L-Cover: Preserving Diversity by Anonymity

Lei Zhang[1], Lingyu Wang[2], Sushil Jajodia[1], and Alexander Brodsky[1]

[1] Center for Secure Information Systems
George Mason University
Fairfax, VA 22030, USA
{lzhang8,jajodia,brodsky}@gmu.edu
[2] Concordia Institute for Information Systems Engineering
Concordia University
Montreal, QC H3G 1M8, Canada
wang@ciise.concordia.ca

Abstract. To release micro-data tables containing sensitive data, generalization algorithms are usually required for satisfying given privacy properties, such as k-anonymity and l-diversity. It is well accepted that k-anonymity and l-diversity are proposed for different purposes, and the latter is a stronger property than the former. However, this paper uncovers an interesting relationship between these two properties when the generalization algorithms are publicly known. That is, preserving l-diversity in micro-data generalization can be done by preserving a new property, namely, l-cover, which is to satisfy l-anonymity in a special way. The practical impact of this discovery is that it may potentially lead to better heuristic generalization algorithms in terms of efficiency and data utility, that remain safe even when publicized.

1 Introduction

The micro-data release problem has attracted much attention due to increasing concerns over personal privacy. Various generalization techniques have been proposed to transform a micro-data table containing sensitive information for satisfying given privacy properties, such as k-anonymity [16] and l-diversity [2]. For example, in the micro-data table shown in Table 1, suppose each patient's medical condition is to be kept confidential. The attributes can thus be classified into three classes, namely, *identity* (Name), *quasi-identifiers* (ZIP, Age), and *sensitive value* (Condition).

Clearly, simply hiding the identity (Name) when releasing the table is not sufficient. A tuple and its sensitive value may still be linked to a unique identity through the quasi-identifiers, if the combination (ZIP, Age) happens to be unique [16]. To prevent such a *linking attack*, the table needs to be generalized to satisfy k-anonymity. For example, if generalization (A) in Table 2 is released, then any linking attack can at best link an identity to a group of two tuples with the same combination (ZIP, Age). We can also see from the example that

Name	ZIP	Age	Condition
Alice	22030	60	flu
Bob	22031	50	tracheitis
Clark	22032	40	cancer
Diana	22035	35	cancer
Ellen	22045	34	pneumonia
Fen	22055	33	gastritis

Table 1. An Example of Micro-Data Table

k-anonymity by itself is not sufficient, since linking an identity to the second group will reveal the condition of that identity to be cancer. This problem is addressed in generalization (B), by satisfying both k-anonymity and l-diversity.

ZIP	Age	Condition
22030∼22031	50∼60	flu tracheitis
22032∼22035	35∼40	cancer cancer
22045∼22055	33∼34	pneumonia gastritis

(A)

ZIP	Age	Condition
22030∼22032	40∼60	flu tracheitis cancer
22035∼22055	33∼35	cancer pneumonia gastritis

(B)

Table 2. Two Potential Table Generalizations

However, the situation is worse in practice. As recently pointed out by Zhang *et al.* [22], an adversary can still deduce both Clark and Diana have cancer, if it is publicly known that generalization (A) is first considered (but not released) before generalization (B) is considered and released. This complication makes the problem of micro-data release more challenging when the algorithm used to compute the disclosed data is assumed to be publicly known. In [22], the authors give a comprehensive study of how a privacy property can be guaranteed in this situation and they also prove it to be an NP-hard problem to optimize data utility while guaranteeing the l-diversity property. Also, how to design heuristic algorithms is discussed and one heuristic generalization algorithms which remains safe when publicized is presented in [22]. However, as shown in [22], the proposed algorithm is not practical due to the data utility it can provide.

In this paper, we uncover an interesting relationship between k-anonymity and l-diversity, which can be used to design better generalization algorithms in terms of efficiency and data utility, while guaranteeing the property of l-diversity when the algorithm itself is publicized. More specifically, our contribution is

two fold. First, we propose a novel strategy for micro-data release. That is, instead of trying to select the generalization with the "best" data utility from all possible generalizations, we first restrict our possible selections to a subset of all possible generalizations, and then optimize the data utility within the subset. Certainly, we guarantee only a local optimality in the restricted set instead of the global optimality. We prove that, as long as the restricted subset satisfies certain properties, the security/privacy of the publicized result will not be affected by whether this applied generalization algorithm is publicized or not. Second, we introduce the property of l-cover, defined on a set of generalizations, which is an anonymity-like property when exchanging the role of identity and sensitive value in a micro-data table. We prove that in order to guarantee the property of l-Diversity on the released data, it is sufficient to have the above subset of generalizations satisfy the property of l-cover. We also show, through examples, that in practice we do not need to compute the entire subset of generalizations that satisfies l-cover. In stead, we only need to construct anonymity groups of size l for sensitive values, which can be done efficiently in advance, and check whether a candidate generalization breaks these groups when optimizing the data utility. Therefore, this technique can be potentially used to design more practical heuristic generalization algorithms compare to the algorithms proposed in [22].

