Anatomy: Simple and Effective Privacy Preservation

Xiaokui Xiao Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong
Sha Tin, New Territories, Hong Kong
{xkxiao, taoyf}@cse.cuhk.edu.hk

ABSTRACT

This paper presents a novel technique, *anatomy*, for publishing sensitive data. Anatomy releases all the quasi-identifier and sensitive values directly in two separate tables. Combined with a grouping mechanism, this approach protects privacy, and captures a large amount of correlation in the microdata. We develop a linear-time algorithm for computing anatomized tables that obey the l-diversity privacy requirement, and minimize the error of reconstructing the microdata. Extensive experiments confirm that our technique allows significantly more effective data analysis than the conventional publication method based on generalization. Specifically, anatomy permits aggregate reasoning with average error below 10%, which is lower than the error obtained from a generalized table by orders of magnitude.

1. INTRODUCTION

Privacy preservation is a serious concern in publication of personal data. Using a popular example in the literature, assume that a hospital wants to release patients' medical records in Table 1, referred to as the *microdata*. Attribute *Disease* is *sensitive*, that is, the hospital must ensure that no adversary can correctly infer the disease of any patient with significant confidence. *Age*, *Sex*, and *Zipcode* are the *quasi-identifier* (QI) attributes, because they may be utilized in combination to reveal the identity of an individual, leading to privacy breach.

Consider an adversary who has the personal details (i.e., age 23 and zipcode 11000) of Bob, and knows that Bob has been hospitalized before. In Table 1, since only tuple 1 matches Bob's QI-values, the adversary asserts that Bob contracted pneumonia.

To avoid this problem, *generalization* [12, 13, 14, 10] divides tuples into *QI-groups*, and transforms their QI-values into less specific forms, so that tuples in the same QI-group cannot be distinguished by their QI-values. Table 2 is a generalized version of Table 1 (e.g., the age 23 and zipcode 11000 of tuple 1 have been replaced with intervals $[21, 60]$ and $[10001, 60000]$, respectively). Here, generalization produces two QI-groups, including tuples 1-4 and 5-8, respectively. As a result, even if an adversary has the exact QI values of Bob, s/he still does not know which tuple in the first QI-group belongs to Bob.

tuple ID	Age	Sex	Zipcode	Disease
1 (Bob)	23	M	11000	pneumonia
2	27	M	13000	dyspepsia
3	35	M	59000	dyspepsia
4	59	M	12000	pneumonia
5	61	F	54000	flu
6	65	F	25000	gastritis
7 (Alice)	65	F	25000	flu
8	70	F	30000	bronchitis

Table 1: The microdata

tuple ID	Age	Sex	Zipcode	Disease
1	[21, 60]	M	[10001, 60000]	pneumonia
2	[21, 60]	M	[10001, 60000]	dyspepsia
3	[21, 60]	M	[10001, 60000]	dyspepsia
4	[21, 60]	M	[10001, 60000]	pneumonia
5	[61, 70]	F	[10001, 60000]	flu
6	[61, 70]	F	[10001, 60000]	gastritis
7	[61, 70]	F	[10001, 60000]	flu
8	[61, 70]	F	[10001, 60000]	bronchitis

Table 2: A 2-diverse table

Two notions, k-*anonymity* and l-*diversity*, have been proposed to measure the degree of privacy preservation. A (generalized) table is k-*anonymous* [12, 13, 14] if each QI-group involves at least k tuples (e.g., Table 2 is 4-anonymous). However, as shown in [10], even with a large k, k-anonymity may still allow an adversary to infer the sensitive value of an individual with extremely high confidence. Hence, we adopt l-diversity [10], which provides stronger privacy protection.

Specifically, a table is l-diverse if, in each QI-group, at most $1/l$ of the tuples possess the most frequent sensitive value[1]. For instance, Table 2 is 2-diverse because, in each QI-group, at most 50% of the tuples have the same *Disease* value. As mentioned earlier, the adversary (targeting Bob's medical record) knows that Bob's tuple must be in the first QI-group, where two tuples are associated with pneumonia, and two with dyspepsia. Hence, the adversary can only make a probabilistic conjecture: Bob could have contracted either disease with the same probability.

1.1 Defects of Generalization in Aggregate Analysis

Although generalization preserves privacy, it often loses considerable information in the microdata, which severely compromises

[1] l-diversity has more complicated requirements, if an adversary's "background knowledge" is taken into account [10]. We will discuss this issue in Section 3.1.

VLDB '06, September 12-15, 2006, Seoul, Korea.

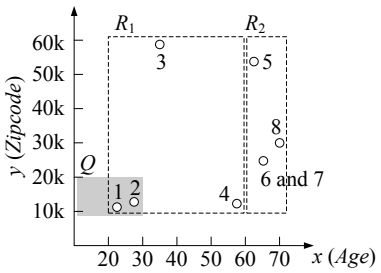


Figure 1: The original and generalized data in the *Age-Zipcode* plane

the accuracy of data analysis. Assume that the hospital releases Table 2, and that a researcher wants to derive from this table an estimate for the following query:

A: `SELECT COUNT(*) FROM` Unknown-Microdata
`WHERE` *Disease* = 'pneumonia' `AND` *Age* <= 30
`AND` *Zipcode* `IN` [10001, 20000]

To illustrate how to process the query, Figure 1 shows a 2D space, where the x-, y-dimensions are *Age* and *Zipcode*, respectively. Each point denotes a tuple in the microdata of Table 1. For example, the x-, y-coordinates of point 1 equal the age and zipcode of tuple 1, respectively. Rectangle R_1 (or R_2) is obtained from the generalized values in the first (or second) QI-group in Table 2. For instance, the x- (y-) projection of R_1 is the generalized age $[20, 60]$ (zipcode [10001, 60000]) of tuples 1-4. Query A is represented as the shaded rectangle Q, whose projection on the x- (y-) dimension is decided by the range condition *Age* ≤ 30 ($10001 \leq$ *Zipcode* ≤ 20000).

Since the researcher sees only R_1 and R_2 (but not the points), s/he answers query A in a way similar to selectivity estimation on a multidimensional histogram [15], as suggested in [9]. Clearly, as R_2 is disjoint with Q, no tuple in the second QI-group can satisfy the query. R_1, however, intersects Q, and hence, is examined as follows.

From the *Disease*-values in Table 2, the researcher knows that 2 tuples in the first QI-group are associated with pneumonia. It remains to calculate the probability p that a tuple in the QI-group qualifies the range predicates of A, or equivalently, the tuple's point representation falls in Q (Figure 1). Once p is available, the query answer can be estimated as $2p$.

Without additional knowledge, the researcher assumes uniform data distribution in R_1, and computes p as $Area(R_1 \cap R_Q)/Area(R_1) = 0.05$. This value leads to an approximate answer 0.1, which, however, is ten times smaller than actual query result 1 (see Table 1).

The gross error is caused by the fact that the data distribution in R_1 significantly deviates from uniformity. Nevertheless, given only the generalized table, we cannot justify any other distribution assumption. This is an inherent problem of generalization: it prevents an analyst from correctly understanding the data distribution inside each QI-group.

1.2 Rationale of Anatomy

To overcome the defects of generalization, we propose an innovative technique, *anatomy*, to achieve privacy-preserving publication that captures the exact QI-distribution.

Specifically, anatomy releases a *quasi-identifier table* (QIT) and a *sensitive table* (ST), which separate QI-values from sensitive values. For example, Tables 3a and 3b demonstrate the QIT and ST obtained from the microdata Table 1, respectively.

row #	Age	Sex	Zipcode	Group-ID
1	23	M	11000	1
2	27	M	13000	1
3	35	M	59000	1
4	59	M	12000	1
5	61	F	54000	2
6	65	F	25000	2
7	65	F	25000	2
8	70	F	30000	2

(a) The quasi-identifier table (QIT)

Group-ID	Disease	Count
1	dyspepsia	2
1	pneumonia	2
2	bronchitis	1
2	flu	2
2	gastritis	1

(b) The sensitive table (ST)

Table 3: The anatomized tables

Construction of the anatomized tables can be (informally) understood as follows. First, we partition the tuples of the microdata into several QI-groups, based on a certain strategy. Here, following the grouping in Table 2, let us place tuples 1-4 (or 5-8) of Table 1 into QI-group 1 (or 2).

Then, we create the QIT. Specifically, for each tuple in Table 1, the QIT (Table 3a) includes all its *exact* QI-values, together with its group membership in a new column *Group-ID*. However, QIT does not store any *Disease* value.

Finally, we produce the ST (Table 3b), which retains the *Disease* statistics of each QI-group. For instance, the first two records of the ST (to avoid confusion, we use 'record', instead of 'tuple', for the data of an ST) indicate that, two tuples of the first QI-group are associated with dyspepsia, and two with pneumonia. Similarly, the next three records imply that, the second QI-group has a tuple associated with bronchitis, two with flu, and one with gastritis.

