Information Leakage in Optimal Anonymized and Diversified Data *

Chengfang Fang Ee-Chien Chang

School of Computing
National University of Singapore
fangchengfang@alumni.nus.edu.sg changec@comp.nus.edu.sg

Abstract. To reconcile the demand of information dissemination and preservation of privacy, a popular approach generalizes the attribute values in the dataset, for example by dropping the last digit of the postal code, so that the published dataset meets certain privacy requirements, like the notions of k-anonymity and ℓ-diversity. On the other hand, the published dataset should remain useful and not over generalized. Hence it is desire to disseminate a database with high "usefulness", measured by a *utility function*. This leads to a generic framework whereby the optimal dataset (w.r.t. the utility function) among all the generalized datasets that meet certain privacy requirements, is chosen to be disseminated. In this paper,we observe that, the fact that a generalized dataset is optimal may leak information about the original. Thus, an adversary who is aware of how the dataset is generalized may able to derive more information than what the privacy requirements constrained. This observation challenges the widely adopted approach that treats the generalization process as an optimization problem. We illustrate the observation by giving counter-examples in the context of k-anonymity and ℓ-diversity.

Key words: Data dissemination, Privacy-preserving, k-anonymity and ℓ-diversity

1 Introduction

Data dissemination and information sharing is required in statistical analysis of a population spreading across different organizations, and is also essential in providing transparency. However, the ease of obtaining and linking different published dataset lead to the concern on the leakage of personal information. To protect privacy we may generalize the attribute values, for example by dropping the last digit of the postal code, before the datasets are released. On the other hand, it is meaningless to disseminate datasets that are over generalized. To achieve the right tradeoff, a widely adopted framework treats the problem

* This is a refined version of the published paper appeared in the proceeding of Information Hiding 2008. The authors were unaware of a similar idea in an earlier work by Wang et al.[19]. This version includes the relevant reference, a new paragraph in Section 2 on related work, and this footnote.

of finding the generalized dataset as an optimization problem. The framework takes a requirement of privacy as the constraint, and a *utility function*, which measures the usefulness of a generalized dataset, as the objective function in the optimization problem. In other words, given a dataset, among all the generalized datasets that meet the privacy requirements, the one that is optimal with respect to the *utility function* is chosen to be disseminated. The framework provides the assurance that the disseminated dataset meets the privacy requirement, and at the same time is useful. The well-known notions on k-anonymity [18] and ℓ-diversity [13] provide notions and requirements of privacy, and both notions are proposed to be employed in the above-mentioned framework. In this paper, we observe that the optimal generalized dataset might no longer satisfy the privacy requirement. Although the disseminated table is chosen from a collection that meets the privacy requirement, the fact that the disseminated dataset is optimal is an additional piece of information. Taking this piece of information into consideration, an adversary may able to derive more information than what the privacy requirement ensured. We will illustrate our observation by investigating the requirements of k-anonymity and ℓ-diversity.

The notion of k-anonymity requires that every record is indistinguishable from at least $(k-1)$ other records for all possible set of attributes. This ensures that at least k tuples share the same generalized identity and thus individual cannot be identified. To illustrate, consider a scenario where a hospital published information of patients in a particular month as shown in Table 1. To protect privacy, values under the attribute "Name" are removed. Elsewhere, an association released information of dentists with information shown in Table 2. It happens that the combination of age, gender and company postal code is unique for Peter in Table 2. Thus, by linking both tables, one may derive Peter's home postal code and he was hospitalized in that month. Table 3 shows a generalization that is 2-anonymized. From Table 3, it is not easy to identify Peter since there are two tuples matching his identity.

Table 1. Released table.

Name	Age	Gender	Home postal	Occupation	Company postal	Illness
	30	F	48546	Crane operator	54832	Anxiety
	41	F	13208	Teacher	11824	Sleeplessness
	43	F	15201	Dentist	11857	Sleeplessness
	32	F	48356	Driver	54832	Anxiety
	26	M	61306	Manager	29054	fever
	22	M	61306	Dentist	29089	fever

The notion of ℓ-diversity is introduced to prevent data inference that is not addressed in k-anonymity. The attributes are classified as *sensitive* and *non-sensitive*, and it is assumed that the publisher knows which attributes are sensitive. Consider the previous scenario where Table 3 is published, and the attribute

Table 2. Public table.

