Minimality Attack in Privacy Preserving Data Publishing [*]

Raymond Chi-Wing Wong[†] Ada Wai-Chee Fu[†] Ke Wang[‡] Jian Pei[‡]
[†]The Chinese University of Hong Kong [‡]Simon Fraser University, Canada
{cwwong,adafu}@cse.cuhk.edu.hk, {wangk,jpei}@cs.sfu.ca

ABSTRACT

Data publishing generates much concern over the protection of individual privacy. Recent studies consider cases where the adversary may possess different kinds of knowledge about the data. In this paper, we show that knowledge of the mechanism or algorithm of anonymization for data publication can also lead to extra information that assists the adversary and jeopardizes individual privacy. In particular, all known mechanisms try to minimize information loss and such an attempt provides a loophole for attacks. We call such an attack a minimality attack. In this paper, we introduce a model called m-confidentiality which deals with minimality attacks, and propose a feasible solution. Our experiments show that minimality attacks are practical concerns on real datasets and that our algorithm can prevent such attacks with very little overhead and information loss.

1. INTRODUCTION

Although data mining is potentially useful, many data holders are reluctant to provide their data for data mining due to the fear of violating individual privacy. In recent years, study has been made to ensure that sensitive information of individuals cannot be identified easily [16, 17, 10, 14, 9]. One well-studied approach is the k-anonymity model which in turn led to other models such as confidence bounding [19], l-diversity [12], (α, k)-anonymity [22], t-closeness [11], (k, e)-anonymity [27] and (c, k)-safety [13].

Generally, the existing models assume that the data in the form of a table T contains (1) a *quasi-identifer (QID)* as a set of attributes (e.g., Date of birth, Zipcode and Sex) which can be used to identify an individual, and (2) *sensitive attributes* which may contain some sensitive values (e.g. HIV of attribute Disease) of individuals. Often, it is also assumed that each tuple in T corresponds to an individual and no two tuples refer to the same individual. All tuples with the same QID value form an *equivalence class* (QID-EC for short). The table T is said to satisfy k-*anonymity* if the size of every equivalence class is greater than or equal to k.

Moreover, in a simplified setting of l-diversity model [12], a QID-EC is said to be l-*diverse* or satisfy l-*diversity* if the proportion of each sensitive value is at most $1/l$. A table satisfies l-diversity (or it is l-diverse) if all QID-EC's in it are l-diverse. In the following discussion, when we refer to l-diversity, we refer to this simplified setting. We shall discuss the complex l-diversity model in Section 5, where we show that our results can be extended to other anonymization models.

In this paper, we study the case where the adversary has some additional knowledge about the mechanism involved in the anonymization, and thus can launch an attack based on this knowledge. We focus on the protection of the relationship between the quasi-identifier and a single sensitive attribute.

1.1 Minimality Attack

In Table 1(a), assume that the QID values of $q1$ and $q2$ can be generalized to Q and assume only one sensitive attribute "disease", in which HIV is a sensitive value. For example, $q1$ may be $\{Nov\ 1930, Z3972, M\}$, $q2$ may be $\{Dec\ 1930, Z3972, M\}$ and Q is $\{Nov/Dec\ 1930, Z3972, M\}$. (Note that $q1$ and $q2$ may also be generalized values.) A tuple associated with HIV is said to be a *sensitive* tuple. For each equivalence class, at most half of the tuples are sensitive. Hence, the table satisfies 2-diversity.

As observed in [10], the existing anonymization approaches for data publishing follow an implicit principle: *"For any anonymization mechanism, it is desirable to define some notion of minimality. Intuitively, a k-anonymization should not generalize, suppress, or distort the data more than it is necessary to achieve k-anonymity."* Based on this minimality principle, Table 1(a) will not be generalized.[1] In fact the above notion of minimality is too strong since almost all known anonymization problems for data publishing are NP-hard, many existing algorithms are heuristical and only attain local minima. We shall later give a more relaxed notion of the minimality principle in order to cover both the optimal

[*] The research of Raymond Chi-Wing Wong and Ada Wai-Chee Fu is supported in part by the RGC Earmarked Research Grant of HKSAR CUHK 4120/05E. The research of Ke Wang and Jian Pei is supported in part by two NSERC Discovery Grants. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

[1] This is the case for each of the anonymization algorithms in [12, 19, 22].

VLDB '07, September 23-28, 2007, Vienna, Austria.
Copyright 2007 VLDB Endowment, ACM 978-1-59593-649-3/07/09.

QID	Disease
$q1$	HIV
$q1$	non-sensitive
$q2$	HIV
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive

(a) good table

QID	Disease
$q1$	HIV
$q1$	HIV
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive

(b) bad table

QID
Q
Q
Q
Q
Q
Q
Q

(c) global

QID
Q
Q
Q
Q
$q2$
$q2$
$q2$

(d) local

Table 1: 2-diversity: global and local recoding

Name	QID
Andre	$q1$
Kim	$q1$
Jeremy	$q2$
Victoria	$q2$
Ellen	$q2$
Sally	$q2$
Ben	$q2$

(a) individual QID

QID
$q1$
$q1$
$q2$
$q2$
$q2$
$q2$
$q2$

(b) multiset

Name	QID
Andre	$q1$
Kim	$q1$
Jeremy	$q2$
Victoria	$q2$
Ellen	$q2$
Sally	$q2$
Ben	$q2$
Tim	$q4$
Joseph	$q4$

(c) individual QID

QID
$q1$
$q1$
$q2$
$q2$
$q2$
$q2$
$q2$
$q4$
$q4$

(d) multiset

Table 2: T^e: external table available to the adversary

as well as the heuristical algorithms. For now, we assume that mimimality principle means that a QID-EC will not be generalized unnecessarily.

Next, consider a slightly different table, Table 1(b). Here, the set of tuples for $q1$ violates 2-diversity because the proportion of the sensitive tuples is greater than 1/2. Thus, this table will be anonymized to a *generalized* table by generalizing the QID values as shown in Table 1(c) by *global recoding* [24, 18]. The tuples in this table contain the generalized values of the QID arranged in the same tuple ordering as the corresponding tuples in Table 1(b). This is a convention we shall use for all examples in this paper showing the anonymization of a table. In global recoding, all occurrences of an attribute value are recoded to the same value. If *local recoding* [16, 1] is adopted, occurrences of the same value of an attribute may be recoded to different values. Such an anonymization is shown in Table 1(d). These anonymized tables satisfy 2-diversity. However, do these tables protect individual privacy sufficiently?

In most previous work (e.g., [17, 9, 10, 24]), the knowledge of the adversary involves an external table T^e such as a voter registration list that maps QIDs to individuals. As in most previous work, we assume that each tuple in T^e maps to one individual and no two tuples map to the same individual. The same is also assumed in the table T to be published. Let us first consider the case when T and T^e are mapped to the same set of individuals. Table 2(a) is an example of T^e.

Assume further that the adversary knows the goal of 2-diversity, s/he also knows whether it is a global or local recoding, and Table 2(a) is available as the external table T^e. With the notion of minimality in anonymization, the adversary reasons as follows: From the published Table 1(c), there are 2 sensitive tuples in total. From T^e, there are 2 tuples with QID=$q1$ and 5 tuples with QID=$q2$. Hence, the equivalence class for $q2$ in the original table *must* already satisfy 2-diversity, because even if both sensitive tuples have QID=$q2$, the proportion of sensitive values in the class for $q2$ is only 2/5. Since *generalization* has taken place, at least one equivalence class in the original table T must have violated 2-diversity, because otherwise no generalization will take place according to minimality. The adversary concludes that $q1$ has violated 2-diversity, and that is possible only if both tuples with QID=$q1$ have a disease value of "HIV". The adversary therefore discovers that Andre and Kim are linked to "HIV".