Anatomy preserves privacy because the QIT does not indicate the sensitive value of any tuple, which must be randomly guessed from the ST. To explain this, consider again the adversary who has the age 23 and zipcode 11000 of Bob. Hence, from the QIT (Table 3a), the adversary knows that tuple 1 belongs to Bob, but does not obtain any information about his disease so far. Instead, s/he gets the id 1 of the QI-group containing tuple 1. Judging from the ST (Table 3b), the adversary realizes that, among the 4 tuples in QI-group 1, 50% of them are associated with dyspepsia (or pneumonia) in the microdata. Note that s/he does not gain any additional hints, regarding the exact diseases carried by these tuples. Hence, s/he arrives at the conclusion that Bob could have contracted dyspepsia (or pneumonia) with 50% probability. This is the same conjecture obtainable from the generalized Table 2, as mentioned earlier.

By announcing the QI values directly, anatomy permits more effective analysis than generalization. Given query A in Section 1.1, we know, from the ST (Table 3b), that 2 tuples carry pneumonia in the microdata, and they are both in QI-group 1. Hence, we proceed to calculate the probability p that a tuple in the QI-group falls in Q (Figure 1). This calculation does not need any assumption about the data distribution in the *Age-Zipcode* plane, *because the distrib-*

ution is precisely released. Specifically, the QIT (Table 3a) shows that tuples 1 and 2 in QI-group 1 appear in Q, leading to the *exact* $p = 50\%$. Thus, we obtain an answer $2p = 1$, which is also the actual query result.

1.3 Contributions

This paper presents a systematic study of the anatomy technique. First, we formalize the new methodology, based on the privacy requirement of l-diversity. Every pair of QIT and ST ensures that the sensitive value of any individual involved in the microdata can be correctly inferred by an adversary with probability at most $1/l$. A larger l leads to stronger privacy protection.

Second, we clarify the theoretical reasoning behind the superiority of anatomy in capturing data correlation. Our results show that anatomy permits a more accurate modeling of each tuple in the microdata than generalization. We provide detailed analysis of the modeling error, and quantify it into a closed formula.

Third, we develop an algorithm that computes anatomized tables in $O(n/b)$ I/Os, where n is the cardinality of the microdata, and b the page size. These tables have provably good quality guarantee, achieving a modeling error deviating from the theoretical lower bound by a factor of at most $1 + 1/n$. Notice that, n is very large in practice (e.g., at the order a million); hence, our algorithm is nearly optimal.

Finally, we prove, through extensive experiments, that anatomy significantly outperforms generalization, in both *effectiveness of data analysis* and *computation cost*. Specifically, the anatomized tables permit highly accurate aggregate search (e.g., query A in Section 1), with average error below 10%, which is lower than the query error obtained from a generalized table by orders of magnitude. The query accuracy of anatomy is unaffected by the dataset dimensionality, whereas the accuracy of generalization decays severely as dimensionality increases. Furthermore, anatomized tables can be computed much faster than generalized tables.

The rest of the paper is organized as follows. Section 2 surveys the previous work on generalization. Section 3 formalizes the anatomy methodology, and clarifies its privacy protection guarantees. Section 4 analyzes correlation preservation. Section 5 develops an algorithm for computing anatomized tables. Section 6 experimentally evaluates the proposed solutions. Section 7 concludes the paper with directions for future work.

2. RELATED WORK

Generalization has been very well studied in the literature [1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18]. LeFevre et al. [8] present an interesting taxonomy to categorize alternative methods based on their "encoding schemes", which impose different constraints in generalizing a QI-value. The highest level of the taxonomy distinguishes *global recoding* from *local recoding*. Specifically, the former requires that, all the tuples with equivalent QI-values must be included in the same QI-group. For instance, tuples 6 and 7 in Table 1 have identical QI-values; hence, they both appear in the second QI-group of Table 2. Local recoding removes this requirement, but has not received considerable attention in the literature (currently, this approach is applied only in several "suppression-based" solutions [8]).

The category of global recoding can be further divided into *Single-dimension encoding* and *multidimension encoding*. Specifically, an encoding is single dimensional, if the generalized forms of two arbitrary QI-groups on the same attribute are either disjoint or equivalent, as is the case in Table 2. When the condition is not satisfied, the encoding is multidimensional. For example, imagine that the *Zipcode*-values of tuples 5-8 in Table 2 were changed to $[20001, 60000]$, which intersects but is not identical to the *Zipcode*-form of tuples 1-4; as a result, the generalization would become multidimensional.

Computing the optimal generalization is harder for encoding schemes with fewer constraints. Unfortunately, it is NP-hard to find the optimal solution, even for simple schemes and quality metrics [2, 9, 11]. Therefore, the existing algorithms rely on heuristics for pruning the search space, in order to discover reasonable generalization within a time limit.

A majority of the literature focuses on k-anonymous generalization. However, Machanavajjhala et al. [10] observe that k-anonymity fails to secure privacy in practice. In particular, they show that, the degree of privacy protection does not really depend on the size of a QI-group, but instead, is determined by the number of *distinct* sensitive values in each QI-group. The observation leads to l-diversity (as will be formalized in Section 3). The analysis of [17] proves that l-diversity always guarantees stronger privacy preservation than k-anonymity.

A serious drawback of generalization is that, when the number d of QI attributes is large, any generalization necessarily loses considerable information in the microdata [1], due to the "curse of dimensionality". Specifically, in high dimensional spaces, each generalized value is always an exceedingly wide interval, in which case the published table is simply useless for research.

This paper is virtually orthogonal to all the previous work. The proposed anatomy technique is a brand-new approach for publishing personal data, which remedies the defects of generalization. Specifically, nearly-optimal anatomized tables can be computed in linear-time with respect to the database cardinality, and capture a significant amount of correlation for any dimensionality.

3. FORMALIZATION OF ANATOMY

Let T be the microdata that needs to be published. T contains d quasi-identifier (QI) attributes A_1^{qi}, A_2^{qi}, ..., A_d^{qi}, and a sensitive attribute A^s. Each A_i^{qi} ($1 \leq i \leq d$) can be either numerical or categorical, but A^s should be categorical, following the assumption of l-diversity [10]. For any tuple $t \in T$, we denote $t[i]$ ($1 \leq i \leq d$) as the A_i^{qi} value of t, and $t[d+1]$ as its A^s value. As a result, t can be regarded as a point in a $(d+1)$-dimensional data space, denoted as DS. In Section 3.1, we first clarify the relevant concepts of anatomy. Then, Section 3.2 explains the privacy guarantees of anatomized tables.

3.1 Concepts

As with generalization, Anatomy requires partitioning the microdata T.

DEFINITION 1. **(Partition/QI-group)** *A* **partition** *consists of several subsets of T, such that each tuple in T belongs to exactly one subset. We refer to these subsets as* **QI-groups**, *and denote them as QI_1, QI_2, ..., QI_m. Namely, $\bigcup_{j=1}^{m} QI_j = T$ and, for any $1 \leq j_1 \neq j_2 \leq m$, $QI_{j_1} \cap QI_{j_2} = \emptyset$.*

We are interested only in l-diverse partitions that can lead to provably good privacy guarantees:

DEFINITION 2. **(l-diverse partition [10])** *A partition with m QI-groups is **l-diverse**, if each QI-group QI_j ($1 \leq j \leq m$) satisfies the following condition. Let v be the most frequent A^s value in QI_j, and $c_j(v)$ the number of tuples $t \in QI_j$ with $t[d+1] = v$; then*

$$c_j(v)/|QI_j| \leq 1/l \quad (1)$$

where $|QI_j|$ is the size (the number of tuples) of QI_j.

Table 1 shows a partition with two QI-groups, where QI_1 contains tuples 1-4, and QI_2 includes tuples 5-8. In QI_1, dyspepsia and pneumonia are equally frequent, i.e., c_1(dyspepsia) = c_1(pneumonia) = 2. In QI_2, the most frequent A^s value is flu, i.e., c_2(flu) = 2. Since $|QI_1| = |QI_2| = 4$, according to Inequality 1, we know that QI_1 and QI_2 constitute a 2-diverse partition.

We are ready to formulate the QIT and ST tables published by anatomy.

DEFINITION 3. **(Anatomy)** *Given an l-diverse partition with m QI-groups,* **anatomy** *produces a* **quasi-identifier table** *(QIT) and a* **sensitive table** *(ST) as follows. The QIT has schema*

$$(A_1^{qi}, A_2^{qi}, ..., A_d^{qi}, \textit{Group-ID}).$$

For each QI-group QI_j ($1 \leq j \leq m$) and each tuple $t \in QI_j$, QIT has a tuple of the form:

$$(t[1], t[2], ..., t[d], j).$$

The ST has schema

$$(\textit{Group-ID}, A^s, \textit{Count}).$$

For each QI-group QI_j ($1 \leq j \leq m$) and each distinct A^s value v in QI_j, the ST has a record of the form:

$$(j, v, c_j(v))$$

where $c_j(v)$ is the number of tuples $t \in QI_j$ with $t[d+1] = v$. Apart from the tuples (or records) defined earlier, the QIT (or ST) does not contain any other data.