Organization

In Section 2, we define our model and examine relevant concepts. In Section 3, we propose a novel strategy for micro-data release. In Section 4, we formalize the concept of l-cover and employs it to compute safe generalizations. We discuss related work in Section 5 and draw conclusions in Section 6.

2 The Model

We first define our notations for micro-data table and generalization. We then discuss privacy properties and how they may disclose information.

2.1 Micro-Data Table and Generalization

A micro-data table is a relation $T(i, q, s)$, where i, q, and s is called the identity, quasi-identifier, and sensitive value, respectively (see Table 3 for a list of important notations used in this paper). Note that both q and s can be a sequence of attributes. We use $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{S}$ for the projection $\Pi_i(T)$, $\Pi_q(T)$, and $\Pi_s(T)$, respectively. Unless explicitly stated otherwise, all projections in this paper preserve duplicates. Therefore, both $\mathcal{Q}$ and $\mathcal{S}$ are actually multisets. Let R_{iq}, R_{qs}, R_{is} denote the three projections $\Pi_{i,q}(T)$, $\Pi_{q,s}(T)$, and $\Pi_{i,s}(T)$, respectively.

$T(i,q,s)$ or T	Micro-data table
$\mathcal{I}$, $\mathcal{Q}$, $\mathcal{S}$	Projections $\Pi_i(T)$, $\Pi_q(T)$, $\Pi_s(T)$
R_{iq}, R_{qs}, R_{is}	Projections $\Pi_{i,q}(T)$, $\Pi_{q,s}(T)$, $\Pi_{i,s}(T)$
GQ	Quasi-identifier generalization
$GT = (GQ, GS, GR_{qs})$	Table generalization
$\mathcal{GQ}_L$	Locally safe set
$\mathcal{GQ}_A$	Candidate set
$\mathcal{T}$	Disclosure set

Table 3. A List of Important Notations

As typically assumed, $\mathcal{I}$, $\mathcal{Q}$ and the relation R_{iq} are considered as public knowledge. We also assume $\mathcal{S}$ as publicly known, since any released generalization will essentially disclose $\mathcal{S}$ any way (we do not consider suppression). On the other hand, the relation R_{is} and R_{qs} both remain secret until a generalization is released. Between them, R_{is} is considered as the private information, whereas R_{qs} is considered as the utility information. The goal of a generalization algorithm is usually to disclose as much information about R_{qs} as possible, while still guaranteeing the secrecy or uncertainty of information about R_{is}.

We need to explicitly distinguish the two stages, in order, of a generalization process, namely, *quasi-identifier generalization* and *table generalization*, as formalized in Definition 1. The key difference is the following. A quasi-identifier generalization GQ only generalizes the publicly known $\mathcal{Q}$, and thus contains information only from the publicly known $\mathcal{Q}$. On the other hand, a table generalization GT generalizes the entire micro-data table, containing information from both the secret relations R_{is} and R_{qs}. This difference will be critical to our further discussion.

Definition 1. (Quasi-Identifier Generalization and Table Generalization) *Given a micro-data table $T(i,q,s)$, we define*

- *a quasi-identifier generalization GQ as any partition on $\mathcal{Q}$, and*
- *a table generalization GT as a triple (GQ, GS, GR_{qs}), where*
  - *GQ is a quasi-identifier generalization,*
  - *GS is a partition on $\mathcal{S}$, and*
  - *$GR_{qs} \subseteq GQ \times GS$ is a one-to-one relation,*

  such that for all $(gq, gs) \in GR_{qs}$, there exists a one-to-one relation $R \subseteq gq \times gs$ satisfying $R \subseteq R_{qs}$.

2.2 Privacy Properties

k-Anonymity The concept of k-anonymity [16] mainly concerns with the size of each group in GQ. More precisely, a table generalization GT satisfies k-anonymity if $\forall gq \in GQ, |gq| \geq k$. Notice that this condition only depends

on GQ. Therefore, if GQ is publicly known, anyone may determine whether a table generalization GT computed based on GQ satisfies k-anonymity, even without knowing the entire GT. As a result, we have the following claim, which is straightforward.

Claim 1 *Given a micro-data table T and a quasi-identifier generalization GQ, to disclose the fact that a table generalization GT computed based on GQ violates k-anonymity does not provide any additional information about T.*

l-Diversity The concept of l-diversity [2] concerns with the diversity of sensitive values that can be linked to each identity. In particular, we shall focus on entropy l-diversity, which requires the entropy of values in a multiset S to be no less than $\log l$, that is, $\sum_{s \in BagToSet(S)} \frac{count(s,S)}{|S|} \log \frac{|S|}{count(s,S)} \geq \log l$, where $count(s, S)$ is the number of appearances of s in S. For a table generalization $GT = (GQ, GS, GR_{qs})$, l-diversity is applied to each group in GS. That is, GT satisfies entropy l-diversity, if for all $gs \in GS$, gs satisfies entropy l-diversity.