Name	Age	Gender	Company postal
...	...	...	...
Peter	22	M	29089
...	...	...	...

Table 3. 2-anonymized table.

Name	Age	Gender	Home postal	Occupation	Company postal	Illness
	3*	F	48***	Outdoor	54832	Anxiety
	3*	F	48***	Outdoor	54832	Anxiety
	4*	F	1****	Indoor	118**	Sleeplessness
	4*	F	1****	Indoor	118**	Sleeplessness
	2*	M	61306	Indoor	290**	fever
	2*	M	61306	Indoor	290**	fever

"Illness" is sensitive. Suppose Alice knows that Peter is hospitalized and she has access to Table 2. Although there are two tuples in Table 3 matching Peter's identity, both of them share the same sensitive value. Thus, Alice can infer that Peter is having fever. To counter this inference, the ℓ-diversity requirement ensures that, among the records with the same identifiers, the sensitive attribute values consist of at least ℓ well-represented values. There are many ways to defined the meaning of what being "well-represented" values, and a natural choice is by requiring the entropy of the attribute values is above certain threshold, say $\log_2 \ell$. If the sensitive values are well represented, then we can have an upper bound on the chances that the adversary can successfully guess the correct value.

As mentioned in the first paragraph, it is meaningless to publish a table that is over generalized. Hence, it is desire to find a generalized table that meets certain requirements on privacy, and optimal with respect to a utility function. There are many choices of utility functions, and typically, they measure the distance of the generalized dataset from the original dataset. An example of utility function counts the number $*$'s in the generalized table. In general, given a dataset, it is not easy to find the optimal. In many interesting settings, the problems are NP-hard [14]. Fortunately, there are extensive works in finding the optimal and many approximation algorithms and effective heuristic are known [11, 6, 15, 17].

The rest of this paper is organized as follows. We discuss the related work in section 2, and give the related background and notations in section 3. In section 4.1 and 4.2 we introduce the formulation and show examples on information leakage of k-anonymized table, then we move on to information leakage of ℓ-diverse table in section 4.3 and give a general theorem in section 4.4. Section 5 gives a conclusion.

2 Related Work

There are extensive works on k-anonymity since Sweeney[18] proposed the notion. The notion of k-anonymity is widely involved in the context of protecting location privacy[7, 10], preserving privacy in communication protocol[20, 21] data mining techniques[2, 9] and many others. There are different way of anonymizing and diversifying a table: achieving via generalization[3, 11], via generalization with suppression[16, 12] and via data swapping and randomization techniques[1, 8]. Meyerson et al. have shown that achieving the optimal generalization is NP-hard for many different settings[14, 4, 5]. Fortunately, there are many practical approximation and heuristic. Sweeney has proposed an heuristic-based approach[17] in 2002. Samarati has also proposed an algorithm of searching a "k-minimal" group that contains the optimal k-anonymizations based on certain preference[15]. Bayardo et al. have proposed a lattice top-down search strategy[6] for optimal k-anonymized tables while LeFevre et al. have proposed a bottom-up searching algorithm[11]. Machanavajjhala et al. have proposed the idea of ℓ-diversity [13] and we follow most of the term and definitions they used.