For instance, based on the 2-diverse partition suggested in Table 2, anatomy produces the QIT and ST in Tables 3a and 3b respectively, as explained in Section 1.2.

When there is no ambiguity, we refer to a pair of QIT and ST collectively as the *anatomized tables*. In Section 4, we will show that anatomized tables capture the correlation in T more accurately than generalized tables. For this purpose, we also need to formalize generalization.

DEFINITION 4. **(Generalization)** *Given a partition of T with m QI-groups, for any tuple $t \in T$, a* **generalized table** *of T contains a tuple of the form*

$$(QI_j[1], QI_j[2], ..., QI_j[d], t[d+1])$$

where QI_j ($1 \leq j \leq m$) is the unique QI-group including t, and $QI_j[i]$ ($1 \leq i \leq d$) is an interval[2] covering $t[i]$. Furthermore, $QI_j[i]$ is identical for all tuples $t \in QI_j$. Apart from the tuples defined earlier, the table does not contain any other data.

[2]If A_i^{qi} is categorical, following a common assumption in the literature, we consider that there is a total ordering on A_i^{qi}.

For instance, let t be tuple 1 in the microdata Table 1. We have $j = 1$, namely, t is contained in the first QI-group. In the generalized Table 2, $QI_1[1] = [21, 60]$ (the generalized age of tuple 1), $QI_1[2]$ = M, and $QI_1[3] = [100001, 60000]$, which, together with $t[4]$ = pneumonia, form the first tuple.

We would like to point out that, although Definition 3 is based on an l-diverse partition, in general, anatomy produces a pair of QIT and ST from any partition (Definition 1) in exactly the same way. In particular, any k-anonymous or l-diverse table has an anatomized counterpart. We concentrate on l-diverse partitions to achieve strong privacy preservation.

It is worth mentioning that Machanavajjhala et al. [10] provide several other "instantiations" of l-diversity to guard against potential "background knowledge" from adversaries. However, as acknowledged in [10], it is impossible to compute a "perfect" l-diverse partition that denies privacy breach from all adversaries, without knowing their background knowledge in advance. Various instantiations apply additional heuristics to enhance the level of privacy protection. For simplicity, we focus on the instantiation of Definition 2 (termed "recursive $(\frac{1}{l-1}, 2)$-diversity" in [10]), but it is straightforward to extend the anatomy formulation to other instantiations.

3.2 Privacy Preservation

A pair of anatomized tables provide a convenient way for the data publisher to find out, for each tuple $t \in T$, all the A^s values that an adversary can associate t with, and the probability of each association. This is formally explained in the next lemma.

LEMMA 1. *If we perform a natural join QIT $\bowtie$ ST, the join result is a table with $d+3$ attributes, containing records of the form*

$$(t[1], t[2], ..., t[d], j, v, c_j(v))$$

where j is the ID of the QI-group including t (i.e., $t \in QI_j$), v an A^s value, and $c_j(v)$ the number of tuples in QI_j with A^s value v. Then, from an adversary's perspective,

$$Pr\{t[d+1] = v\} = c_j(v)/|QI_j| \quad (2)$$

where $|QI_j|$ denotes the size of QI_j.

PROOF. Consider any tuple $t \in T$, which is contained in QI-group QI_j (in the underlying l-diverse partition) for some $j \in [1, m]$. The adversary, who attempts to find out $t[d+1]$, can obtain j from the QIT which, however, does not have A^s data. Hence, the adversary can only conjecture that $t[d+1]$ equals one of the A^s values (pertinent to QI_j) summarized the ST. Without any other information, the adversary assumes that every tuple in QI_j has an equal chance to carry any A^s value relevant to QI_j, which leads to Equation 2. □

We explain the lemma using Table 4, which demonstrates part of the result of the natural join between Tables 3a and 3b (only the join results related to QI-group 1 are shown). QI-group 1 has 4 tuples. Hence, from the first record of Table 4, an adversary knows that

Age	Sex	Zipcode	Group ID	Disease	Count
23	M	11000	1	dyspepsia	2
23	M	11000	1	pneumonia	2
27	M	13000	1	dyspepsia	2
27	M	13000	1	pneumonia	2
35	M	59000	1	dyspepsia	2
35	M	59000	1	pneumonia	2
59	M	12000	1	dyspepsia	2
59	M	12000	1	pneumonia	2
...	...	...	...	...	...

Table 4: Partial result of the natural join between Tables 3a and 3b (only results pertinent to QI-group 1 are shown)

tuple 1 in the QIT (Table 3a) has probability 2 / 4 = 50% to carry dyspepsia in the microdata, according to Equation 2. Similarly, the second record implies that tuple 1 has 50% probability to be associated with pneumonia. On the other hand, the QI-values of tuple 1 are not combined with any other disease such as flu, meaning that tuple 1 cannot have flu as its real *Disease*-value.

COROLLARY 1. *Given a pair of QIT and ST, an adversary can correctly re-construct any tuple $t \in T$ with a probability at most $1/l$.*

PROOF. Tuple t is correctly re-constructed, if and only if the adversary precisely obtains its real A^s value v_{real}. By Equation 2, we know that $Pr\{t[d+1] = v_{real}\} = c_j(v_{real})/|QI_j|$, where QI_j is the unique QI-group containing t. Recall that a pair of anatomized tables is obtained from an l-diverse partition (Definition 2). Hence, by Equation 1, $c_j(v_{real})/|QI_j| \leq 1/l$. □

Corollary 1 gives the privacy protection guarantee at the *tuple level*. It is also necessary to discuss the corresponding guarantee at the *individual level*, since in practice multiple individuals may have the same QI-values, thus complicating the privacy-attack process performed by an adversary.

To explain this, consider that an adversary has the age 65 and zipcode 25000 of Alice (the "owner" of tuple 7 in Table 1), and wants to infer the medical record of Alice from the QIT and ST in Tables 3a and 3b, respectively. S/he consults the QIT, and sees that, in QI-group 2 (denoted as QI_2), both tuples 6 and 7 match the QI-values of Alice. Hence, s/he examines two scenarios.

First, assuming that tuple 6 belongs to Alice, the adversary uses Lemma 1 to derive the probability distribution for the tuple's disease value. According to Equation 2, tuple 6 has probability $c_2(\text{flu})/|QI_2| = 2/4 = 50\%$ to carry flu. Notice that, in the microdata, tuple 6 does not really belong to Alice. However, it does not matter — *the adversary may "happen to" use a wrong tuple to infer the correct sensitive value of Alice!* From tuple 6, the adversary actually has 50% probability to figure out that Alice contracted flu.

In the second scenario, the adversary assumes that tuple 7 belongs to Alice, through which (similar to tuple 6) s/he also has 50% probability to obtain the real disease of Alice. Finally, (without further knowledge) the adversary assumes that the two scenarios occur with the same likelihood $\frac{1}{2}$. Therefore, the overall breach probability should be calculated as $\frac{1}{2} \cdot 50\% + \frac{1}{2} \cdot 50\%$, where $\frac{1}{2}$ and 50% have the same semantics as in the above discussion.

In fact, Lemma 1 shows that tuple 7 (the real tuple of Alice) can be re-constructed with 50% likelihood. Namely, the breach probability at the individual level coincides with that at the tuple level. This happens because tuples 6 and 7 appear in the same QI-group. In general, as long as tuples with identical QI-values always end up in the same QI-group (as is true for global-recoding generalization reviewed in Section 2), the probabilities of the two levels are always equivalent. In this case, it suffices to discuss only the (simpler) tuple level; as a result, the individual level has not been addressed before (all the existing generalization schemes adopt global recoding).

Anatomy, however, allows high flexibility in forming QI-groups such that tuples with the same QI-values do not always belong to the same QI-group. Therefore, we must provide a formal result regarding the individual-level breach probability.

THEOREM 1. *Given a pair of QIT and ST, an adversary can correctly infer the sensitive value of any individual with probability at most $1/l$.*

PROOF. Consider any individual o whose QI-values are equivalent to those of totally f tuples t_1, t_2, ..., t_f in the microdata. Assume that tuple t_i ($1 \leq i \leq f$) belongs to QI-group QI_{j_i} ($1 \leq j_i \leq m$, where m is the total number of QI-groups). Let v_{real} be the real A^s value of o.

The adversary infers v_{real} in two steps. First, s/he guesses that each of t_1, ..., t_f belongs to o with probability $1/f$. Then, for each scenario where t_i ($1 \leq i \leq f$) belongs to o, by Lemma 1, s/he figures out that v_{real} is the A^s value of o with probability $c_{j_i}(v_{real})/|QI_{j_i}|$. Hence, the overall probability that the A^s value of o is inferred equals

$$\sum_{i=1}^{f} c_{j_i}(v_{real})/(f \cdot |QI_{j_i}|)$$

Recall that, by the property of l-diverse partition (Definition 2), $c_{j_i}(v_{real})/|QI_{j_i}| \leq 1/l$. Hence, the above formula is at most $\sum_{i=1}^{f}(\frac{1}{f} \cdot \frac{1}{l}) = 1/l$. □

3.3 Comparison with Generalization

We would like to emphasize that our intention is not to eliminate generalization; there is no doubt that generalization is an important technique, partly proved by the fact that it has received much attention in the literature. Instead, our goal is to present an alternative option for privacy preservation, which has its own advantages, since it can retain a larger amount of data characteristics (as shown in the subsequent sections). Indeed, anatomy is not an all-around winner. Intuitively, by releasing the QI-values directly, anatomy may allow a higher breach probability than generalization. Nevertheless, such probability is always bounded by $1/l$, as long as the background knowledge of an adversary is not stronger than the level allowed by the l-diversity model. Next, we will explain these observations in detail.

