

Anatomy: Privacy and Correlation Preserving Publication

Xiaokui Xiao

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

Yufei Tao

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

Abstract

This article presents the *anatomy* technique for anonymized publication of sensitive data. Anatomy releases all the quasi-identifier and sensitive values directly in two separate tables. Combined with a grouping mechanism, this approach effectively protects privacy, and captures a large amount of correlation in the microdata. We propose an efficient algorithm for computing anatomized tables that fulfill the l -diversity anonymity requirement, and minimize the error of reconstructing the microdata, according to any L_p norm, the KL-divergence, and the discernability metrics. The algorithm is accompanied by optional heuristics that continuously enhance the data utility of anatomy, until a user-specified time limit has been reached. We also provide detailed explanations about how to leverage anatomized tables to understand the characteristics of the microdata. Extensive experiments confirm that anatomy allows significantly more accurate data analysis than conventional anonymization methods based on generalization and data swapping.

The short version of this article appeared in VLDB 06. The current submission improves our preliminary work by (i) including a thorough discussion of the previous methods, (ii) extending the analysis of anatomy to several other metrics of information loss (i.e., generic L_p norm, KL-divergence, and discernability), (iii) elaborating how to deploy the anonymized data for statistical studies, (iv) presenting a new algorithm for computing anatomized tables, and (v) featuring a more comprehensive experimental evaluation.

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 a *microtable* (whose tuples are 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. To illustrate this, suppose that the table is released directly; consider an adversary who has the personal details (i.e., age 21 and zipcode 10001) 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 dyspepsia. This process is called a *linking attack* [35, 36].

To protect privacy, *generalization* [35, 36] 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 21 and zipcode 10001 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 5-7, 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.

A generalized table is deemed "adequately anonymized", if it satisfies a certain *anonymization principle*. *k-anonymity* and *l-diversity* are the two principles that have received most attention in the literature. Specifically, a table is *k-anonymous* [35, 36] if each QI-group involves at least k tuples (e.g., Table 2 is 3-anonymous). However, even with a large k , *k-anonymity* may still allow an adversary to infer the sensitive value of an individual with high confidence [25]. *l-diversity* overcomes this problem by demanding that, in each QI-group, at most $1/l$ of the tuples possess an identical sensitive value. For instance, Table 2 is 3-diverse because, in each QI-group, at most a third of the tuples have the same disease. Consider again the adversary targeting Bob's privacy. S/he can make only a probabilistic conjecture: Bob could have contracted one of the four diseases in the first QI-group with the same probability.

1.1 Defects of Generalization in Data Analysis

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

tuple ID	Age	Sex	Zipcode	Disease
1 (Bob)	21	M	10001	dyspepsia
2	27	M	13000	pneumonia
3	35	M	60000	flu
4	60	M	12000	gastritis
5	61	M	10001	bronchitis
6	70	F	60000	pneumonia
7 (Alice)	70	F	60000	gastritis

Table 1: The microdata

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

Table 2: A 3-diverse table

A: `SELECT COUNT(*) FROM Unknown-Microdata`

`WHERE Disease is a stomach disease AND Age <= 30 AND Zipcode ∈ [10k, 20k].`

By observing Table 2, the researcher learns that no tuple in the second QI-group satisfies the query, since their ages fall in [61, 70], and hence, cannot satisfy the predicate $Age \leq 30$. On the other hand, in the first QI-group, tuples 1 and 4 have stomach diseases, but their QI values (i.e., age [21, 60] and zipcode [10001, 60000]) may or may not fulfill the query conditions on the QI attributes. Thus, the researcher needs to derive the probability p that each of tuples 1 and 4 qualifies the predicates on the QI attributes, after which the query answer can be estimated as $2p$.

Without additional knowledge on the QI values of tuples 1 and 4, the researcher assumes that their *Age* and *Zipcode* are uniformly distributed in the intervals [21, 60] and [10001, 60000], respectively. Within the intervals, there are 40 (50000) distinct *Age* (*Zipcode*) values, leading to totally $40 \times 50000 = 2 \times 10^6$ combinations of age and zipcode. 10^5 of these combinations satisfy the predicates $Age \leq 30$ and $Zipcode \in [10k, 60k]$. Consequently, the probability p is computed as $\frac{10^5}{2 \times 10^6} = 0.05$, resulting in an estimated answer $2p = 0.1$ for query A. This answer, however, is ten times smaller than the actual query result 1 (see Table 1).