Wang et al. proposed an attack based on the similar assumption that adversaries know the generalization algorithm[19]. They observed that, if an algorithm follows the *minimality principle*, then an adversary with such knowledge can carry out the *minimality attack*. Under the minimality principle, an algorithm will output a table T^* which, (1) meets the privacy requirements, and (2) there is no other table $\hat{T}$ that meets the privacy requirements and $\hat{T} \geq T^*$ (w.r.t. a partial order to be defined below). For two tables T_1, and T_2 where T_1 can be generalized to T_2, we define $T_1 \geq T_2$. Since the output is only required to be optimal among any chain to the original, it is not guaranteed to be optimal among all possible generalizations of the original. Note the subtle difference from our assumption, where we consider the optimal among all possible generalizations. Intuitively, our assumption is stronger (i.e. the adversaries exploit more information), since an optimal algorithm must follow the minimality principle. On the other hand, since computing the optimal is difficult, most known methods satisfy only the minimality principle.

3 Background and Notations

In this section, we describe k-anonymity and ℓ-diversity, based on definitions in [13] with slight modifications. The definition given in [13] requires the classification of attributes into sensitive and non-sensitive. However, this requirement is not enforced in some known works and the classification may not be trivial in some real life datasets. Hence, for definition of k-anonymity, we do not classify the attribute.

Each dataset $T = \{t_1, t_2, \ldots, t_n\}$ is a set of tuples, and can be viewed as a table as shown in Table 1, where a row corresponds to a tuple, and a column corresponds to an attribute. For a $t \in T$ and an attribute A, let us denote $t[A]$ the value of attribute A of the tuple t.

3.1 Generalization

Let D to be the domain of an attribute value, for example, D can be {0, 1}, set of integers or set of strings. We say that D^* is a generalization of D if D^* is a partition of D. That is, D^* is a collection of non-intersecting subsets of D, whose union is D. We say that a $c^* \in D^*$ is the *generalized value* of $c \in D$ if $c \in c^*$. If every value in a table T is replaced by its generalized value for some D^*, then we say that the new table T^* is a generalized table of T. For two generalized domain D_0^* and D_1^*, we say that D_1^* is a generalization of D_0^* if, for any $c_0^* \in D_0^*$, there exists a $c_1^* \in D_1^*$ such that $c_0^* \subseteq c_1^*$. Similarly, if D_0^* and D_1^* is the domain of T_0^* and T_1^* respectively, then we say that T_1^* is a generalization of T_0^*.

For example, the domain of "Home postal" in Table 3 is the set of 5-digits strings. Replacing the string 13205 to 1320* is a generalization. The generalized D_0^* domain contains the set { 13200, 13201, ..., 13209 }.

If the string is further replaced by 132**, the new generalized domain D_1^* contains a set {13200, 13201, ..., 13299 }. Furthermore, D_1^* is a generalization of D_0^*. Since a generalized domain is a partition of the original domain, it is not possible to have both 13*** and 132** appeared in a column of the table.

3.2 *k*-anonymity

A set of attributes $\{A_1, A_2, \ldots, A_w\}$ of a table is called a *quasi-identifier*. Let $\mathcal{QI}$ be a collection of quasi-identifiers[1]. We say that a tuple t_1 is k-anonymized, if for any quasi-identifier $C \in \mathcal{QI}$, there exist $k-1$ other tuples $t_2, \ldots, t_k$ such that $t_1[C] = t_2[C] = \ldots = t_k[C]$. A table T is k-anonymized if every tuple is k-anonymized.

If a table T is k-anonymized, given any quasi-identifier in $\mathcal{QI}$, each tuple cannot be distinguished from at least $k-1$ tuples. For example, in a 2-anonymized table shown in Table 3, even if an adversary knows Table 2, he is unable to identify Peter's tuple in Table 3, since the third and fourth tuple has the same generalized value.

3.3 ℓ-diversity

Under the notion of ℓ-diversity, each attribute is classified as either *sensitive* or *non-sensitive* but not both. Furthermore, a quasi-identifier contains only non-sensitive attributes. Hence, only the non-sensitive attributes can be linked with other public tables. The publisher is assumed to know which attributes are sensitive before the table is generalized. Note that such classification of attributes is not enforced in k-anonymity.