Clearly, unlike k-anonymity, l-diversity depends on not only GQ but also GS and GR_{qs}. Therefore, without knowing a table generalization GT, it is impossible to check whether GT satisfies l-diversity simply based on the knowledge about the corresponding GQ. In another word, the fact that a table generalization GT violates l-diversity may provide additional information about T, even though GT itself is not known. Such a disclosure is in the form of knowledge about *unsafe groups*, which is formalized in Definition 2.

Definition 2. (Unsafe Group) *Given a micro-data table T, a multiset of quasi-identifiers $Q' \subseteq \mathcal{Q}$ is said to be an unsafe group with respect to entropy l-diversity, if $|Q'| \geq l$ and $S' = \{s : (q, s) \in R_{qs}, q \in Q'\}$ does not satisfy entropy l-diversity.*

Clearly, no multiset S' of size less than l can ever satisfy entropy l-diversity. The following claim is then straightforward.

Claim 2 *Given a micro-data table T and a quasi-identifier generalization GQ, to disclose the fact that a table generalization GT computed based on GQ violates entropy l-diversity will*

- *not provide additional information about T, if GT also violates l-anonymity.*
- *provide the additional information about T that there exists at least one unsafe group in GQ, if GT satisfies l-anonymity.*

P-safety As illustrated in Section 1, when a generalization algorithm is publicly known, enforcing k-anonymity and l-diversity on the released table generalization is not sufficient. More specifically, if an algorithm is known to have considered $i-1$ table generalizations computed based on $GQ_1, GQ_2, \ldots, GQ_{i-1}$ before it finally releases $GT = (GQ_i, GS_i, GR_{qs_i})$, then l-diversity can no longer be evaluated on each group $gs \in GS_i$.

For instance, for generalization (B) in Table 2, when we evaluate l-diversity on each group of three conditions, we are actually assuming that an adversary can only guess the secret micro-data table (that is, Table 1) from generalization (B) alone. Therefore, any table not in a conflict with generalization (B) will be a valid guess. However, if it is a known fact that generalization (A) has been considered but not released, the adversary can drop any guessed table if it can make generalization (A) satisfy the required l-diversity.

More generally, the concept of *disclosure set* depicts the set of all possible guesses about a secret micro-data table T, when the generalization algorithm is publicly known [22]. The concept *P-safety* (P can be any privacy property, such as l-diversity) then defines the correct way for evaluating any privacy property based on the disclosure set. We repeat the proposed definition of *disclosure set* and the property of *P-safety* as follows.

Definition 3. (Disclosure Set and P-Safe) *Given a micro-data table T and a generalization algorithm that will consider the quasi-identifier generalizations $GQ_1, GQ_2, \ldots, GQ_n$ in the given order for satisfying a given privacy property P, we say*

- *the disclosure set $\mathcal{T}$ of a table generalization GT is the set of all micro-data tables for which the generalization algorithm will also output GT.*
- *a table generalization GT is P-safe, if for all identities $i' \in \mathcal{I}$, P is satisfied on the multiset $S_{i'} = \{s' : (i', s') \in \Pi_{i,s}(T'), T' \in \mathcal{T}\}$, where $\mathcal{T}$ is the disclosure set of GT.*

The concept of P-safety guarantees the desired privacy property to be satisfied even when the applied generalization algorithm is publicly known. However, the cost is high. To find an optimal table generalization that satisfies a privacy property, such as entropy l-diversity, is generally a NP-hard problem [22]. Also, as discussed in [22], it is even hard to have an efficient heuristic algorithm that provide practical data utility. In the rest of this paper, we will propose a different but more efficient strategy to address this issue.

3 A Novel Strategy for Micro-Data Release

We now consider a different strategy for micro-data release that decouples privacy preservation from data utility optimization. Roughly speaking, in stead of

optimizing the data utility in all possible quasi-identifier generalizations, it will first find a subset of the generalizations that satisfies two conditions: (1) every quasi-identifier generalization in the subset will yield a table generalization satisfying the given privacy property; (2) the given privacy property will still hold, even if the whole subset of the quasi-identifier generalizations is known to satisfy the first condition. Once such a subset of quasi-identifier generalizations is found, any data utility optimization can be done freely inside this collection without worrying about whether the generalization algorithm is publicized or not. Note that, a subset of generalizations of the above form can have a large size so that can computation of it is not practical. In the next section, we will show that, in practice, we can replace such computations by verifying whether a candidate generalization satisfies a proposed new property, which can be don efficiently.