The derivation in Section 3.2 implicitly makes two assumptions:

- A1: the adversary has the QI-values of the target individual (i.e., Alice);

Name	Age	Sex	Zipcode
Ada	61	F	54000
Alice	65	F	25000
Bella	65	F	25000
Emily	*67*	*F*	*33000*
Stephanie	70	F	30000
...	...	...	...

Table 5: The voter registration list (publicly accessible)

- A2: the adversary also knows that the individual is definitely involved in the microdata.

In fact, usually both assumptions are satisfied in practical privacy-attacking processes. For example, in her pioneering paper [14], Sweeney shows how to reveal the medical record of the governor of Massachusetts from the data released by the Group Insurance Commission, after obtaining the governor's QI-values from public sources. The revelation is possible because Sweeney knew in advance that the record of the governor must be present in the microdata. Otherwise, no inference could be drawn against the governor because the "privacy-leaking" record could as well just belong to a person who happens to share the same QI-values as the governor.

In general, if both Assumptions A1 and A2 are true, anatomy provides as much privacy control as generalization, that is, the privacy of a person is breached with a probability at most $1/l$. For instance, if an adversary is sure that Alice has been hospitalized before, from Alice's QI-values, s/he can assert that Alice must be described by one of tuples 5-8 in the generalized Table 2. Then, s/he carries out the rest of her/his probabilistic conjecture (about the disease of Alice) in the same way as s/he would do after identifying Alice to be in Group 2 of the anatomized Table 3a.

Now, consider the case where A1 holds, but A2 does not. Accordingly, the overall breach probability of Alice has a Bayes form:

$$Pr_{A2}(\text{Alice}^{qi}) \cdot Pr_{breach}(\text{Alice}^{s}|A2) \quad (3)$$

where $Pr_{A2}(\text{Alice}^{qi})$ is the chance for Alice to be involved in the microdata, and $Pr_{breach}(\text{Alice}^{s}|A2)$ the likelihood for the adversary to correctly guess the disease of Alice on condition that Alice appears in the microdata. As analyzed earlier, anatomy and generalization give the same $Pr_{breach}(\text{Alice}^{s}|A2)$, which is simply the preach probability when both A1 and A2 are valid.

To compute $Pr_{A2}(\text{Alice}^{qi})$, an adversary typically needs to consult another external database [17], which relates QI-values to concrete personal identities for all the persons in the microdata, perhaps together with some other people. An example of such an external source is a voter registration list, partially demonstrated in Table 5, where the record of Emily is italicized to indicate that she is not involved in the microdata of Table 1. In this scenario, generalization and anatomy make a difference. Specifically, judging from (the QI-values of tuples 5-8 in) the generalized Table 2, the adversary sees that each person shown in Table 5 could be involved in the microdata with equal likelihood, and hence, calculates $Pr_{A2}(\text{Alice}^{qi})$ as 4/5. On the other hand, given the anatomized Table 3, the adversary concludes that $Pr_{A2}(\text{Alice}^{qi}) = 1$ (here s/he can figure out that Emily is definitely absent from the microdata). As a result, generalization provides a stronger overall privacy-preserving guarantee. Nevertheless, since anatomy ensures $Pr_{breach}(\text{Alice}^{s}|A2) \leq 1/l$, it also secures the same upper bound $1/l$ for Formula 3.

Although generalization has the above advantage over anatomy, *the advantage cannot be leveraged in computing the published data*. This is because the publisher cannot predict or control the external database to be utilized by an adversary, and therefore, must guard against an "accurate" external source that does not involve any person absent in the microdata. For instance, if Table 5 did not contain Emily, the voter list would produce $Pr_{A2}(\text{Alice}^{qi}) = 1$ in attacking the privacy of Alice from Table 2 (instead of 4/5 as discussed earlier). In other words, to ensure a maximum breach probability p using generalization, we must still set l to $\lceil 1/p \rceil$, i.e., same as in applying anatomy.

Finally, if neither assumption A1 nor A2 is satisfied, the breach probability of Alice becomes

$$\sum_{\forall x} Pr_{A1}(x) \cdot Pr_{A2}(x|A1) \cdot Pr_{breach}(\text{Alice}^{s}|A1, A2) \quad (4)$$

where x is a vector representing a possible set of QI-values of Alice, and $Pr_{A1}(x)$ equals the probability that x captures Alice's real QI-values, whereas Pr_{A2} and Pr_{breach} follow the same semantics as in Formula 3, but on condition that x is real. The comparison results between anatomy and generalization are analogous to those discussed for the previous case where A1 is true and A2 is not.

4. PRESERVING CORRELATION

A good publication method should preserve both privacy and data correlation (between QI- and sensitive attributes). Using a concrete query, we have shown in Section 1.1 that anatomy allows more effective aggregate analysis than generalization. Next, we provide the underlying theoretical rationale.

Obviously, for any tuple $t \in T$, every publication method will lose certain information of t (if not, it is equivalent to disclosing t directly, contradicting the goal of privacy). On the other hand, the method should permit development of an approximate modeling of t (otherwise, the published table is useless for research). Hence, the quality of correlation preservation depends on how accurate the re-constructed modeling is.

Intuition. Let us first examine the correlation between *Age* and *Disease* in the microdata of Table 1. The two attributes define a 2D space $DS_{A,D}$. Every tuple in the table can be mapped to a point in $DS_{A,D}$. For example, tuple 1, denoted as t_1, corresponds to point $(t_1[A], t_1[D])$, where $t_1[A]$ is the age 23 of t_1, and $t_1[D]$ its disease 'pneumonia'.

We can model t_1 using a probability density function (pdf) $\mathcal{G}_{t_1} : DS_{A,D} \rightarrow [0, 1]$. Specifically:

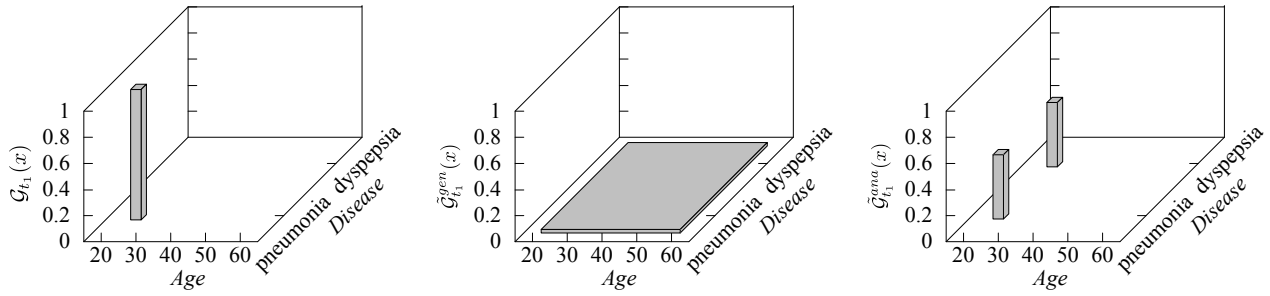
$$\mathcal{G}_{t_1}(x) = \begin{cases} 1 & \text{if } x = (t_1[A], t_1[D]) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where x is a 2D random variable in $DS_{A,D}$. Figure 2a demonstrates the pdf.

Assume that a researcher wants to re-construct an approximate pdf $\tilde{\mathcal{G}}^{gen}_{t_1}$ of t_1 from the generalized Table 2. From her/his perspective, $t_1[A]$ can be any value in the interval $[21, 60]$ with equality probability $1/40$, but $t_1[D]$ must be pneumonia. Hence,

$$\tilde{\mathcal{G}}^{gen}_{t_1}(x) = \begin{cases} 1/40 & \text{if } x[A] \in [21, 60] \text{ and} \\ & x[D] = \text{pneumonia} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

which is illustrated in Figure 2b.



(a) Original (b) Approximated from generalization (c) Approximated from anatomy

Figure 2: Original/re-constructed pdf of tuple 1 in Table 1

Instead, suppose that the researcher re-constructs a pdf $\tilde{\mathcal{G}}^{ana}_{t_1}$ from the QIT and ST in Tables 3a and 3b. This time, s/he knows that $t_1[A]$ must be 23 (since age is published directly), but $t_1[D]$ can be pneumonia or dyspepsia with 50% probability (the ST shows that half of the tuples in QI-group 1 are associated with these two diseases, respectively). Therefore,

$$\tilde{\mathcal{G}}^{ana}_{t_1}(x) = \begin{cases} 1/2 & \text{if } x = \text{(23, pneumonia) or} \\ & x = \text{(23, dyspepsia)} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

as shown in Figure 2c. Obviously, the pdf approximated from the anatomized tables is more accurate than that (Figure 2b) from the generalized table.