Given q^*, a value of a quasi-identifier, let us define the q^*-*block* to be the set of tuples with value q^*. Let $n(q^*, s, T)$ denote the number of tuples that has value q^* and value s for a sensitive attribute. For example, in Table 7, the

[1] It is not necessary that $\mathcal{QI}$ contains all possible quasi-identifiers. Some previous works restrict $\mathcal{QI}$ to quasi-identifiers that can be linked with other tables.

block of tuples with values (`130**`, M, `A-`) has two sensitive value, "Anxiety" and "Cancer".

In general, a table is said to be ℓ-diverse if, for every q^*-block, the values of any sensitive attribute is "well-represented" by ℓ values. There are a number of ways to quantify how "well-representative" a block is. A simple requirement is to have at least ℓ sensitive values in every q^*-block. In this paper, we adopt the notion of *entropy ℓ-diverse* as defined in [13].

Entropy ℓ-diverse. A table T is said to be entropy ℓ-diverse if, for every q^*-block, and any sensitive attribute with domain S,

$$-\sum_{s\in S} P(q^*,s,T)\log(P(q^*,s,T)) \geq \log(\ell), \quad (1)$$

$$\text{where } P(q^*,s,T) = n(q^*,s,T) \;/\; \sum_{s'\in S} n(q^*,s',T)\ .$$

$P(q^*,s,T)$ is the ratio of tuples that has the sensitive value s among the tuples in the q^*-block.

Suppose an adversary has the value q of a quasi-identifier, and the generalized table T^*. Let q^* be the corresponding generalized value of q in the table T^*. Let us assume that, in the original table T, q is unique, and each tuple in the q^*-block (in the table T^*) is equally likely to be the actual tuple with quasi-identifier q. Hence, if he predicts that the tuple has sensitive value s, his chance of success is the ratio $P(q^*,s,T)$. This can be viewed as the *posterior belief* of the tuple having sensitive value s. Let us write it as,

$$\beta_{q,s}\ .$$

Hence, the left hand side in inequality (1) is the entropy of the posterior belief.

3.4 Utility Function

Ideally, a utility function measures the amount of information retained in a generalized table T^*. Generally, its value increase as the "distance" between T^* and the original T decreases. Here is an example of a simple utility function which counts the number of `*`'s in the generalized table T^*.

$$U(T^*) = -\sum_{t\in T^*}\sum_{q\in\mathcal{QI}_{T^*}} f(t,q)\ , \quad (2)$$

where $f(t,q) = k$ is the number of `*`'s contained in $t[q]$.
There are many choices of utility function, we uses the above function (2), which is widely adopted, in our discussions.

3.5 Optimal Generalized Table

Given a table T, let $\mathcal{C}(T)$ be the collection of all possible generalizations of T. Given a privacy requirement, which can be k-anonymity, and/or ℓ-diversity, let $\mathcal{P}$ to be the set of all tables that satisfy the requirement[2]. Let $\mathcal{G}(T)$ be the table in $(\mathcal{C}(T) \cap \mathcal{P})$ that is optimal with respect to a given utility function. Conversely, given a generalized table T^*, we write $\mathcal{G}^{-1}(T)$ to be:

$$\mathcal{G}^{-1}(T) = \{T \mid T^* = \mathcal{G}(T)\} \ . \tag{3}$$

That is, it is the inverse of the function $\mathcal{G}$.

Remarks. Note that the definition of $\mathcal{G}(\cdot)$ relies on the definition of the privacy requirement, and the utility function. Also, note that the set $\mathcal{G}^{-1}(T)$ does not contain generalized tables.

We assume that the optimal is unique, and the generalization process is deterministic. Our main observation can be extended to generalization algorithms that are probabilistic. However, for clarity, we choose to handle deterministic algorithms in this paper.

4 Information Leakage

This section gives the formulation of information leakage (Section 4.1 and 4.3). Examples of information leakage in k-anonymized and ℓ-diversified table will be given in Section 4.2 and 4.4 respectively.

4.1 Formulation of Leakage in k-anonymized Tables

Given a T^*, we say that it can be *inverted* if

1. $|\mathcal{G}^{-1}(T^*)| = 1$, and
2. the table in $\mathcal{G}^{-1}(T^*)$ is not k-anonymized.