QID	Disease
$q1$	HIV
$q1$	Lung Cancer
$q2$	Gallstones
$q2$	HIV
$q2$	Ulcer
$q2$	Alzheimer
$q2$	Diabetes
$q4$	Ulcer
$q4$	Alzheimer

(a) good table

QID	Disease
$q1$	HIV
$q1$	HIV
$q2$	Gallstones
$q2$	Lung Cancer
$q2$	Ulcer
$q2$	Alzheimer
$q2$	Diabetes
$q4$	Ulcer
$q4$	Alzheimer

(b) bad table

QID
Q
Q
Q
Q
Q
Q
Q
$q4$
$q4$

(c) global

QID
Q
Q
Q
Q
$q2$
$q2$
$q2$
$q4$
$q4$

(d) local

Table 3: 2-diversity (where all values in Disease are sensitive): global and local recoding

In some previous work, it is assumed that the set of individuals in the external table T^e can be a superset of that for the published table. Table 2(c) shows such a case, where there is no tuple for Tim and Joseph in Table 1(a) and Table 1(b). If it is known that $q4$ cannot be generalized to Q (e.g. $q4$=$\{Nov$ 1930, $Z3972, F\}$ and Q=$\{Jan/Feb$ 1990, $Z3972$, $M\}$), then the adversary can be certain that the tuples with QID=$q4$ are not in the original table. Thus, the extra $q4$ tuples in T^e do not have any effect on the above reasoning of the adversary and, therefore, the same conclusion can be drawn. We call such an attack based on the minimality principle a *minimality attack.*

Observation 1. *If a table T is anonymized to T^* which satisfies l-diversity, it can suffer from a minimality attack. This is true for both global and local recoding and for the cases when the set of individuals related to T^e is a superset of that related to T.*

In the above example, some values in the sensitive attribute Disease are not sensitive. Would it help if all values in the sensitive attributes are sensitive? In the tables in Table 3, we assume that all values for Disease are sensitive. Table 3(a) satisfies 2-diversity but Table 3(b) does not. Suppose anonymization of Table 3(b) results in Table 3(c) by global recoding and Table 3(d) by local recoding. The adversary is armed with the external table Table 2(c) and the knowledge of the goal of 2-diversity, s/he can launch an attack by reasoning as follows: with 5 tuples for QID=$q2$ and each sensitive value appearing at most twice, there cannot be any violation of 2-diversity for the tuples with QID=$q2$. There must have been a violation for QID=$q1$. For a violation to take place, both tuples with QID=$q1$ must be linked to the same disease. Since HIV is the only disease that appears twice in the table, Andre and Kim must have contracted HIV.

Observation 2. *Minimality attack is possible no matter the sensitive attribute contains non-sensitive values or not.*

The intended *objective* of 2-diversity is to make sure that an adversary cannot deduce with a probability above 1/2 that an individual is linked to any sensitive value. Thus, the published tables violate this objective.

Some previous studies [23, 27, 13] propose the bucketization technique. However, it is easy to show that minimality attacks still may happen. For example, the multisets in Tables 2(b) and (d) are inherently available in the methods using the bucketization techniques. However, the above minimality attacks to Andre would also be successful if the knowledge of the external table Table 2(a) is replaced by that of a multiset of the QID values as shown in Table 2(b) plus the QID value of Andre; or if Table 2(c) is replaced by the multiset in Table 2(d) plus the QID value of Andre.

1.2 Contributions

In this paper, we introduce the problem of minimality attacks in privacy preservation for data publishing. Our contributions include the following.

First, to the best of our knowledge, this is the first work to study the attack by minimality in privacy preserving data publishing. We propose an m-confidentiality model to capture the privacy preserving requirement under the additional adversary knowledge of the minimality of the anonymization mechanisms.

Second, since almost all known anonymization methods for data publishing attempt to minimize information loss, we show in Section 5 how minimality attack can be a practical concern in various known anonymization models.

Third, we propose a solution to generate a published data set which satisfies m-confidentiality. Our method makes use of the existing mechanisms for k-anonymity with additional precaution steps. Interestingly, although it has been discovered that k-anonymity is incapable of handling sensitive values in some cases, it is precisely this feature that makes it a useful component in our method to counter attacks by minimality for protecting sensitive data. Since k-anonymization does not consider the sensitive values, its result is not related to whether some tuples need to be anonymized due to the sensitive values. Without this relationship, an attack by minimality becomes infeasible.

Last, we have conducted a comprehensive empirical study to show that minimality attacks are on a practical concern on real data sets. Compared to a most competent existing algorithms for k-anonymity, our method introduces very minor computation overhead but achieves comparable information loss.

The rest of the paper is organized as follows. In Section 2, we review the related work. We formulate the problem in Section 3, and characterize the nature of minimality attacks in Section 4. We show that minimality attacks are practical concerns in various anonymization models in Section 5. We give a simple yet effective solution in Section 6. An empirical study is reported in Section 7. The paper is concluded in Section 8.

2. RELATED WORK

Since the introduction of k-anonymity, there have been a number of enhanced models such as confidence bounding [19], l-diversity [12], (α, k)-anonymity [22], t-closeness [11], (k, e)-anonymity [27] and personalized privacy [24], which additionally consider the privacy issue of disclosure of the *relationship* between the quasi-identifier and the sensitive attributes. Confidence bounding is to bound the confidence by which a QID can be associated with a sensitive value. T is said to satisfy (α, k)-*anonymity* if T is k-anonymous and the proportion of each sensitive value in every equivalence class is at most α, where $\alpha \in [0, 1]$ is a user parameter. If we set $\alpha = \frac{1}{l}$ and $k = 1$, then the (α, k)-anonymity model becomes the simplified model of l-diversity.

An adversary may also have some additional knowledge about the individuals in the dataset or some knowledge about the data involved [12, 7, 13]. [12] considers the possibility that the adversary can exclude some sensitive values. For example, Japanese have an extremely low incidence of heart disease. Thus, the adversary can exclude heart disease in a QID-EC for a Japanese individual. [7] considers that additional information may be available in terms of some statistics on some of the attributes, such as age statistics and zip code statistics. More recently, [13] tries to protect sensitive data against background knowledge in the form of implications, e.g., if an individual A has HIV then another individual B also has HIV, and proposes a model called (c, k)-*safety* to protect against such attacks. However, none of the above work considers the knowledge of the anonymization mechanism discussed in this paper. In Section 5, we shall show that the above previous studies are vulnerable to minimality attacks. Other than generalization, more general distortion can be applied to data before publishing. The use of distortion has been proposed in earlier work such as [15, 4].

The idea of attack by minimality has been known for some time in cryptographic attacks where the adversary makes use of the knowledge of the underlying cryptographic algorithm. In particular, a timing attack [8] in a public-key encryption system, such as RSA, DSS and SSL, is a practical and powerful attack that exploits the timing factor of the implemented algorithm, with the assumption that the algorithm will not take more time than necessary. Measuring response time for a specific query might give away relatively large amounts of information. To defend timing attack, the same algorithm can be implemented in such a way that every execution returns in exactly x seconds, where x is the maximum time it ever takes to execute the routine. In this extreme case, timing does not give an attacker any helpful information. In 2003, Boneh and Brumley [3] demonstrated a practical network-based timing attack on SSL-enabled web servers which recovered a server private key in a matter of hours. This led to the widespread deployment and use of blinding techniques in SSL implementations.

3. PROBLEM DEFINITION

Let T be a table. We assume that one of the attributes is a sensitive attribute where some values in this attribute should not be linkable to any individual. A *quasi-identifier* (QID) is a set of attributes of T that may serve as identifications for some individuals.

ASSUMPTION 1. *Each tuple in the table T is related to one individual and no two tuples are related to the same individual.*

We assume that each attribute has a corresponding conceptual *taxonomy* $\mathcal{T}$. A lower level domain in the taxonomy $\mathcal{T}$ provides more details than a higher level domain. For example, Figure 1 shows a generalization taxonomy of "Education" in the "Adult" dataset [2]. Values "undergrad" and "postgrad" can be generalized to "university".[2] Generaliza-

[2] Such hierarchies can also be created for numerical attributes

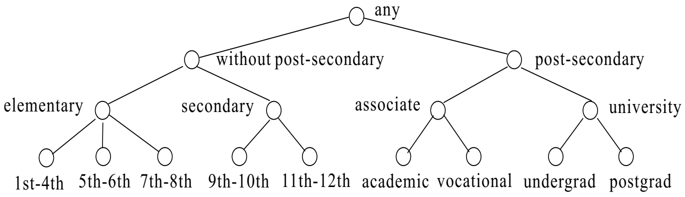


Figure 1: Generalization taxonomy of "Education" in the "Adult" dataset

tion replaces lower level domain values in the taxonomy with higher level domain values.

Some previous studies consider taxonomies only for QID attributes while some other also consider taxonomies for the sensitive attributes. In some earlier studies on anonymization, the taxonomy for an attribute in the QID or the sensitive attribute is a tree. However, in general, a taxonomy may be a directed acyclic graph (DAG). For example, "day" can be generalized to "week", or via "month" to "year", or via "season" to "year". Therefore, we extend the meaning of a taxonomy to any *partially ordered set* with a partial order. An attribute may have more than one taxonomy, where a certain value can belong to two or more taxonomies.[3]

Let $\mathcal{T}$ be a taxonomy for an attribute in QID. We call the leaf nodes of the taxonomy $\mathcal{T}$ the *ground values*.

In Figure 1, values "1st-4th", "undergrad" and "vocational" are some ground values in $\mathcal{T}$. As "university" is an ancestor of "undergrad", we obtain "undergrad" $\prec$ "university".

When a record contains the sensitive value of "lung cancer", it can be generalized to either "respiratory disease" or "cancer". While "cancer" and "lung cancer" are sensitive, "respiratory disease" as a category in general may not be. Therefore, we can assume the following property.