3.1 Locally Safe Set

First, we consider a collection of quasi-identifier generalizations of which each can generalize the given micro-data table into a safe table generalization, as formalized in Definition 4.

Definition 4. (Locally Safe Set) *Given a micro-data table T and a desired entropy l-diversity, a* locally safe set *of quasi-identifier generalizations is the set $\mathcal{GQ}_L = \{GQ$: the table generalization of T computed based on GQ satisfies entropy l-diversity$\}$.*

Consider an example shown in Figure 1, which depicts the multisets of quasi-identifiers $\mathcal{Q}$ and sensitive values $\mathcal{S}$ of a micro-data table. Assume entropy 2-diversity is the desired privacy property. We can then compute the locally safe set $\mathcal{GQ}_L$. For example, $\mathcal{GQ}_L$ includes $GQ = \{\{q_1, q_3\}, \{q_2, q_4\}, \{q_5, q_6\}\}$.

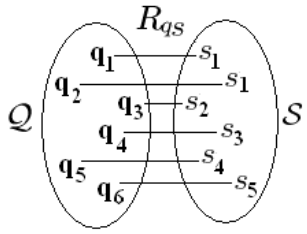

Fig. 1. An example of $\mathcal{Q}$ and $\mathcal{S}$

Next, assume an adversary has full knowledge about $\mathcal{GQ}_L$ itself (note that the knowledge about $GQ \in \mathcal{GQ}_L$ is different from that about the table gener-

alization computed based on GQ). If this knowledge does not violate the desired entropy l-diversity, then any optimization of generalization function for best data utility will never violate the desired entropy l-diversity. The reason is the optimization process is now simulatable. That is, the adversary, with the knowledge about $\mathcal{GQ}_L$ and the publicly known data utility metric, can repeat the optimization process and obtain the same result. In other words, we have the following claim which is straightforward:

Claim 3 *If disclosing the locally safe set $\mathcal{GQ}_L$ does not violate the desired entropy l-diversity, then any optimization of generalization function for best utility within $\mathcal{GQ}_L$ will not violate entropy l-diversity.*

However, the knowledge about $\mathcal{GQ}_L$ may indeed violate entropy l-diversity. First of all, by Claim 1, we know that to disclose all quasi-identifier generalizations whose corresponding table generalizations satisfy l-anonymity will not disclose any information, We call this the *candidate set* of quasi-identifier generalizations.

Definition 5. (Candidate Set) *Given a micro-data table T and a desired entropy l-diversity, a* candidate set *of quasi-identifier generalizations is the set $\mathcal{GQ}_A = \{GQ$: the table generalization of T computed based on GQ satisfies l-anonymity$\}$*

Therefore, the knowledge about $\mathcal{GQ}_L$ is equivalent to knowing about $\mathcal{GQ}_A \setminus \mathcal{GQ}_L$, which is the set of quasi-identifier generalizations whose corresponding table generalizations satisfy l-anonymity but violate entropy l-diversity. By Claim 2, the knowledge about $\mathcal{GQ}_L$ may therefore violate entropy l-diversity.

For example, in Figure 1, if we disclose $\mathcal{GQ}_L$, anyone can notice that any partition of $\mathcal{Q}$ that contains the subset $\{q_1, q_2\}$ or $\{q_1, q_2, q_x\}$, for any $q_x \in \mathcal{Q}$), will not appear in $\mathcal{GQ}_L$. On the other hand, for any other two-element set $\{q_x, q_y\}(q_x, q_y \in \mathcal{Q})$, there always exists at least one $GQ \in \mathcal{GQ}_L$ such that $\{q_x, q_y\} \in GQ$. Since the multiset of sensitive values $\mathcal{S}$ is public knowledge, anyone knows there is only one value s_1 that appears twice in $\mathcal{S}$. Therefore, anyone can determine the facts: $(q_1, s_1), (q_2, s_1) \in R_{qs}$.

3.2 Globally Safe Set

Now we study the condition for the knowledge about $\mathcal{GQ}_L$ to be safe. Consider the example shown in Figure 2 and assume entropy 2-diversity.

Clearly, there are two sets of quasi-identifiers, each of which contains two elements, that will not appear in the locally safe set $\mathcal{GQ}_L$. These are $\{q_1, q_2\}$ and $\{q_3, q_6\}$. Interestingly, at this time, the knowledge about $\mathcal{GQ}_L$ will only indicate that one of the following two facts holds:

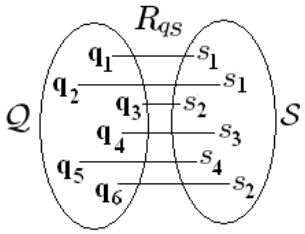

Fig. 2. Another example of $\mathcal{Q}$ and $\mathcal{S}$

- $(q_1, s_1), (q_2, s_1), (q_3, s_2), (q_6, s_2) \in R_{qs}$
- $(q_1, s_2), (q_2, s_2), (q_3, s_1), (q_6, s_1) \in R_{qs}$

Since the above two facts are equally likely to be true, the knowledge about $\mathcal{GQ}_L$ will not violate entropy 2-diversity by means of Definition 6. In the definition, the set of tables $\mathcal{T}$ can be regarded as the disclosure set (see Section 2.2) of $\mathcal{GQ}_L$. That is, $\mathcal{T}$ is the set of micro-data tables not in conflict with the fact that $\mathcal{GQ}_L$ is a locally safe set.

