for the adversary.

We also measured *the average absolute difference between $1/r$ and the probability by which individual privacy is breached among all sensitive tuples.* When $r = 2$, it is equal to 0.3917. When r is increased to 4, it becomes 0.2140. This is because, when r is larger, the size of an QI-group is larger and thus it is more difficult to have privacy breaches. Thus, the average absolute difference is smaller. Similarly, we measured *the average square difference between $1/r$ and the probability by which individual privacy is breached among all sensitive tuples.* The trend is similar. When r is 2, it is equal to 0.1810. When r becomes 4, it is equal to 0.0675.

(2) Variation of the QI size: When the QI size is equal to 3 with default settings where $r = 2$, the average probability causing privacy breaches is 0.80307. When the size is increased to 8, it becomes 0.943526. This is because when there are more QI attributes, it is more likely that a QI attribute (or attribute set) gives a global distribution which can lead to privacy breaches.

We also have a case study on the published data generated by *Anatomy.* Suppose the QI attributes chosen are Age, Marital Status and Occupation and the sensitive attribute is Education. In the original data, there are the following 2 tuples.

Age	Marital Status	Occupation	Education
39	Never-married	Adm-clerical	Bachelors
20	Married-civ-spouse	Craft-repair	5th-6th

Suppose the objective of *Anatomy* is 2-diversity. Since “5th-6th” is a sensitive value, *Anatomy* forms an QI-group containing these two tuples. However, from the global distribution derived from the published data with respect to attribute Occupation, the probability that an individual with Occupation=“Adm-clerical” is linked to a low education is only 0.02 but the probability that an individual with Occupation="Craft-repair" is linked to a low education is 0.04. Since there is a significant difference in global distribution of attribute Occupation, the probability that the second tuple above is linked to a low education is 0.67 (which is greater than 0.5).

It is noted that the global distribution derived from the published data matches the real situation that "Adm-clerical" jobs require higher educations than "Craft-repair". In other words, the foreground knowledge can help the adversary to breach individual privacy. More specifically, let us check whether the *real* global distribution derived from the original table is similar to the global distribution derived from the published data. From the original table, the probability that an individual with Occupation="Adm-clerical" is linked to a low education is only 0.01 but the probability that an individual with Occupation="Craft-repair" is linked to a low education is 0.04. We observe that this global distribution is similar to that derived from the published data.

With our default experimental setting using sufficient sample size, for 2-diversity, the average relative error of the global probabilities derived from the published data=0.7% which achieves 99.3% accuracy. For 10-diversity, the error increases to 5.26% where the accuracy is 94.74%. It shows that statistically the accuracy is very high. In other words, the foreground knowledge derived from the published data is quite accurate compared with the knowledge derived from the original table.

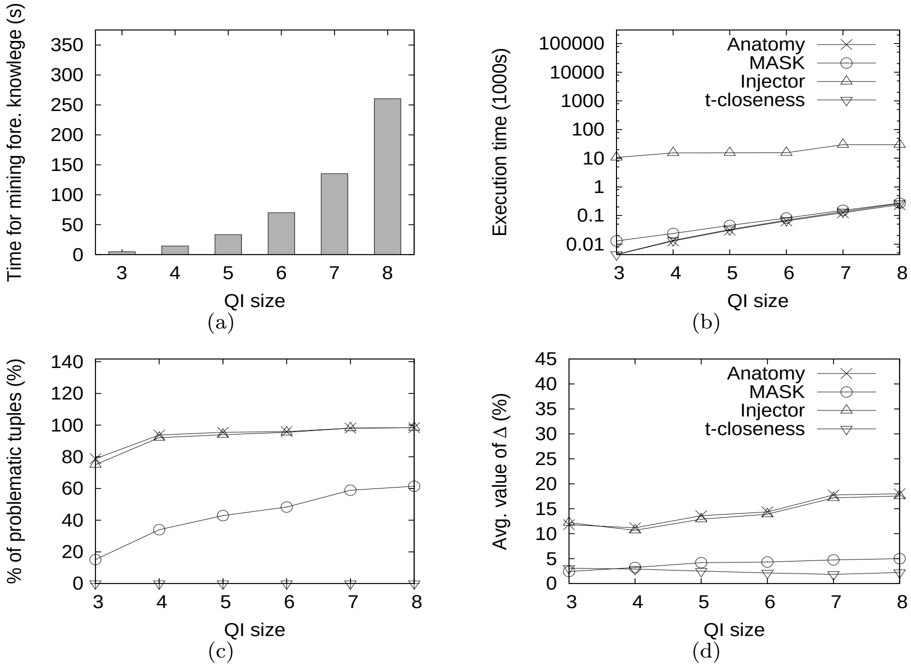
In all our experiments, privacy breaches can be found within 12 minutes, which shows that foreground knowledge attack can easily be realized.