Assumption 2. *(Taxonomy property): In a taxonomy for a sensitive attribute, the ancestor nodes of a non-sensitive node are also non-sensitive. The ancestor of a sensitive node may be either sensitive or non-sensitive.*

In a faithful anonymization, a value can be generalized to any ancestor. For example, "lung cancer" may be generalized to "cancer" or "respiratory disease". With the above assumption, if a node is sensitive, all ground values in its descendants are sensitive.

With a taxonomy for the sensitive attribute, such as the one in Figure 1, in general, the protection is not targeting on a single ground value. In Figure 1, all the values under "elementary" may be sensitive in the sense that there should not be linkage between an individual and the set of values {1st-4th, 5th-6th, 7th-8th}. That is, the adversary must not be able to deduce with confidence that an individual has education between 1st to 8th grade. In general, a group of sensitive values may not be under one subtree. For example, for diseases, it is possible that cancer and HIV are both considered sensitive. So, a user should not be linked to the set {HIV, cancer} with a high probability. However, HIV and Cancer are not under the same category in the taxonomy. For this more general case, we introduce the *sensitive value set*, which is a set of ground values in the taxonomy for the sensitive attribute. In such a taxonomy, there can be multiple sensitive value sets.

A major technique used in the previous studies is to recode the QID values in such a way that a set of individuals will be matched to the same generalized QID value and, in the set, the occurrence of values in any sensitive value set is not frequent. Hence, the records with the same QID value (which could be a generalized value) is of interest. In a table T, the equality of the QID values determines an equivalence relation on the set of tuples in T. A QID equivalence class, or simply QID-EC, is a set of tuples in T with identical QID value. For simplicity, we also refer to a QID-EC by the identical QID value.

Definition 1 (Anonymization). *Anonymization is a a one-to-one mapping function f from a table T to an anonymized table T^*, such that f maps each tuple t in T to a tuple $f(t) = t^*$ in T^*. Let $t^*.A$ (or $f(t).A$) be the value of attribute A of tuple t^* (or $f(t)$). Given a set of taxonomies $\tau = \{\mathcal{T}_1, ..., \mathcal{T}_u\}$, an anonymization defined by f conforms to τ iff $t.A \preceq f(t).A$ holds for any t and A.*

For instance, Table 1(b) is anonymized to Table 1(c). The mapping function f maps the tuples with $q1$ and $q2$ to Q.

Let K_{ad} be the knowledge of the adversary. In most previous work [17, 9, 10, 24], in addition to the published data set T^*, K_{ad} involves an external table T^e such as Voter registration list that maps QIDs to individuals. In the literature, two possible cases of T^e have been considered: (1) *Worst Case:* the set of individuals in the external table T^e is equal to the set of individuals in the original table T; (2) *Superset Case:* the set of individuals in the external table T^e is a proper superset of the set of individuals in the original table T. Assuming the worst case scenario is the safest stance and it has been the assumption in most previous studies. We have shown in our first two examples that, in either of the above two cases, minimality attacks are possible.

The objective of privacy preservation is to limit the probability of the linkage from any individual to any sensitive value set s in the sensitive attribute. We define this probability or *credibility* as follows.

Definition 2 (Credibility). *Let T^* be a published table which is generated from T. Consider an individual $o \in O$ and a sensitive value set s in the sensitive attribute. Credibility(o, s, K_{ad}) is the probability that an adversary can infer from T^* and background knowledge K_{ad} that o is associated with s.*

The background knowledge particularly addressed here is about the minimality principle as formulated below.

Definition 3 (Minimality Principle). *Suppose $\mathcal{A}$ is an anonymization algorithm for a privacy requirement $\mathcal{R}$ which follows the* minimality principle. *Let table T^* be a table generated by $\mathcal{A}$ and T^* satisfies $\mathcal{R}$. Then, for any QID-EC X in T^*, there is no specialization (reverse of generalization) of the QID's in X which results in another table T' which also satisfies $\mathcal{R}$.*

by generalizing values to value range and to wider value ranges. The ranges can be determined by users or a machine learning algorithm [5].

[3] Note that a taxonomy may not be a lattice. For example, consider attribute disease. "Nasal cancer" and "lung cancer" may both be under two parents of "cancer" and "respiratory disease".

Note that this minimality principle holds for both global recoding and local recoding. If $\mathcal{A}$ is for global recoding (local recoding), both T^* and T' are global recoding (local recoding). So far we focus on the privacy requirement of l-diversity. However, in Section 5, we shall consider cases where $\mathcal{R}$ is other requirements.

Assumption 3. *(Adversary knowledge K_{ad}^{min}) In the definition of $Credibility(o, s, K_{ad})$, we consider the cases where K_{ad} includes T^*, the multiset T^q containing all QID occurrences in the table T, the QID values of a target individual in T, a set of taxonomies τ and whether the anonymization $\mathcal{A}$ conforms to the taxonomies τ, the target privacy requirement $\mathcal{R}$, and whether $\mathcal{A}$ follows the minimality principle. We refer to this knowledge as K_{ad}^{min}.*

If Table 1(a) is the result generated from an anonymization mechanism (e.g., the adapted Incognito algorithm in [12]) for l-diversity that follows the minimality principle, suppose the multiset in Table 2(b) is known and the QID value of individual o is known to be $q1$, then $Credibility(o, \{HIV\}, K_{ad}^{min}) = 1/2$. When the same K_{ad}^{min} is applied to Table 1(c), then $Credibility(o, \{HIV\}, K_{ad}^{min}) = 1$. Section 4 describes how to compute the credibility.

The above minimality principle is very general and does not demand that $\mathcal{A}$ minimizes the overall information loss, nor does it depend on how the information loss is defined. Almost all known anonymization algorithms (including Incognito based methods [10, 12, 13, 11] and top-down approaches [6, 24, 22, 18]) try to reduce information loss of one form or another, and they all follow the above principle.

In the examples in Section 1, the value of l (for l-diversity) is used by the adversary. However, l is not included in K_{ad}^{min}. This is because, in many cases, it can be deduced from the published table T^*. For example, for the anonymization in Table 1(d), the adversary can deduce that l must be 2.

Definition 4 (m-confidentiality). *A table T is said to satisfy m-confidentiality (or T is m-confidential) if, for any individual o and any sensitive value set s, $Credibility(o, s, K_{ad})$ does not exceed $1/m$.*

For example, Tables 1(a) satisfies 2-confidentiality.

When a table T is anonymized to a more generalized table T^*, it is of interest to measure the information loss that is incurred. There are different ways to define information loss. Since we shall measure the effectiveness of our method based on the method in [24], we also adopt a similar measure of information loss. The idea is similar to the normalized certainty penalty [26].

Definition 5 (Coverage and Base). *Let $\mathcal{T}$ be the taxonomy for an attribute in QID. The coverage of a generalized QID value v^*, denoted by $coverage[v^*]$, is given by the number of ground values v' in $\mathcal{T}$ such that $v' \prec v^*$. The base of the taxonomy $\mathcal{T}$, denoted by $base(\mathcal{T})$, is the number of ground values in the taxonomy.*

For example, in Figure 1, $coverage$[“university”] = 2 since “undergrad” and “postgrad” can be generalized to “university”, $base(\mathcal{T}) = 9$.

A weighting can be assigned for each attribute A, denoted by $weight(A)$, to reflect the users' opinion on the significance of information loss in different attributes. Let $t.A$ denote the value of A in tuple t.

QID	Disease
$q1$	HIV
$q1$	HIV
$q2$	HIV
$q2$	non-sensitive
$q3$	HIV
$q3$	HIV
$q3$	non-sensitive
$q3$	non-sensitive
$q3$	non-sensitive
...	...
$q3$	non-sensitive

Table 4: A table which violates 2-diversity

QID	Disease
Q	HIV
Q	HIV
Q	HIV
Q	non-sensitive
Q	HIV
Q	HIV
Q	non-sensitive
Q	non-sensitive
Q	non-sensitive
...	...
Q	non-sensitive

Table 5: A 2-diverse table by global recoding of Table 4

Definition 6 (Information Loss). *Let table T^* be an anonymization of table T by means of a mapping function f. Let $\mathcal{T}_A$ be the taxonomy for attribute A which is used in the mapping and v^* be the nearest common ancestor of $t.A$ and $f(t).A$ in $\mathcal{T}_A$. The information loss of a tuple t^* in T^* introduced by f is given by*

$$\mathcal{IL}(t^*) = \sum_{A \in QID} \left\{ \frac{coverage[v^*] - 1}{base(\mathcal{T}_A) - 1} \times weight(A) \right\}$$

The information loss is given by $Dist(T, T^*) = \frac{\sum_{t^* \in T^*} \mathcal{IL}(t^*)}{|T^*|}$

If $f(t).A = t.A$, then $f(t).A$ is a ground value, the nearest common ancestor v^*=$t.A$, and $coverage[v^*] = 1$. If this is true for all A's in QID, then $\mathcal{IL}(t^*)$ is equal to 0, which means there is no information loss. If $t.A$ is generalized to the root of taxonomy $\mathcal{T}_A$, then the nearest common ancestor v^* = the root of $\mathcal{T}_A$. Thus, $coverage[v^*] = base(\mathcal{T}_A)$ and, if this is the case for all A's in QID, then $\mathcal{IL}(t^*) = 1$. Note that we have modified the definition in [24] in order to achieve the range of [0,1] for $\mathcal{IL}(t^*) = 1$ and also for $Dist(T, T^*)$.

