Anonymization-Based Attacks in Privacy-Preserving Data Publishing

RAYMOND CHI-WING WONG
The Hong Kong University of Science and Technology
ADA WAI-CHEE FU
The Chinese University of Hong Kong
and
KE WANG and JIAN PEI
Simon Fraser University

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 article, 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 article, 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.

Categories and Subject Descriptors: H.2.0 [**Database Management**]: General—*Security, integrity, and protection*; H.2.8 [**Database Management**]: Database Applications—*Data mining*; K.4.1 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms: Security, Algorithms, Experimentation

R. C.-W. Wong was supported in part by the HKSAR RGC Direct Allocation Grant DAG08/09.EG01. J. Pei was supported in Part by an NSERC Discovery grant and an NSERC Discover Accelerator Supplements grant. All opinions, findings, conclusions and recommendations in this article are those of the authors and do not necessarily reflect the reviews of the funding agencies.

Authors' addresses: R. C.-W. Wong, Computer Science and Engineering Department, the Hong Kong University of Science and Technology, Hong Kong; email: raywong@cse.ust.hk; A. W.-C. Fu, Computer Science and Engineering Department, the Chinese University of Hong Kong, Hong Kong; email: adafu@cse.cuhk.edu.hk; K. Wang, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada; email: wangk@cs.sfu.ca; J. Pei, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada; email: jpei@cs.sfu.ca.

DOI 10.1145/1538909.1538910 http://doi.acm.org/10.1145/1538909.1538910

Additional Key Words and Phrases: Privacy preservation, data publishing, k-anonymity, l-diversity, minimality attack

ACM Reference Format:

Wong, R. C.-W., Fu, A. W.-C., Wang, K., and Pei, J. 2009. Anonymization-based attacks in privacy-preserving data publishing. ACM Trans. Database Syst. 34, 2, Article 8 (June 2009), 46 pages. DOI = 10.1145/1538909.1538910 http://doi.acm.org/10.1145/1538909.1538910.

1. INTRODUCTION

Although data mining is potentially useful, many data holders are reluctant to provide their data for data mining for fear of violating individual privacy. In recent years, study has been made to ensure that the sensitive information of individuals cannot be identified easily [Samarati 2001; Sweeney 2002a, 2002b; LeFevre et al. 2006, 2005; Meyerson and Williams 2004]. One well-studied approach is the k-anonymity model [Ciriani et al. 2007] which in turn led to other models such as confidence bounding [Wang et al. 2006], l-diversity [Machanavajjhala et al. 2006], (α, k)-anonymity [Wong et al. 2006], t-closeness [Li and Li 2007], (k, e)-anonymity [Zhang et al. 2007], and (c, k)-safety [Martin et al. 2007]. These models assume that the data or table T contains: (1) a *quasi-identifer (QID)*, which is a set of attributes (e.g., a QID may be {Date of birth, Zipcode, Sex}) in T which can be used to identify an individual, and (2) *sensitive attributes*, attributes in T 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*, which we call QID-EC. The table T is said to satisfy *k-anonymity* if the size of every equivalence class is greater than or equal to k. The intuition of k-anonymity is to make sure that each individual is indistinguishable from other $k - 1$ individuals.

In this article, we study the case where the adversary has some additional knowledge about the mechanism involved in the anonymization and launches an attack based on this knowledge. We focus on the protection of the relationship between the quasi-identifier and a single sensitive attribute. In a simplified setting of l-diversity model [Machanavajjhala et al. 2006], 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. The intended objective is to make sure that an adversary cannot deduce with a probability above $1/l$ that an individual is linked to any sensitive value. In the following discussion, when we refer to l-diversity, we refer to this simplified setting.[1] The complex l-diversity model is discussed in Section 5, in which we show that our results can be extended to other anonymization models.

[1]This simplified model is a special case of the confidence bounding in Wang et al. [2006] and is the same as (α, k)-anonymity [Wong et al. 2006] when $k = 1$ and $\alpha = 1/l$.

Table I. 2-Diversity: Global and Local Recoding

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	Disease
Q	HIV
Q	HIV
Q	non-sensitive
Q	non-sensitive
Q	non-sensitive
Q	non-sensitive
Q	non-sensitive

(c) global

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

(d) local

1.1 Minimality Attack

In Table I(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 LeFevre et al. [2005], existing approaches of anonymization for data publishing have 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 I(a) will not be generalized.[2] In fact the aforesaid 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 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 I(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 I(c) by *global recoding* [Xiao and Tao 2006b; Wang and Fung 2006]. In global recoding, all occurrences of an attribute value are recoded to the same value. If *local recoding* [Sweeney 2002a; Aggarwal et al. 2005a, 2005b] is adopted, occurrences of the same value of an attribute may be recoded to different values. Such an anonymization is shown in Table I(d). These anonymized tables satisfy 2-diversity. The question we are interested in is whether these tables really protect individual privacy.

In most previous works [Sweeney 2002b; LeFevre et al. 2006, 2005; Xiao and Tao 2006b], the knowledge of the adversary involves an external table T^e

[2]This is the case for each of the anonymization algorithms in Machanavajjhala et al. [2006], Wang et al. [2006], and Wong et al. [2006].

Table II. T^e: External Table Available to the Adversary

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

such as a voter registration list that maps QIDs to individuals.[3] As in many previous works, 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 II(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 II(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 I(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".

In some previous works, 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 II(c) shows such a case, where there is no tuple for Tim and Joseph in Table I(a) and Table I(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 tuples with QID $= q4$ in T^e do not have any effect on the previous 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*.

[3]There are many sources of such an external table T^e. Most municipalities sell population registers that include the identifiers of individuals along with basic demographics; examples include local census data, voter lists, city directories, and information from motor vehicle agencies, tax assessors, and real estate agencies [Samarati and Sweeney 1998]. In the voter list, 97% of the voters were identifiable with just the full postal code and birth date [Sweeney 1997]. From Sweeney [2002b], it is reported that a city's voter list in two diskettes was purchased for twenty dollars, and was used to reidentify medical records.

Table III. 2-Diversity (where all values in Disease are sensitive): Global and Local Recoding

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	Disease
Q	HIV
Q	HIV
Q	Gallstones
Q	Lung Cancer
Q	Ulcer
Q	Alzheimer
Q	Diabetes
$q4$	Ulcer
$q4$	Alzheimer

(c) global

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

(d) local

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 preceding 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 III, we assume that all values for Disease are sensitive. Table III(a) satisfies 2-diversity but Table III(b) does not. Suppose anonymization of Table III(b) results in Table III(c) by global recoding and Table III(d) by local recoding. The adversary is armed with the external table Table II(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, Andre and Kim must have contracted HIV.

Observation 2. Minimality attack is possible whether the sensitive attribute contains nonsensitive values or not.

Recall that 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.

The previous attacks to Andre would also be successful if the knowledge of the external table Table II(a) is replaced by that of a multiset of the *QID* values as shown in Table II(b) plus the QID value of Andre; or if Table II(c) is replaced by the multiset in Table II(d) plus the QID value of Andre. Note that the multisets in Tables II(b) and (d) are inherently available in the published data if the bucketization technique as in Xiao and Tao [2006a], Zhang et al. [2007], or Martin et al. [2007] is used.

Observation 3. The minimality attacks to an individual t would also be successful if the knowledge of the external table T^e (which is either a superset of individuals of the published table or not) is replaced by that of a multiset of the *QID* values of the external table T^e plus the QID value of t.

Table IV. Illustration of Near to Minimality Principle

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

(a) good table

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

(b) bad table

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

(c) global

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

(d) local

A strong requirement of 3-diversity is used to achieve the original intended requirement of 2-diversity.

It is natural to ask whether there is a privacy breach if the data publisher generalizes the table a little more than minimal. In this case, we say that the anonymization algorithm follows a *near to* minimality principle. Suppose the intended objective is to generate a table which satisfies a privacy requirement of 2-diversity. Under the near to minimality principle, the publisher generates a table which satisfies a stronger privacy requirement of 3-diversity. Again we assume that the adversary knows that the algorithm adopted guarantees 3-diversity while minimizing the information loss. Does a published table which satisfies 3-diversity guarantee that the probability that an individual is linked to a sensitive value is at most 1/2? The answer is interestingly no.

Consider Table IV. Suppose our original intended privacy requirement is 2-diversity because we want to guarantee that the probability that an individual is linked to a sensitive value is at most 1/2. Based on the near to minimality principle, a stronger 3-diversity is attained instead. Table IV(a) satisfies 3-diversity but Table IV(b) does not. Thus, Table IV(c) and Table IV(d) are generated by global recoding and local recoding, respectively. By similar arguments, with the knowledge of a strong requirement 3-diversity and Table IV(c), the adversary can also deduce that the probability that an individual with QID value $= q1$ is equal to 2/3 which is greater than the intended maximum disclosure probability of 1/2. This is because the two HIV values must be linked to the tuples with QID $= q1$. Otherwise, there will be no violation of 3-diversity and there is no need for generalization. Similar arguments can be made to Table IV(d). We call this kind of attack the *near-to-minimality attack*.

Observation 4. Near-to-minimality attack is possible when the anonymization algorithm follows the near to minimality principle.

From the preceding discussion, we described the attack by minimality and the attack by near-to-minimality are successful under the principles of minimality principle and near-to-minimality principles used in the anonymization algorithm. Both are based on some knowledge about the algorithm, let us call an attack based on such knowledge an *attack by mechanism*. Hence minimality or near-minimality attack are under this bigger class of attack.

1.2 Contributions

In this article, 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, we are the first to study the anonymization-based attack, including the attack by minimality, the attack by near-to-minimality, and the attack by mechanism, in privacy-preserving data publishing. A preliminary version of this article was published in Wong et al. [2007]. Wong et al. [2007] focuses on the study about the attack by minimality but does not study the attack by near-to-minimality and the attack by mechanism. We propose an m-confidentiality model to capture the privacy-preserving requirement under the additional adversary knowledge of the minimality (or the near-to-minimality) of the anonymization mechanisms. Zhang et al. [2007] also considers a privacy model under the consideration of the mechanism based on some minimality principle. However, as we will show in Section 2, our model is more general than Zhang et al. [2007].

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 successful in a variety of known anonymization models. We also show that the attack by mechanism is also possible in a well-known recognized algorithms, Incognito-like algorithm [LeFevre et al. 2005; Samarati 2001; Machanavajjhala et al. 2006; Li and Li 2007; Wong et al. 2006], Mondrian-like algorithm [LeFevre et al. 2006], and Zhang's algorithm [Zhang et al. 2007], in this section.

Third, we propose a solution to generate a published dataset which satisfies m-confidentiality. Our method makes use of the existing mechanisms for k-anonymity with additional precaution steps. The existing mechanisms can adopt either the bucketization technique or the generalization technique over the QID attributes while the additional precaution steps are to distort some values in the sensitive attribute for m-confidentiality. Interestingly, although it has been discovered by recent research works 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 and attacks by near-to-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 (and an attack by near-to-minimality) becomes infeasible.

Fourth, we have conducted a comprehensive empirical study on both the problem and our method. We show how such a minimality attack can succeed on a real dataset in our experiment. Compared to the most competent existing algorithms for k-anonymity, our method introduces very minor computation overhead. The information loss generated by our method is also comparable to those resulting from known algorithms for k-anonymity.

The rest of the article 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 the attacks by minimality and mechanism 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 work 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 [Wang et al. 2006], l-diversity [Machanavajjhala et al. 2006], (α, k)-anonymity [Wong et al. 2006], t-closeness [Li and Li 2007], (k, e)-anonymity [Zhang et al. 2007], personalized privacy [Xiao and Tao 2006b], and workload-aware anonymization model [LeFevre et al. 2008], 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 [Machanavajjhala et al. 2006; Kifer and Gehrke 2006; Martin et al. 2007; Li and Li 2008]. Machanavajjhala et al. [2006] 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. Kifer and Gehrke [2006] 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. Martin et al. [2007] tries to protect sensitive data against background knowledge in the form of implications, for example, 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 aforesaid works consider the knowledge of the anonymization mechanism discussed in this article and a preliminary version of this article published in Wong et al. [2007]. In Section 5 we shall show that the aforesaid previous works 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 works such as Agrawal and Srikant [2000] and Evfimievski et al. [2002].

The idea of attack by minimality has been known for some time in cryptographic attack where the adversary makes use of the knowledge of the underlying cryptographic algorithm. In particular, a timing attack [Kocher 1996] 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 attacks, 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, Brumley and Boneh [2003] 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.

Recently, Wong et al. [2007] and Zhang et al. [2007] pointed out that a privacy breach is possible when the nature of the anonymization process is considered, where Wong et al. [2007] is a preliminary version of this article.

Zhang et al. [2007] proposes a privacy model which considers the underlying anonymization process. Consider a privacy requirement $\mathcal{R}$ such as l-diversity. Suppose the adversary knows the external table T^e containing the QID attribute of all individuals in the original table T. In the model proposed by Zhang et al. [2007], according to T^e, the adversary lists all possible tables generalized from T^e and orders them in a table sequence sorted by the ascending order of the information loss of the tables (after generalization), namely $T_1, T_2, \ldots, T_{\mathcal{N}}$, where $\mathcal{N}$ is the total number of possible tables generalized from T^e. If the published table T^* is equal to T_i in the table sequence, the adversary deduces that $T_1, T_2, \ldots, T_{i-1}$ must violate the privacy requirement $\mathcal{R}$. According to this observation, the adversary can trigger an attack to deduce that an individual is linked to a sensitive value. However, Zhang et al. [2007] is different from our work as follows.

Our model is much more general than Zhang et al. [2007] and covers the model proposed by Zhang et al. [2007]. In this article, we consider any mechanism that follows the minimality principle. Zhang et al. [2007] considers one particular global recoding anonymization mechanism that follows the minimality principle. The sequence of tables considered in Zhang et al. [2007] is based on the information loss of the tables. However, an anonymization algorithm in general is not restricted to the direct consideration of information loss. Since the problem is in general NP-hard, other heuristics are often used. Our model is very general and does not assume a specific criterion used in the algorithm. Most existing algorithms arrive at a local minima as defined in our model. Detailed analysis about this can be found in Section 5.2. In addition, we consider the near-to-minimality principle but Zhang et al. [2007] does not contain this study.

Zhang et al. [2007] studies the attack on the published table which is generalized by a global-recoding algorithm. It is not clear how their model can be adapted to the local-recoding case. Zhang et al. [2007] proposes an algorithm resistant to attacks when the target is l-diversity. However, there is no formal analysis of this attack. The attack is exemplified by a single scenario in the Introduction of Zhang et al. [2007]. The analysis for k-anonymity is not trivially extended to that with l-diversity. In comparison, our work gives a comprehensive analysis of the attack with both local recoding and global recoding when l-diversity is considered. The algorithm of evaluating the attack can be computed in polynomial time in the total number of tuples and the total number of QID attributes.