5.2 Privacy Breach in Other Privacy Models

We studied privacy breaches with four algorithms, *Anatomy* [Xiao and Tao 2006], *MASK* [Wong et al. 2007], *Injector* [Li and Li 2008] and *t-closeness* [Li and Li 2007]. They are selected because they consider l-diversity or similar privacy requirements, so we need only set $l = r$. For *Anatomy*, we set $l = r$. For *MASK*, the parameters k and m used in MASK are set to r. For *Injector*, the parameters $minConf$, $minExp$ and l are set to 1, 0.9 and r, respectively, which are the default settings in [Li and Li 2008]. For *t-closeness*, similar to [Li and Li 2007], we set $t = 0.2$, and as in [Li and Li 2007], algorithm Incognito [LeFevre et al. 2005] is adopted in the computation. We evaluate the algorithms in terms of four measurements: (1) *time for mining foreground knowledge*, (2) *execution time*, (3) the *proportion of problematic tuples among all sensitive tuples*, (4) the average of the greatest difference in the global probabilities in each QI-group (In our figures, we label this as *average value of* $\triangle$), and (5) the *relative error ratio* in answering an aggregate query as in [Xiao and Tao 2006; Wong et al. 2007; Li and Li 2008] by the published data. For each measurement, we conducted the experiments 100 times and took the average.

We do not report the time for finding privacy breaches because the time is very short (within a few minutes). For the sake of space, since the proportion of non-sensitive tuples identified wrongly for privacy breaches is small (less than 10%), we do not report here.

Let us explain measurements (4) and (5). (4) Consider a QI-group QI_k contains two tuples matching signatures s_i and s_j, respectively. Suppose $p(s_i : x)$ is the greatest global probabilities and $p(s_j : x)$ is the smallest in the QI-group. The



(a) (b) (c) (d)

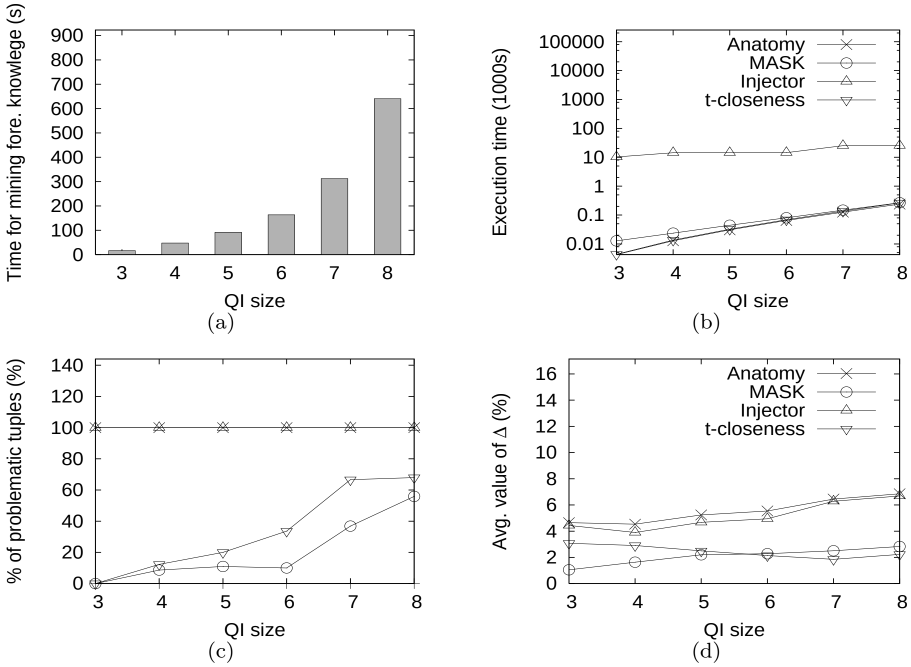
Fig. 1. Effect of QI size ($r = 2$)

value of $\triangle$ in QI_k is equal to $p(s_i : x) - p(s_j : x)$. The average value of $\triangle$ is taken among all QI-groups and all attribute sets $\mathcal{A}$ with sufficient samples. (5) The relative error ratio measures the utility of the published data. We adopt all query parameters in [Xiao and Tao 2006; Wong et al. 2007; Li and Li 2008]. For each evaluation, we performed 10,000 queries and reported the average relative error ratio.