That is, from T^*, there is only one table T whose optimal generalized data is T^*. Note that it is more interesting to include the second condition since a table that is already k-anonymized will be published as it is.

In cases where the inverse is not unique, all tables in $\mathcal{G}^{-1}(T^*)$ may still able to be generalized to a single table that is not k-anonymized. Given a generalized T^*, we say that it can be *partially inverted* if there is a T_0^* such that

1. For all $T \in \mathcal{G}^{-1}(T^*)$, T can be generalized to T_0^*, and
2. T_0^* is not k-anonymized.

Hence, if a table can be partially inverted, by linking with certain tables, there exists a tuple t_0 and quasi-identifier Q, such that t_0 shares the same identity (with respect to Q) with at most $(k-2)$ tuples. Thus, the original assurance of k-anonymity is compromised.

[2] To simplify notations, we do not parameterized $\mathcal{P}$ with the requirements. In the paper, it is always clear from the context which privacy requirement is referred to.

4.2 Examples for k-anonymized Table

Example 1: Inverting a table. This section gives an optimal generalized table T^* that can be inverted. This simple example provides a simple form that can be extended to larger examples. The original table T contains one attribute Att_1 whose domain is binary strings of length 2. The 2-anonymized table is shown in Table 4 (a).

The original value for `0*` can be either `00` or `01`. Due to symmetry, there are only 5 possible tables that can be generalized to T^*: either it contains four `00`'s, three `00`'s, two `00`'s, one `00`, or none. Let us examine these cases.

1. Four `00`'s, two `00`'s and none: In each case, the table already satisfies 2-anonymity, and thus its optimal generalized table is itself. Hence, they are not in $\mathcal{G}^{-1}(T^*)$.
2. Three `00`'s: In this case, there is only one `01`. Hence, it does not satisfy 2-anonymity. However, its optimal anonymized table is not T^*. Instead, the table with three `00`'s and three `*1`'s attains optimal.
3. One `00`: Table 4 (b) shows this case. This table does not satisfy 2-anonymity, and it is easy to verify that T^* is its optimal 2-anonymized table.

Therefore, $\mathcal{G}^{-1}(T^*)$ contains only one table and it does not satisfy 2-anonymity.

Table 4. (a) An optimal 2-anonymized table. (b) The only possible original.

(a)

Att_1
11
11
0*
0*
0*
0*

(b)

Att_1
11
11
01
01
01
00

Example 2: Partially inverting a table. We now give an optimal 2-anonymized table T^* that can be partially inverted. The original table T contains one attribute $Attr_1$ whose domain is binary strings of length 4. The anonymized table T^* is shown in Table 5.

There are two generalized values, `000*` and `**01`. The original value for `000*` can be either `0000` or `0001`. Similar to the previous example, by examine each case, we can deduce that the original table has two `0011`'s, three `0001`'s, one `0000`, for the four `000*`'s.

Now, let us consider the two tuples with `**01`. The `*`'s appear in the first and second position. If, the values are the same at either the first or the second

position, then we can have a generalized table with lower utility. For example, {0001, 0101} can be generalized to 0*01, which requires only two *'s. Thus, their optimal is not T^* and the choices for $\mathcal{G}^{-1}(T^*)$ are narrowed to {0010, 1110} and {1010, 0110}. It is easy to check that both cases have Table 5 as its optimal generalization. Therefore, $\mathcal{G}^{-1}(T^*)$ contains two tables as shown in Table 6 (a) & (b), which can be generalized to the table T_0^* as shown in Table 6 (c). Hence, T^* can be partially inverted.

Table 5. An optimal 2-anonymized table.

Att_1
0011
0011
000*
000*
000*
000*
**10
**10

Table 6. (a) & (b) The two possible original of Table 5. (c) A generalized table T_0^* that does not satisfy 2-anonymity.