Definition 6. (Globally Safe Set) *Given a micro-data table T and a set of quasi-identifier generalizations $\mathcal{GQ}$, let $\mathcal{T}$ be the set of tables satisfying that for all $T' \in \mathcal{T}$,*

- *T' has the same $\mathcal{I}$, $\mathcal{Q}$, $\mathcal{S}$, and R_{iq} as T does, and*
- *the table generalization of T' computed based on every $GQ \in \mathcal{GQ}$ satisfies entropy l-diversity,*

we say $\mathcal{GQ}$ is a globally safe set of quasi-identifier generalizations, if $\forall i' \in \mathcal{I}$, l-diversity is satisfied on the multiset $S_{i'} = \{s' : (i', s') \in \Pi_{i,s}(T'), T' \in \mathcal{T}\}$.

By Claim 3, if a locally safe set of quasi-identifier generalizations $\mathcal{GQ}_L$ happens to be also a globally safe set, then any optimization generalization function for best data utility within $\mathcal{GQ}_L$ will not violate entropy l-diversity. However, we also know from above discussions that $\mathcal{GQ}_L$ is not always globally safe . Therefore, we need to further restrict the optimization of generalization function to be within subsets of $\mathcal{GQ}_L$ that are globally safe.

4 l-Cover

We introduce the concept of *l-cover* for finding globally safe sets of quasi-identifier generalizations. Recall that our strategy has two stages:

1. Find a globally safe set $\mathcal{GQ} \subseteq \mathcal{GQ}_L$.

2. Compute a table generalization GT based on $GQ \in \mathcal{GQ}$ such that GT has the optimal data utility.

Correspondingly, we have two forms of l-cover, weak l-cover and l-cover.

4.1 Weak l-Cover

In Figure 2, we can observe that two values, s_1 and s_2, both appear twice in the multiset $\mathcal{S} = \{s_1, s_1, s_2, s_2, s_3, s_4\}$. Therefore, the corresponding two sets of quasi-identifiers $\{q_1, q_2\}$ and $\{q_3, q_6\}$ provide a *cover* for each other in the sense that they cannot be distinguished based on the knowledge about $\mathcal{GQ}_L$. In this case, sensitive values having exactly the same number of appearances in $\mathcal{S}$ cover each other. However, this is not a necessary condition. Consider another example shown in Figure 3 and assume entropy 2-diversity.

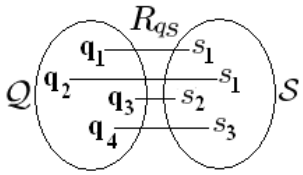

Fig. 3. *cover* from sensitive values with different number of appearances

We have the locally safe set $\mathcal{GQ}_L = \{GQ_1, GQ_2, GQ_3\}$ where $GQ_1 = \{\{q_1, q_3\}, \{q_2, q_4\}\}$, $GQ_2 = \{\{q_1, q_4\}, \{q_2, q_3\}\}, GQ_3 = \{\{q_1, q_2, q_3, q_4\}\}$. We can observe that $\{q_1, q_2\}$ never appears in any of the quasi-identifier generalizations in $\mathcal{GQ}_L$ due to their identical sensitive value s_1. Moreover, the set $\{q_3, q_4\}$ never appears, either. Therefore, $\{q_3, q_4\}$ becomes a cover of $\{q_1, q_2\}$ even though their corresponding sensitive values have different number of appearances. More generally, we define the concept of cover in the following.

Definition 7. (Cover) *Given a set of quasi-identifier generalizations $\mathcal{GQ}$ on $\mathcal{Q}$, we say $Q \subseteq \mathcal{Q}$ and $Q' \subseteq \mathcal{Q}$ provide cover for each other, if*

- *$Q' \cap Q = \phi$, and*
- *there exists a bijection $f_{cover} : Q \rightarrow Q'$ satisfying the following. For any $Q_x \in GQ, GQ \in \mathcal{GQ}$, there exists $GQ' \in \mathcal{GQ}$ such that $(Q_x \setminus (Q \cup Q')) \cup f_{cover}(Q_x \cap Q) \cup f^{-1}_{cover}(Q_x \cap Q') \in GQ'$.*

Note that, if $Q_x \cap (Q \cup Q') = \phi$, $GQ' = GQ$ naturally exists. Interestingly, the way we provide a cover for a set of quasi-identifiers Q is similar to providing

"anonymity" to sensitive values. In another word, the concept of cover is similar to anonymity if we exchange the role of identity and sensitive value in the micro-data table. Therefore, analogous to k-anonymity, we have the metric of l-cover in Definition 8.