We have conducted the experiments by varying two factors: (1) the QI size, and (2) r.

Figure 1 and Figure 2 show the results when r is set to 2 and 10, respectively. Figure 1(a) shows that the time for mining foreground knowledge increases with the QI size because the algorithm needs to process more attribute sets. Figure 1(b) shows that the execution time increases with the QI size because the algorithms have to process more QI attributes.

Figure 1(c) shows that the proportion of problematic tuples among sensitive tuples increases with QI size. With a larger QI size, there is a higher chance that individual privacy breaches due to more attributes which can be used to construct the global distributions. *MASK* has fewer privacy breaches compared with *Anatomy* and *Injector* because the side-effect of the minimization of QI values in each QI-group adopted in *MASK* makes the difference in the global distribution among all tuples in each QI-group smaller. Thus, the number of individual with privacy breaches is smaller. It is noted that there is no violation in *t-closeness*. The reason why t-closeness has no privacy breaches is due to the large QI-groups formed by global recoding with respect to value $r(= 2)$. The average size of the



(a)

(b)

(c)

(d)

Fig. 2. Effect of QI size ($r = 10$)

QI-group in the table satisfying t-closeness is at least 4000 and the utility of the table is low. It is noted that parameter t is independent of parameter r. We will show that t-closeness has privacy breaches when $r = 10$.

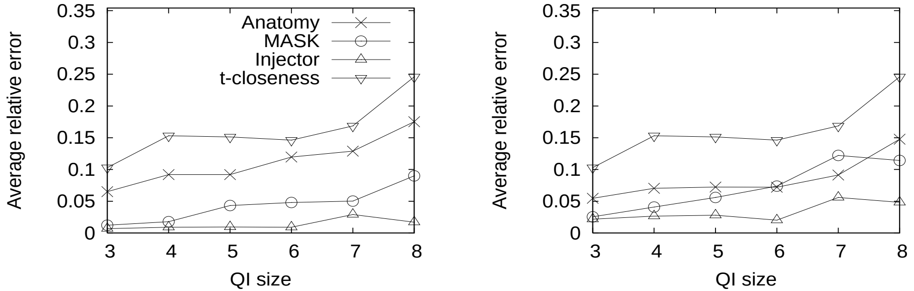
In Figure 1(d), when the QI size increases, the average value of $\triangle$ with respect to every attribute set increases, as shown in Figure 1(d). The average value of $\triangle$ is the largest in *Anatomy* and *Injector*, and the third largest in *MASK*. This is because *Anatomy* and *Injector* does not take the global distribution directly into the consideration for merging but *MASK* does indirectly during the minimization of QI values.

Figure 3(a) shows that the average relative error of t-closeness is the largest since it forms large QI-groups by global recoding which introduce a lot of errors and thus reduce the utility of the published data.

We have also conducted experiments when $r = 10$ as shown in Figure 2. The results are also similar. But, the time for mining foreground knowledge is larger. Since r is larger and thus $1/r$ is smaller, the average value of $\triangle$ is smaller when $r = 10$. Also, when $r = 10$, there are privacy breaches for *t-closeness* in Figure 2(c) because there is a higher privacy requirement when $r = 10$ and thus the size of the QI-group is not large enough for protection.

6. RELATED WORK

With respect to attribute types considered for data anonymization, there are two branches of studying. The first branch is anonymization according to the QI attributes. A typical model is k-anonymity [Aggarwal et al. 2005; LeFevre et al.



(a) $r = 2$

(b) $r = 10$

Fig. 3. Effect of QI size on average relative error

2005]. The other branch is the consideration of both quasi-identifier attributes and sensitive attributes. Some examples are [Machanavajjhala et al. 2006], [Wong et al. 2006], [Li and Li 2007], [Li and Li 2008] and [Brickell and Shmatikov 2008]. In this paper, we focus on this branch. We want to check whether the probability that each individual is linked to any sensitive value is at most a given threshold.

l-diversity [Machanavajjhala et al. 2006] proposes a model where l is a positive integer and each QI-group contains l "well-represented" sensitive values. For t-closeness [Li and Li 2007], the distribution in each QI-group in T^* with respect to the sensitive attribute is roughly equal to the distribution of the entire table T^*. Given a real number $\alpha \in [0, 1]$ and a positive integer k, (α, k)-anonymity [Wong et al. 2006] maintains that, for each QI-group QI, the number of tuples in QI is at least k and the frequency (in fraction) of each sensitive value in QI is at most α.