(a)

Att_1
0011
0011
0001
0001
0001
0000
0010
1110

(b)

Att_1
0011
0011
0001
0001
0001
0000
0110
1010

(c)

Att_1
0011
0011
0001
0001
0001
0000
**10
**10

4.3 Formulation of Leakage in ℓ-diversified Tables

Recall the definition of posterior belief $\beta_{q,s}$ in Section 3.3. If the fact that the table T^* is optimal is taken into consideration, the probability that a tuple in the q^*-block having the sensitive value s may change and is not longer $P(q^*, s, T)$. Let us call this probability the *enhanced belief* and write it as:

$$\gamma_{q,s} \ .$$

Consider a table T, and its optimal ℓ-diversified table T^*, in addition, let S be a sensitive attribute, Q a quasi-identifier. We say that T^* suffers *partial disclosure* if there exist some $q \in Q$ and $s \in S$ such that:

$$\beta_{q,s} < \gamma_{q,s} \ .$$

Furthermore, we say that T^* suffers *total disclosure* if

$$\beta_{q,s} < \gamma_{q,s} = 1 \ .$$

4.4 Results for ℓ-diversified Tables

Table 7. An optimal 2-anonymized and 2-diversified table.

	Postal code	Gender	Blood group	Condition
1	130**	M	A+	Heart Disease
	130**	M	A+	Viral Infection
2	130**	M	A-	Anxiety
	130**	M	A-	Cancer
3	130**	M	B*	Cancer
	130**	M	B*	Fever
	130**	M	B*	Cough
	130**	M	B*	Diabetes

Example. We now give an optimal diversified table T^* that suffers total disclosure in this section, and we will extend this example to a more general form later in this section.

The table T^* is shown in Table 7. The attribute "Condition" is the only sensitive attribute and the others are non-sensitive. Similar to previous examples, the utility function is based on the number of *'s. The $\mathcal{QI}$ contains the set of all non-sensitive attributes. The leftmost column indicates different blocks and is not part of the table. This table is an optimal 2-anonymized and entropy 2-diversified table.

It is entropy 2-diversified because for block 1 and block 2 we have the following for the inequality (1),

$$-2 \cdot \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = \log_2 (2) \ .$$

and for block 3, we have

$$-4 \cdot \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right) > \log_2 (2) \ .$$

Consider an adversary who has a quasi-identifier value (13021, M, `A+`), which can be identified with block 3. This block has 4 different sensitive values. Thus, the posterior belief for this identity having "cancer" is $\frac{1}{4}$.

Now, using the fact that the table is optimal, the probability changes. The original value for the attribute "Blood group" for each tuple in block 3 can be either `B+` or `B-`. There are 8 different cases for block 3.

1. The non-sensitive value of the first tuple in block 3 is (`130**`, M, `B+`) and the number of the other three tuples in block 3 having original value (`130**`, M, `B+`) is:
   - 1 or 3: In these two cases, block 3 is already 2-anonymized and 2-diversified. This is against the assumption that T^* is optimal.
   - 2: In this case, block 3 contains only 1 tuple having the non-sensitive value (`130**`, M, `B-`). This tuple can be generalized with tuples in block 2 as "`*-`" to achieve one less `*`. Thus, this case can be eliminated.
   - 0: Only the first tuple in block 3 has the non-sensitive value (`130**`, M, `B+`) with can be generalized with tuples in block 1 to achieve one less `*`. Thus, this case can also be eliminated.
2. The non-sensitive value of the first tuple in block 3 is (`130**`, M, `B-`) and the number of the other three tuples in q^*-block 3 having original value (`130**`, M, `B-`) is:
   - 1 or 3: In both cases, the original table is already 2-diversify.
   - 2: In this case, the optimal generalized table is not T^*
   - 0: It is easy to verify that its optimal generalized table is T^*.

A generalization of the original table is as shown in Table 8. In addition, for the identity (`13021`, M, `B-`) having "cancer", the enhanced belief is 1 which is higher than the posterior belief of $\frac{1}{4}$. Thus, this table suffers total disclosure for tuple (`13021`, M, `B-`). Furthermore, this table also suffers partial disclosure for (`13021`, M, `B+`) (details omitted).