Definition 8. (Weak l-Cover) *Given a micro-data table T and a set of quasi-identifier generalizations $\mathcal{GQ}$, $\mathcal{GQ}$ is said to satisfy* weak l-cover *if for any $Q \subseteq \mathcal{Q}$ satisfying $\exists s' \in \mathcal{S}, Q = \{q' : (q', s') \in R_{qs}\}$, we have*

- *there exist at least $l-1$ covers of Q: $Q_1, Q_2, \ldots, Q_{l-1}$, and*
- *$\forall j \neq j', Q_j \cap Q_{j'} = \phi$.*

Claim 4 states that weak l-cover is a sufficient condition for a globally safe set (this condition is also suspected to be necessary). Intuitively, each sensitive value (and its number of appearances) is blended into the sensitive values of its $l-1$ or more covers. The knowledge about the quasi-identifier generalizations $\mathcal{GQ}$ thus will not violate l-diversity.

Claim 4 *A set of quasi-identifier generalizations $\mathcal{GQ}$ is a globally safe set with respect to entropy l-diversity if $\mathcal{GQ}$ satisfies weak l-cover.*

Proof Sketch: Consider the set of tables $\mathcal{T}$ and the multiset of sensitive values $S_{i'}$ as defined in Definition 6. Let s' be the most frequent element in $S_{i'}$. Based on the definition of weak l-cover, the set of quasi-identifier $Q = \{q' : q' \in \mathcal{Q}, (q', s') \in R_{qs}\}$ has $l-1$ covers, $Q_1, \ldots, Q_{l-1}$ each of which has a corresponding bijection $f_i : Q \rightarrow Q_i (1 \leq i \leq l-1)$. Therefore, for any table $T' \in \mathcal{T}$ satisfying $(i', q', s') \in T'$, there must exist $T_1 \in \mathcal{T}$ such that $(i', q', s_1) \in T_1$ and $s' \neq s_1$, where s_1 satisfies $(i'', f_i(q), s_1) \in T'$. Similarly, we can have $s_2, \ldots, s_{l-1}$ corresponding to each cover of Q. Therefore, there exist at least $l-1$ other different sensitive values that have the same number of appearances as s' does in $S_{i'}$. The property of entropy l-diversity is thus satisfied. □

From Claim 3 and Claim 4, we immediately have the following.

Claim 5 *A generalization algorithm will not violate entropy l-diversity while optimizing data utility within a set of quasi-identifier generalizations $\mathcal{GQ}$ that satisfies weak l-cover.*

Note that, among the previous examples, those shown in Figure 2 and Figure 3 satisfy l-cover, whereas the one in Figure 1 does not. Therefore, a (deterministic) generalization algorithm may violate entropy l-diversity, when it attempts to disclose a quasi-identifier generalization with optimal data utility for the micro-data table shown in Figure 1, even if the corresponding table generalization is not yet disclosed.

4.2 l-Cover

From the previous discussions, we will optimize data utility within a globally safe set of quasi-identifier generalizations. Once this optimization process finishes, we will need to compute and release a table generalization based on the optimal quasi-identifier generalization. However, such a disclosure introduces additional knowledge about the secret micro-data table, and may violate the desired entropy l-diversity.

First, consider the example shown in Figure 3, which has a locally safe set $\mathcal{GQ}_L$ that is also globally safe. Assume the optimization of generalization function has found that inside $\mathcal{GQ}_L$, the quasi-identifier generalization $GQ_1 = \{\{q_1, q_3\}, \{q_2, q_4\}\}$ is optimal. We can thus compute the table generalization shown in Table 4.

Quasi-Identifier	Sensitive Value
q_1	s_1
q_3	s_2
q_2	s_1
q_4	s_3

Table 4. Table Generalization for Figure 3

With this table generalization disclosed, the set of quasi-identifiers $\{q_3, q_4\}$ is still a cover of $\{q_1, q_2\}$. That is, an adversary still cannot tell which of them is associated with both appearances of the sensitive value s_1. Therefore, in this particular case, the table generalization in Table 4 can be safely released.

However, this is not always the case. Releasing a table generalization may violate the privacy property that has been satisfied in the process of finding a globally safe set and optimizing data utility. Consider the example shown in Figure 2. Assume that $GQ_1 = \{\{q_1, q_3\}, \{q_2, q_4\}, \{q_5, q_6\}\}$ is the optimal quasi-identifier generalization. Based on GQ_1, we can compute the following table generalization.