Towards a more rigorous comparison, given an approximate pdf $\tilde{\mathcal{G}}_{t_1}$ (Equation 6 or 7), a natural way of quantifying its approximation quality is to calculate its "L_2 distance" from the actual pdf $\mathcal{G}_{t_1}$ (Equation 5):

$$\sum_{x \in DS_{A,D}} \left(\tilde{\mathcal{G}}_{t_1}(x) - \mathcal{G}_{t_1}(x)\right)^2. \quad (8)$$

The distance of $\tilde{\mathcal{G}}^{ana}_{t_1}$ is 0.5, indeed significantly lower than the distance 22.5 of $\tilde{\mathcal{G}}^{gen}_{t_1}$.

Although we focused on t_1, in the same way, it is easy to verify that the anatomized tables permit better re-construction of the pdfs of all tuples in Table 1.

General Results and Quality Metric. As defined in Section 3, each tuple t in the microdata T can be regarded as a point in a $(d+1)$-dimensional space DS (including all the QI- and sensitive dimensions). Next, we generalize the above discussion to DS.

We model t as a pdf $\mathcal{G}_t(x) : DS \rightarrow [0, 1]$:

$$\mathcal{G}_t(x) = \begin{cases} 1 & \text{if } x = t \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where x is a random variable in DS. Note that the condition $x = t$ implies $x[i] = t[i]$ for all $i \in [1, d+1]$, where $x[i]$ and $t[i]$ are the i-th coordinates of x and t, respectively.

In a generalized table, let t belong to a QI-group QI. As stated in Definition 4, the generalized form of t is $(QI[1], QI[2], ..., QI[d], t[d+1])$, where $QI[i]$ $(1 \leq i \leq d)$ is an interval enclosing $t[i]$. Denote the length of $QI[i]$ as $L(QI[i])$ (if A^{qi}_i is discrete, $L(QI[i])$ should be interpreted as the number of different values in $QI[i]$). Then, the reconstructed pdf $\tilde{\mathcal{G}}^{gen}_t(x)$ of t is

$$\tilde{\mathcal{G}}^{gen}_t(x) = \begin{cases} \frac{1}{\prod_{i=1}^{d} L(QI[i])} & \text{if } x[i] \in QI[i] \ \forall i \in [1,d] \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Next we discuss anatomized tables. Also assume QI as the QI-group containing t (in the underlying l-diverse partition). Let v_1, v_2, ..., v_λ be all the distinct A^s values in QI (e.g., for QI-group 1 in Table 3a, $\lambda = 2$, whereas for QI-group 2, $\lambda = 3$). Denote $c(v_h)$ $(1 \leq h \leq \lambda)$ as the *Count* value in the ST corresponding to v_h. The reconstructed pdf $\tilde{\mathcal{G}}^{ana}_t(x)$ of t is

$$\tilde{\mathcal{G}}^{ana}_t(x) = \begin{cases} c(v_1)/|QI| & \text{if } x = (t[1], ..., t[d], v_1) \\ ... & ... \\ c(v_\lambda)/|QI| & \text{if } x = (t[1], ..., t[d], v_\lambda) \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where $|QI|$ is the number of tuples in QI, and the QI-values $t[1]$, ..., $t[d]$ of t are directly released in the QIT.

Notice that $\tilde{\mathcal{G}}^{ana}_t(x)$ is greater than 0, only when x lies at one of the λ points in DS, as described in the if-conditions of Equation 11. That is, $\tilde{\mathcal{G}}^{ana}_t(x)$ consists of λ "spikes" at these points ($\lambda = 2$ in Figure 2c). On the other hand, in practice, $\tilde{\mathcal{G}}^{gen}_t(x)$ typically takes a small value when x distributes across a large region. Namely, the occurrence probability of t is "smeared" onto all the points in that region (see Figure 2b), thus deviating significantly from the actual $\mathcal{G}_t(x)$.

Given an approximate pdf $\tilde{\mathcal{G}}_t$ (Equation 10 or 11), we quantify its error from the actual $\mathcal{G}_t$ (Equation 9) as

$$Err_t = \int_{x \in DS} \left(\tilde{\mathcal{G}}_t(x) - \mathcal{G}_t(x)\right)^2 \mathbf{d}x. \quad (12)$$

Naturally, taking into account all tuples $t \in T$, a good publication method should minimize the following *re-construction error* (RCE):

$$RCE = \sum_{\forall t \in T} Err_t. \quad (13)$$

5. A NEARLY-OPTIMAL ANATOMIZING ALGORITHM

We propose an efficient algorithm for computing anatomized tables that (almost) minimize the RCE (Equation 13). In particular, the RCE of the resulting QIT and ST achieve an RCE that deviates from the theoretical lower bound by only a factor less than $1 + 1/n$ where n is the size of T. Furthermore, our algorithm has linear I/O complexity $O(n/b)$, where b denotes the page size.

5.1 Lower Bound of Reconstruction Error

The following theorem establishes the lower bound of the RCE achievable by any anatomized tables.

THEOREM 2. *RCE (Equation 13) is at least $n(1-1/l)$, for any pair of QIT and ST, where n is the cardinality of the microdata T.*

PROOF. Anatomized tables (Definition 3) are computed from an l-diverse partition. Let the partition contain QI groups QI_1, ..., QI_m. For each $j \in [1, m]$, use α_j to denote the average Err_t (Formula 12) for all tuples $t \in QI_j$. Thus, RCE can be rewritten as

$$RCE = \sum_{j=1}^{m} (|QI_j| \cdot \alpha_j).$$

The rest of the proof will show that $\alpha_j \geq 1-1/l$, for all $j \in [1, m]$. As a result, the above equation leads to

$$RCE \geq \sum_{j=1}^{m} (|QI_j| \cdot (1-1/l)) = n(1-1/l),$$

thus completing the proof (notice $\sum_{j=1}^{m} |QI_j| = n$).

By symmetry, it suffices to prove $\alpha_j \geq 1-1/l$ for any QI_j. Hence, we omit the subscript j in the sequel. Without loss of generality, assume that QI contains λ distinct A^s values v_1, ..., v_λ. In particular, there are $c(v_h)$ ($1 \leq h \leq \lambda$) tuples in QI with A^s value v_h.

Consider an arbitrary tuple $t \in QI$ with A^s value v_h (for some $h \in [1, \lambda]$). The actual pdf $\mathcal{G}_t$ and approximate $\tilde{\mathcal{G}}_t^{DZ}$ are given in Equations 9 and 11, respectively. Thus, by Equation 12, we have

$$Err_t = \left(1 - \frac{c(v_h)}{|QI|}\right)^2 + \sum_{h'=1 \wedge h' \neq h}^{\lambda} \frac{c(v_{h'})^2}{|QI|^2}.$$

For computing the average α of Err_t for all $t \in QI$, we combine the above formula with the fact that $c(v_h)$ tuples have A^s value v_h:

$$\alpha = \frac{\sum_{h=1}^{\lambda} c(v_h) \cdot \left(\left(1 - \frac{c(v_h)}{|QI|}\right)^2 + \sum_{\substack{h'=1 \\ h' \neq h}}^{\lambda} \frac{c(v_{h'})^2}{|QI|^2}\right)}{|QI|}.$$

Thus, it remains to solve the minimum α subject to the constraints

$$\sum_{h=1}^{\lambda} c(v_h) = |QI|, \text{ and } c(v_h) \leq \frac{|QI|}{l} \text{ for all } h \in [1, \lambda]$$

(the second constraint is due to Definition 2).

Let us ignore the second constraint temporarily. Then, minimization of α subject to the first constraint is a standard problem tackled by the *Lagrange multiplier method* [3]. Application of the method results in $\alpha \geq (1-1/\lambda)$, where the equality holds only when $c(v_1) = ... = c(v_h) = |QI|/l$.

Now, we take into account the second constraint, which leads to $\sum_{h=1}^{\lambda} c(v_h) \leq \lambda \cdot |QI|/l$. The left side of the inequality equals $|QI|$. Hence, the inequality indicates that $\lambda \geq l$.

Therefore, $\alpha \geq (1-1/\lambda) \geq (1-1/l)$, where the equality holds when $c(v_1) = ... = c(v_h) = |QI|/l$, and $\lambda = l$. □

5.2 The Algorithm

Figure 3 presents the algorithm *Anatomize* which, given a microdata table T and a parameter l, obtains a pair of QIT and ST for publication. *Anatomize* first computes an l-diverse partition of T (Lines 1-12), and then, produces the QIT and ST (Lines 13-18) from the partition. Since populating the QIT and ST is already clarified in Definition 3, we concentrate on finding the partition.