Since the proposed algorithm in Zhang et al. [2007] is a global-recoding algorithm, in general it incurs more information loss and its utility is lower. Furthermore, its complexity increases exponentially with the number of domains

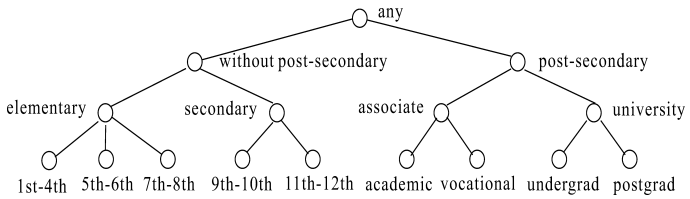

Fig. 1. Generalization taxonomy of "Education" in the "Adult" dataset.

of QID attributes. In comparison, we propose an algorithm which is resistant to minimality attacks and the algorithm can be either a local recoding or a global recoding. The complexity of our proposed algorithm is roughly equal to the complexity of any existing algorithm for k-anonymity. If we adopt a polynomial-time algorithm for k-anonymity in our algorithm MASK, MASK also computes in polynomial time. While there is no experimental result reported in Zhang et al. [2007], in this article, we have conducted empirical studies to show the practicality of the attack and the feasibility of our proposed algorithm.

Finally there are some problems with the proposed methods in Zhang et al. [2007]. (1) The major focus of Zhang et al. [2007] is finding the privacy breach for k-anonymity. A basic assumption is that the adversary knows the QID attributes of all individuals in the table. However, the analysis in Zhang et al. [2007] is on the risk of disclosure of the QID values of some individuals, and this contradicts the assumption of the QID knowledge. (2) Zhang et al. [2007] claims that their proposed algorithm is resistant to attack when l-diversity is considered. However, it still suffers from the minimality attack. Details can be found in Section 5.2.

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 [Blake and Merz 1998]. Values "undergrad" and "postgrad" can be generalized to "university".[4] Generalization replaces lower-level domain values in the taxonomy with higher-level domain values.

Some previous studies consider taxonomies only for QID attributes while some others also consider taxonomies for the sensitive attributes. In some earlier studies on anonymization, the taxonomy for an attribute in the QID or the

[4]Such taxonomies can also be created for numerical attributes by generalizing values to value range and to wider value ranges. The ranges can be determined by users or a machine learning algorithm [Fayyad and Irani 1993].

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.[5]

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 nonsensitive node are also nonsensitive. The ancestor of a sensitive node may be either sensitive or nonsensitive.

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 preceding 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.

[5]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".

Definition 1 (*Anonymization*). Anonymization is 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 $\tau = \{\mathcal{T}_1, \ldots, \mathcal{T}_u\}$, an anonymization defined by f conforms to τ iff $t.A \preceq f(t).A$ holds for any t and A.

For instance, Table I(b) is anonymized to Table I(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 works [Sweeney 2002b; LeFevre et al. 2006, 2005; Xiao and Tao 2006b], in addition to the published dataset T^*, K_{ad} involves an external table T^e such as a 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) ssuperset 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 aforesaid 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 [Wong et al. 2007].

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 next [Wong et al. 2007].

Definition 3 (*Minimality Principle*). Suppose $\mathcal{A}$ is an anonymization algorithm for a privacy requirement $\mathcal{R}$. Let table T^* be a table generated by $\mathcal{A}$ and T^* satisfies $\mathcal{R}$. $\mathcal{A}$ is said to satisfy the minimality principle if, 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 $\mathcal{R}$.

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 $\mathcal{R}$ can be another requirement.

It is also noted that the minimality principle applies to not only the traditional taxonomies but also any arbitrary acyclic graphs because the principle only requires specialization steps which both traditional taxonomies and arbitrary acyclic graphs can uniquely provide.

The minimality principle can be extended to the near-to-minimality principle as described in Section 1. The near-to-minimality principle is formally defined as follows.

Definition 4 (*Near-to-Minimality Principle*). Suppose $\mathcal{A}$ is an anonymization algorithm for a privacy requirement $\mathcal{R}$. Let the intended privacy requirement be $\mathcal{R}$. Let table T^* be a table generated by $\mathcal{A}$ and T^* satisfies a stronger privacy requirement $\mathcal{R}'$. $\mathcal{A}$ is said to satisfy the near-to-minimality principle if, 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 $\mathcal{R}'$.

For example, if the intended privacy requirement $\mathcal{R}$ considered is 2-diversity, algorithm $\mathcal{A}$ may generate a table T^* which satisfies 3-diversity. This is an example of near-to-minimality principle. The attack based on this principle is called near-to-minimality attack.

For the sake of illustration, we focus on the minimality attack only. However, it should be noted that the proposed algorithm in Section 6 can handle both the minimality attack and the near-to-minimality attack.

Assumption 3 (*Adversary Knowledge* K_{ad}^{min}). For *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 I(a) is the result generated from an anonymization mechanism (e.g., the adapted Incognito algorithm in Machanavajjhala et al. [2006]) for l-diversity that follows the minimality principle, suppose the multiset in Table II(b) is known and the *QID* value of individual o is known to be $q1$, then *Credibility*$(o, \{HIV\}, K_{ad}^{min}) = 1/2$. For Table I(c), *Credibility*$(o, \{HIV\}, K_{ad}^{min}) = 1$.

Section 4 will describe how to compute *Credibility*(o, s, K_{ad}^{min}) when $\mathcal{A}$, the anonymization algorithm, adopts the minimality principle.

The aforesaid minimality principle is very general and does not demand that $\mathcal{A}$ achieves a global minima in terms of the information loss, nor does it depend on how the information loss is defined. Almost all known anonymization algorithms (including Incognito-based methods [LeFevre et al. 2005; Machanavajjhala et al. 2006; Martin et al. 2007; Li and Li 2007] and top-down approaches [Fung et al. 2005; Xiao and Tao 2006b; Wong et al. 2006; Wang and Fung 2006]) try to reduce information loss of one form (like Definition 7 to be described later) or another, and they all follow the previous principle. It is important to define the problem based on local minima rather than global minima since most problems of optimal anonymization are shown to be NP-hard and many heuristical methods are therefore proposed which may not jump out of a local minima. For example, algorithms such as Fung et al. [2005], Xiao and Tao [2006b], and Wang and Fung [2006] may not minimize the information loss globally. Instead, such an algorithm may return a locally-optimized table.

One key observation is that as long as the data anonymization algorithm is deterministic, an adversary can perform minimality attacks. This is because from the output, s/he can exclude certain output possibilities since different

inputs give different outputs. However, probabilistic algorithms like Anatomy [Xiao and Tao 2006a] do not suffer from the minimality attack.

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 I(d), the adversary can deduce that l must be 2.

Definition 5 (*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, Table I(a) satisfies 2-confidentiality. This is because since there is no need to generalize Table I(a), we know that every QID-EC in Table I(a) must satisfy 2-diversity. As we mentioned before, an exclusion of some possibilities occurs when there are some QID-EC which satisfy 2-diversity but some which do not satisfy 2-diversity. In this table, since all QID-EC's satisfy 2-diversity, there is no exclusion of any certain output possibilities. Thus, we deduce that $Credibility(o, s, K_{ad})$ is at most 1/2.

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 Xiao and Tao [2006b], we also adopt a similar measure of information loss. The idea is similar to the normalized certainty penalty [Xu et al. 2006].

Definition 6 (*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.

Definition 7 (*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_A^* 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,

$$\mathcal{IL}(t^*) = \sum_{A \in QID} \{\mathcal{IL}(t^*, A) \times weight(A)\}$$

where

$$\mathcal{IL}(t^*, A) = \begin{cases} \dfrac{coverage[v_A^*] - 1}{base(\mathcal{T}_A) - 1} & \text{if } base(\mathcal{T}_A) > 1 \\ 0 & \text{if } base(\mathcal{T}_A) = 1 \end{cases}$$

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 $v_A^* = t.A$, and $coverage[v_A^*] = 1$. 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_A^* = the root of $\mathcal{T}_A$. Thus, $coverage[v_A^*] = base(\mathcal{T}_A)$ 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 Xiao and Tao [2006b] 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, it is not possible to 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 8 (*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. We first describe the derivation of the formula of credibility for global recoding in Section 4.1. Then, we describe the derivation for local recoding, which is more complicated, in Section 4.2.

4.1 Global Recoding

In this section, we consider the evaluation of the credibility for global recoding. First we give an example to illustrate the main intuitive ideas in Section 4.1.1. Next we shall describe the general formulation in Section 4.1.2.

4.1.1 *An Example.* The derivation of credibility for global recoding is better illustrated with the example as shown in Table VI which is a global recoding of Table V to achieve 2-diversity. In Table VI containing 14 tuples with QID = Q, {HIV} is the only sensitive value set and the goal is 2-diversity. Assume that T^* and T^e have matching cardinality on Q.

Definition 9. 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^*.

From T^e, the adversary can determine that there are two tuples in $q1$, two tuples in $q2$, and 10 tuples in $q3$. Suppose $q1, q2$, and $q3$ can be generalized to Q. The total number of tuples in T^e with QID that can be generalized to Q is equal to 14 (which is the total number of tuples with QID = Q in T^*). Thus, T^* and T^e have matching cardinality on Q.

Table V. A Table which Violates 2-Diversity

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 VI. A 2-Diverse Table by Global Recoding of Table V

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

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 VI) 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 VII. There are only five possible combinations as shown. We call this table the sensitive tuple distribution table for Q.

Definition 10 (*Sensitive Tuple Distribution Table*). Let Q be a QID value in T^* and $\mathcal{C}$ be a set of classes in T^e which can be generalized to Q. Suppose $\mathcal{C}$ is equal to $\{C_1, C_2, \ldots, C_u\}$. Let n_i be the total number of tuples in C_i for $i = 1, 2, \ldots, u$ and n_s be the total number of tuples with values in sensitive value set s in the data set T^*.

A possible combination of the number of sensitive tuples among all classes in $\mathcal{C}$ is defined to be represented by u-tuple $(m_1, m_2, \ldots, m_u)$ where $m_i \in [1, n_s]$ for $i = 1, 2, \ldots, u$ such that:

(1) $\sum_{i=1}^{u} m_i = n_s$,
(2) $m_i \leq n_i$ for $i = 1, 2, \ldots, u$, and

Table VII. Possible Combinations of Number of Sensitive Tuples

	Number of sensitive tuples			Total number of cases
	$q1$	$q2$	$q3$	
(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

(3) there exists one $C_i \in \mathcal{C}$ such that $m_i/n_i > 1/l$ (i.e., C_i violates l-diversity in this combination).

A sensitive tuple distribution table for Q contains a number of rows, each of which corresponds to a possible combination of the number of sensitive tuples among all classes in $\mathcal{C}$ associated with a constant which is the total number of cases (or assignments) to denote the total number of possible assignments for this corresponding possible combination (which will be defined next).

Consider a possible combination $(m_1, m_2, \ldots, m_u)$. Intuitively, given the total number of tuples with values in sensitive value set s in the dataset T^* denoted by n_s, we *distribute* these n_s sensitive values into u different classes in $\mathcal{C}$ where m_i corresponds to the total number of sensitive values distributed to C_i. A possible combination can be regarded as one of the possible ways of the distribution. The first condition corresponds to that the sum of all sensitive values distributed to all classes in $\mathcal{C}$ is equal to the total number of sensitive tuples for Q in T^*. The second condition means that, for each class C_i in $\mathcal{C}$, the total number of sensitive values in C_i is at most the size of the class C_i. The third condition means that some of the classes in $\mathcal{C}$ must violate l-diversity.

For example, the first row in Table VII corresponds to the fact that there are two sensitive tuples in $q1$, no sensitive tuple in $q2$, and three sensitive tuples in $q3$ (which is denoted by 3-tuple (2, 0, 3)) where $q1$ violates 2-diversity.

In scenario (a), there are $C_2^2 \times C_0^2 \times C_3^{10} = 120$ different possible ways to assign the sensitive values to the tuples. In scenario (b), there are $C_2^2 \times C_1^2 \times C_2^{10} = 90$ different *assignments* or *cases*.[6] 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 $120 + 90 + 10 + 90 + 120 = 430$. Without any additional knowledge about the assignments, we must assume that each of these cases occurs with the same probability 1/430. Consider the credibility that an individual o with value $q1$ is linked to HIV given K_{ad}^{min}. There are two possible cases.

—*Case 1.* 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_{ad}^{min} is equal to $220/430 = 0.5116$.

—*Case 2.* 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_{ad}^{min} is equal to $90/430 = 0.2093$.

[6]In the following, we refer assignments as cases.

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

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

$$Prob(\text{Case 1}) \times Prob(\ q1 \text{ is linked to HIV in Case 1}) + Prob(\text{Case 2}) \times Prob(\ q1 \text{ is linked to HIV in Case 2})$$

$Prob(\ q1$ 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 $1/2 = 0.5$. we have

$$Credibility\big(o, \{HIV\}, K_{ad}^{min}\big) \ = \ 0.5116 \times 1 + 0.2093 \times 0.5 = 0.616,$$

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

4.1.2 *General Formula.* The general formula of the computation of the credibility is based on the idea illustrated before. 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_{ad}^{min}, 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 $x \in \mathcal{G}'$, $P(\{x\}) = 1/|\mathcal{G}'|$.

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 $C_1, C_2, \ldots C_u$ 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 $n_1 \leq n_2 \leq \cdots \leq n_u$. Let n_s be the total number of tuples with values in sensitive value set s in the dataset.

In Table V, 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 dataset 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, $\mathcal{C} = \{C_1, C_2, \ldots, C_p\}$ and $\mathcal{C}' = \{C_{p+1}, C_{p+1}, \ldots, C_u\}$.

LEMMA 1. *If the classes $\mathcal{X} = \{C_1, \ldots C_u\}$ have been 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 $Credibility(o, s, K_{ad}^{min})$, where $o \in C_i$, for $i = 1, 2, \ldots, p$. 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 $Prob(|C_i(s)| = j|K_{ad}^{min})$ be the probability that there are exactly j occurrences of tuples with s in C_i given K_{ad}^{min}. 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

$$\begin{aligned} &Credibility(o, s, K_{ad}^{min}), \text{ where } o \in C_i, 1 \le i \le p \\ &= Prob(o \text{ is linked to } s \text{ in } C_i \mid K_{ad}^{min}) \\ &= \sum_{j=1}^{n_i} Prob(|C_i(s)| = j \mid K_{ad}^{min}) \times \frac{j}{n_i}. \end{aligned}$$

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

However, constructing the table like Table VII directly is inefficient. Suppose the size of each QID-EC in T is known. It is easy to verify that the maximum number of rows (or scenarios) in such a table is $O((n_p)^p)$ where n_p is the number of tuples in class C_p (which has the greatest size in $\mathcal{C}$). Besides, computing the credibility by using the techniques discussed in Section 4.1.1 takes $O((n_p)^p)$ time. In Appendix A, we propose a dynamic programming method which is much more efficient and takes only $O(p^2 n_s + p n_s^3 l)$ time. The general principle used in Section 4.1.1 is still kept in the dynamic programming approach. However, since the dynamic programming approach reuses the results of many subproblems during the computation of larger problems, a lot of redundant computations can be saved. For example, imagine that Table VII (currently representing three QID-EC's, namely $q1, q2$, and $q3$) is extended to the table representing four QID-EC's, namely $q0, q1, q2$, and $q3$. Suppose there are five tuples with values in sensitive value set s. The extended table is shown in Table VIII. It is easy to see that the computation of the total number of cases for each of the first three scenarios (or rows) involves the same computation for the number of all possible combinations for $q0$ and $q1$ only. This is true also for the scenarios (d), (e), and (f). It is inefficient if these redundant computations are computed from scratch each time. Thus, we propose a dynamic programming approach to keep track of the results of these (redundant) computations so that these kinds of redundant computations can be avoided. Details of the dynamic programming approach can be found in Appendix A in the ACM Digital Library.