Quasi-Identifier	Sensitive Value
q_1	s_1
q_3	s_2
q_2	s_1
q_4	s_3
q_5	s_2
q_6	s_4

Table 5. Table Generalization for Figure 2

Recall that during the discussion about Figure 2, we have shown that the locally safe set of quasi-identifier generalizations $\mathcal{GQ}_L$ is also globally safe. More specifically, $\{q_1, q_2\}$ and $\{q_3, q_6\}$ provide cover for each other. An adversary thus cannot tell which of these is associated to s_1 and which to s_2. However, if the table generalization in Table 5 is disclosed, then clearly, since $\{q_5, q_6\}$ is associated with $\{s_2, s_4\}$, q_6 must not be associated with s_1 in the micro-data table. Therefore, the following must be true: $(q_1, s_1), (q_2, s_1), (q_3, s_2), (q_6, s_2) \in R_{qs}$, which violates entropy l-diversity.

In the above example, the table generalization in Table 5 contains extra information that is not part of the knowledge about $\mathcal{GQ}_L$. Therefore, the table generalization computed based on GQ_1 cannot be safely released, even though $\mathcal{GQ}_L$ is globally safe. To prevent such cases, we should not consider quasi-identifier generalizations like GQ_1 for the optimization of data utility. Instead, the optimization process should be confined to a subset of the globally safe set $\mathcal{GQ}_L$ that satisfies a stronger condition, as formalized in Definition 9.

Definition 9. (l-Cover) *Given a micro-data table T and a set of quasi-identifier generalizations $\mathcal{GQ}$, $\mathcal{GQ}$ is said to satisfy l-*cover *if for any $Q \subseteq \mathcal{Q}$ satisfying $\exists s' \in \mathcal{S}, Q = \{q' : (q', s') \in R_{qs}\}$, we have*

- *there exist at least $l-1$ covers of Q: $Q_1, Q_2, \ldots, Q_{l-1}$,*
- *$\forall j \neq j', Q_j \cap Q_{j'} = \phi$, and*
- *$\forall GQ \in \mathcal{GQ}, \forall Q_x \in GQ, |Q_x \cap Q| = |Q_x \cap Q_j|$ $(j = 1, 2, \ldots, l-1)$.*

The property of l-cover basically requires a set of quasi-identifier generalizations $\mathcal{GQ}$ to satisfy both weak l-cover and an additional conditions, that is, the disclosure of a table generalization computed based on any $GQ \in \mathcal{GQ}$ will not include any extra information that is not part of the knowledge about $\mathcal{GQ}$. Any such table generalization can thus be safely released. More formally, we have the following.

Claim 6 *Given a micro-data table T, and a set of quasi-identifier generalizations $\mathcal{GQ}$ satisfying l-*cover*, and any $GQ \in \mathcal{GQ}$, let $GT = (GQ, GS, GR_{qs})$ be the table generalization of T computed based on GQ. Also, let $\mathcal{T}$ be the set of all tables having the same $\mathcal{I}, R_{iq}, \mathcal{Q}, \mathcal{S}$ and the same table generalization GT when computed based on GQ. We then have that entropy l-diversity is satisfied on the multiset $s_{i'} = \{s' : (i', s') \in \Pi_{i,s}(T'), T' \in \mathcal{T}\}$ for all $i' \in \mathcal{I}$.*

Proof Sketch: The proof of this claim is similar to that of Claim 4, except that the disclosure of the table generalization GT does not allow an adversary to disregard any quasi-identifier generalization in $\mathcal{GQ}$ by Definition 9. □

In addition, we can have another interesting observation about those sensitive values that appear exactly once in $\mathcal{S}$. That is, as long as each group in the table generalization has more than l different sensitive values, those sensitive values that appear only once will be protected with l-cover. Therefore, in practice we only need to be concerned with those sensitive values that appear multiple times. From Claim 5 and Claim 6, the following argument is now straightforward.

Claim 7 *A generalization algorithm will not violate entropy l-diversity while disclosing a table generalization with the optimal data utility, if the optimization process is confined to a set of quasi-identifier generalizations that satisfies l-cover.*

Since the locally safe set of quasi-identifier generalizations $\mathcal{GQ}_L$ does not always satisfy l-cover. To preserve the property of entropy l-diversity, we may need to find a subset of $\mathcal{GQ}_L$ that does so. Note that, to avoid the huge complexity to compute the entire $\mathcal{GQ}_L$, We can: (1) in advance construct l covers for any sensitive values that appears more than once; (2) check whether a given generalization is contained in a $\mathcal{GQ}_L$ that satisfies l-cover by checking whether the property of L-cover can be violated by the given generalization, based on the previously constructed l-covers and the generalizations that have already been considered. We shall leave detailed methods and the study of the corresponding performances to our future work. Nonetheless, by following this approach, we will not face the NP-hard problem of preserving entropy l-diversity with publicized algorithms as pointed out in [22], and expect to have "better" heuristic algorithms that guarantees entropy l-diversity when publicized, in terms of data utility and efficiency.