Although minimizing information loss poses a loophole for attack by minimality, one cannot completely ignore information loss since, without such a notion, we allow for complete distortion of the data which will also render the published data useless.

Definition 7 (PROBLEM). *Optimal m-confidentiality: Given a table T, generate an anonymized table T^* from T which satisfies m-confidentiality where the information loss $Dist(T, T^*)$ is minimized.*

4. CREDIBILITY: SOURCE OF ATTACK

In this section, we characterize the nature of minimality attack. Minimality attack is successful if the adversary can compute the credibility values and find a violation of m-confidentiality when the privacy requirement is l-diversity. This computation depends on a combinatorial analysis on the possibilities given the knowledge of K_{ad}^{min}. In particular, the adversary attacks by excluding some possible scenarios, tilting the probabilistic balance towards privacy disclosure.

4.1 Global Recoding

The derivation of credibility is better illustrated with the example as shown in Table 5 which is a global recoding of Table 4 to achieve 2-diversity. In Table 4, {HIV} is the only sensitive value set and the goal is 2-diversity. Assume that T and T^e have *matching cardinality* on Q. From T^e, the

	Number of sensitive tuples			Total number
	$q1$	$q2$	$q3$	of cases
(a)	2	0	3	120
(b)	2	1	2	90
(c)	2	2	1	10
(d)	1	2	2	90
(e)	0	2	3	120

Table 6: Possible combinations of number of sensitive tuples

adversary can determine that there are two tuples in $q1$, two tuples in $q2$ and 10 tuples in $q3$. Since there are 10 tuples with a QID value of $q3$, and there are in total 5 sensitive tuples, $q3$ trivially satisfies 2-diversity. As T^* (Table 5) is generalized, the adversary decides that at least one of the QID-EC's $q1$ and $q2$ contains two sensitive tuples. With this in mind, the adversary lists all the possible combinations of the number of sensitive tuples among the three classes $q1$, $q2$ and $q3$ in which either $q1$ or $q2$ or both contain 2 sensitive tuples as shown in Table 6. There are only five possible combinations as shown. We call this table as the *sensitive tuple distribution table*.

In scenario (a), there are $C^2_2 \times C^2_0 \times C^{10}_3 = 120$ different possible ways to assign the sensitive values to the tuples. In scenario (b), there are $C^2_2 \times C^2_1 \times C^{10}_2 = 90$ different assignments or cases. Similarly, there are 10 cases, 90 cases and 120 cases in scenarios (c), (d) and (e), respectively. The total number of cases is equal to $120+90+10+90+120 = 430$. Consider the credibility that an individual o with value $q1$ is linked to HIV given K^{min}_{ad}. There are two possible cases.

In the first case, there are two sensitive tuples in $q1$. The total number of cases where there are two sensitive tuples in $q1$ is equal to $120+90+10 = 220$. The probability that Case 1 occurs given K^{min}_{ad} is equal to $220/430 = 0.5116$.

In the second case, there is one sensitive tuple in $q1$. The total number of cases where there is one sensitive tuple in $q1$ is equal to 90. The probability that Case 2 occurs given K^{min}_{ad} is equal to $90/430 = 0.2093$.

In the following, we use $Prob(E)$ to denote the probability that event E occurs.

Thus, the credibility that an individual o with QID value $q1$ is linked to HIV given K^{min}_{ad} is equal to

$$\begin{aligned}&Prob(\text{Case 1}) \times Prob(\text{ } q1 \text{ is linked to HIV in Case 1})\\ &\quad + Prob(\text{Case 2}) \times Prob(\text{ } q1 \text{ is linked to HIV in Case 2})\end{aligned}$$

$Prob($ $q1$ is linked to HIV in Case 1) is equal to 2/2=1.

$Prob($ $q1$ is linked to HIV in Case 2) is equal to 1/2=0.5.

$$\begin{aligned}Credibility(o, \{HIV\}, K^{min}_{ad}) &= 0.5116 \times 1 + 0.2093 \times 0.5\\ &= 0.616,\end{aligned}$$

which is greater than 0.5. This result shows that the published table violates 2-confidentiality.

General Formula

The general formula of the computation of the credibility is based on the idea illustrated above. We have a probability space $(\Omega, \mathcal{F}, P)$, where Ω is the set of all possible assignments of the sensitive values to the tuples, $\mathcal{F}$ is the power set of Ω, and P is a probability mass function from $\mathcal{F}$ to the real numbers in [0,1] which gives the probability for each element in $\mathcal{F}$. Given K^{min}_{ad}, there will be a set of assignments $\mathcal{G}$ in Ω which are impossible or $P(\mathcal{G}) = 0$ and if $x \in \mathcal{G}$ then $P(\{x\}) = 0$. Without any other additional knowledge, we assume that the probability of the remaining assignments are equal. That is, $\mathcal{G}' = \Omega - \mathcal{G}$, $P(\mathcal{G}') = 1$ and for $x \in \mathcal{G}'$, $P(\{x\}) = 1/|\mathcal{G}'|$.

Definition 8. *Let Q be a QID-EC in T^*. Tables T^* and T^e have* matching cardinality *on Q if the number of tuples in T^e with QID that can be generalized to Q is the same as that in T^*.*

Let $\mathcal{X}$ be a maximal set of QID-EC's in T which are generalized to the same QID-EC Q in the published table T^*. Suppose T^* and T^e have matching cardinality on Q. Let $C_1, C_2, ...C_u$ be the QID-EC's in $\mathcal{X}$ sorted in ascending order of the size of the QID-EC's. Let n_i be the number of tuples in class C_i. Hence, $n_1 \leq n_2 \leq ... \leq n_u$. Let n_s be the total number of tuples with values in sensitive value set s in the data set.

In Table 4, there are three classes, namely $q1, q2$ and $q3$. Thus, $u = 3$. C_1 corresponds to $q1$, C_2 corresponds to $q2$ and C_3 corresponds to $q3$. Also, $n_1 = 2, n_2 = 2$ and $n_3 = 10$.

Suppose the published table is generalized in order to satisfy the l-diversity requirement.

If $n_s \leq \lfloor \frac{n_i}{l} \rfloor$, then C_i in the original data set must satisfy the l-diversity requirement without any generalization. Class C_i may violate the l-diversity requirement only if $n_s > \lfloor \frac{n_i}{l} \rfloor$. Let $\mathcal{C}$ be the set of all classes C_i where $n_s > \lfloor \frac{n_i}{l} \rfloor$. Let $\mathcal{C}'$ be the set of the remaining classes. Let p be the total number of classes in $\mathcal{C}$. Since the classes are sorted, $\mathcal{C} = \{C_1, C_2, ..., C_p\}$ and $\mathcal{C}' = \{C_{p+1}, C_{p+1}, ..., C_u\}$.

Lemma 1. *If a set of classes $\mathcal{X} = \{C_1, ...C_u\}$ are generalized to their parent class in $\mathcal{T}$, the adversary can deduce that at least one class (in the original table) violates l-diversity among $\mathcal{C}$ and all classes in $\mathcal{C}'$ (in the original table) do not violate l-diversity.*

Obviously, the credibility of individuals in a class in $\mathcal{C}'$ is smaller than or equal to $\frac{1}{l}$. However, the credibility of individuals in a class in $\mathcal{C}$ may be greater than $\frac{1}{l}$. Thus, the adversary tries to compute $Credibility(o, s, K^{min}_{ad})$, where $o \in C_i$, for $i = 1, 2, ..., p$. Suppose there are j tuples with the sensitive value set s in C_i. Let $|C_i(s)|$ denote the number of occurrences of the tuples with s in C_i. The probability that o is linked to a sensitive value set is $\frac{j}{n_i}$, where n_i is the class size of C_i. Let $Prob(|C_i(s)| = j|K^{min}_{ad})$ be the probability that there are exactly j occurrences of tuples with s in C_i given K^{min}_{ad}. By considering all possible number j of occurrences of tuples with s from 1 to n_i in C_i, the general formula for credibility is given by:

$$\begin{aligned}&Credibility(o, s, K^{min}_{ad}), \text{ where } o \in C_i, 1 \leq i \leq p\\ = \ &\text{Prob}(o \text{ is linked to } s \text{ in } C_i \mid K^{min}_{ad})\\ = \ &\textstyle\sum_{j=1}^{n_i} Prob(|C_i(s)| = j \mid K^{min}_{ad}) \times \dfrac{j}{n_i}\end{aligned}$$

In the above formula, $Prob(|C_i(s)| = j|K^{min}_{ad})$ can be calculated by considering all possible cases. Conceptually, a table such as Table 6 will be constructed, in which some possible combinations will be excluded due to the minimality notion in K^{min}_{ad}.