Summary. Let us summarize the derivation of the general formula as follows. In the published table T^*, consider a QID-EC P. Suppose we have u classes in the original dataset T, namely $C_1, C_2, \ldots, C_u$, which can be generalized to P.

Table VIII. Possible Combinations of Number of Sensitive Tuples

	Number of sensitive tuples				Total number of cases
	$q0$	$q1$	$q2$	$q3$	
(a)	0	2	0	3	...
(b)	0	2	1	2	...
(c)	0	2	2	1	...
(d)	1	1	0	3	...
(e)	1	1	1	2	...
(f)	1	1	2	1	...
...	...	...	...	...	...

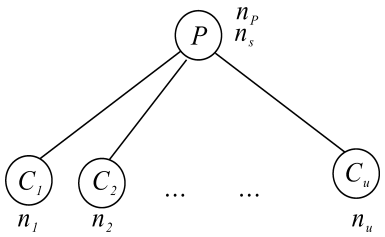

Fig. 2. Generalization taxonomy of classes $C_1, C_2, \ldots, C_u$.

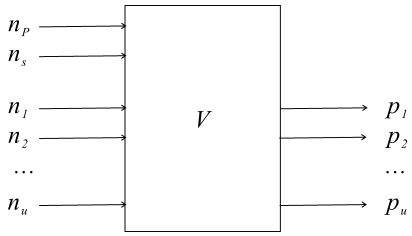

Fig. 3. Function V.

The number of tuples in QID-EC C_i is equal to n_i (see Figure 2). The adversary possesses the following information.

n_P	the number of tuples in P
n_s	the number of sensitive tuples in P
n_i	the number of tuples in class C_i (which is a child of P)

The credibility for each class C_i, denoted by p_i, can be computed according to the formula we derived in the last section. We call the computation of the credibility in this simple scenario function V (see Figure 3). Note that, with a dynamic programming approach described in Appendix A, function V also runs in polynomial time in p, n_s, and l.

4.2 Local Recoding

In this section, we will describe the formula for local recoding. The major idea is also similar to the case for global recoding described in Section 4.1. Specifically, given a QID-EC Q in T^*, we find all QID-EC's in T^e which can be generalized to

Table IX. Another Table which Violates 2-Diversity

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

Q. Then, we generate the sensitive tuple distribution table, among these QID-EC's, which involves all possible combinations of the number of sensitive tuples. It is noted that some combinations are excluded because they cannot occur according to the minimality principle. According to this table, we calculate the total number of cases or assignments for each combination. Finally, according to these calculated values, we compute the probability that an individual is linked to a sensitive value.

However, due to the different natures between local recoding and global recoding, there are some differences in the derivation of the formula. Specifically, since occurrences of the same value such as $q1$ of an attribute may be recoded to different values such as the original value $q1$ and a generalized value Q in local recoding, we have to consider two cases. Case 1: The tuples for $q1$ do not undergo any generalization. Case 2: The tuples for $q1$ are generalized to Q. By considering these two different cases, we compute the credibility accordingly. In the following, we illustrate this idea with a simple example first to give an intuition of the derivation of the formula in Section 4.2.1. Next, we describe how we generalize the formula in Section 4.2.2.

4.2.1 *An Example.* An example is shown in Table IX 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 IX 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 IX is generalized to Table X, 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 nonsensitive 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.

Table X. A 2-Diverse Table of Table IX by Local Recoding

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

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

—*Case 1.* 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.

—*Case 2.* 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

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

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,

$$Prob(o \text{ is linked to HIV in Case 1} \mid K_{ad}^{min}) = \frac{2}{4} = \frac{1}{2}.$$
$$Prob(o \text{ is linked to HIV in Case 2} \mid K_{ad}^{min}) = 1.$$
$$Credibility(o, \{HIV\}, K_{ad}^{min}) = \frac{4}{5} \times \frac{1}{2} + \frac{1}{5} \times 1 = \frac{3}{5},$$

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

4.2.2 *General Formula.* The previous example shows the basic idea of the derivation of the general formula.

Suppose there are u QID-EC's in the original dataset, namely $C_1, C_2, \ldots, C_u$, which can be generalzied to the same value $C_{\mathcal{G}}$. After the generalization, some tuples in some C_i are generalized to $C_{\mathcal{G}}$ 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 X, C_i corresponds to $q1$ and $\mathcal{C}_\mathcal{G}$ 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 $\mathcal{C}_\mathcal{G}$. 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, we have the following.

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

The term $Prob(o \text{ is linked to } s \text{ in } \mathcal{C}_\mathcal{G} \text{ in } T^* | K_{ad}^{min})$ 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 value, the preceding analysis is extended to include more possible combinations of outcomes. However, the basic ideas remain similar.

More specifically, the question is how to compute Prob(o is linked to s in $\mathcal{C}_\mathcal{G}$ in $T^* | K_{ad}$). We can also make use of function V. We can regard this generalized dataset as follows. We just consider $\mathcal{C}_\mathcal{G}$ but do not consider C_i in T^*.

As $\mathcal{C}_\mathcal{G}$ has the same generalized value in the published dataset for different values of i, P (in V) is mapped to $\mathcal{C}_\mathcal{G}$, and n_P (in V) is mapped to the total number of generalized tuples in T^* (i.e., $\sum_{i=1}^{u} n_{i,g}$). n_s is mapped to the total number of sensitive tuples in P in T^*. Then, n_1 in V is mapped to $n_{1,g}$, n_2 in V is mapped to $n_{2,g}$ and so on. Thus, we have the following mapping.

Parameter	Mapping Value
n_P	$\sum_{i=1}^{u} n_{i,g}$
n_s	total number of sensitive tuples in P in T^*
$n_1, n_2, \ldots, n_u$	$n_{1,g}, n_{2,g}, \ldots, n_{u,g}$

Then, we can use function V to obtain the credibility.

Similarly, the computation of credibility for local recoding also runs in polynomial time in $|T|$, p, n_s, and l. This is because the computation of the term Prob(o is linked to s in $\mathcal{C}_\mathcal{G}$ in $T^* | K_{ad}$) runs in polynomial time in $|T|$, p, n_s, and l (by Theorem 6) and evaluating $n_{i,g}$, $n_{i,u}$, $n_{i,u(s)}$, and n_i only requires one scan of database.

4.3 Attack Conditions

We have seen previously 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.

Table XI. Anonymization for (3,3)-Diversity

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	Disease
Q	Diabetics
Q	HIV
Q	HIV
Q	Lung Cancer
Q	Ulcer
Q	Alzhema
Q	Gallstones

(c) global

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

(d) local

THEOREM 1. *If there is no exclusion of any scenario from the table, then the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC; that is, an attack by minimality is not possible.*

From the preceding theorem, it is interesting to know whether it is a must that the a minimality attack occurs when there are some excluded combinations(s) in the sensitive tuple distribution table. The answer is an interesting "no."

LEMMA 2. *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}.

5. GENERAL MODEL

In Section 5.1, we show that the minimality attack is possible in a variety of privacy models. We will describe an attack which is based on the process of anonymization in Section 5.2.

5.1 Minimality Attack

In this subsection, we show that the minimality attack can be successful on a variety of anonymization models. In Tables XI to XV, 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.

The major idea of minimality attacks is similar to the case for l-diversity discussed previously. Given a published table T^* satisfying a privacy requirement $\mathcal{R}$, the adversary tries to "guess" the original table. Without having any knowledge of the minimality principle, s/he generates all possible tables by enumerating all possible combinations of sensitive tuples among the QID-EC's. However, in fact, with the knowledge the minimality principle used in the anonymization algorithm, by excluding some possibilities, s/he lists all possible original tables which only violate $\mathcal{R}$ and can be generalized to T^* by the anonymization algorithm. From the remaining possible tables, it is possible for the adversary to breach individual privacy. In the following, we describe how the attacks are possible in a variety of models.

Table XII. 0.5-Closeness Anonymization

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

(b) bad table

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

(c) global

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

(d) local

Table XIII. (k, e)-Anonymity for $k = 2$ and $e = 5k$

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	Income
Q	30k
Q	30k
Q	20k
Q	10k
Q	40k

(c) global

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

(d) local

Table XIV. Anonymization for (0.6, 2)-Safety

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

(a) good table

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

(b) bad table

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

(c) global

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

(d) local

Table XV. Anonymization for Personalized Anonymity

QID	Education	Guarding Node
$q1$	1st-4th	elementary
$q2$	undergrad	none
$q2$	undergrad	none

(a) bad table

QID	Education
Q	1st-4th
Q	undergrad
Q	undergrad

(b) global

QID	Education
Q	1st-4th
Q	undergrad
$q2$	undergrad

(c) local

In general, similar to l-diversity, the minimality attack in each of the following models can also be achieved in polynomial time in terms of $|T|$, p, n_s, and θ where p is the total number of QID-EC's which may violate the privacy requirement $\mathcal{R}$, n_s is the total number of tuples with sensitive value set s and θ is a parameter of some privacy requirements $\mathcal{R}$. Thus, it is not difficult for the adversary to breach individual privacy in the following models.

5.1.1 *Recursive* (c, l)*-Diversity.* With recursive (c, l)-diversity [Machanavajjhala et al. 2006], in each QID-EC, let v be the most frequent sensitive value, and 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 XI(c) is a global recoding for Table XI(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 XI(d).

5.1.2 t*-Closeness.* Recently, t-closeness [Li and Li 2007] 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 Li and Li [2007].

$$D[\mathbb{P}, \mathbb{Q}] = 1/2 \sum_{i=1}^{m} |p_i - q_i|$$

Consider Table XII(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 0.5. 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 XII(d).

5.1.3 *(*k, e*)-Anonymity.* The model of (k, e)-anonymity [Zhang et al. 2007] 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. Consider the tables in Table XIII (where Income is a sensitive numeric attribute). From Table XIII(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 XIII(d) for local recoding.

5.1.4 (c, k)*-Safety.* (c, k)-safety [Martin et al. 2007] considers the worst-case background knowledge. The background knowledge defined in Martin et al. [2007] is a set of k implications whose simplest form is: If individual p_1 is linked to sensitive value v_1, then individual p_2 is linked to sensitive value v_2, where v_1 may equal v_2. This model requires that the probability that any individual is

linked to a sensitive value given a background knowledge containing any k implications is at most c. Consider the tables in Table XIV. Table XIV(c) is a global recoding of Table XIV(b). From Table (c), the adversary can eliminiate the cases where there is 0 or 1 HIV value in the bucket $q1$, since in such cases Table XIV(b) would have been (0.6,2)-safe. Since Table XIV(b) is anonymized to Table XIV(c), there must be 2 HIV occurrences in the bucket for $q1$ in Table XIV(b). Hence an individual with $q1$ is linked to HIV with a probability of 2/3, higher than the intended threshold of 0.6. Similar attacks can be launched against the local recoding in Table XIV(d).

5.1.5 *Personalized Privacy.* Xiao and Tao [2006b] proposed a personalized privacy model where each individual can provide his/her preference on the protection of his/her sensitive value, denoted by a guarding node. For example, an individual o with a value "1st–4th" may specify "elementary" as a guarding node in order that any QID-EC that may contain o should contain at most $1/l$ tuples with "elementary" values. For $l = 2$, in the tables in Table XV, Tables XV(b) and XV(c) are global and local recodings for Table XV(a), respectively. Suppose the adversary knows that everyone with an undergraduate degree does not mind to disclose his/her education. Based on the minimality principle, if the "1st–4th" belongs to a $q2$ tuple, then Table XV(a) will not be anonymized, so the tuple with QID $= q1$ must be linked to "1st–4th". Similar attack will be successful on Table XV(c).

5.1.6 *Sequential Release.* Wang and Fung [2006], Xiao and Tao [2007], and Bu et al. [2008] proposed sequential releases of the table. Since the models of sequential releases also require a privacy requirement $\mathcal{R}$ and the released tables are anonymized based on the principle of minimality, the minimality attack is still possible. For the interest of space, we skip the details here.

5.1.7 *General Attack by Minimality.* In the proposed anonymization mechanism for each of the aforesaid 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.

5.2 Attack by Mechanism

In this section, we consider that the adversary knows the process of anonymization. We shall see that with this additional knowledge, s/he can breach privacy of individuals. We call this kind of attack an attack by mechanism. In the following, we study three well-known frameworks, namely the Incognito-like algorithm [LeFevre et al. 2005; Samarati 2001; Machanavajjhala et al. 2006; Li and Li 2007; Wong et al. 2006], Mondrian-like algorithm [LeFevre et al. 2006], and Zhang's algorithm [Zhang et al. 2007].

Table XVI. A Dataset Illustrating the Attack by Mechanism

Gender	Education	Disease
Male	postgraduate	HIV
Female	undergraduate	HIV
Female	postgraduate	non-sensitive
Female	undergraduate	non-sensitive
Female	undergraduate	non-sensitive

The common major idea of the attacks for these kinds of algorithms is based on the following. Suppose the adversary is given a published table T^*. S/he tries to enumerate all possible tables generalized from the correspondence external table T^e and sort them in a certain ordering which is dependent on the anonymization algorithm. One of the ordering criteria is the information loss of the generalized table, which is used in the Incognito-like algorithm and Zhang's algorithm. Another ordering criterion is the heuristic used in the algorithm. For example, the Mondrian-like algorithm uses a heuristic to determine the ordering of choosing dimensions for partitioning the data. With this ordering assumed in the anonymization algorithm, all tables ordered before the published table T^* must violate the privacy requirement. (Note: if the privacy requirement satisfies the generalization property described in Section 5.2.1 or the monotonicity property and the anonymization algorithm performs this process in the reverse ordering, the adversary can deduce that, conversely, all tables ordered after T^* must satisfy the privacy requirement.) The adversary makes use of this information to breach individual privacy.

It is noted that, with respect to attack by mechanism, our work is much more general compared with Zhang's algorithm [Zhang et al. 2007]. Firstly, Zhang's algorithm assumes only one particular model of anonymization algorithms. However, we do not assume any particular model for privacy breaches. Secondly, Zhang's algorithm makes use of the information loss to sort all possible tables. However, our model covers not only the ordering according to the information loss but also the ordering according to other criteria including some heuristics used in the anonymization algorithm.