5 Related Work

The initial works [1, 3, 7, 9, 10] were concerned with conducting data census, while protecting the privacy of sensitive information in disclosed tables. Two approaches, data swapping [6, 14, 19] and data suppression [11] were suggested to protect data, but could not quantify how well the data is protected. The work [5] gave a formal analysis of the information disclosure in data exchange. The work [16] showed that publishing data sets even without identifying attributes can cause privacy breaches and suggested a new notion of privacy called k-anonymity. Achieving k-anonymity with the best data utility was proved to be NP-hard [13]. A similar measure, called *blending in a crowd* was proposed by [18]. The work [21] proposed a new generalization framework based on the concept of "personalized anonymity." In addition, many works, e.g., [4, 15, 16, 12,

17, 8], proposed efficient algorithms for k-anonymity. The work [2] discussed deficiency of k-anonymity as a measure of privacy, and proposed an alternative property of l-diversity to ensure privacy protection in the micro-data disclosure, and demonstrated that algorithms developed for k-anonymity can also be used for l-diversity. The above works, however, did not take into account that the disclosure algorithm and sequence may be known to the adversary. The work [22] provide an comprehensive analysis of both safety and complexity for the disclosure algorithm for micro-data disclosure under such assumption. Another work [20] tackles a similar issue but in a more specific problem setting.

6 Conclusion

We have uncovered the similarity between k-anonymity and l-diversity under a novel strategy for micro-data release. More specifically, we have proposed to confine the optimization of generalization function for best data utility to a globally safe subset of all possible quasi-identifier generalizations. This approach decoupled privacy preservation from data utility optimization, which essentially simplified both. To find a globally safe set, we have provided the concept of l-cover and shown that to satisfy this novel property is basically to satisfy l-anonymity in a special way. This result may lead to "better" heuristic algorithms than existing solutions in terms of data utility and efficiency, while guaranteeing the data privacy with publicized algorithms. Our future work will focus on the algorithm design and performance study.

Acknowledgment

Lei Zhang and Sushil Jajodia were partially supported by the National Science Foundation under grants CT-0716567, CT-0716323, and CT-0627493, and by the Air Force Office of Scientific Research under grants FA9550-07-1-0527 and FA9550-08-1-0157. We thank the anonymous reviewers for their valuable comments to improve this paper.

References

1. A.Dobra and S.E.Feinberg. Bounding entries in multi-way contingency tables given a set of marginal totals. In *Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000*. Springer Verlag, 2003.
2. A.Machanavajjhala, J.Gehrke, D.Kifer, and M.Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, 2006.
3. A.Slavkovic and S.E.Feinberg. Bounds for cell entries in two-way tables given conditional relative frequencies. *Privacy in Statistical Databases*, 2004.

4. 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.
5. G.Miklau and D.Suciu. A formal analysis of information disclosure in data exchange. In *SIGMOD*, 2004.
6. G.T.Duncan and S.E.Feinberg. Obtaining information while preserving privacy: A markov perturbation method for tabular data. In *Joint Statistical Meetings*. Anaheim,CA, 1997.
7. I.P.Fellegi. On the question of statistical confidentiality. *Journal of the American Statistical Association*, 67(337):7–18, 1993.
8. K.LeFevre, D.DeWitt, and R.Ramakrishnan. Incognito: Efficient fulldomain k-anonymity. In *SIGMOD*, 2005.
9. L.H.Cox. Solving confidentiality protection problems in tabulations using network optimization: A network model for cell suppression in the u.s. economic censuses. In *Proceedings of the Internatinal Seminar on Statistical Confidentiality*, 1982.
10. L.H.Cox. New results in disclosure avoidance for tabulations. In *International Statistical Institute Proceedings*, 1987.
11. L.H.Cox. Suppression, methodology and statistical disclosure control. *J. of the American Statistical Association*, 1995.
12. L.Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.
13. A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *ACM PODS*, 2004.
14. P.Diaconis and B.Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 1998.
15. P.Samarati. Protecting respondents' identities in microdata release. In *IEEE TKDE*, 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. R.J.Bayardo and R.Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, 2005.
18. S.Chawla, C.Dwork, F.McSherry, A.Smith, and H.Wee. Toward privacy in public databases. In *Theory of Cryptography Conference*, 2005.
19. T.Dalenius and S.Reiss. Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6:73–85, 1982.
20. R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, pages 543–554, 2007.
21. X.Xiao and Y.Tao. Personalized privacy preservation. In *SIGMOD*, 2006.
22. L. Zhang, S. Jajodia, and A. Brodsky. Information disclosure under realistic assumptions: Privacy versus optimality. In *ACM Conference on Computer and Communications Security (CCS) 2007*.