QID	Disease
$q1$	HIV
$q1$	HIV
$q1$	non-sensitive
$q1$	non-sensitive
$q1$	HIV
$q2$	non-sensitive
$q2$	non-sensitive
...	...
$q2$	non-sensitive
$q2$	HIV

Table 7: Another table which violates 2-diversity

QID	Disease
$q1$	HIV
$q1$	HIV
$q1$	non-sensitive
$q1$	non-sensitive
Q	HIV
Q	non-sensitive
$q2$	non-sensitive
...	...
$q2$	non-sensitive
$q2$	HIV

Table 8: A 2-diverse table of Table 7 by local recoding

4.2 Local Recoding

An example is shown in Table 7 to illustrate the derivation of the credibility with local recoding for l-diversity. For the QID, assume that only $q1$ and $q2$ can be generalized to Q. Assume that Table 7 and the corresponding T^e have matching cardinality on Q. The proportion of the sensitive tuples in the set of tuples with $q1$ is equal to $3/5 > 1/2$. Thus, the set of tuples with $q1$ does not satisfy 2-diversity. Table 7 is generalized to Table 8, which satisfies 2-diversity, while the distortion is minimized.

Assume the adversary has knowledge of K_{ad}^{min}. From the external table T^e, there are 5 tuples with $q1$ and 8 tuples with $q2$. These are the only tuples with QID that can be generalized to Q. The adversary reasons in this way. There are four sensitive tuples in T^*. Suppose they all appear in the tuples containing $q2$, $q2$ still satisfies 2-diversity. The generalization in T^* *must* be caused by the set of tuples in $q1$. In T^*, the QID-EC for Q contains one sensitive tuple and one non-sensitive tuple. The sensitive tuple should come from $q1$ because if this sensitive tuple does not come from $q1$, there will have been no need for the generalization.

Consider the credibility that an individual o with QID $q1$ is linked to HIV given K_{ad}. There are two cases, too.

In the first case, the tuple of o appears in the QID-EC of $q1$ in T^*. There are four tuples with value $q1$ in T^*. From T^e, there are five tuples with $q1$. The probability that Case 1 occurs is $4/5$.

In the second case, the tuple of o appears in the QID-EC of Q in T^*. There are totally five tuples with $q1$ and there are four tuples with value $q1$ in T^*. Hence, one such tuple must have been generalized and is now in the QID-EC of Q in T^*. The probability of Case 2 is $1/5$.

$Credibility(o, \{HIV\}, K_{ad}^{min})$ is equal to

$$\begin{aligned} = \; & Prob(\text{Case 1}) \times \; Prob(o \text{ is linked to HIV in Case 1 } | K_{ad}^{min}) \\ & + Prob(\text{Case 2}) \times Prob(o \text{ is linked to HIV in Case 2 } | K_{ad}^{min}) \end{aligned}$$

Since 2 out of 4 tuples in the QID-EC of $q1$ in T^* contain HIV, and the HIV tuple in the QID-EC of Q in T^* must be from $q1$, Thus,

$$Prob(o \text{ is linked to HIV in Case 1 } | \; K_{ad}^{min}) = \tfrac{2}{4} = \tfrac{1}{2}.$$

$$Prob(o \text{ is linked to HIV in Case 2 } | \; K_{ad}^{min}) = 1.$$

$$Credibility(o, \{HIV\}, K_{ad}^{min}) = \tfrac{4}{5} \times \tfrac{1}{2} + \tfrac{1}{5} \times 1 = \tfrac{3}{5},$$

which is greater than 0.5. Thus, the anonymized table violates 2-confidentiality.

General Formula

Suppose there are u QID-EC's in the original data set, namely $C_1, C_2, ..., C_u$, which can be generalzied to the same value $\mathcal{C_G}$. After the generalization, some tuples in some C_i are generalized to $\mathcal{C_G}$ while some are not. We define the following symbols which will be used in the derivation of the credibility.

n_i number of tuples with class C_i in T^e

$n_{i,g}$ number of generalized tuples in T^* whose original QID is C_i

$n_{i,u}$ number of ungeneralized tuples in T^* with QID $= C_i$

$n_{i,u(s)}$ number of sensitive ungeneralized tuples in T^* with QID $= C_i$

The value of $n_{i,u}$ can be easily obtained by scanning the tuples in T^*. $n_{i,g}$ can be obtained by subtracting $n_{i,u}$ from n_i. Similarly, it is easy to find $n_{i,u(s)}$. For example, in Table 8, C_i corresponds to $q1$ and $\mathcal{C_G}$ corresponds to Q. Thus, $n_{i,u} = 4$, $n_i = 5$, $n_{i,g} = 1$ and $n_{i,u(s)} = 2$.

In order to calculate $Credibility(o, s, K_{ad}^{min})$, where o has QID of C_i, the adversary needs to consider two cases. The first case is that the tuple of o is generalized to $\mathcal{C_G}$. The second case is that the tuple of o is not generalized in T^*. Let $t^*(o)$ be the tuple of individual o in T^*. By considering these two cases,

$$\begin{aligned} & Credibility(o, s, K_{ad}^{min}), \text{ where } o \in C_i \\ = \; & Prob(o \text{ is linked to } s \text{ in } T^* | K_{ad}^{min}) \\ = \; & Prob(t^*(o) \in \mathcal{C_G} \text{ in } T^*) \\ & \quad \times Prob(o \text{ is linked to } s \text{ in } \mathcal{C_G} \text{ in } T^* | K_{ad}^{min}) \\ & + Prob(t^*(o) \in C_i \text{ in } T^*) \\ & \quad \times Prob(o \text{ is linked to } s \text{ in } C_i \text{ in } T^* | K_{ad}^{min}) \\ = \; & \frac{n_{i,g}}{n_i} \times Prob(o \text{ is linked to } s \text{ in } \mathcal{C_G} \text{ in } T^* | K_{ad}^{min}) \\ & + \frac{n_{i,u}}{n_i} \times \frac{n_{i,u(s)}}{n_{i,u}} \end{aligned}$$

The term $Prob(o \text{ is linked to } s \text{ in } \mathcal{C_G} \text{ in } T^* | K_{ad}^{min})$ can be computed by using the formula in global-recoding, which takes into account of the minimality of the anonymization.

For the case when a set of QID-EC's are generalized to more than one values, the above analysis is extended to include more possible combinations of outcomes. Details can be found in [21]. The basic ideas remain similar.

4.3 Attack Conditions

We have seen in the above that a minimality attack is always accompanied by some exclusion of some possibilities by the adversary because of the minimality notion. We can characterize this attack criterion in the following.

THEOREM 1. *An attack by minimality is possible only if the adversary can exclude some possible combinations of the number of sensitive tuples among the QID-EC's in the sensitive tuple distribution table based on the knowledge of* K_{ad}^{min}.

Proof sketch. If there is no exclusion from the table, then the credibility as computed by the formulae is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. □

An attack by minimality is not always successful even when there are some excluded combination(s) in the sensitive tuple distribution table based on K_{ad}^{min}. To illustrate, consider

QID	Disease
$q1$	Diabetics
$q1$	HIV
$q1$	Lung Cancer
$q2$	HIV
$q2$	Ulcer
$q2$	Alzhema
$q2$	Gallstones

(a) good table

QID	Disease
$q1$	Diabetics
$q1$	HIV
$q1$	HIV
$q2$	Lung Cancer
$q2$	Ulcer
$q2$	Alzhema
$q2$	Gallstones

(b) bad table

QID
Q
Q
Q
Q
Q
Q
Q

(c) global

QID
Q
Q
Q
Q
$q2$
$q2$
$q2$

(d) local

Table 9: Anonymization for (3,3)-diversity

an example where 2 QID's $q1$ and $q2$ are generalized to Q. There are 4 tuples of $q1$ and 2 tuples of $q2$. In total, there are 3 occurrences of the sensitive value set s in the 6 tuples. If 2-diversity is the goal, then we can exclude the case of 2 sensitive $q1$ tuple and 1 sensitive $q2$ tuple. After the exclusion, the credibility of any linkage between any individual to s still does not exceed 0.5.

5. GENERAL MODEL

In this section, we show that minimality attacks can be successful on a variety of anonymization models. In Tables 9 to 11, we show *good tables* that satisfy the corresponding privacy requirements in different models, *bad tables* that do not, and *global* and *local* recodings of the bad tables which follow the minimality principle and unfortunately suffer from minimality attacks.