5.2.1 *Incognito-Like Algorithm.* We first consider a well-known generalization-based framework which is adopted by algorithm Incognito [LeFevre et al. 2005] and Samarati's algorithm [Samarati 2001]. Algorithm Incognito has been applied for handling many other privacy models such as l-diversity [Machanavajjhala et al. 2006], t-closeness [Li and Li 2007], and (α, k)-anonymity [Wong et al. 2006]. Let us focus on l-diversity on the privacy requirement $\mathcal{R}$ for illustration. We first illustrate the anonymization process of the framework.

Table XVI shows a dataset containing two attributes (Gender and Education) and one sensitive attribute Disease, where HIV is the sensitive value. Figures 4(a) and 4(b) show the generalization taxonomies of attributes Gender and Education, respectively. Each node in a generalization taxonomy of attribute A corresponds to a generalization domain with respect to A. The generalization domain in the lower level has more detailed information than the

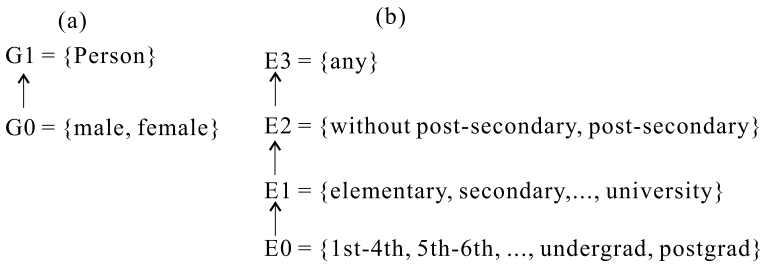

Fig. 4. Generalization taxonomy.

Table XVII. $< G0, E1 >$

Gender	Education	Disease
Male	university	HIV
Female	university	HIV
Female	university	non-sensitive
Female	university	non-sensitive
Female	university	non-sensitive

Table XVIII. Illustration of Generalization Property

Gender	Education	Disease
Person	postgraduate	HIV
Person	undergraduate	HIV
Person	postgraduate	non-sensitive
Person	undergraduate	non-sensitive
Person	undergraduate	non-sensitive

(a) $< G1, E0 >$

Gender	Education	Disease
Person	university	HIV
Person	university	HIV
Person	university	non-sensitive
Person	university	non-sensitive
Person	university	non-sensitive

(b) $< G1, E1 >$

higher level. For example, in Figure 4(a), generalization domain G0 (with respect to Gender) has more detailed information than G1. Domains of multiple attributes can be combined to more complex generalization domains such as $< G0, E1 >$.

(GENERALIZATION PROPERTY). *Let T be a table and let Q be an attribute set in T. Let G and G' be generalization domains with respect to Q, where G' is more general than G. If the table T generalized with the generalization domain G with respect to Q is l-diverse, then the table T generalized with G' with respect to Q is also l-diverse.*

For example, consider generalization of the dataset in Table XVI, let us set $l = 2$. Table XVIII(a), the table generalized with $< G1, E0 >$, satisfies l-diversity. As $< G1, E1 >$ is more general than $< G1, E0 >$, the table generalized with domain $< G1, E1 >$ should also be l-diverse (Table XVIII(b)).

With the algorithm in the framework, first, all possible generalization domains are generated. For example, Figure 5 shows a subset of all possible generalization domains. Then, we test whether the table generalized with a generalization domain G satisfies l-diversity from bottom to top. By the generalization property, if G satisfies l-diversity, we do not need to test generalization domains above it. This is because all generalization domains above it must satisfy l-diversity. Finally, the algorithm chooses the table with the minimum information loss among all tables satisfying l-diversity.

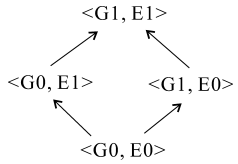

Fig. 5. A subset of all possible generalization domains.

For example, suppose our privacy requirement is 2-diversity. We know that $< G0, E0 >$ and $< G0, E1 >$ do not satisfy 2-diversity but $< G1, E0 >$ and $< G1, E1 >$ satisfy 2-diversity. Since $< G1, E0 >$ generates a table with the least information loss, Table XVIII(a) will be published.

Suppose the adversary knows the anonymization mechanism (i.e., the aforesaid algorithm) including the formula of computation of the information loss. Without loss of generality, we assume that the information loss used in the algorithm is defined as Definition 7. If the table as shown in Table XVIII(a) is released, the adversary will reason as follows.

From Table XVIII(a), the generalization domain $< G1, E0 >$ is used for publishing the table. According to Figure 5, $< G1, E1 >$ must also satisfy 2-diversity (by the generalization property). Since the algorithm chooses the table with the least information loss, $< G0, E0 >$ should violate 2-diversity. Besides, $< G0, E1 >$ must not satisfy 2-diversity. It is noted that the information loss of the table generalized with $< G0, E1 >$ (as shown in Table XVII) is equal to 0.0625 and the information loss of the table generalized with $< G1, E0 >$ (as shown in Table XVIII(a)) is equal to 0.5. Thus, suppose $< G0, E1 >$ satisfies 2-diversity, the table will be generalized with $< G0, E1 >$ instead of $< G1, E0 >$.

Thus, the adversary knows that the table generalized with $< G0, E1 >$ (i.e., Table XVII) must violate 2-diversity. In the published table XVIII(a), there are two HIV values. Since Table XVII violates 2-diversity, one of the HIV values must be linked to the set of tuples with QID values = (Male, university). Hence if the attack target is an individual o with QID values = (Male, university), then the adversary can confirm that o is linked to HIV.

In conclusion, attack by mechanism is possible. It is further noted that the adversary cannot breach the privacy in this example if we just consider the minimality attack defined in Definition 3 and regard the projected Table XVI on attributes Gender and Education as the external table T^e and Table XVIII(a) as the published table.

Suppose we are given a published table T^*. The adversary lists all possible tables such that it can be generalized from the correspondence external table T^e and the information loss of these tables is less than that of T^*. Let R be the running time of the Incognito-like algorithm. It is easy to verify that the step of enumerating all possible tables takes $O(R)$. For each of these tables, the adversary can also adopt similar techniques in Section 4 to breach individual privacy, which requires $O(p^2 n_s + p n_s^3 l)$. Thus, the total running time of performing such attack is $O((p^2 n_s + p n_s^3 l)R)$.

5.2.2 *Mondrian-Like Algorithm.* Let us we first describe algorithm Mondrian [LeFevre et al. 2006] and analyze how attack by mechanism is feasible.

Table XIX. A Generalized Table Illustrating Privacy Breaches from Algorithm Mondrian

Zipcode	Age	Disease
51102	36	HIV
51104	36	HIV
51101	31	non-sensitive
51102	31	non-sensitive
51104	31	non-sensitive
51105	31	non-sensitive

(a) A raw table

Zipcode	Age	Disease
[51101-51105]	[31-36]	HIV
[51101-51105]	[31-36]	HIV
[51101-51105]	[31-36]	non-sensitive
[51101-51105]	[31-36]	non-sensitive
[51101-51105]	[31-36]	non-sensitive
[51101-51105]	[31-36]	non-sensitive

(b) A fully generalized table

Zipcode	Age	Disease
[51101-51105]	[32-36]	HIV
[51101-51105]	[32-36]	HIV
[51101-51105]	31	non-sensitive
[51101-51105]	31	non-sensitive
[51101-51105]	31	non-sensitive
[51101-51105]	31	non-sensitive

(c) A table after the partition of Table (b) according to attribute Age

Zipcode	Age	Disease
[51101-51103]	[31-36]	HIV
[51104-51105]	[31-36]	HIV
[51101-51103]	[31-36]	non-sensitive
[51101-51103]	[31-36]	non-sensitive
[51104-51105]	[31-36]	non-sensitive
[51104-51105]	[31-36]	non-sensitive

(d) A table after the partition of Table (b) according to attribute Zipcode

In LeFevre et al. [2006], algorithm Mondrian originally is designed for k-anonymity. It is easy to extend the model to l-diversity, which will be described next.

We illustrate the algorithm with Table XIX(a) where the QID attributes are two numeric attributes, Zipcode and Age, and the sensitive attribute is Disease. Suppose the privacy requirement is 2-diversity. Algorithm Mondrian is a kd-tree-based algorithm. Firstly, it fully generalizes the table as shown in Table XIX(b) such that each tuple is in the same QID-EC. Then, it iteratively chooses one dimension according to a simple heuristic for partitioning the data, and the iterations continue until there is no feasible partitioning. In LeFevre et al. [2006], the heuristic chooses the dimension with the widest (normalized) range of values. In this example, without loss of generality, we assume the domain of attribute Zipcode is much larger than the domain of attribute Age. According to the heuristic, algorithm Mondrian chooses attribute Age as the first dimension for partitioning. The median of attribute Age is equal to 31. Thus, range [31-36] is partitioned into range [31-31] (or value 31) and range [32-36]. The resulting table is shown in Table XIX(c). However, it is easy to verify that Table XIX(c) violates 2-diversity. So, the partitioning according to attribute Age is infeasible. Then, algorithm Mondrian chooses the remaining attribute (i.e., attribute Zipcode) for partitioning. Since the median value of attribute Zipcode is 51103, we obtain the resulting table as shown in Table XIX(d). It is also easy to verify that Table XIX(d) satisfies 2-diversity. Thus, Table XIX(b) is partitioned to Table XIX(d). Similarly, algorithm Mondrian performs similar steps in each partition in Table XIX(d). However, there are no feasible partitionings. Thus, Table XIX(d) is published as the final output.

Suppose the adversary knows the anonymization algorithm together with the heuristic used. Since s/he knows the external table T^e and the published Table XIX(d) T^*, s/he deduces that the partitioning from Table XIX(b) according to attribute Age (i.e., Table XIX(c)) is infeasible. In other words, one of the

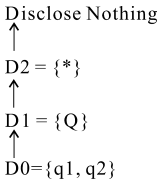

Fig. 6. Generalization taxonomy illustrating privacy breaches from Zhang's algorithm.

Table XX. A Generalized Table Illustrating Privacy Breaches from Zhang's Algorithm

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

(a) A table generalized with $< D1 >$

QID	Disease
*	HIV
*	HIV
*	non-sensitive
*	non-sensitive
*	non-sensitive
*	non-sensitive
*	non-sensitive

(b) A table generalized with $< D2 >$

two partitions (or QID-ECs) in Table XIX(c) violates 2-diversity. By similar arguments, s/he can deduce that the first two individuals in the table must be linked to HIV. In conclusion, attack by mechanism is possible when the published table is generated by algorithm Mondrian.

The complexity of performing the attack for Mondrian-like algorithm is also similar to that for Incognito-like algorithm. Suppose there are N possible binary partitions for Mondrian-like algorithm. The total running time is $O((p^2 n_s + p n_s^3 l)N)$.

5.2.3 *Zhang's Algorithm.* First, we describe how Zhang's algorithm[7] generates a published table. We illustrate this by our motivating example as shown in Table I(b). Suppose we are given the generalization taxonomy as shown in Figure 6 where "Disclose Nothing" is a special node representing that the table should not be disclosed in Zhang's algorithm. Consider the privacy requirement $\mathcal{R}$ is l-diversity. Zhang's algorithm involves three phases. In the first phase, Zhang's algorithm lists a table sequence of all possible generalized tables in the ascending order of information loss of the tables. In this example, the sequence is T_0 : Table I(b) (the table generalized with $< D0 >$), T_1 : Table XX(a) (the table generalized with $< D1 >$), T_2 : Table XX(b) (the table generalized with $< D2 >$), and the table T_3 denoting that nothing should be disclosed. In the second phase, Zhang's algorithm checks whether each table satisfies l-diversity in the sequence order until it finds a table satisfying l-diversity. Since T_0 violates 2-diversity, it checks the next table T_1 which satisfies 2-diversity. After finding a table (i.e., T_1) which satisfies 2-diversity, it switches to the third phase where it checks whether each T_j of the remaining tables in the sequence

[7]For the sake of illustration, we skip some steps of Zhang's algorithm, but this does not affect the illustration of the attack by mechanism.

has at least l possible specializations which violates l-diversity in the reverse order of the table sequence. Once there is a table T_j which satisfies the aforesaid criteria, T_j is published. Otherwise, T_1 is published. For instance, T_3 is first checked. Since there is only one possible specialization of T_3 (i.e., T_2) satisfying 2-diversity and no specialization violating 2-diversity, Zhang's algorithm switches to check the second last table T_2. In this case, T_2 has also only one possible specialization (i.e., T_1) satisfying 2-diversity and no specialization violating 2-diversity. Since all tables checked in the third phase do not satisfy the criteria, T_1 (i.e., Table XX(a)) is published.

Suppose the adversary knows the anonymization mechanism and the published table T_1. From the perspective of the adversary, T_1 has only one specialization (i.e., Table I(b)) which violates 2-diversity. One interesting question is: Will T_1 still be generated if the original table like Table I(a) satisfies 2-diversity? The answer is trivially no. Without loss of generality, we assume the original table is Table I(a). The following analysis still holds in the other table which is one of the specializations of T_1 which satisfies 2-diversity. Consider Table I(a). It is easy to follow Zhang's algorithm with this table and generate the published table as the original table (i.e., Table I(a)). In other words, among all possible specializations which either satisfies 2-diversity or not, there is only one specialization (i.e., Table I(b)) which triggers Zhang's algorithm to generate the published Table XX(a). Thus, the adversary deduces that the set of tuples with QID $= q1$ is linked to HIV. In conclusion, Zhang's algorithm suffers from minimality attack defined in this article.

Similar to the Incognito-like algorithm, the running time of the attack for Zhang's algorithm is $O((p^2 n_s + p n_s^3 l)R)$ where R is the running time of Zhang's algorithm.

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 I(c) or further generalize $q2$ to Q in Table I(d), it does not deter the minimality attack. The result still reveals the linkage of $q1$ to HIV as before. We show later that optimal m-confidentiality is NP-complete for both global recoding and local recoding in Theorem 2 and Theorem 3, respectively.

Optimal global m-confidentiality: given a table T and a nonnegative cost e, can we generate a table T^* from T by global recoding (local 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-complete.*

THEOREM 3. *Optimal m-confidentiality under local recoding is NP-complete.*

From the preceding, optimal m-confidentiality under both global and local recoding are difficult problems, and known heuristical methods based on incremental generalization may not help in defending the attack. 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 or suppress 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.

Another naive method can be to generalize or suppress the table randomly in order to avoid any minimality notion. Similarly, the information loss of this published table is high.

Here we propose a feasible solution for the m-confidentiality problem which can give low information loss. Although some problems are uncovered that question privacy protection by k-anonymity [Machanavajjhala et al. 2006], k-anonymity has been successful in some practical applications. This indicates that when a dataset 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.

In addition to the blinding effect for privacy protection, k-anonymity has its advantage of giving low information loss of the published table. A typical algorithm for k-anonymity tries to minimize the QID values within each QID-EC and thus generate the published table with low information loss. It is noted that this algorithm is not targeting for l-diversity (although it tries to minimize the information loss).