Anatomize starts (Line 1) by initiating an empty QIT and ST, and variable $gcnt$, which counts the number of QI-groups created. Then, it hashes the tuples of T into buckets by A^s, so that each bucket includes the tuples with the same A^s value (Line 2). The subsequent execution involves a *group-creation* step, followed by a *residue-assignment* phase.

Group-Creation. This step is performed in iterations, and continues as long as there are at least l non-empty buckets (Line 3). Each iteration yields a new QI-group QI_{gcnt} (Line 4) as follows. First, *Anatomize* obtains a set S consisting of the l hash buckets that *currently* have the largest number of tuples (Line 5). Note that the content of S may vary in different iterations. Then, from each bucket in S (Line 6), a random tuple is selected (Line 7), and added to QI_{gcnt} (Line 8). Therefore, QI_{gcnt} contains l tuples with distinct A^s values.

PROPERTY 1. *At the end of the group-creation phase, each non-empty bucket has only one tuple.*

PROOF. An l-diverse partition exists, if and only if T satisfies an *eligibility condition*[3] [10]: at most n/l tuples are associated with the same A^s value, where n is the cardinality of T. We will prove that, Property 1 always holds under this condition.

Assume, on the contrary, after the first (group-creation) phase, a set of *bad buckets* have sizes at least 2. Obviously, there are at most $l-1$ bad buckets (otherwise, the group-generation phase could not have terminated). Since each iteration moves l tuples from buckets into a QI-group, the first phase executes $\lfloor n/l \rfloor$ iterations, denoted as I_1, I_2, ..., $I_{\lfloor n/l \rfloor}$, respectively.

Before iteration $I_{\lfloor n/l \rfloor}$ starts, at most $l-1$ buckets (termed *sizable $\lfloor n/l \rfloor$-buckets*) have sizes at least 2 (otherwise, there would be at least l non-empty buckets after $I_{\lfloor n/l \rfloor}$, contradicting the fact that $I_{\lfloor n/l \rfloor}$ is the last iteration). On the other hand, we already know that, *after* $I_{\lfloor n/l \rfloor}$, all the bad buckets have sizes at least 2. Hence, every bad bucket is a sizable $\lfloor n/l \rfloor$-bucket, and must belong to S (retrieved at Line 5) in $I_{\lfloor n/l \rfloor}$. Thus, each bad bucket loses a tuple in $I_{\lfloor n/l \rfloor}$, meaning that, *before* $I_{\lfloor n/l \rfloor}$, the bucket has size at least 3.

Similarly, before $I_{\lfloor n/l \rfloor - 1}$, at most $l-1$ buckets (termed *sizable $(\lfloor n/l \rfloor - 1)$-buckets*) have sizes at least 3 (otherwise, there would be at least l sizable $\lfloor n/l \rfloor$-buckets, contradicting our earlier analysis). On the other hand, we already know that, *after* $I_{\lfloor n/l \rfloor} - 1$, all the bad buckets have sizes at least 3. Hence, every bad bucket is a sizable $(\lfloor n/l \rfloor - 1)$-bucket, and must belong to S in $I_{\lfloor n/l \rfloor - 1}$. Thus, each bad bucket loses a tuple in $I_{\lfloor n/l \rfloor - 1}$, meaning that, *before* $I_{\lfloor n/l \rfloor - 1}$, the bucket has size at least 4.

[3] If this condition is violated, neither k-anonymity nor l-diversity can prevent an adversary from correctly inferring a tuple in T with a probability at least $1/l$.

Algorithm Anatomize (T, l)
1. QIT = $\emptyset$; ST = $\emptyset$; $gcnt$ = 0
2. hash the tuples in T by their A^s values (each bucket per A^s value)

/ Lines 3-8 are the group-creation step */*

3. while there are at least l non-empty hash buckets

   / Lines 4-8 form a new QI-group */*

4. $gcnt = gcnt + 1$; $QI_{gcnt} = \emptyset$
5. S = the set of l largest buckets
6. for each bucket in S
7. remove an arbitrary tuple t from the bucket
8. $QI_{gcnt} = QI_{gcnt} \cup \{t\}$

/ Lines 9-12 are the residue-assignment step */*

9. for each non-empty bucket

   / this bucket has only one tuple; see Property 1 */*

10. t = the only residue tuple of the bucket
11. S' = the set of QI-groups that do not contain the A^s value $t[d+1]$

    / S' has at least one QI-group; see Property 2 */*

12. assign t to a random QI-group in S'

/ Lines 13-18 populate QIT and ST */*

13. for j = 1 to $gcnt$
14. for each tuple $t \in QI_j$
15. insert tuple $(t[1], ..., t[d], j)$ into QIT
16. for each distinct A^s value v in QI_j
17. $c_j(v)$ = the number of tuples in QI_j with A^s value v
18. insert record $(j, v, c_j(v))$ into ST
19. return QIT and ST

Figure 3: The anatomizing algorithm

Carrying out the same discussion to the other iterations, we arrive at a fact that each bucket in S_{bad} has size at least $\lfloor n/l \rfloor + 1$ at the beginning of *Anatomize*. The fact violates the eligibility condition, because $\lfloor n/l \rfloor + 1 > n/l$. □

We use the term *residue tuple* to refer to a tuple remaining in a bucket, at the end of the group-creation phase. Clearly, there are at most $l - 1$ such tuples.

Residue-Assignment. For each residue tuple t, *Anatomize* collects a set S' of QI-groups (produced from the previous step), where no tuple has the same A^s value as t (Lines 8-11). Interestingly, as proved shortly, S' includes at least one QI-group. Then, at Line 12, t is assigned to an arbitrary group in S'.

PROPERTY 2. *The set S' (computed at Line 11 of Figure 3) always includes at least one QI-group.*

PROOF. Assume, on the contrary, that S' is empty when processing tuple t (at Line 11). As explained in the previous proof, the number of QI-groups is $\lfloor n/l \rfloor$. Since S' is empty, each QI-group has at least a tuple whose A^s value equals $t[d+1]$. It follows that the number of tuples in T with A^s value $t[d+1]$ is at least $1 + \lfloor n/l \rfloor$, which is larger than n/l. This contradicts the eligibility condition mentioned in the proof of Property 1. □

Correctness. Since Lines 13-19 of Figure 3 essentially implement Definition 3, *Anatomize* is correct, if and only if Lines 1-12 produce an l-diverse partition of T. We establish this in the following property, which actually shows a stronger fact.

PROPERTY 3. *After the residue-assignment phase, each QI-group has at least l tuples. Furthermore, all tuples in each QI-group have distinct A^s values.*

PROOF. After the group-creation step, every QI-group has l tuples with distinct A^s values (these tuples are obtained from different hash buckets). In the residue-assignment phase, the assignment of a tuple into a QI-group ensures that all tuples in the group still have distinct A^s values. Hence, Property 3 is correct. □

5.3 Analysis

In this section, we analyze the efficiency and effectiveness of *Anatomize* (Figure 3). First, Theorem 3 provides the space and time complexities of *Anatomize*. In particular, the proof of the theorem describes an efficient way to implement the algorithm. Then, Theorem 4 explains the quality of the resulting QIT and ST.

THEOREM 3. Anatomize *requires $O(\lambda)$ memory, and $O(n/b)$ I/Os, where λ is the number of distinct A^s values in T, n is the cardinality of T, and b is the disk page size.*

PROOF. The hashing at Line 1 of Figure 3 consumes $O(\lambda)$ memory, and performs $O(n/b)$ I/Os.

During the first phase, we can keep in memory an array with λ elements, where the i-th ($1 \leq i \leq \lambda$) element maintains the size of the i-th bucket. Therefore, at Line 5, set S can be decided with no I/O overhead. To implement Line 7, for each bucket, we allocate a buffer page for reading its content. All the QI-groups are sequentially into a *QI-group file*, in the order they are created. For this purpose, we allocate an output buffer page. In this way, the group-creation step requires $O(\lambda)$ memory and $O(n/b)$ I/Os.

At the beginning of the residue-assignment phase, we read all the (at most $l - 1$) residue residue tuples into memory. Next, we perform a single scan of the QI-group file, and assign these tuples to appropriate QI-groups during the scan. This step needs $O(l)$ memory ($l \leq \lambda$, for satisfying the eligibility condition in the proof of Property 1), and performs $O(n/b)$ I/Os.

Each QI-group so far has $O(l)$ tuples. Thus, populating the QIT and ST (Lines 13-18) can be easily achieved with $O(l)$ memory, and $O(n/b)$ I/Os. Therefore, the overall space and I/O complexities of *Anatomize* are $O(\lambda)$ and $O(n/b)$, respectively. □

THEOREM 4. *If the cardinality n of T is a multiple of l, the QIT and ST computed by* Anatomize *achieve the lower bound of RCE in Theorem 2. Otherwise, the RCE of the anatomized tables is higher than the lower bound by a factor at most $1 + \frac{1}{n}$.*

PROOF. Let $r = n \mod l$. Depending on whether n is a multiple of l, there are two cases.