Recursive (c, l)-diversity: With recursive (c, l)-diversity [12], in each QID-EC, let v be the most frequent sensitive value, if we remove the next $l - 2$ most frequent sensitive values, the frequency of v must be less than c times the total count of the remaining values. Table 9(c) is a global recoding for Table 9(b). With the knowledge of minimality in the anonymization, the adversary deduces that the QID-EC for $q2$ must satisfy (3, 3)-diversity and that the QID-EC for $q1$ must contain two HIV values. Thus, the intended obligation that an individual should be linked to at least 3 different sensitive values is breached. Similar arguments can be applied to Table (d).

t-closeness: Recently, t-closeness [11] was proposed. If table T satisfies t-closeness, the distribution $\mathbb{P}$ of each equivalence class in T is roughly equal to the distribution $\mathbb{Q}$ of the whole table T with respect to the sensitive attribute. More specifically, the difference between the distribution of each equivalence class in T and the distribution of the whole table T, denoted by $D[\mathbb{P}, \mathbb{Q}]$, is at most t. Let us use the definition in [11]: $D[\mathbb{P}, \mathbb{Q}] = 1/2 \sum_{i=1}^{m} |p_i - q_i|$. Consider Table 10(c). For each possible sensitive value distribution $\mathbb{P}$ for QID-EC $q2$, the adversary computes $D[\mathbb{P}, \mathbb{Q}]$. S/he finds that $D[\mathbb{P}, \mathbb{Q}]$ is always smaller than 0.2. Hence the anonymization is due to $q1$. S/he concludes that both tuples with QID=$q1$ are sensitive. Similar arguments can also be made to Table (d).

(k, e)-anonymity: The model of (k, e)-anonymity [27] considers the anonymization of tables with numeric sensitive attributes. It generates a table where each equivalence class is of size at least k and has a range of the sensitive values at least e. In the tables in Table 11, we show the bucketization in terms of QID values, the individuals with the same QID value are in the same bucket. Consider the tables in Table 11 (where Income is a sensitive numeric attribute). From Table (c), the adversary deduces that the tuples with QID=$q1$ must violate (k, e)-anonymity and must be linked with two $30k$ incomes. We obtain a similar conclusion from Table (d) for local recoding.

QID	Disease
$q1$	HIV
$q1$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	HIV
$q2$	HIV

(a) good table

QID	Disease
$q1$	HIV
$q1$	HIV
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	non-sensitive
$q2$	HIV

(b) bad table

QID
Q
Q
Q
Q
Q
Q

(c) global

QID
Q
Q
Q
$q2$
$q2$
$q2$

(d) local

Table 10: 0.2-closeness anonymization

QID	Income
$q1$	30k
$q1$	20k
$q2$	30k
$q2$	20k
$q2$	40k

(a) good table

QID	Income
$q1$	30k
$q1$	30k
$q2$	20k
$q2$	10k
$q2$	40k

(b) bad table

QID
Q
Q
Q
Q
Q

(c) global

QID
Q
Q
Q
$q2$
$q2$

(d) local

Table 11: (k, e)-anonymity for $k = 2$ and $e = 5k$

We also have examples to show the feasibility of minimality attacks on the algorithms for (c, k)-safety in [13], Personalized Privacy in [24], and sequential releases in [18] and [25]. In the proposed anonymization mechanism for each of the above cases in the respective references, the Minimality Principle in Definition 3 holds if we set $\mathcal{R}$ to the objective at hand, such as recursive (c, l)-diversity, t-closeness and (k, e)-anonymity. By including the knowledge related to minimality attack to the background knowledge, the adversary can derive the probabilistic formulae for computing the corresponding credibility in each case, where the idea of eliminating impossible cases as shown in Section 4 is a key to the attack.

6. ALGORITHM

The problem of optimal m-confidentiality is a difficult problem. In most data anonymization methods, if a generalization step does not reach the privacy goal, further generalization can help. However, further generalizations will not solve the problem of m-confidentiality. If we further generalize Q to $*$ in Table 1(c) or further generalize $q2$ to Q in Table 1(d), it does not deter the minimality attack. The result still reveals the linkage of $q1$ to HIV as before. We show below optimal m-confidentiality is NP-hard for global recoding.

Optimal global m-confidentiality: Given a table T and a non-negative cost e, can we generate a table T^* from T by global recoding which satisfies m-confidentiality and where the information loss of $Dist(T, T^*)$ is less than or equal to e?

THEOREM 2. *Optimal m-confidentiality under global recoding is NP-hard.*

Limited by space, we leave the proof in [21].

However, as the adversary relies on the minimality assumption, we can tackle the problem at its source by removing the minimality notion from the anonymization. The main idea is that, even if some QID-EC's in a given table T originally do not violate l-diversity, we can still generalize the QID. Since the anonymization does not play according to the minimality rule, the adversary cannot launch the minimality attack directly. However, a question is: how much shall we generalize or anonymize? It is not desirable to lose on data utility.

A naive method to generalize everything in an excessive manner would not work well, since the information loss will also be excessively large. From the formula for information loss, if every QID attribute value must go at least one level up the taxonomies, then for typical taxonomies, the information loss will be a sizeable fraction.

Here we propose a feasible solution for the m-confidentiality problem. Although some problems are uncovered that questions the utility of k-anonymity in protecting sensitive values, k-anonymity has been successful in some practical applications. This indicates that when a data set is k-anonymized for a given k, the chance of a large proportion of a sensitive value set s in any QID-EC is very likely reduced to a safe level. Since k-anonymity does not try to anonymize based on the sensitive value set, it will anonymize a QID-EC even if it satisfies l-diversity. This is the blinding effect we are targeting for. However, there is no guarantee of m-confidentiality by k-anonymity alone, where $m = l$.

Hence, our solution is based on k-anonymity, with additional precaution steps taken to ensure m-confidentiality. Let us call our solution Algorithm MASK (Minimality Attack Safe K-anonymity), which involves four steps.

Algorithm 1 – MASK

```
1: From the given table $T$, generate a $k$-anonymous table $T^k$
   where $k$ is a user parameter.
2: From $T^k$, determine the set $\mathcal{V}$ containing all QID-EC's which
   violate $l$-diversity in $T^k$, and a set $\mathcal{L}$ containing QID-EC's
   which satisfy $l$-diversity in $T^k$. How to select $\mathcal{L}$ will be
   described below.
3: For each QID-EC $Q_i$ in $\mathcal{L}$, find the proportion $p_i$ of tuples
   containing values in the sensitive value set $s$. The distribution
   $\mathcal{D}$ of the $p_i$ values is determined.
4: For each QID-EC $E \in \mathcal{V}$, randomly pick a value of $p_E$ from
   the distribution $\mathcal{D}$. The sensitive values in $E$ are distorted in
   such a way that the resulting proportion of the sensitive value
   set $s$ in $E$ is equal to $p_E$.
```

Step 1 anonymizes a given table to satisfy k-anonymity. After Step 1, some QID-EC's may not satisfy l-diversity. Steps 2 to 4 ensure that all QID-EC's in the result are l-diverse. In Step 2, we select a QID-EC set $\mathcal{L}$ from T^k. The purpose is to disguise the distortion so that the adversary cannot tell the difference between a distorted QID-EC and many undistorted QID-EC's. We set the size of $\mathcal{L}$, denoted by u, to $(l-1) \times |\mathcal{V}|$. Among all the QID-EC's in T^k that satisfies l-diversity, we pick u QID-EC's with the highest proportions of the sensitive value set s.

THEOREM 3. *Algorithm MASK generates m-confidential data sets.*

The above holds because MASK does not follow the minimality principle. It is easy to find an l-diverse table T^* generated by MASK with a QID-EC X in T^* so that a specialization of the QID's in X results in another table T' which also satisfies l-diversity.

The use of $\mathcal{L}$ for the distortion of $\mathcal{V}$ is to make the distribution of s proportions in $\mathcal{V}$ look indistinguishable from that of a large QID-EC set ($\mathcal{L}$). This is an extra safeguard for the algorithm in case the adversary knows the mechanism of anonymization. If the QID-EC's in $\mathcal{V}$ simply copy the s proportion from an l-diverse QID-EC in T_k with the greatest s proportion, the repeated pattern may become a source of attack. In our setting, the probability that some QID-EC in $\mathcal{V}$ has the same s proportion as a QID-EC in $\mathcal{L}$ is $1/l$. Therefore, for l repeated occurrences of an s proportion, the probability that any one belongs to a QID-EC in $\mathcal{V}$ is only $1/l (= 1/m)$.

	Attribute	Distinct Values	Generalizations	Height
1	Age	74	5-, 10-, 20-year ranges	4
2	Work Class	7	Taxonomy Tree	3
3	Martial Status	7	Taxonomy Tree	3
4	Occupation	14	Taxonomy Tree	2
5	Race	5	Taxonomy Tree	2
6	Sex	2	Suppression	1
7	Native Country	41	Taxonomy Tree	3
8	Salary Class	2	Suppression	1
9	Education	16	Taxonomy Tree	4

Table 12: Description of Adult Data Set

Generation of Two Tables - Bucketization

Conventional anonymization methods produce a single generalized table T as shown in Table 5. Recently [23] proposed to generate two separate tables from T with the introduction of an attribute called GID that is shared by the two tables. The first table T_{QID} contains the attributes of QID and GID, and the second table T_{sen} contains GID and the sensitive attribute(s). The two tables are created from T^* by assigning each QID-EC in T^* a unique GID. The advantage is that we can keep the original values in T of the QID in T_{QID} and hence reduce information loss. However, the single table T has the advantage of clarity and requiring no extra interpretation on the data receiver's part. In our experiments, we shall try both the approach of generating a single table T and the approach of generating two tables (also known as bucketization) as in [23, 27, 13].