Let us call our solution Algorithm MASK (Minimality Attack Safe K-anonymity), which involves four steps as shown in Algorithm 1.

It is noted that Algorithm MASK is not limited to a particular anonymization/recoding strategy. In step 1, any existing algorithm for k-anonymity can be adopted. If the adopted algorithm is a global (local) recoding algorithm, then Algorithm MASK generates the table by global (local) recoding.

After Step 1, some QID-EC's may not satisfy l-diversity. Steps 2 to 4 in the algorithm will ensure that all QID-EC's in the result are l-diverse. In Step 2 as given, 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.

Algorithm. 1 – MASK

1: From the given table T, generate a minimal 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.
   (a) We set the size of $\mathcal{L}$, denoted by u, to $(l-1)\times|\mathcal{V}|$. If the total number of QID-EC's which satisfy l-diversity in T^k is smaller than $u(=(l-1)\times|\mathcal{V}|)$, we report that the table cannot be published. (Note: This case is rare because since typically there are rare sensitive values in a table, there are a sufficient number of QID-EC's which satisfy l-diversity in T^k.) Otherwise, we do the following: among all the QID-EC's in T^k that satisfy l-diversity, we pick u QID-EC's with the highest proportions of the sensitive value set s and insert them into $\mathcal{L}$.
   (b) If $\mathcal{V}=\emptyset$, then we can return T^k as our published table. Otherwise, then we do continue the following steps.

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. We name the QID-EC E in which sensitive values are distorted as a *distorted QID-EC*.

Table XXI. An Illustration of Algorithm MASK

QID	Disease
$q1$	HIV
$q2$	HIV
$q3$	HIV
$q3$	non-sensitive
$q4$	non-sensitive
$q4$	non-sensitive

(a) A raw table T

QID	Disease
Q	HIV
Q	HIV
$q3$	HIV
$q3$	non-sensitive
$q4$	non-sensitive
$q4$	non-sensitive

(b) 2-anonymous table of T

QID	Disease
Q	HIV
Q	non-sensitive
$q3$	HIV
$q3$	non-sensitive
$q4$	non-sensitive
$q4$	non-sensitive

(c) 2-confidential table of T

Example 1. We illustrate algorithm MASK with the raw table as shown in Table XXI(a). Suppose we set $k=2$ and $m=2$. The first step is to generate a 2-anonymous table. Suppose $q1$ and $q2$ can be generalized to Q. Firstly, we adopt an algorithm for k-anonymous which generates table T^k as shown in Table XXI(b) where all QID-EC's are of size at least 2. Secondly, from Table XXI(b), we obtain $\mathcal{V}=\{Q\}$. This is because the set of tuples with QID $=Q$ violates 2-diversity. We also obtain $\mathcal{L}=\{q3\}$. This is because both the set of tuples with QID $=q3$ and the set of tuples with QID $=q4$ satisfy 2-diversity. More specifically, the proportion of the tuples with QID $=q3$ is 0.5 and the proportion of the tuples with QID $=q4$ is 0. Since we set the size of $\mathcal{L}$ to be equal to $(l-1)\times|\mathcal{V}|=(2-1)\times 1=1$, we choose the QID-EC with QID $=q3$ for $\mathcal{L}$ because it has the greatest proportion of HIV among the QID-EC's in T^k that satisfy 2-diversity. Thirdly, we determine the distribution $\mathcal{D}$ of the p_i values. Since there is only one QID-EC (with HIV proportion equal to 0.5) in $\mathcal{L}$, we create a distribution $\mathcal{D}$ such that 0.5 must be returned whenever we draw from distribution $\mathcal{D}$. (Suppose $\mathcal{L}$ is equal to $\{q3,q4\}$. Since the proportions of

the QID-EC with QID $= q3$ and the QID-EC with QID $= q4$ are equal to 0.5 and 0, respectively, $\mathcal{D}$ is the distribution where the probability that 0.5 is returned is equal to 0.5 and the probability that 0 is returned is equal to 0.5.) Fourthly, we randomly pick a value from distribution $\mathcal{D}$. Suppose 0.5 is picked. Since $\mathcal{V}$ contains the QID-EC E with QID $= Q$ only, the sensitive values in E are distorted such that the resulting proportion of HIV in E is equal to 0.5. For example, we change one of the sensitive values in E to a nonsensitive value. Finally, Table XXI(c) is generated and is ready to be published. □

6.1 Analysis

We first prove that algorithm MASK generates an m-confidential table in Section 6.1.1. Then, we show some possible attacks from the perspective of the adversary in Section 6.1.2. The complexity of algorithm MASK is analyzed in Section 6.1.3. Finally, Section 6.1.4 describes that algorithm MASK can also be extended to other cases such as near-to-minimality attack easily.

6.1.1 *m-confidentiality.*

THEOREM 4. *Algorithm MASK generates m-confidential datasets.*

6.1.2 *Possible Potential Attack.* Let us analyze some possible potential attacks if the adversary obtains the table generated from algorithm MASK. We will show that the adversary cannot breach individual privacy from these possible attacks.

First, the adversary may breach individual privacy according to the frequency of the tuples with a sensitive value set s in a QID-EC. Suppose individual o is a target individual for the attack and o is in the QID-EC E in the published table T^*. Since the anonymization algorithm does not follow the minimality principle, the adversary may deduce that the probability that o is linked to s is equal to the frequency of the tuples with s in E. Since the published table is l-diverse and thus E is also l-diverse, the probability is at most $1/l$.

Second, s/he can deduce that the probability that an individual is linked to a sensitive value set s is greater than $1/l$ according to the distribution of the sensitive values among the QID-EC's in the published table. Suppose the size of $\mathcal{L}$ is equal to 1 only, instead of $(l-1) \times |\mathcal{V}|$. We will show that s/he may breach individual privacy with a higher chance. In this case, it is very likely that algorithm MASK will distort each QID-EC $E \in \mathcal{V}$ such that the frequency of the sensitive value set s is equal to $1/l$. In other words, the published table T^* contains many repeated occurrences of the QID-EC's which have $1/l$ tuples with s. From the adversary's perspective, it is likely that these QID-EC's originally violate l-diversity in the original table T. Thus, this may breach individual privacy.

Thus, if we set the size of $\mathcal{L}$ to be $(l-1)\times|\mathcal{V}|$, we can protect individual privacy. More specifically, 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 $1/l(= 1/m)$.

6.1.3 *Complexity.* The first step of algorithm MASK takes $O(R)$ where R is the running time of an algorithm which generates a k-anonymous table. Since the number of sensitive values is fixed, it is easy to verify that the second step and the third step take $O(|T|)$ where $|T|$ is the number of tuples in table T. If the number of QID-EC's in $\mathcal{V}$ is $|\mathcal{V}|$, the fourth step takes $O(|\mathcal{V}|)$. Thus, the total running time of algorithm MASK is $O(R + |T| + |\mathcal{V}|) = O(R + |T|)$.

THEOREM 5. *Algorithm MASK takes $O(R+|T|)$ time where R is the running time of the algorithm adopted for k-anonymity.*

Suppose each QID attribute has a generalization taxonomy of height h. Let $|A|$ be the total number of QID attributes. If we adopt a global-recoding algorithm, Incognito [LeFevre et al. 2005], in our first step to generate a k-anonymous table, the total running time of algorithm MASK is $O(|T| \times h \times \sum_{i=0}^{|A|\times h} C_i^{|A|\times h} + |T|) = O(|T| \times h \times \sum_{i=0}^{|A|\times h} C_i^{|A|\times h})$. If a local-recoding algorithm, (α, k)-anonymity [Wong et al. 2006], is adopted, the total running time becomes $O(|T|^2 \times h + |T|) = O(|T|^2 \times h)$.

In our experiment, the first step of k-anonymization (which takes $O(R)$ time) occupies over 98% of the whole execution time of algorithm MASK. Thus, the remaining steps are insignificant with respect to the whole execution time.

6.1.4 *Extension.* It is noted that algorithm MASK can also generate m-confidential datasets when the near-to-minimality attack, instead of the minimality attack, is considered. This is because MASK does not follow the near-to-minimality principle.

6.2 Generation of Two Tables—Bucketization

Conventional anonymization methods produce a single generalized table T as shown in Table VI. Recently Xiao and Tao [2006a] 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 will try both the approach of generating a single table T and the approach of generating two tables (also known as bucketization) as in Xiao and Tao [2006a], Zhang et al. [2007], and Martin et al. [2007].

Table XXII. Description of Adult Dataset

	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

7. EMPIRICAL STUDY

A Pentium IV 2.2GHz PC with 1 GM RAM was used to conduct our experiment. The algorithm was implemented in C/C++. In our experiment, we adopted the publicly available dataset, Adult Database from the UCIrvine Machine Learning Repository [Blake and Merz 1998]. This dataset (5.5 MB) was also adopted by LeFevre et al. [2005], Machanavajjhala et al. [2006], Wang et al. [2004], and Fung et al. [2005]. We used a configuration similar to LeFevre et al. [2005], and Machanavajjhala et al. [2006]. The records with unknown values were first eliminated resulting in a dataset with 45,222 tuples (5.4 MB). Nine attributes were chosen in our experiment, as shown in Table XXII. By default, we chose the first eight attributes and the last attribute in Table XXII 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 dataset with existing minimality-based anonymization algorithms. We adopted the Adult dataset and the selected algorithm was the (α, k)-anonymity algorithm [Wong et al. 2006]. 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 the minimality attack successfully uncovered QID-EC's which violates m-confidentiality, where $m = l$. We use m and l interchangeably in the following. Let us call the tuples in such QID-EC's the problematic tuples. Figure 7(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 7(b), when m increases, it is obvious that the average credibility of problematic tuples decreases. When m increases, $1/m$ decreases. Thus, each

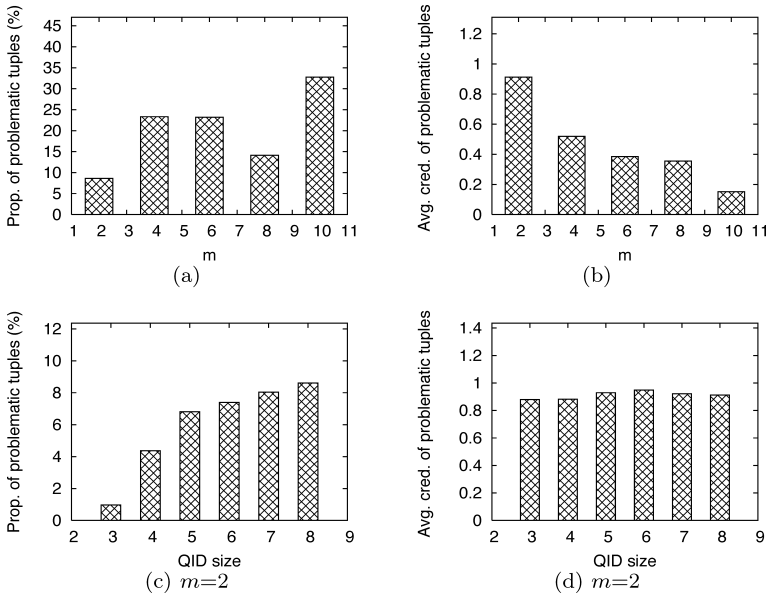

Fig. 7. Proportion of problematic tuples and average credibility of problematic tuples against m and QID size.

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 7(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 7(d) shows that the average credibility of problematic tuples remains 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 7(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" as follows. Thus, the published table contains the following tuples.

Age	Workclass	Martial Status	Education
80	with-pay	married-spouse-absent	7th-8th
80	with-pay	married-spouse-absent	HS-grad
80	private	married-spouse-absent	HS-grad

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.

7.2 Analysis of the Proposed Algorithm

We compared our proposed algorithm with a local recoding algorithm for (α, k)-anonymity [Wong et al. 2006] ((α, k)-A). Let us refer to our proposed algorithm MASK described in Section 6 by m-$conf$.

Since we are the first to propose the privacy model with the consideration of the minimality principle, there is no existing algorithm which can protect individual privacy when the minimality principle is considered. As shown in Section 5, all existing works which rely on the minimality principle are also vulnerable to the minimality attack. For the sake of interest, we compare our proposed algorithm with k-anonymity or l-diversity because our proposed algorithm involves two components: (1) k-anonymity in step 1 (considering QID attributes only) and (2) l-diversity in steps 2–4 (considering both QID attributes and the sensitive attribute together). (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 7. 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

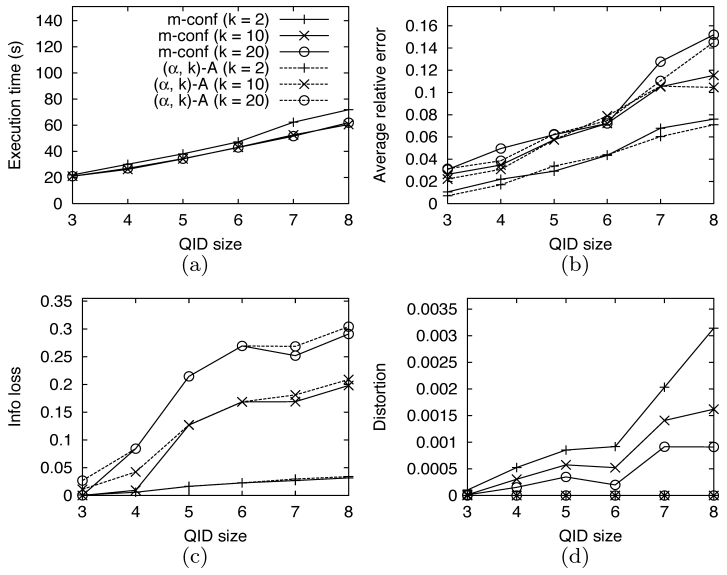

Fig. 8. Performance vs QID size ($m = 2$).

paragraph in Section 6.2). 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 Xiao and Tao [2006a], 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 Xiao and Tao [2006a]. 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 8 and Figure 9. One important observation is that the results are little affected by the values of k, which varies from 2 to 10 to 20, and this is true for the execution time, the relative error, the information loss, and for 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

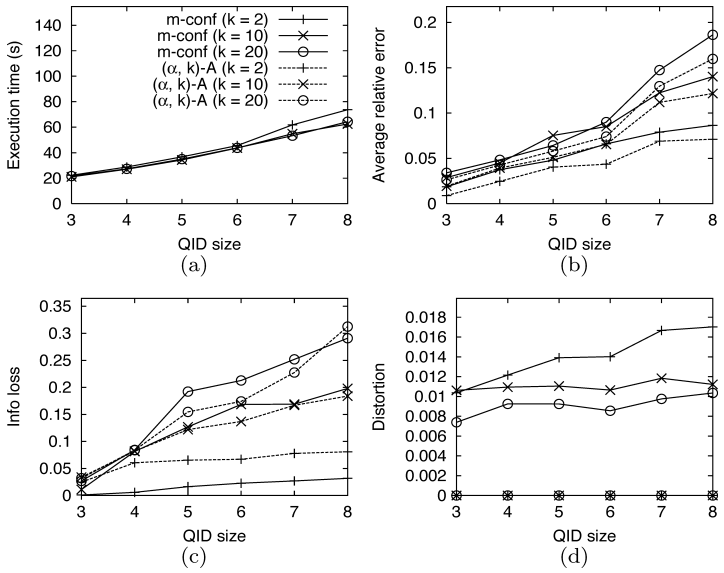

Fig. 9. Performance vs. QID size ($m = 10$).

to fend off the 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 8(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 8(b) and 8(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 8(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.