Case 1 ($r = 0$): *Anatomize* terminates directly after the group-creation phase. Each QI-group has exactly l tuples with distinct A^s values. Combining Equations 9, 11, and 12, we have, for each tuple $t \in T$,

$$Err_t = \left(1 - \frac{1}{l}\right)^2 + \frac{l-1}{l^2} = 1 - \frac{1}{l}.$$

By Equation 13, $RCE = n(1 - \frac{1}{l})$.

Case 2 ($r \neq 0$): Consider the moment when the group-creation phase finishes. So far, totally $n - r$ (a multiple of l) tuples have

been added into QI-groups. According to the analysis of Case 1, the current RCE (with respect to the tuples already in QI-groups) is $(n-r)(1-\frac{1}{l})$.

Next, we show that, after assigning a residue tuple t at Line 12 of Figure 3, the overall RCE increases by 1. With out loss of generality, assume that t is assigned to a QI-group QI with β tuples, all of which have distinct A^s values, and their A^s values are different from that of t (see Property 3). Before the assignment, following the derivation of Case 1, the RCE of QI[4] equals $\beta(1-\frac{1}{\beta})$. After the assignment, the RCE of QI becomes $(\beta+1)(1-\frac{1}{\beta+1})$, so that the overall RCE (of all the tuples in QI-groups) increases by

$$(\beta+1)\left(1-\frac{1}{\beta+1}\right)-\beta\left(1-\frac{1}{\beta}\right)=1.$$

As mentioned earlier, before the assignment step starts, the overall RCE equals $(n-r)(1-\frac{1}{l})$. Therefore, after assigning all r residue tuples, the RCE becomes

$$(n-r)\left(1-\frac{1}{l}\right)+r=n\left(1-\frac{1}{l}\right)\left(1+\frac{r}{n(l-1)}\right).$$

which is greater than the lower bound $n(1-\frac{1}{l})$ by a factor of $1+\frac{r}{n(l-1)}$. Given that $r \leq l-1$, we complete the proof. □

Note that, for a large T, $1+\frac{1}{n} \approx 1$, namely, the RCE of the tables output by *Anatomize* is extremely close to the lower bound.

6. EXPERIMENTS

This section experimentally evaluates the effectiveness and efficiency of anatomy. For this purpose, we utilize a real dataset CENSUS[5] containing personal information of 500k American adults. The dataset has 9 discrete attributes as summarized in Table 6.

From CENSUS, we create two sets of microdata tables, in order to examine the influence of dimensionality and sensitive-value distribution. The first set has 5 tables, denoted as OCC-3, ..., OCC-7, respectively. Specifically, OCC-d ($3 \leq d \leq 7$) treats the first d attributes in Table 6 as the QI-attributes, and *Occupation* as the sensitive attribute A^s. For example, OCC-3 is 4D, and contains QI-attributes *Age*, *Gender*, and *Education*. The second set also has 5 tables SAL-3, ..., SAL-7, where SAL-d ($3 \leq d \leq 7$) has the same QI-attributes as OCC-d, but includes *Salary-class* as the A^s.

To study the impact of cardinality, we generate datasets with various cardinalities n, by randomly sampling n tuples from the "full" OCC-d or SCC-d ($3 \leq d \leq 7$) with 500k tuples.

We compare anatomy against (l-diverse) generalization on two aspects: (i) usefulness of the resulting publishable tables for data analysis, and (ii) cost of computing these tables. For generalization, we employ the state-of-the-art algorithm in [9], which adopts multi-dimension recoding (explained in Section 2). The value of l is fixed to 10, i.e., the sensitive value of each individual can be correctly inferred by an adversary with at most 10% probability.

As stated in Definition 4, each generalized value is an interval. The last column of Table 6 describes the details of generalization on each QI-attribute. Specifically, "free interval" means that the end points of a generalized interval can fall on any value in the domain of the corresponding attribute. "Taxonomy tree (x)", on the other hand, indicates that the end points must lie on particular values, conforming to a taxonomy with height x (see [8] for more details of generalization based on a taxonomy).

[4]The RCE of QI equals the sum of Err_t of all tuples $t \in QI$.

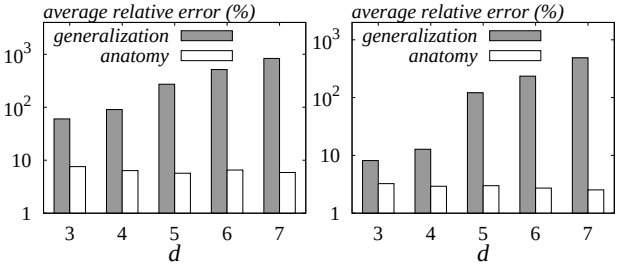
[5]Downloadable at *http://www.ipums.org*.

Attribute	Number of distinct values	Generalization method (inapplicable to anatomy)
Age	78	Free interval
Gender	2	Taxonomy tree (2)
Education	17	Free interval
Marital	6	Taxonomy tree (3)
Race	9	Taxonomy tree (2)
Work-class	10	Taxonomy tree (4)
Country	83	Taxonomy tree (3)
Occupation	50	NA (sensitive)
Salary-class	50	NA (sensitive)

Table 6: Summary of attributes

Parameter	Values
l	**10**
cardinality n	100k, 200k, **300k**, 400k, 500k
number of QI-attributes d	3, 4, **5**, 6, 7
query dimensionality qd	1, 2, ..., $\boldsymbol{d}$
expected selectivity s	1%, ..., **5%**, ..., 10%

Table 7: Parameters and tested values



(a) OCC-d (b) SAL-d

Figure 4: Query accuracy vs. the number d of QI-attributes

6.1 Effectiveness for Aggregate Reasoning

We consider queries of the form:

SELECT COUNT(*) FROM Unknown-Microdata
WHERE $pred(A_1^{qi})$ AND ... AND $pred(A_{qd}^{qi})$ AND $pred(A^s)$

Specifically, a query involves qd random QI-attributes A_1^{qi}, ..., A_{qd}^{qi} (in the underlying microdata), and the sensitive attribute A^s, where qd is a parameter called *query dimensionality*. For instance, if the microdata is OCC-3 and qd = 2, then $\{A_1^{qi}, A_2^{qi}\}$ is a random 2-sized subset of {*Age*, *Gender*, *Education*}. For any attribute A, the predicate $pred(A)$ has the form

$$(A = x_1 \text{ OR } A = x_2 \text{ OR ... OR } A = x_b)$$

where $x_i (1 \leq i \leq b)$ is a random value in the domain of A (recall that all attributes are discrete). The value of b depends on the *expected query selectivity* s:

$$b = \left\lceil |A| \cdot s^{1/(qd+1)} \right\rceil \quad (14)$$

where $|A|$ is the domain size of A. A higher s leads to more selection conditions in $pred(A)$.

Table 7 summarizes the parameters of our experiments, as well as their values examined. The values in bold are the defaults. Unless