In the literature, different kinds of background knowledge are considered [Machanavajjhala et al. 2006; Kifer and Gehrke 2006; Martin et al. 2007; Wong et al. 2007; Li et al. 2009; Wong et al. 2010; Ganta et al. 2008; Li and Li 2008; Aggarwal et al. 2006]. [Kifer and Gehrke 2006] proposes the statistics of some attributes such as age and zipcode can be also available to the public. [Martin et al. 2007] considers another background knowledge in form of implications. [Wong et al. 2007] discovers that the minimality principle of the anonymization algorithm can also be used as a background knowledge. [Li et al. 2009] proposes to use the kernel estimation method to mine the background knowledge from the original table. [Wong et al. 2010] studies how to use probabilistic distribution-based background knowledge for anonymization. [Ganta et al. 2008] describes that there are many tables published from different sources containing overlapping individuals.

[Li and Li 2008] finds that association rules can be mined from the *original* table and thus can be used for privacy protection during anonymization. In [Aggarwal et al. 2006], the problem of privacy attack by adversarial association rule mining is investigated. Hence, the association rules are the foreground knowledge. However, as pointed out in [Silverstein et al. 1997], association rules used in [Li and Li 2008] and [Aggarwal et al. 2006] can contradict the true statistical properties. Also the solution in [Aggarwal et al. 2006] is to invalidate the rules, but this will violate the data mining objectives of data publication.

A recent work [Aggarwal 2008] proposes to generate a table in form of an uncertain data model. However, this work considers only k-anonymity, which ignores any sensitive attribute.

Recently, [Kifer 2009] proposes to use the concepts of exchangeability and de-Finetti's theorem to reason about privacy attacks. [Kifer 2009] proposes to use a machine learning model, Naive Bayes, to model the foreground knowledge derived from the published table. The work in [Kifer 2009] is different from ours in the following ways. Firstly, [Kifer 2009] makes an independence assumption among QI attributes when the foreground knowledge is considered. We do not have such an assumption because the foreground knowledge is expressed in the form of global distributions with respect to different attribute sets $\mathcal{A}$ where $\mathcal{A}$ is of any size and can express the correlations among attributes in $\mathcal{A}$. Secondly, the modeling of foreground knowledge in [Kifer 2009] is different from ours. [Kifer 2009] adopt the Naive Bayes model to express the foreground knowledge, where the probability of a certain assignment of an individual to a sensitive value is based on the assumption that all distributions of the linkage probabilities are the same. However, this assumption has no basis. We also adopt Bayesian probability, but our prior knowledge in the Bayesian probability is the linkage probabilities based on the given table, and the posterior probability is the linkage probability conditioned on that the given tuple is within a certain anonymized group. Our prior knowledge is hence not based on an unrealistic assumption. Thirdly, [Kifer 2009] shows the attack in the published table generated by a particular algorithm, Anatomy, for a particular privacy requirement, l-diversity. But, we show empirically the attacks in published tables generated by not only *Anatomy* [Xiao and Tao 2006] but also other existing algorithms (e.g., *MASK* [Wong et al. 2007] and *Injector* [Li and Li 2008]) for many privacy requirements like l-diversity, m-confidentiality and t-closeness. It is important to show how privacy breaches occur in other algorithms in the literature.

Instead of publishing an anonymized table, [Zhu et al. 2009] considers publishing the data mining results in forms of association rules which can be obtained from the original table. In [Zhu et al. 2009], the main focus is to derive sensitive information from a set of association rules. [Zhu et al. 2009] is different from ours since it does not study how to derive sensitive information from the anonymized table.

7. CONCLUSION

In this paper, we point out a fundamental privacy breach problem which has been overlooked in the past. With the consideration of the utility of the anonymized table, group based anonymization suffers from privacy breaches. Our experiments show that existing well-known privacy models *Anatomy*, *MASK*, *Injector* and *t-closeness* suffer from serious privacy breaches in a benchmark dataset. For future work, we plan to study how to anonymize the data to defend against foreground knowledge attack. In our experiment, we observe that the chance of privacy breaches is lower if each group contains tuples with "similar" global probabilities. Thus, forming QI-groups with "similar" tuples is one possible strategy. Another future work is to study the effect of background knowledge that may be possessed by the adversary.

Acknowledgement

We are grateful to the anonymous reviewers for their constructive comments on this paper. The research of Raymond Chi-Wing Wong is supported by HKRGC GRF 621309. The research of Philip S. Yu is supported by US NSF through grants IIS-0914934, OISE-0968341 and OIA-0963278.