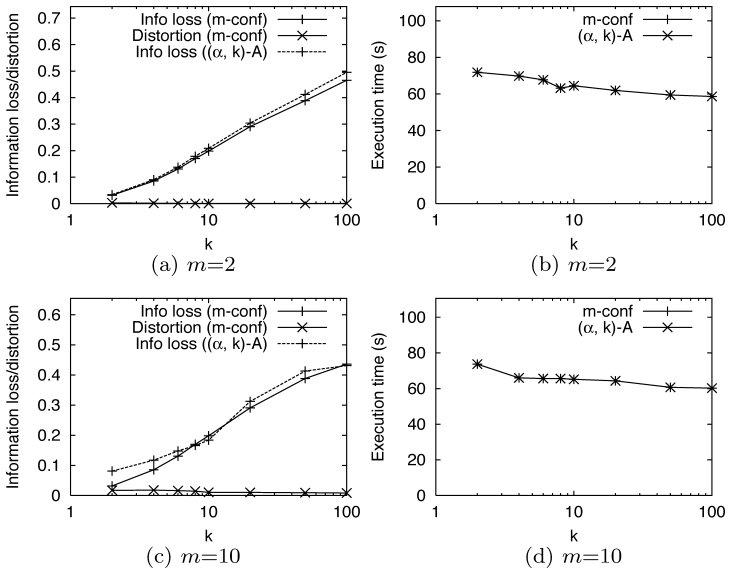

Fig. 10. Two Tables case: effect of varying m and k.

7.2.2 *The Two Tables Approach.* Our next set of experiments analyze the performance of the two table approach under various conditions.

Effect of k. Figure 10 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 11(a) shows that 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 11(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 are 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 overhead and pays a very minimal price in information loss in the exchange for m-confidentiality.

8. CONCLUSION

In existing privacy preservation methods for data publishing, minimality in information loss is an underlying principle. In this article, 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 also show that a near-to-minimality attack is still possible. We propose the m-confidentiality model which deals with attack

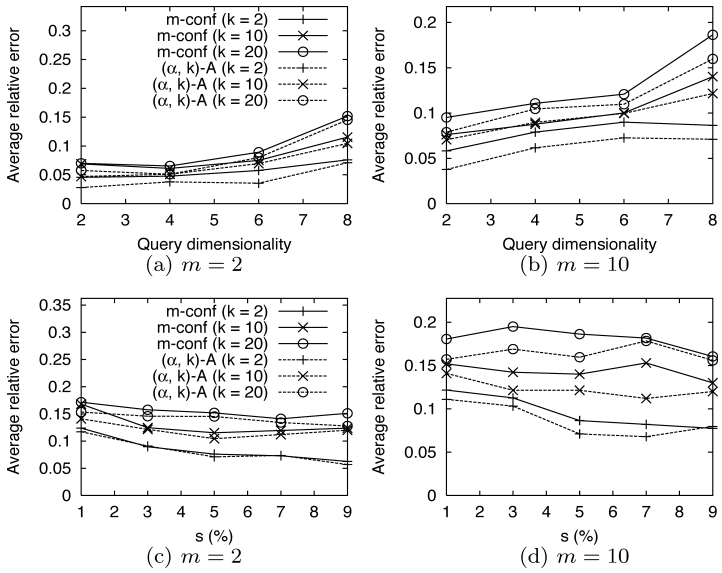

Fig. 11. Two tables case: effects of varying query dimensionality and selectivity.

by minimality and attack by near-to-minimality. We also propose an algorithm which generates an m-confidential dataset. We conducted experiments to show that our proposed algorithm requires little overhead both in terms of execution time and information loss. For future work we are interested to determine any other kinds of attack that can be related to the nature of the anonymization process.

REFERENCES

AGGARWAL, G., FEDER, T., KENTHAPADI, K., MOTWANI, R., PANIGRAHY, R., THOMAS, D., AND ZHU, A. 2005a. Anonymizingtables. In *Proceedings of the International Conference on Database Theory (ICDT'05),* 246–258.

AGGARWAL, G., FEDER, T., KENTHAPADI, K., MOTWANI, R., PANIGRAHY, R., THOMAS, D., AND ZHU, A. 2005b. Approximation algorithms for k-anonymity. *J. Privacy Technol.*

AGRAWAL, R. AND SRIKANT, R. 2000. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. ACM Press, 439–450.

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

BRUMLEY, D. AND BONEH, D. 2003. Remote timing attacks are practical. In *Proceedings of the USENIX Security Symposium.*

BU, Y., FU, A. W.-C., WONG, R. C.-W., CHEN, L., AND LI, J. 2008. Privacy preserving serial data publishing by role composition. In *Proceedigs of the International Conference on Very Large Databases*.

CIRIANI, V., VIMERCATI, S. D. C. D., FORESTI, S., AND SAMARATI, P. 2007. k-Anonymity. In *Security in Decentralized Data Management*.

EVFIMIEVSKI, A., SRIKANT, R., AND RAKESH AGRAWAL, J. G. 2002. Privacy preserving mining of association rules. In *Proceedings of the International SIGKDD Conference on Knowledge Discovery and Data Mining*.

FAYYAD, U. M. AND IRANI, K. B. 1993. Multi-Interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI'93)*. Morgan Kaufmann.

FUNG, B. C. M., WANG, K., AND YU, P. S. 2005. Top-down specialization for information and privacy preservation. In *Proceedings of the International Conference on Data Engineering (ICDE*), 205–216.

GAREY, M. AND JOHNSON, D. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* Freeman.

HOLYER, I. 1981. The np-completeness of some edge-partition problems. *SIAM J. Comput. 10*, 4, 713–717.

KIFER, D. AND GEHRKE, J. 2006. Injecting utility into anonymized datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data.*

KOCHER, P. C. 1996. Timing attacks on implementations of Diffe-Hellman RSA, DSS, and other systems. In *Proceedings of the International Cryptology Conference (CRYPTO'96)*, 104–113.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2005. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 49–60.

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006. Mondrian multidimensional k-anonymity. In *Proceedings of the International Conference on Data Engineering (ICDE).*

LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2008. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Datab. Syst. 33*, 3.

LI, N. AND LI, T. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the International Conference on Data Engineering (ICDE).*

LI, T. AND LI, N. 2008. Injector: Mining background knowledge for data anonymization. In *Proceedings of the International Conference on Data Engineering (ICDE).*

MACHANAVAJJHALA, A., GEHRKE, J., AND KIFER, D. 2006. l-Diversity: Privacy beyond k-anonymity. In *Proceedings of the International Conference on Data Engineering (ICDE).*

MARTIN, D. J., KIFER, D., MACHANAVAJJHALA, A., AND GEHRKE, J. 2007. Worst-case background knowledge for privacy-preserving data publishing. In *Proceedings of the International Conference on Data Engineering (ICDE).*

MEYERSON, A. AND WILLIAMS, R. 2004. On the complexity of optimal k-anonymity. In *Proceedings of the Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'04),* 223–228.

SAMARATI, P. 2001. Protecting respondents' identities in micro-data release. *Trans. Knowl. Data Eng. 13*, 6, 1010–1027.

SAMARATI, P. AND SWEENEY, L. 1998. Generalizing data to provide anonymity when disclosing information (abstract). In *Proceedings of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems (PODS)*.

SWEENEY, L. 1997. Weaving technology and policy together to maintain confidentiality. *J. Law, Med. Ethics 25*, 2–3, 98–110.

SWEENEY, L. 2002a. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertainess, Fuzziness Knowl.-based Syst. 10*, 5, 571–588.

SWEENEY, L. 2002b. k-Anonymity: A model for protecting privacy. *Int. J. Uncertainty, Fuzziness Knowl.-based Syst. 10*, 5, 557–570.

WANG, K. AND FUNG, B. 2006. Anonymizing sequential releases. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

WANG, K., FUNG, B. C. M., AND YU, P. S. 2006. Handicapping attacker's confidence: An alternative to k-anonymization. *Knowl. Inf. Syst. Int. J.*

WANG, K., YU, P. S., AND CHAKRABORTY, S. 2004. Bottom-up generalization: A data mining solution to privacy protection. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'04),* 249–256.

WONG, R., FU, A., WANG, K., AND PEI, J. 2007. Minimality attack in privacy preserving data publishing. In *Proceedings of the International Conference on Very Large Databases (VLDB).*

WONG, R., LI, J., FU, A., AND WANG, K. 2006. (Alpha, k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishing. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

XIAO, X. AND TAO, Y. 2006a. Anatomy: Simple and effective privacy preservation. In *Proceedings of the International Conference on Very Large Databases (VLDB).*

XIAO, X. AND TAO, Y. 2006b. Personalized privacy preservation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data.*

XIAO, X. AND TAO, Y. 2007. m-Invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*.

XU, J., WANG, W., PEI, J., WANG, X., SHI, B., AND FU, A. 2006. Utility-based anonymization using local recoding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

ZHANG, L., JAJODIA, S., AND BRODSKY, A. 2007. Information disclosure under realistic assumptions: Privacy versus optimality. In *Proceedings of the 14th ACM Conference on Computer and Communications Security.*

ZHANG, Q., KOUDAS, N., SRIVASTAVA, D., AND YU, T. 2007. Aggregate query answering on aononymized tables. In *Proceedings of the International Conference on Data Engineering (ICDE'07).*

Received March 2008; revised October 2008; accepted November 2008

Online Appendix to: Anonymization-Based Attacks in Privacy-Preserving Data Publishing

RAYMOND CHI-WING WONG
The Hong Kong University of Science and Technology
ADA WAI-CHEE FU
The Chinese University of Hong Kong
and
KE WANG and JIAN PEI
Simon Fraser University

A. DYNAMIC PROGRAMMING FOR EFFICIENT CREDIBILITY COMPUTATION

In this section, we will describe how we compute the credibility efficiently by dynamic programming. Specifically, Prob($|C_i(s)| = j \mid K_{ad}^{min}$) can be calculated by a dynamic programming approach. Before describing how to make use of a dynamic programming approach, we define the following events. Let F_i be the event that $0 \leq |C_i(s)| \leq \lfloor \frac{n_i}{l} \rfloor$. Let G_i be the event that $\lfloor \frac{n_i}{l} \rfloor + 1 \leq |C_i(s)| \leq n_i$. Let H_i be the event that $0 \leq |C_i(s)| \leq n_i$.

We illustrate the events in Figure 12. We can see that $F_i \cup G_i = H_i$.

The aim is to evaluate Prob($|C_i(s)| = j \mid K_{ad}^{min}$).

$$\begin{aligned} & Prob(|C_i(s)| = j \mid K_{ad}^{min}) \\ = \; & Prob(|C_i(s)| = j \mid \text{at least one } C_k \text{ among } C_1, C_2, \ldots, C_p \text{ violates } l\text{-diversity}) \end{aligned}$$

Since the event that at least one C_k among $C_1, C_2, \ldots, C_p$ violates l-diversity is equal to the event that at least one G_k occurs among $G_1, G_2, \ldots, G_p$, we have

$$\begin{aligned} & Prob(|C_i(s)| = j \mid K_{ad}^{min}) \\ = \; & Prob(|C_i(s)| = j \mid \text{at least one } G_k \text{ occurs among } G_1, G_2, \ldots, G_p) \\ = \; & \frac{\text{total no. of cases that } |C_i(s)| = j \text{ and at least one } G_k \text{ occurs among } G_1, \ldots, G_p}{\text{total no. of cases that at least one } G_k \text{ occurs among } G_1, G_2, \ldots, G_p}. \end{aligned}$$

Let A be the numerator (i.e., total number of cases that $|C_i(s)| = j$ and at least one G_k occurs among $G_1, \ldots, G_p$). Let B be the denominator (i.e., total number of cases that at least one G_k occurs among $G_1, G_2, \ldots, G_p$).

In the following, we consider the number of cases in the records in $C_1, C_2, \ldots, C_p$ only. Let there be x sensitive values in $C_1, C_2, \ldots, C_p$. Suppose that from dynamic programming, the total number of cases in the records in $C_1, C_2, \ldots, C_p$ is equal to Q. We can easily obtain the total number of cases in the records in all classes (i.e., $C_1, C_2, \ldots, C_p, C_{p+1}, \ldots, C_u$) by multiplying Q by $C_{n_s - x}^{N}$, where N is the total number of records in $C_{p+1}, C_{p+2}, \ldots, C_u$ and n_s is the total number of sensitive values in the total dataset.

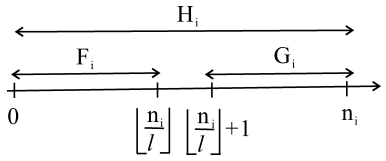

Fig. 12. Illustration of some events.

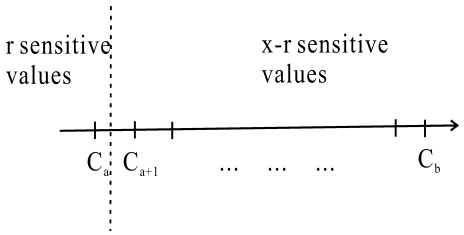

Fig. 13. Illustration of $n([a, b], x)$, $m([a, b], x)$, and $u([a, b], x)$.

For dynamic programming, we make use of three variables for the computation of A and B.

(1) $n([a, b], x)$ is the number of cases where at least one G_k occurs among G_a, G_{a+1}, ..., G_b when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.
(2) $m([a, b], x)$ is the number of cases where H_a, H_{a+1}, ... and H_b occur when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.
(3) $u([a, b], x)$ is the number of cases where F_a, F_{a+1}, ... and F_b occur when there are x sensitive values in C_a, C_{a+1}, ..., C_b, for $a, b = 1, 2, \ldots, p$ and $x = 1, 2, \ldots, n_s$.