7. EMPIRICAL STUDY

A Pentium IV 2.2GHz PC with 1GM RAM was used to conduct our experiment. The algorithm was implemented in C/C++. In our experiment, we adopted the publicly available data set, Adult Database from the UCIrvine Machine Learning Repository [2]. This data set (5.5MB) was also adopted by [10, 12, 20, 6]. We used a configuration similar to [10, 12]. The records with unknown values were first eliminated resulting in a data set with 45,222 tuples (5.4MB). Nine attributes were chosen in our experiment, as shown in Table 12. By default, we chose the first eight attributes and the last attribute in Table 12 as the quasi-identifer and the sensitive attribute, respectively. As discussed in the previous sections, attribute "Education" contains a sensitive value set containing all values representing the education levels before "secondary" (or "9th-10th") such as "1st-4th", "5th-6th" and "7th-8th".

7.1 Analysis of the minimality attack

We are interested to know how successful the minimality attack can be in a real data set with existing minimality-based anonymization algorithms. We adopted the Adult data set and the selected algorithm was the (α, k)-anonymity algorithm [22]. We set $\alpha = 1/l$ and $k = 1$, so that it corresponds to the simplified l-diversity. We have implemented an algorithm based on the general formulae in Section 4 to compute the credibility values. We found that minimality attack successfully uncovered QID-EC's which violates m-confidentiality, where $m = l$. We use m and l exchangeably in the following. Let us call the tuples in such QID-EC's the *problematic tuples*. Figure 2(a) shows the proportion of problematic tuples among all sensitive tuples under the variation of m, where the total number of sensitive tuples is 1,566. The general trend is that the proportion increases when m increases. When m increases, there is higher chance that problematic tuples are generalized with more generalized tuples. Also, it is more likely that those generalized tuples are easily uncovered for the minimality attack.

In Figure 2(b), when m increases, it is obvious that the average credibility of problematic tuples decreases. When m increases, $1/m$ decreases. Thus, each QID-EC contains at most $1/m$ occurrences of the sensitive value set. Thus, this lowers the credibility of the tuples in QID-ECs.

Figure 2(c) shows that the proportion of problematic tuples increases with QID size. This is because, when QID size is larger, the size of each QID-EC is smaller. It is more likely that a QID-EC violates the privacy requirement. Thus, more tuples are vulnerable for the minimality attack. Figure 2(d) shows that the average credibility of problematic tuples remain nearly unchanged when the QID size increases. This is because the credibility is based on m. It is noted that the average credibility in Figure 2(d) is about 0.9, which is greater than 0.5 (=1/2).

We also examined some cases obtained in the experiment. Suppose we adopt the QID attributes as (age, workclass, martial status) with sensitive attribute Education. The original table contains one tuple with QID=(80, self-emp-not-inc, married-spouse-absent) and two tuples with QID=(80, private, married-spouse-absent).

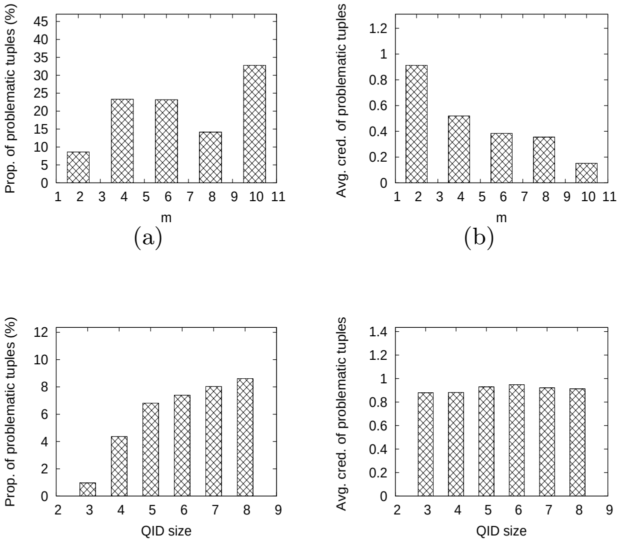
Age	Workclass	Martial Status	Education
80	self-emp-not-inc	married-spouse-absent	7th-8th
80	private	married-spouse-absent	HS-grad
80	private	married-spouse-absent	HS-grad

Suppose $m = 2$. Recall that "7th-8th" is in the sensitive value set. Since the first tuple violates 2-diversity, the Workclass of tuple 1 and tuple 2 are generalized to "with-pay". In this case, it is easy to check that the credibility for an individual with QID= (80, self-emp-not-inc, married-spouse-absent) is equal to 1.

Another uncovered case involves more tuples. The original table contains one tuple with QID=(33, self-emp-not-inc, married-spouse-absent) and 17 tuples with QID=(33, private, married-spouse-absent).

Similarly, when $m = 2$, the first tuple violates 2-diversity. Thus, Workclass of tuple 1 and tuple 2 are generalized to "with-pay" in the published table. Similarly, the adversary can deduce that the individual with QID=(33, self-emp-not-inc, married-spouse-absent) is linked with a low education (i.e., Education="1st-4th") since this credibility is equal to 1.

Consider the default QID size = 8. When $m = 2$, the execution time of the computation of the credibility of each QID-ECs in the original table is about 173s. When $m = 10$, the execution time is 239s. It is not costly for an adversary to launch a minimality attack.



(a) (b)

(c) m=2 (d) m=2

Figure 2: Proportion of problematic tuples and average credibility of problematic tuples against m and QID size

7.2 Analysis of the proposed algorithm

We compared our proposed algorithm with a local recoding algorithm for (α, k)-anonymity [22] ((α, k)-A). Let us refer to our proposed algorithm MASK described in Section 6 by m-$conf$. (α, k)-A does not guarantee m-confidentiality, but it is suitable for comparison since it considers both k-anonymity and l-diversity, where $l = m$. We are therefore interested to know the overhead required in our approach in order to achieve m-confidentiality. When we compared our algorithm with (α, k)-anonymity, we set $\alpha = 1/m$ and the k value is the same as that use in our algorithm. We evaluated the algorithms in terms of four measurements: *execution time*, *relative error ratio*, *information loss* of QID attributes and *distortion* of sensitive attribute. The distortion of sensitive attribute is calculated by the information loss formula in Definition 6. We give it a different name for the ease of reference. By default, the weighting of each attribute used in the evaluation of information loss is equal to $1/|QID|$, where $|QID|$ is the QID size. For each measurement, we conducted the experiments 100 times and took the average.

We have implemented two different versions of Algorithm MARK: (A) one generalized table is generated and (B) two tables are generated (see the last paragraph in Section 6). For Case (A), we may generalize the QID attributes of the data and distort the sensitive attribute of the data. Thus, we measured these by information loss and distortion, respectively. For Case (B), since the resulting tables do not generalize QID, there is no information loss for QID. The distortion of the sensitive attribute is the same as in Case (A). Hence in the evaluation of information loss and distortion, we only report the results for Case (A).

For case (B) with the generation of two ungeneralized tables, T_{QID} and T_{sen}, as in [23], we measure the error by the *relative error ratio* in answering a aggregate query. We adopt both the form of the aggregate query and the parameters of the *query dimensionality qd* and the *expected query selectivity s* from [23]. For each evaluation in the case of two anonymized tables, we performed 10,000 queries and then reported the average relative error ratio. By default, we set $s = 0.05$ and qd to be the QID size.

We conducted the experiments by varying the following factors: (1) the QID size, (2) m, (3) k, (4) query dimensionality qd (in the case of two anonyzmied tables), and (5) selectivity s (in the case of two anonymized tables).

7.2.1 The single table approach

The results for the single table case are shown in Figure 3 and Figure 4. One important observation is that the results are little affected by the values of k which varies from 2 to 10 to 20, this is true for the execution time, the relative error, the information loss and the distortion. This is important since k is a user parameter and the results indicate that the performance is robust against different choices of the value of k.

A second interesting observation is that the information loss of (α, k)-A is greater than m-$conf$ in some cases. This seems surprising since m-$conf$ has to fend off minimality attack while (α, k)-A does not. The explanation is that in some cases, more generalization is required in (α, k)-A to satisfy l-diversity. However, the first step of m-$conf$ only considers k-anonymity and not l-diversity. Thus, the generalization in m-$conf$ is less compared to (α, k)-A, leading to less information loss. For compensation, the last two steps of m-$conf$ ensure l-diversity and incur distortion, while (α, k)-A has no such steps.