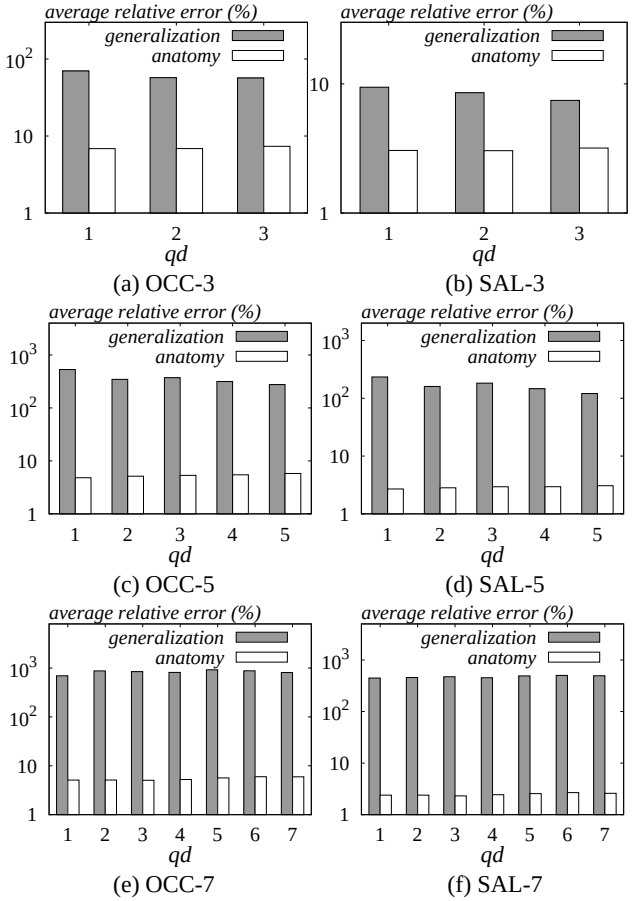


Figure 5: Query accuracy vs. query dimensionality qd

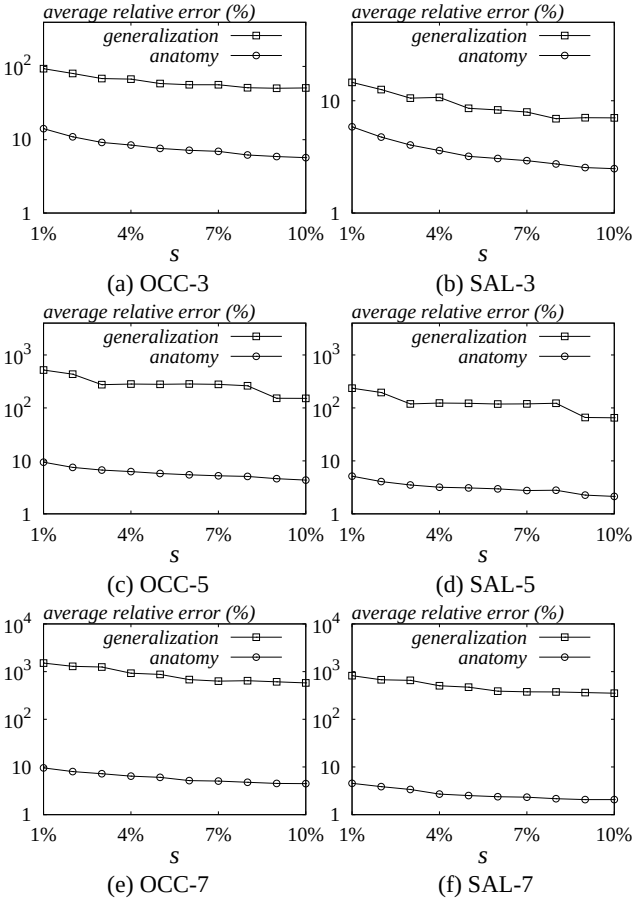


Figure 6: Query accuracy vs. selectivity s

specifically stated, each parameter is set to its default value in the following experiments.

Given a microdata relation, we compute the corresponding anatomized and generalized tables. Then, we process a *workload* of 10000 queries (with the same qd and s) on the resulting tables, using the algorithms explained in Sections 1.1 (for generalized tables) and 1.2 (for anatomized tables), respectively. The effectiveness of anatomy/generalization is measured as its *average relative error* in answering a query. Specifically, for each query, its relative error equals $|act - est|/act$, where act is its actual result derived from the microdata, and est the estimate computed from the anatomized/generalized table.

The first set of experiments investigates the effect of d on query accuracy. Figure 4a (4b) plots the error of anatomy and generalization as a function of d, for dataset OCC-d (SAL-d). As expected, anatomy permits significantly more accurate aggregate analysis, since it captures a larger amount of correlation in the microdata than generalization, as discussed in Section 4. Furthermore, the effectiveness of anatomy is not affected by d (its error is always below 10%), whereas the error of generalization grows exponentially with d. In particular, for d = 7, the error of anatomy is lower by two orders of magnitude.

Next, we concentrate on 3 values of d = 3, 5, and 7. For each d, we measure the accuracy of anatomy and generalization using workloads of different query dimensionalities qd. Figures 5a and 5b illustrate the results for OCC-3 and SAL-3 (i.e., d = 3), respectively. Interestingly, the error of generalization decreases as qd grows higher. To explain this, recall that all queries have the same (expected) selectivity s = 5%. Hence, when qd becomes larger, the number b (Equation 14) of values queried on each attribute increases considerably, leading to a more sizable search region, which in turn reduces error.

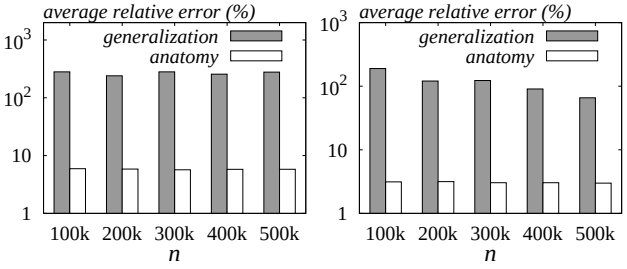
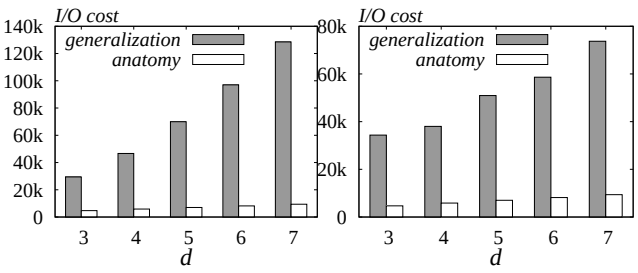


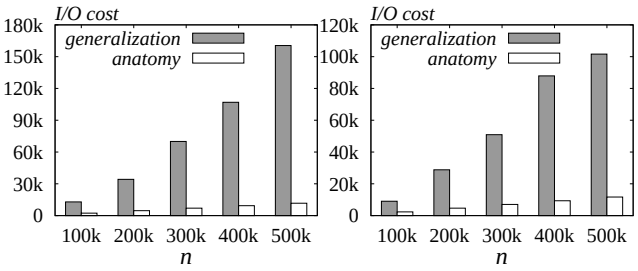
Figure 7: Accuracy vs. dataset cardinality n

Figures 5c, 5d repeat the above experiments on OCC-5 and SAL-5 respectively, validating similar observations. Figures 5e and 5f demonstrate the results on the microdata with $d = 7$. Notice that, here the effectiveness of generation no longer improves with qd, which indicates that all the generalized values have become exceedingly-wide intervals under $d = 7$. As a result, the generalized tables are useless for analysis. In contrast, regardless of d and qd, anatomy is consistently more accurate than generalization by at least an order of magnitude.



(a) OCC-d (b) SAL-d

Figure 8: I/O cost vs. the number d of QI-attributes



(a) OCC-5 (b) SAL-5

Figure 9: I/O cost vs. dataset cardinality n

To study the impact of query selectivity s, we again examine the microdata with d = 3, 5, and 7. Figures 6a-6f present the error of both techniques as a function of s, for the 6 microdata tables used in Figure 5, respectively. The precision of both anatomy and generalization improves as s increases, with anatomy being the clear winner. Finally, Figure 7 examines how the accuracy of each method scales with the dataset cardinality. Again, Anatomy achieves significantly lower error in all cases.

In summary, we showed that anatomy allows very accurate aggregate analysis. Its error is usually smaller than that of generalization by an order of magnitude. Furthermore, the effectiveness of anatomy is not affected by the dimensionalities of datasets and queries.

6.2 Computation Overhead

In the sequel, we compare anatomy against generalization on the I/O cost of computing publishable tables, with the page size set to 4096 bytes, and a memory capacity of 50 pages. Figure 8 presents the comparison results as d varies from 3 to 7. Evidently, anatomy incurs significantly fewer I/Os. Figure 9 plots the I/O overhead as a function of n. As predicted by Theorem 3, the cost of anatomy scales linearly with n, as opposed to the super-linear behavior of generalization. For large d or n, anatomy is 10 times faster than generalization.

7. CONCLUSIONS

Although generalization is a common methodology for protecting privacy, it loses considerable information in the microdata, and thus, prohibits effective data analysis. This paper developed anatomy, an innovative technique which preserves both privacy and correlation in the microdata, and hence, overcomes the drawbacks of generalization. Extensive experiments confirm that anatomy permits researchers to derive, from the published tables, highly accurate aggregate information about the unknown microdata, with an average error below 10% (as opposed to over 100% error of generalization).

As another important fact, anatomized tables can be computed in I/O cost linear to the database cardinality. In particular, these tables have nearly optimal quality guarantees in correlation preserving. Furthermore, despite its rigorous theoretical justification, our anatomizing algorithm is simple, and can be easily implemented in an existing database system.

This work also initiates several directions for future investigation. For example, in this paper, we focused on the case where there is a single sensitive attribute. Extending our technique to multiple sensitive attributes is an interesting topic. As another direction, it would be highly useful to study how anatomized tables can be utilized for effective mining of interesting patterns in the microdata, perhaps through minimization of other metrics of measuring information loss (e.g., KL-divergence [7] and discernibility [4, 9]).

Acknowledgements

This work was done when the authors were with the City University of Hong Kong, and supported by Grant CityU 1163/04E from the Research Grant Council of the HKSAR government. We would like to thank the anonymous reviewers for their insightful comments.

REFERENCES

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

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

[3] G. Arfken and H. Weber. *Mathematical Methods for Physicists*. Academic Press, 1995.

[4] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.

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

[6] V. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD*, pages 279–288, 2002.

[7] D. Kifer and J. E. Gehrke. Injecting utility into anonymized datasets. *To appear in SIGMOD 2006*.

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

[9] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, 2006.

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

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

[12] P. Samarati. Protecting respondents' identities in microdata release. *TKDE*, 13(6):1010–1027, 2001.

[13] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, page 188, 1998.

[14] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowlege-Based Systems*, 10(5):557–570, 2002.

[15] N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. In *SIGMOD*, pages 428–439, 2002.

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

[17] X. Xiao and Y. Tao. Personalized privacy preservation. *To appear in SIGMOD*, 2006.

[18] C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. In *VLDB*.