Consider $m([a, b], x)$. Among $C_a, C_{a+1}, \ldots, C_b$, we divide the classes into two parts, $\{C_a\}$ and $\{C_{a+1}, \ldots, C_b\}$. See Figure 13. Suppose we allocate r sensitive values to C_a and $x - r$ sensitive values to $C_{a+1}, \ldots, C_b$. The number of cases where there are r sensitive values in class C_a of size n_a is equal to $C_r^{n_a}$. The number of cases where $G_{a+1}, \ldots, G_b$ occur when the number of sensitive values allocated to them is equal to $x - r$ is equal to $m([a + 1, b], x - r)$. Thus, for a given r, the total number of cases is equal to $C_r^{n_a} \times m([a + 1, b], x - r)$.

$$m([a, b], x) = \sum_{r=0}^{n_a} C_r^{n_a} \times m([a + 1, b], x - r)$$

We define the base cases of $m([a, b], x)$ as follows. The base case happens when $a = b$. It is impossible that the number of sensitive values allocated to C_a is greater than the class size of C_a or smaller than 0. Thus the term should be set to 0 in both cases. If the number of sensitive values allocated to C_a ranges from 0 to n_a, the term is the number of possible combinations where there are

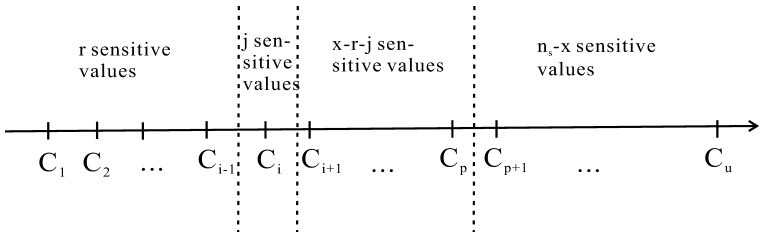

Fig. 14. Illustration of $\mathcal{A}(x)$.

x sensitive values in class C_a of size n_a (i.e., $C_x^{n_a}$).

$$m([a,a],x) = \begin{cases} 0 & \text{if} \quad x > n_a \\ 0 & \text{if} \quad x < 0 \\ C_x^{n_a} & \text{if } 0 \le x \le n_a \end{cases}$$

The term $u([a,b],x)$ is the same as $m([a,b],x)$ except that the upper boundary of term F_i is equal to $\lfloor \frac{n_a}{l} \rfloor$, instead of n_a. Similarly, we have the following formula.

$$u([a,b],x) = \sum_{r=0}^{\lfloor \frac{n_a}{l} \rfloor} C_r^{n_a} \times u([a+1,b],x-r)$$

$$u([a,a],x) = \begin{cases} 0 & \text{if } x \ge \lfloor \frac{n_a}{l} \rfloor + 1 \\ 0 & \text{if} \quad x < 0 \\ C_x^{n_a} & \text{if } 0 \le x \le \lfloor \frac{n_a}{l} \rfloor \end{cases}$$

Next consider $n([a,b],x)$. Let r be the number of tuples with s in C_a. We can also derive $n([a,b],x)$ as follows similarly by considering two cases: (1) $\lfloor \frac{n_a}{l} \rfloor + 1 \le r \le n_a$ and (2) $0 \le r \le \lfloor \frac{n_a}{l} \rfloor$.

$$n([a,b],x) = \sum_{r=\lfloor \frac{n_a}{l} \rfloor + 1}^{n_a} C_r^{n_a} \times m([a+1,b],x-r) + \sum_{r=0}^{\lfloor \frac{n_a}{l} \rfloor} C_r^{n_a} \times n([a+1,b],x-r)$$

The base cases of $n([a,b],x)$ can also be easily derived as follows.

$$n([a,a],x) = \begin{cases} 0 & \text{if} \quad x > n_a \\ 0 & \text{if} \quad 0 \le x \le \lfloor \frac{n_a}{l} \rfloor \\ C_x^{n_a} & \text{if } \lfloor \frac{n_a}{l} \rfloor + 1 \le x \le n_a \end{cases}$$

Now, consider A. Recall that A is the total number of cases that; (1) at least one G_k occurs among $G_1, G_2, \ldots, G_p$ and (2) there are j sensitive values in C_i.

Let $\mathcal{A}(x)$ be the total number of aforesaid cases provided that there are x sensitive values in $C_1, C_2, \ldots, C_p$.

We consider the number of cases involving all classes (i.e., $C_1, \ldots, C_p$, $C_{p+1}, \ldots, C_u$). Suppose we allocate x sensitive values in $C_1, C_2, \ldots, C_p$ and $n_s - x$ sensitive values in $C_{p+1}, C_{p+2}, \ldots, C_u$. Recall that C_i contains j sensitive values. Within $C_1, C_2, \ldots, C_p$, we further allocate: (1) r sensitive values to $C_1, C_2, \ldots, C_{i-1}$, (2) j sensitive values to C_i, and (3) $x - r - j$ sensitive values to $C_{i+1}, C_{i+2}, \ldots, C_p$. See Figure 14.

There are two cases.

Case 1. $\lfloor \frac{n_i}{l} \rfloor + 1 \leq j \leq n_i$, that is, G_i occurs. This means that the number of sensitive values in a class C_k ($|C_k(s)|$) of $C_1, C_2, \ldots, C_{i-1}$ or $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k.

The number of cases that $|C_k(s)|$ for a class C_k of $C_1, C_2, \ldots, C_{i-1}$ ranges from 0 to n_k is equal to $m([1, i-1], r)$. Similarly, the number of cases that $|C_k(s)|$ for a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k is equal to $m([i+1, p], x-r-j)$. Thus, if we consider all possible values of r from 0 to $x-j$, the total number of these cases is equal to $\sum_{r=0}^{x-j} m([1, i-1], r) \times m([i+1, p], x-r-j)$.

Note that the number of cases that there are j sensitive values in C_i of size n_i is equal to $C_j^{n_i}$. Also, the number of cases that there are $n_s - x$ sensitive values in N tuples in classes $C_{p+1}, C_{p+2}, \ldots, C_u$ is equal to $C_{n_s-x}^N$. Thus, $\mathcal{A}(x) = C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} m([1, i-1], r) \times m([i+1, p], x-r-j)$ in this case.

Case 2. $0 \leq j \leq \lfloor \frac{n_i}{l} \rfloor$, that is, G_i does not occur. There are the following subcases.

Case 2(a). At least one G_k occurs among $G_1, G_2, \ldots, G_{i-1}$, In this case, the number of sensitive values in a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k.

The number of cases that at least one G_k occurs among $G_1, G_2, \ldots, G_{i-1}$ is equal to $n([1, i-1], r)$. The number of cases that the number of sensitive values in a class C_k of $C_{i+1}, C_{i+2}, \ldots, C_p$ ranges from 0 to n_k is equal to $m([i+1, p], x-r-j)$. Thus, the total number of these cases is equal to $n([1, i-1], r) \times m([i+1, p], x-r-j)$.

Case 2(b). All G_k among $G_1, G_2, \ldots, G_{i-1}$ does not occur. In other words, all $F_1, F_2, \ldots, F_{i-1}$ occur. Besides, we should also know that there is at least one G_k occuring among $G_{i+1}, G_{i+2}, \ldots, G_p$.

The number of cases that all $F_1, F_2, \ldots, F_{i-1}$ occur is equal to $u([1, i-1], r)$. The number of cases that at least one G_k occurs among $G_{i+1}, G_{i+2}, \ldots, G_p$ is equal to $n([i+1, p], x-r-j)]$. Thus, the total number of these cases is equal to $u([1, i-1], r) \times n([i+1, p], x-r-j)$.

By combining Case 2(a) and Case 2(b) and considering all possible values r from 0 to $x-j$, we obtain the total number of cases equal to $\sum_{r=0}^{x-j} [n([1, i-1], r) \times m([i+1, p], x-r-j) + u([1, i-1], r) \times n([i+1, p], x-r-j)$.

Similarly, the number of cases that there are j sensitive values in C_i of size n_i is equal to $C_j^{n_i}$. Also, the number of cases that there are $n_s - x$ sensitive values in N tuples in classes $C_{p+1}, C_{p+2}, \ldots, C_u$ is equal to $C_{n_s-x}^N$. Thus, the total number of cases in Case (2) is equal to $C_{n_x-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} [n([1, i-1], r) \times m([i+1, p], x-r-j) + u([1, i-1], r) \times n([i+1, p], x-r-j)]$.

We obtain $\mathcal{A}(x)$ as follows.

$$\mathcal{A}(x) = \begin{cases} C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} m([1, i-1], r) \\ \quad \times m([i+1, p], x-r-j) & \text{if } \lfloor \frac{n_i}{l} \rfloor + 1 \leq j \leq n_i \\ C_{n_s-x}^N \times C_j^{n_i} \times \sum_{r=0}^{x-j} [n([1, i-1], r) \\ \quad \times m([i+1, p], x-r-j) & \text{if } 0 \leq j \leq \lfloor \frac{n_i}{l} \rfloor \\ \quad + u([1, i-1], r) \times n([i+1, p], x-r-j)] \end{cases}$$

By considering all possible values of x from $\lfloor \frac{n_1}{l} \rfloor + 1$ to n_s, A is equal to the following. (Note that it is impossible that $x < \lfloor \frac{n_1}{l} \rfloor + 1$ because it means that there is no need for generalization.)

$$A = \sum_{x=\lfloor \frac{n_1}{l} \rfloor + 1}^{n_s} \mathcal{A}(x)$$

Consider B where B is the total number of cases where at least one G_k occurs among $G_1, G_2, \ldots, G_p$. By considering all possible values x from $\lfloor \frac{n_1}{l} \rfloor + 1$ to n_s, we obtain the following formula.

$$B = \sum_{x=\lfloor \frac{n_1}{l} \rfloor}^{n_s} n([1, p], x) \times C_{n_s - x}^{N}$$

Algorithm. Algorithm 2 shows the computation of the credibility by dynamic programming. It involves two phases. In phase 1, we compute the variables $n([a, b], x)$, $m([a, b], x)$ and $u([a, b], x)$. In phase 2, we compute $Credibility(o, s, K_{ad}^{min})$ where $o \in C_i$ for $i = 1, 2, \ldots, p$ by using the variables used in phase 1, namely $n([a, b], x)$, $m([a, b], x)$, and $u([a, b], x)$.

Let $|T|$ be the number of tuples in T. Algorithm 2 runs in polynomial time in $|T|$, p, n_s, and l. It is easy to verify that phase 1 takes $O(|T| + p^2 n_s)$. Consider phase 2. Computing one instance of A and computing one instance of B take $O(n_s^2)$ and $O(n_s)$, respectively. Since there are $O(pn_p)$ iterations, phase 2 takes $O(pn_p n_s^2)$. Since $n_p = O(n_s l)$, the complexity of phase 2 becomes $O(pn_s^3 l)$. Thus, the running time of Algorithm 2 is $O(|T| + p^2 n_s + pn_s^3 l)$.

THEOREM 6. *Algorithm 2 runs in $O(|T| + p^2 n_s + pn_s^3 l)$ time.*

The previous theorem means that computing the credibility of an individual only takes polynomial time in $|T|$, p, n_S, and l. In other words, this kind of attack is highly feasible.

B. PROOF OF LEMMAS/THEOREMS

PROOF OF THEOREM 1. We will prove that the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC by first considering a class Q in T^* where only two QID values in T^e, namely q_1 and q_2, are generalized to Q. Then, we relax the proof by considering a class Q where multiple QID values are generalized to Q.

Consider a QID value Q in T^*. Suppose q_1 and q_2 (in T^e) are generalized to Q in T^*. Let n_1 and n_2 be the number of tuples with value q_1 and q_2, respectively. Let x be the total number of sensitive tuples in Q.

Consider four cases. *Case* 1: $x \leq n_1$ and $x \leq n_2$. Without loss of generality, we consider $Credibility(o, s, K_{ad}^{min})$ where o has a QID value on q_1. We further consider a number of subcases. *Case* (a): $x = 1$. We have the sensitive

Algorithm. 2 Algorithm for Computing Credibility

// Phase 1(a): Initialization
obtain n_i for all i
for $a = 1$ to p **do**
  for $x = 0$ to n_s **do**
    initialize $n([a,a],x)$, $m([a,a],x)$ and $u([a,a],x)$
  end for
end for
// Phase 1(b): Recursion
// Compute $m([a,b],x)$ and $u([a,b],x)$
for $x = 0$ to n_s **do**
  for $a = p$ downto 1 **do**
    for $b = a+1$ to p **do**
      $m([a,b],x) \leftarrow 0$
      for $r = 0$ to n_a **do**
        if $x - r \geq 0$ **then**
          $m([a,b],x) \leftarrow m([a,b],x) + C_r^{n_a} \times m([a+1,b],x-r)$
        end if
      end for
      $u([a,b],x) \leftarrow 0$
      for $r = 0$ to $\lceil n_a/l \rceil$ **do**
        if $x - r \geq 0$ **then**
          $u([a,b],x) \leftarrow u([a,b],x) + C_r^{n_a} \times u([a+1,b],x-r)$
        end if
      end for
    end for
  end for
end for
// Compute $n([a,b],x)$
for $x = 0$ to n_s **do**
  for $a = p$ downto 1 **do**
    for $b = a+1$ to p **do**
      $n([a,b],x) \leftarrow 0$
      for $r = \lfloor n_a/l \rfloor + 1$ to n_a **do**
        if $x - r \geq 0$ **then**
          $n([a,b],x) \leftarrow n([a,b],x) + C_r^{n_a} \times m([a+1,b],x-r)$
        end if
      end for
      for $r = 0$ to $\lfloor n_a/l \rfloor$ **do**
        if $x - r \geq 0$ **then**
          $n([a,b],x) \leftarrow n([a,b],x) + C_r^{n_a} \times n([a+1,b],x-r)$
        end if
      end for
    end for
  end for
end for
// Phase 2: Computing credibility $Credibility(o, s, K_{ad}^{min})$
Let $cred_i$ be $Credibility(o, s, K_{ad}^{min})$ where $o \in C_i$
for $i = 1$ to p **do**
  $cred_i \leftarrow 0$
  for $j = 1$ to n_i **do**
    calculate A and B according to $n([a,b],x)$, $m([a,b],x)$ and $u([a,b],x)$
    $cred_i \leftarrow cred_i + \frac{A}{B} \times \frac{j}{n_i}$
  end for
end for

Table XXIII. Possible Combinations of Number of Sensitive Tuples when $x = 1$

	Number of sensitive tuples		Total number of cases
	$q1$	$q2$	
(a)	0	1	$C_0^{n_1} \times C_1^{n_2}$
(b)	1	0	$C_1^{n_1} \times C_0^{n_2}$

Table XXIV. Possible Combinations of Number of Sensitive Tuples when $x = 2$

	Number of sensitive tuples		Total number of cases
	$q1$	$q2$	
(a)	0	2	$C_0^{n_1} \times C_2^{n_2}$
(b)	1	1	$C_1^{n_1} \times C_1^{n_2}$
(c)	2	0	$C_2^{n_1} \times C_0^{n_2}$

tuple distribution table as shown in Table XXIII. It is easy to see that

$$\begin{aligned} \mathit{Credibility}(o, s, K_{ad}^{min}) &= \frac{\text{total number of cases for Scenario (b)}}{\text{total number of all possible cases}} \times \frac{1}{n_1} \\ &= \frac{C_1^{n_1} \times C_0^{n_2}}{C_0^{n_1} \times C_1^{n_2} + C_1^{n_1} \times C_0^{n_2}} \times \frac{1}{n_1} \\ &= \frac{n_1}{n_2 + n_1} \times \frac{1}{n_1} \\ &= \frac{1}{n_1 + n_2} \end{aligned}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

Case (b): $x = 2$. Similarly, we have the sensitive tuple distribution table as shown in Table XXIV. We have

$$\begin{aligned} \mathit{Credibility}(o, s, K_{ad}^{min}) &= \frac{\text{total number of cases for Scenario (b)}}{\text{total number of all possible cases}} \times \frac{1}{n_1} \\ &\quad + \frac{\text{total number of cases for Scenario (c)}}{\text{total number of all possible cases}} \times \frac{2}{n_1} \\ &= \frac{C_1^{n_1} \times C_1^{n_2}}{C_0^{n_1} \times C_2^{n_2} + C_1^{n_1} \times C_1^{n_2} + C_2^{n_1} \times C_0^{n_2}} \times \frac{1}{n_1} \\ &\quad + \frac{C_2^{n_1} \times C_0^{n_2}}{C_0^{n_1} \times C_2^{n_2} + C_1^{n_1} \times C_1^{n_2} + C_2^{n_1} \times C_0^{n_2}} \times \frac{2}{n_1} \\ &= \frac{n_1 n_2}{\frac{n_2(n_2-1)}{2} + n_1 n_2 + \frac{n_1(n_1-1)}{2}} \times \frac{1}{n_1} \\ &\quad + \frac{\frac{n_1(n_1-1)}{2}}{\frac{n_2(n_2-1)}{2} + n_1 n_2 + \frac{n_1(n_1-1)}{2}} \times \frac{2}{n_1} \\ &= \frac{2(n_1 + n_2 - 1)}{n_1^2 + n_2^2 + 2n_1 n_2 - n_1 - n_2} \end{aligned}$$

$$= \frac{2(n_1 + n_2 - 1)}{(n_1 + n_2)(n_1 + n_2 - 1)}$$
$$= \frac{2}{n_1 + n_2}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

Case (c): $x > 2$. Inductively, we can also derive that

$$Credibility(o, s, K_{ad}^{min}) = \frac{x}{n_1 + n_2}$$

which is equal to the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC Q.