The execution times of the two algorithms are similar because the first step of m-$conf$ occupies over 98% of the execution time on average and the first step is similar to (α, k)-A.

In Figure 3(a), the execution time increases with the QID size, since greater QID size results in more QID-EC's. When k is larger, the execution time is smaller, this is because the number of QID-EC's will be smaller.

Figures 3(b) and (d) show that the average relative error and the distortion of the algorithms increase with the QID size. This is because the number of QID-EC's increases and the average size of each equivalence class decreases. For m-$conf$, the probability that a QID-EC violates l-diversity (after the k-anonymization step) will be higher. Thus, there is a higher chance for the distortion and higher average relative error. When k is larger, the average relative error of the two algorithms increases. This is because the QID attribute will be generalized more, giving rise to more querying errors. If k is larger, the QID-EC size increases, the chance that a QID-EC violates l-diversity is smaller, so the distortion will be less.

In Figure 3(c), when the QID size increases, the information loss of the QID attributes increases since the probability that the tuples in the original table have different QID values is larger. Thus, there is a higher chance for QID generalization leading to more information loss. Similarly, when k is larger, the information loss is larger.

7.2.2 The two tables approach

Our next set of experiments analyze the performance of the two table approach under various conditions.

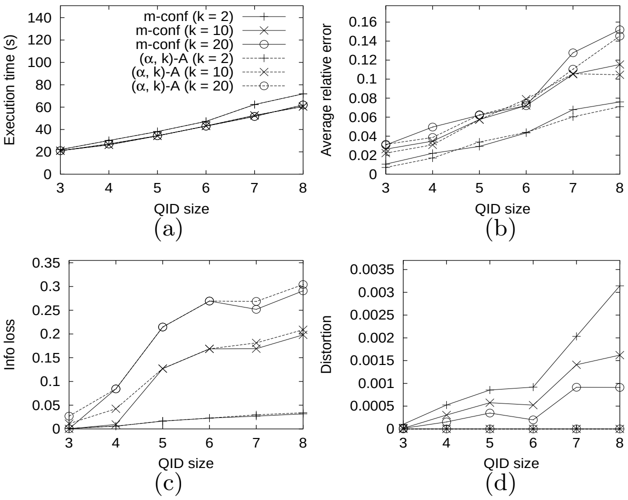


Figure 3: Performance vs QID size ($m = 2$)

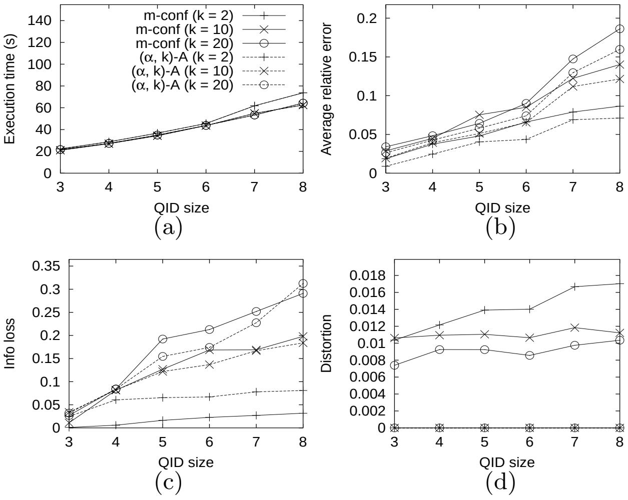


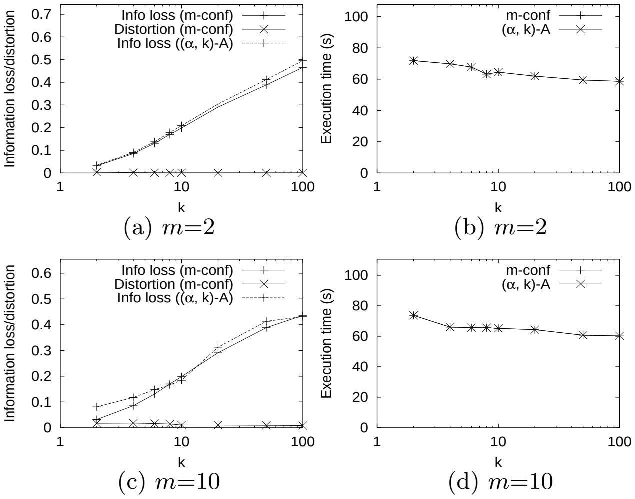
Figure 4: Performance vs QID size ($m = 10$)

Effect of k: Figure 5 shows the experimental results when k is varied. The trends are similar to the single table case, and can be explained similarly.

Effect of Query Dimensionality qd: For $m = 2$, Figure 6(a) shows the average relative error increases when the query dimensionality increases. As the query will match fewer tuples, fewer tuples in an equivalence class will match the query, resulting in more relative error. If k is larger, the average relative error is larger because we generalize more data with larger k. Similar trends can also be observed when $m = 10$.

Effect of Selectivity s: In Figure 6(c), the average relative error decreases when s increases. This is because, if s is larger, more tuples will be matched with a given query, and more tuples in an equivalence class is matched with a given query. Similarly, when k is larger, there is more generalization, and the average relative error is larger. We observe similar trends when m=10. Similarly, the average relative error is larger when m=10.

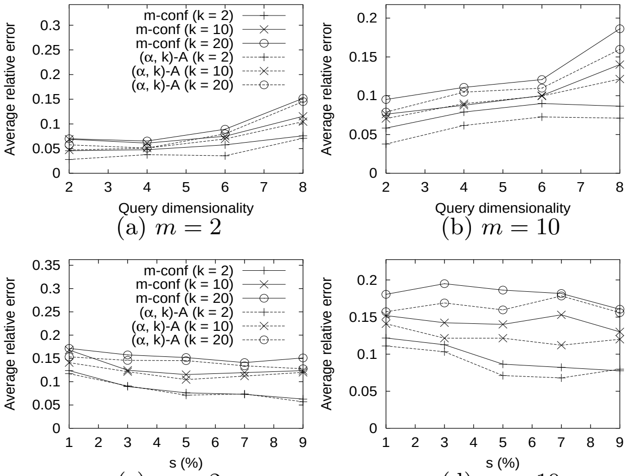
In conclusion, we find that our algorithm creates very little



(a) m=2 (b) m=2

(c) m=10 (d) m=10

Figure 5: Two Tables case : effect of varying m and k



(a) $m = 2$ (b) $m = 10$

(c) $m = 2$ (d) $m = 10$

Figure 6: Two Tables Case : effects of varying query dimensionality and selectivity

overhead and pays a very minimal price in information loss in the exchange for m-confidentiality.

8. CONCLUSIONS

In existing privacy preservation methods for data publishing, minimality in information loss is an underlying principle. In this paper, we show how this can be used by an adversary to launch an attack on the published data. We call this a minimality attack. We propose the m-confidentiality model which deals with attack by minimality and also a solution for this problem. For future work we are interested in determining any other kinds of attacks related to the nature of the anonymization process.

9. REFERENCES

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

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

[3] D. Brumley and D. Boneh. Remote timing attacks are practical. In *USENIX Security Symposium*, 2003.

[4] Alexandre Evfimievski, Ramakrishnan Srikant, and Johannes Gehrke Rakesh Agrawal. Privacy preserving mining of association rules. In *KDD*, 2002.

[5] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*. Morgan Kaufmann, 1993.

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

[7] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, 2006.

[8] Paul C. Kocher. Timing attacks on implementations of Diffe-Hellman RSA, DSS, and other systems. In *CRYPTO*, pages 104–113, 1996.

[9] K. LeFevre, D. DeWitt, , and R. Ramakrishnan. Multidimensional k-anonymity. In *M. Technical Report 1521, University of Wisconsin*, 2005.

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

[11] N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.

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

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

[14] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228, 2004.

[15] Ramakrishnan Srikant Rakesh Agrawal. Privacy-preserving data mining. In *SIGMOD*, 2000.

[16] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International journal on uncertainty, Fuzziness and knowldege based systems*, 10(5):571 – 588, 2002.

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

[18] K. Wang and B. Fung. Anonymizing sequential releases. In *SIGKDD*, 2006.

[19] K. Wang, B. C. M. Fung, and P. S. Yu. Handicapping attacker's confidence: An alternative to k-anonymization. In *Knowledge and Information Systems: An International Journal*, 2006.

[20] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM*, pages 249–256, 2004.

[21] R.C.W. Wong, A. Fu, A. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *Technical Report, Chinese University of Hong Kong*, 2007.

[22] R.C.W. Wong, J. Li, A. Fu, and K. Wang. (alpha, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *SIGKDD*, 2006.

[23] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, 2006.

[24] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, 2006.

[25] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, 2007.

[26] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. Fu. Utility-based anonymization using local recoding. In *SIGKDD*, 2006.

[27] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on aononymized tables. In *ICDE*, 2007.