REFERENCES

Aggarwal, C. C. 2008. On unifying privacy and uncertain data models. In *ICDE*. 386–395.

Aggarwal, C. C., Pei, J., and Zhang, B. 2006. On privacy preservation against adversarial data mining. In *KDD*. 510–516.

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. 2005. Anonymizing tables. In *ICDT*. 246–258.

Antova, L., Koch, C., and Olteanu, D. 2007. 10^{10^6} worlds and beyond: Efficient representation and processing of incomplete information. In *ICDE*. 606–615.

Blake, E. K. C. and Merz, C. J. 1998. UCI repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html.

Brickell, J. and Shmatikov, V. 2008. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. In *KDD*. 70–78.

Burdick, D., Deshpande, P., Jayram, T., Ramakrishnan, R., and Vaithyanathan, S. 2005. Olap over uncertain and imprecise data. In *VLDB*. 123–144.

Burdick, D., Doan, A., Ramakrishnan, R., and Vaithyanathan, S. 2007. Olap over imprecise data with domain constraints. In *VLDB*. 39–50.

Chapra, S. C. and Canale, R. P. 2002. Numerical methods for engineers. In *McGraw-Hill, 4th ed.*

Cheng, R., Chen, J., Mokbel, M., and Chow, C. 2008. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In *ICDE*. 973–982.

Fung, B. C. M., Wang, K., and Yu, P. S. 2005. Top-down specialization for information and privacy preservation. In *ICDE*. 205–216.

Ganta, S. R., Kasiviswanathan, S. P., and Smith, A. 2008. Composition attacks and auxiliary information in data privacy. In *KDD*. 265–273.

Imielinski, T. and Jr., W. L. 1984. Incomplete information in relational databases. In *Journal of ACM*. 761–791.

Kifer, D. 2009. Attacks on privacy and definetti's theorem. In *SIGMOD*. 127–138.

Kifer, D. and Gehrke, J. 2006. Injecting utility into anonymized datasets. In *SIGMOD*. 217–228.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. 2005. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*. 49–60.

Li, N. and Li, T. 2007. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*. 106–115.

Li, T. and Li, N. 2008. Injector: Mining background knowledge for data anonymization. In *ICDE*. 446–455.

Li, T., Li, N., and Zhang, J. 2009. Modeling and integrating background knowledge in data anonymization. In *ICDE*. 6–17.

Machanavajjhala, A., Gehrke, J., and Kifer, D. 2006. l-diversity: privacy beyond k-anonymity. In *ICDE*. 24.

Martin, D. J., Kifer, D., Machanavajjhala, A., and Gehrke, J. 2007. Worst-case background knowledge for privacy-preserving data publishing. In *ICDE*. 126–135.

Nergiz, M. E. and Clifton, C. 2007. Thoughts on k-anonymization. In *Data & Knowledge Engineering*. 622–645.

Silverstein, C., Motwani, R., and Brin, S. 1997. Beyond market baskets: Generalizing association rules to correlations. In *SIGMOD*. 265–276.

SWEENEY, L. 2002. k-anonymity: a model for protecting privacy. *International journal on uncertainty, Fuzziness and knowldege based systems 10(5)*, 557–570.

TOIVONEN, H. 1996. Sampling large databases for association rules. In *VLDB*. 134–145.

WANG, K., YU, P. S., AND CHAKRABORTY, S. 2004. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*. 249–256.

WONG, R., FU, A., WANG, K., AND PEI, J. 2007. Minimality attack in privacy preserving data publishing. In *VLDB*. 543–554.

WONG, R., LI, J., FU, A., AND WANG, K. 2006. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *KDD*. 754–759.

WONG, R. C.-W., FU, A. W.-C., WANG, K., XU, Y., PEI, J., AND YU, P. 2010. Probabilistic inference protection on anonymized data. In *ICDM*. 1127–1132.

XIAO, X. AND TAO, Y. 2006. Anatomy: Simple and effective privacy preservation. In *VLDB*. 139–150.

XIAO, X. AND TAO, Y. 2007. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*. 689–700.

ZHANG, Q., KOUDAS, N., SRIVASTAVA, D., AND YU, T. 2007. Aggregate query answering on aononymized tables. In *ICDE*. 116–125.

ZHU, Z., WANG, G., AND DU, W. 2009. Deriving private information from association rule mining results. In *ICDE*. 18–29.