We consider the other three cases. *Case* 2: $x \leq n_1$ and $x > n_2$, *Case* 3: $x > n_1$ and $x \leq n_2$, and *Case* 4: $x > n_1$ and $x > n_2$. With similar arguments, we also conclude that

$$Credibility(o, s, K_{ad}^{min}) = \frac{x}{n_1 + n_2}.$$

Now, we consider the class Q where multiple QID values are generalized to Q. Since the idea is similar and the key idea is no exclusion of any scenarios in the sensitive tuple distribution table, we obtain that the credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. □

PROOF OF LEMMA 2. To prove this lemma, we give 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. □

PROOF OF THEOREM 2. We shall transform the problem of Exact Cover by 3-Sets (X3C) [Holyer 1981] to the m-confidentiality anonymization problem. X3C is defined by: Given a set X with $|X| = 3q$ and a collection C of 3-element subsets of X. Does C contain an exact cover for X, namely a subcollection $C' \subseteq C$ such that every element of X occurs in exactly one member of C'?

Given an instance of X3C, we transform it to an instance of optimal m-confidentiality under global recoding as follows. Create a table T with two attributes $\mathcal{Q}$ and S, where $\mathcal{Q}$ is a QID attribute and S is a sensitive attribute that may contain sensitive values. For S, there is only one sensitive value s_v and one nonsensitive value s_n. We set $weight(\mathcal{Q}) = 1$. For each element x in X, create a tuple with $\mathcal{Q} = x$ and $S = s_v$. Hence, each value of x appears in exactly one tuple. Let the elements in C be $c_1, \ldots, c_N$. For each element $c_i = (x, y, z)$ in C, create a taxonomy $\mathcal{T}_i$. $\mathcal{T}_i$ contains ground elements of x, y, z, n_{i1}, n_{i2}, and n_{i3}, which are children of a root node r_i. Create 3 tuples with $\mathcal{Q} = n_{ij}$ and $S = s_n$, for $j = 1, 2, 3$.

The remaining of the proof is to show: C contains an exact cover for X if and only if there is a solution T^* for the 2-confidentiality problem with

$Dist(T, T^*) = e$ where $e = \frac{2q}{q+N}$. Firstly, we prove that if C contains an exact cover for X, then there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$. Let C' be the exact cover for X. We know that every element of X occurs in exactly one member of C'. Then, for each $c_i = (x, y, z) \in C'$, the correspondence taxonomy $\mathcal{T}_i$ is used for the generalization of x, y, and z together with n_{i1}, n_{i2}, and n_{i3} because with global recoding all occurrences of an attribute value are recoded to the same value. Thus, for each generalization from $\mathcal{T}_i$, the information loss of these six tuples in T^* are 6. Since $|C'| = q$, the total information loss among all tuples (i.e., $\sum_{t^* \in T^*} \mathcal{IL}(t^*)$) is equal to $6q$. Since $Dist(T, T^*) = \frac{\sum_{t^* \in T^*} \mathcal{IL}(t^*)}{|T^*|}$ and the total number of tuples in T^* is equal to $3q + 3N$, we have $Dist(T, T^*) = \frac{6q}{3q+3N} = \frac{2q}{q+N}$. Besides, note that the adversary cannot launch a minimality attack since each QID value appears only in one tuple in the set of tuples. The adversary cannot exclude any possible combination of the table of sensitive tuple distribution. From Theorem 1, minimality attack is not possible. Besides, the frequency of each QID-EC with s_v in T^* is at most 0.5. Thus, there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$.

Now, we prove that if there is a solution T^* for the 2-confidentiality problem with $Dist(T, T^*) = \frac{2q}{q+N}$, then C contains an exact cover for X. Similarly, since each QID value appears only in one tuple, it is impossible for the adversary to exclude any possible combination of the table of sensitive tuple distribution. From Theorem 1, minimality attack is not possible. Since the frequency of each QID-EC with s_v in T^* is at most 0.5 and $Dist(T, T^*) = \frac{2q}{q+N}$, T^* is a result of the generalizations by using exactly q taxonomies $\mathcal{T}_i$ containing disjoint ground values. Otherwise, either the frequency of some QID-EC's in T^* is greater than 0.5 or $Dist(T, T^*) > \frac{2q}{q+N}$, that is, each tuple with value s is generalized by exactly one generalization taxonomy. In other words, each element in X occurs in exactly one member of the set $C' \subseteq C$ such that $|C'| = q$ and each $c \in C'$ corresponds to $\mathcal{T}_i$ used for generalization. Thus, C contains an exact cover C' for X.

Besides, it is easy to see that the reduction runs in polynomial time. From Theorem 6 (in Section 4), we know that we can compute the credibility of each individual in polynomial time. Thus, we can verify problem optimal m-confidentiality in polynomial time. So, problem optimal m-confidentiality under global recoding is NP-complete. □

Proof of Theorem 3. We shall transform the problem of Partition into 4-Cliques [Holyer 1981] to the m-confidentiality anonymization problem. Partition into 4-Cliques is defined by: Given a simple graph $G = (V, E)$, with $|E| = 6k$ for some integer k, can the edges of G be partitioned into k edge-disjoint 4-cliques?

Given an instance of Edge Partition into 4-Cliques. Set $m = 6$. For each vertex $v \in V$, construct a QID attribute. For each edge $e \in E$, where $e = (v_1, v_2)$, create a record r_{v_1,v_2} in which the QID attribute values v_1 and v_2 are equal to 1 and all other QID attribute values equal to 0. Besides, we associate each record with a sensitive attribute S. We generate sensitive attribute values of all records as follows. If two edges share a common vertex, the sensitive attribute

values of their corresponding records are different. The aforesaid principle can be accomplished with the following steps. Firstly, the sensitive values of all records are set to 0. Then, we randomly find a record r where the corresponding edge is e. Let A be the set of all records where the corresponding vertices have a common vertex with e. Then, we can obtain a set A' containing the sensitive values of all records in A. Find the smallest positive value which does not occur in A'. Assign this value as the sensitive attribute value of record r. Repeat the previous steps for each of the remaining records with sensitive value $= 0$. It is noted that the preceding process resembles a process of edge coloring. However, since the aforesaid process does not require that the edges are colored with a limited (or optimal) number of colors, it can be done in polynomial time.

We define the cost in the 6-confidentiality problem to be the number of suppressions applied in the dataset. We show that this cost is at most $24k$ if and only if E can be partitioned into a collection of k edge-disjoint 4-cliques.

Suppose E can be partitioned into a collection of k disjoint 4-cliques. Consider a 4-clique Q with vertices v_1, v_2, v_3, and v_4. If we suppress the attributes v_1, v_2, v_3, and v_4 in the 6 records corresponding to the edges in Q, then a cluster of these 6 records are formed where each modified record has four *'s. Note that the the frequency of each sensitive value in this cluster is at most 1/6. Similar to Theorem 2, the adversary cannot launch a minimality attack since each QID value appears only in one tuple in the cluster. Thus, the dataset satisfies 6-confidentiality. The cost of the 6-confidentiality is equal to $6 \times 4 \times k = 24k$.

Suppose the cost for the 6-confidentiality problem is at most $24k$. As G is a simple graph, any six records should have at least four different attributes. So each record should have at least four *'s in the solution of 6-confidentiality. Then, the cost of 6-confidentiality is at least $6 \times 4 \times k = 24k$. Combining with the proposition that the cost is at most $24k$, we find that the cost is exactly equal to $24k$ and thus each record should have exactly four *'s in the solution. Each cluster should have exactly 6 records (with different sensitive values). Suppose the six modified records contain four *'s in attributes v_1, v_2, v_3, and v_4, the records contain 0's in all other nonsensitive attributes. This corresponds to a 4-clique with vertices v_1, v_2, v_3 and v_4. Thus, we conclude that the solution corresponds to a partition into a collection of k edge-disjoint 4-cliques.

Similar to Theorem 2, it is easy to see that the reduction runs in polynomial time. From Theorem 6 (in Section 4), we know that we can compute the credibility of each individual in polynomial time. Thus, we can verify problem optimal m-confidentiality in polynomial time. We conclude that optimal m-confidentiality under local recoding is NP-complete. □

PROOF OF THEOREM 4. In order to prove that T^* generated by algorithm MASK is m-confidential, we analyze how the adversary performs an attack, given that the adversary knowledge contains not only the knowledge described in Assumption 3 but also the mechanism of algorithm MASK. In the following, we show that the credibility computed is at most $1/m$. Let the privacy requirement considered be $\mathcal{R}$ (i.e., m-confidentiality). Let the privacy requirement for k-anonymity be $\mathcal{R}^k$. From T^*, the adversary knows that, in T^*, for each

QID-EC Q_i, the size of Q_i is at least k and the frequency of each sensitive value (in fraction) is at most $1/m$. We consider two cases.

Case 1. The published table T^* is equal to the minimal k-anonymous table T^k generated in step 1 of algorithm MASK; that is, $|\mathcal{V}|$ is equal to $\emptyset$. We know that algorithm MASK generates T^* such that the information loss of T^* is minimal with respect to the QID attributes. Note that T^* is a result of generalization for privacy requirement $\mathcal{R}^k$ (instead of privacy requirement $\mathcal{R}$). However, at the same time, T^* also satisfies m-diversity.

We prove that T^* also satisfies $\mathcal{R}$ in the following. Similar to Section 4, we can also compute the credibility by constructing the sensitive tuple distribution table with condition (1) and condition (2) (but not condition (3)) of Definition 10 accordingly in this case. It is noted that we do not need to consider condition (3) since the generalization step for generating T^* is caused by "unequal" QID values in the original table T (without the consideration of the sensitive attribute). In other words, the generalization is performed for $\mathcal{R}^k$ (instead of m-diversity). Since condition (3) is not considered, there is no exclusion of any combination of the number of sensitive tuples in the sensitive tuple distribution table in the adversary's analysis of this case. (However, the existence of the exclusion used in Section 4 is due to the fact that the generalization is caused by the consideration of both QID attributes and the sensitive attribute; that is, the generalization is performed for m-diversity instead of $\mathcal{R}^k$ where condition (3) is involved during the generation of the sensitive tuple distribution table.)

Since there is no exclusion of any combination in the sensitive tuple distribution table in this case, by Theorem 1, the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. Besides, in T^*, for each QID-EC Q_i, the frequency of each sensitive value (in fraction) is at most $1/m$. We deduce that $Credibility(o, s, K_{ad}^{min})$ is at most $1/m$ for any individual o and any sensitive value set s. So, T^* satisfies $\mathcal{R}$.

Case 2. The published table T^* is not equal to T^k; that is, $|\mathcal{V}|$ is not equal to $\emptyset$. Then, the adversary knows that step 2(a) of algorithm MASK is performed. We consider two subcases.

Subcase (a). The total number of QID-EC's which satisfy m-diversity in T^k is smaller than $u(= (m-1) \times |V|)$. This case is impossible because T^* is already published.

Subcase (b). The total number of QID-EC's which satisfy m-diversity in T^k is equal to or greater than u. The analysis of the credibility in this case is different from Case 1. This analysis involves two major steps. The first step is that the adversary deduces that some of the nonsensitive values in T^* originally come from sensitive values in T. This is because some sensitive values are distorted or modified to become nonsensitive values in step 4 of algorithm MASK. The second step is similar to Section 4 and Case (1). Specifically, according to the sensitive values deduced in the first step, the adversary can compute the credibility by the sensitive tuple distribution table with condition (1) and condition (2) but not condition (3) of Definition 10 accordingly. Again, it is noted that we do not need to consider condition (3) since the generalization step for generating

T^* is caused by "unequal" QID values in the original table T (without the consideration of the sensitive attribute).

In the following, we will show that it is difficult for the adversary to achieve the first step. Then, with this result, we assume that we only need to consider the sensitive values in T^* to calculate the credibility in the second step.

For the first step, it is infeasible for the adversary to figure out what are the original sensitive values because the adversary does not have the knowledge about: (1) the size of $\mathcal{V}$, (2) the original frequency of the sensitive tuples in each QID-EC $\in \mathcal{V}$, and (3) which QID-EC's in T^* come from $\mathcal{V}$. It may be argued that the adversary can first consider all possible choices of the aforesaid knowledge, compute the credibility for each choice with the second step, and finally compute the final credibility with all choices. This approach does not work because without sufficient knowledge, we do not know the probability that each choice occurs. Assuming a random world assumption (i.e., all such probabilities have the same values) is also not reasonable. This is because, for example, the adversary cannot tell whether the probability that $|\mathcal{V}| = 1$ occurs is equal to the probability that $|\mathcal{V}| = 2$ occurs or not. With this reasoning, the deduction of the original sensitive values is impossible.

For the second step, we assume that the sensitive values considered come from T^*. By similar arguments as Case 1, since there is no exclusion of any combination, the credibility as computed by the formulae for credibility is exactly the ratio of the sensitive tuples to the total number of tuples in the generalized QID-EC. Besides, in T^*, for each QID-EC Q_i, the frequency of each sensitive value (in fraction) is at most $1/m$. $Credibilityx(o, s, K_{ad}^{min})$ is at most $1/m$ for any individual o and any sensitive value set s. So, T^* satisfies $\mathcal{R}$. □