General Result. The previous example is for $k, \ell = 2$. We now show that total disclosure can occurred for any k, ℓ where $k \geq \ell \geq 2$.

THEOREM 1 *For any k, ℓ such that $k \geq \ell \geq 2$, there exists an optimal k-anonymized and ℓ-diversified table T^* that suffers total disclosure.*

Table 8. Generalization of the original table for Table 7.

	Postal code	Gender	Blood group	Condition
1	130**	M	A+	Heart Disease
	130**	M	A+	Viral Infection
2	130**	M	A-	Anxiety
	130**	M	A-	Cancer
3	130**	M	B-	Cancer
4	130**	M	B+	Fever
	130**	M	B+	cough
	130**	M	B+	diabetes

Proof:

Let $m = \lceil \frac{k}{\ell} \rceil$ and let n be a large number greater than $3k$. Consider a table containing a non-sensitive attribute $Attr_1$ whose domain is bit string of length 2, and a sensitive attribute $Attr_2$ whose domain is the set $\{A_1, A_2, ..., A_n\}$ and its k-anonymized, ℓ-diversified table T^* as shown in Table 9. For abbreviation, the right-most column indicates the number of tuples with the same values. For example, in the first row, the "m" indicates that T^* contains m tuples with value $(11, A_1)$.

This table T^* is entropy ℓ-diverse because for block 1 and 2 we have $-\ell \cdot (\frac{1}{\ell}) \log_2(\frac{1}{\ell}) = \log_2(\ell) = \log_2(\ell)$ and for block 3 we have $-(n-\ell)\cdot(\frac{1}{n-\ell}) \log_2(\frac{1}{n-\ell}) = \log_2(n-\ell) > \log_2(\ell)$.

Suppose an adversary wants to guess the sensitive value of non-sensitive value 00, which is generalized to block 3. His posterior belief for this tuple has sensitive value $A_{\ell+1}$ is $\beta_{0*,A_{\ell+1}} = \frac{1}{n-\ell}$.

Now, let us consider the scenario where the adversary knows the fact that T^* is optimal.

Let us introduce the following lemma.

LEMMA 2 *Given a q^*-block Q of exactly ℓ different sensitive values, it is entropy ℓ-diverse only if all these ℓ different sensitive values have the same number of tuples in this q^* block.*

This lemma holds because only when all sensitive values are of same number, the entropy $-(\ell) \cdot \frac{1}{\ell} \log_2(\frac{1}{\ell})$ is equal to $\log_2(\ell)$.

In this scenario, we should consider the enhanced belief with the above lemma. Note that the original value for 0* can only be either 00 or 01. We divide the possible original tables to the following cases:

1. The first tuple of block 3 is 01 and the number of other tuples having 01 as their non-sensitive attribute is:

Table 9. Released table in k-anonymity and ℓ-diversity.

	$Attr_1$	$Attr_2$	number of tuples
1	11	A_1	m
	11	A_2	m
	11	A_3	m
	...	...	...
	11	A_ℓ	m
2	10	A_2	m
	10	A_3	m
	10	A_4	m
	...	...	...
	10	$A_{\ell+1}$	m
3	0*	$A_{\ell+1}$	1
	0*	$A_{\ell+2}$	1
	...	...	...
	0*	A_{n-1}	1
	0*	A_n	1

(a) More than $k-2$ but less than $n-k$. tuples with 01 and 00 are more than k and they all have different sensitive value. Therefore, the original table is already k-anonymized and ℓ-diversify without generalizing the $Attr_1$. This is against the assumption that T^* is optimal.
(b) More than $n-k-1$. The number of tuples having non-sensitive value 00 is less than k and hence the original table is not k-anonymized. However, we can generalize those tuples having 00 with block 1 and reduce the number of $*$'s. Therefore, this case can be eliminated.
(c) Less than $k-1$. We can generalize these tuples with block 2 to reduce the number of $*$'s. Thus, T^* is not optimal.

2. The first tuple of block 3 is 00 and the number of other tuples having 00 as their non-sensitive attribute is:
(a) More than $k-2$ but less than $n-k$. This case can be eliminated as the original table is already k-anonymized and ℓ-diversified.
(b) More than $n-k-1$. T^* is not optimal in this case.
(c) Less than $k-1$ but more than 0. This case can still be eliminated. As long as there are more than one tuple having the non-sensitive value 00, we can still combine these tuples with block 2 (Lemma 2).
(d) Zero. T^* is optimal as we cannot add the first tuple of block 3 alone to block 2 (Lemma 2).

Thus, $\mathcal{G}^{-1}(T^*)$ contains only a unique table to as shown in Table 10, and$\gamma_{00,A_{\ell+1}}$ is 1. □

Table 10. Original table of Table 9.

	$Attr_1$	$Attr_2$	number of tuples
1	11	A_1	m
	11	A_2	m
	11	A_3	m
	...	...	...
	11	A_ℓ	m
2	10	A_2	m
	10	A_3	m
	10	A_4	m
	...	...	...
	10	$A_{\ell+1}$	m
3	00	$A_{\ell+1}$	1
4	01	$A_{\ell+2}$	1
	...	...	...
	01	A_{n-1}	1
	01	A_n	1

5 Conclusion

In this paper, we have showed that the framework of choosing an optimal (w.r.t an objective function) table from a collection of candidates that satisfies certain privacy requirements, does not provide the assurance that the chosen table will retain the privacy requirements. This is because the fact that the table is optimal is a piece of additional information, which can be exploited by the adversaries. This observation is demonstrated by counter-examples of optimal anonymized and diversified tables. It is interesting to find out whether such framework has been followed in other formulation of privacy, or other security requirements. On the other hand, it is also interesting to find out whether there is a choice of utility function and privacy requirement that can be securely applied in this framework.

References

1. N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, pages 515–556, 1989.
2. C. C. Aggarwal. On k-anonymity and the curse of dimensionality. *31st International Conference on Very Large Data Bases*, pages 901–909, 2005.
3. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. k-anonymity: Algorithms and hardness. *Technical report, Stanford University*, 2004.

4. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. *10th International Conference on Database Theory*, pages 246–258, 2005.
5. G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy Technology*, 2005.
6. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. *International Conference on Data Engineering*, pages 217–228, 2005.
7. C. Bettini, X. S. Wang, and S. Jajodia. Protecting privacy against location-based personal identification. *Secure Data Management*, pages 185–199, 2005.
8. G. T. Duncan and S. E. Feinberg. Obtaining information while preserving privacy: A markov perturbation method for tabular data. *Joint Statistical Meetings*, pages 351–362, 1997.
9. B. Fung, K. Wang, and P. Yu. Top-down specialization for information and privacy preservation. *International Conference on Data Engineering*, pages 205–216, 2005.
10. B. Gedik and L. Liu. A customizable k-anonymity model for protecting location privacy. *25th International Conference on Distributed Computing Systems*, 2005.
11. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. *SIGMOD*, 2005.
12. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. *International Conference on Data Engineering*, 2006.
13. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *International Conference on Data Engineering*, page 24, 2006.
14. A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. *23rd ACM Symposium on the principles of Database Systems*, pages 223–228, 2004.
15. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, pages 1010–1027, 2001.
16. P. Samarati and L. Sweeney. protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *Technical report, CMU, SRI*, 1998.
17. L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, pages 571–588, 2002.
18. L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based System*, pages 557–570, 2002.
19. R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. *Very Large Data Bases*, pages 543–554, 2007.
20. S. Xu and M. Yung. k-anonymous secret handshakes with reusable credentials. *11th ACM Conference on Computer and Communications Security*, pages 158–167, 2004.
21. G. Yao and D. Feng. A new k-anonymous message transmission protocol. *5th International Workshop on Information Security Applications*, pages 388–399, 2004.