The gross estimation error is caused by the fact that, the age and zipcode distributions in the first QI-group significantly deviate from uniformity. Nevertheless, given only the generalized table, the researcher 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 the *anatomy*¹ technique to achieve both privacy- and correlation-preserving publication. Given a microtable T , anatomy first divides the tuples in T into several QI-groups, and assigns a unique ID to each group. After that, anatomy creates a *quasi-identifier table (QIT)*

¹The name is chosen, since our technique “anatomizes” the microdata into two tables for publication.

tuple ID	Age	Sex	Zipcode	Group-ID
1	21	M	10001	1
2	27	M	13000	1
3	35	M	60000	1
4	60	M	12000	1
5	61	M	10001	2
6	70	F	60000	2
7	70	F	60000	2

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

(a) The quasi-identifier table (QIT) (b) The sensitive table (ST)

Table 3: The anatomized tables

that contains all QI attributes in T , as well as an attribute *Group-ID*. Each tuple t in T gives rise to a tuple t^{qi} in the QIT, such that (i) t^{qi} and t have identical QI values, and (ii) the *Group-ID* of t^{qi} equals the ID of the QI-group that includes t . Finally, anatomy generates a *sensitive table (ST)*, which contains the sensitive attribute of T , along with two attributes *Group-ID* and *Count*. For each QI-group QI , and for each sensitive value v appearing in QI , anatomy inserts in the ST a tuple t^s , such that (i) the sensitive value of t^s is v , (ii) the *Group-ID* of t^s equals the ID of QI , and (iii) the *Count* of t^s equals the number of tuples in QI that have sensitive value v .

For example, assume that we divide the tuples in Table 1 into two QI-groups, following the partitioning in Table 2, i.e., QI-group 1 (2) contains tuples 1-4 (5-7). Given these QI-groups, the QIT and ST created by anatomy are shown in Table 3. In particular, the first row in Table 3a captures the fact that, there is a male patient involved in QI-group 1, with an age 21 and a zipcode 10001. On the other hand, the first row in Table 3b indicates that, one of the tuples in QI-group 1 has a sensitive value dyspepsia.

Anatomy preserves privacy because QIT does not indicate the sensitive value of any tuple, which must be randomly guessed from ST. To explain this, consider again the adversary who has the age 21 and zipcode 10001 of Bob. From 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 Table 3b, the adversary realizes that, the 4 tuples in QI-group 1 carry different diseases. Note that s/he does not gain any additional hints, regarding the exact disease of each tuple. Hence, s/he arrives at the conclusion that Bob could have contracted dyspepsia, flu, gastritis, or pneumonia with equal likelihood. 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 QI values in Table 3a, that no tuple in QI-group 2 can possibly qualify the query predicates. Furthermore, from Table 3b, it is clear that two tuples of QI-group 1 carry stomach diseases in the microdata. Hence, we proceed to calculate the probability p that a tuple in QI-group

1 satisfies the predicates in query A . This calculation does not need any assumption about the age or zipcode distribution of the tuples, *because the distribution is precisely released*. Specifically, Table 3a shows that tuples 1 and 2 in QI-group 1 appear in Q , leading to $p = 50\%$. Thus, we obtain an answer $2p = 1$, which is also the actual query result.

1.3 Contributions

This article presents a systematic study of the anatomy technique. Our contributions can be summarized as follows.

- We formalize anatomy to satisfy the anonymity requirement of l -diversity, and theoretically justify the superiority of anatomy against generalization, in capturing data correlation.
- We develop an algorithm, *Anatomize*, which computes a pair of QIT and ST in $O(n \log \lambda + n \cdot l)$ time, where n is the number of tuples in the microdata, and λ is the domain size of the sensitive attribute. Furthermore, the QIT and ST generated by *Anatomize* incur provably small information loss.
- We elaborate how anatomized tables can be used to perform statistical analysis. In particular, we establish a crucial connection between anatomized tables and probabilistic relations [10], which facilitates the derivation of various forms of answers: the expectation (of the actual query result), lower and upper bounds, or even the actual distribution. Our results hold for a wide range of count queries.
- Based on our modeling of anatomized tables as probabilistic relations, we propose *Anatomize**, a time-responsive version of *Anatomize*, that continuously improves the utility of the anatomized tables for count queries, until a user-specified time limit is reached.
- We show, through extensive experiments, that anatomy significantly outperforms the existing anonymization techniques (i.e., generalization, data swapping, and randomized response), in the effectiveness of data analysis.

The rest of the article is organized as follows. Section 2 surveys the previous research that is related to ours. Section 3 formalizes anatomy, and clarifies its privacy guarantees. Section 4 analyzes correlation preservation. Section 5 develops algorithm *Anatomize* for computing anatomized tables, and proves its quality under several metrics. Section 6 explains the details of statistical analysis using anatomized relations. Section 7 presents algorithm *Anatomize** for enhancing the utility of publication. Section 8 experimentally evaluates the proposed solutions. Section 9 concludes the article with directions for future work.

2 Related Work

A short version of this article appeared in [40]. Compared with that preliminary work, this article features three additional major contributions. First, in Section 5, we investigate the information loss of anatomy with respect to any L_p norm, KL-divergence, and the discernability metric (only L_2 norm was considered in [40]). Second, Section 6 elaborates the theory behind employing anatomized tables for statistical analysis. Third, in Section 7, we propose a new algorithm for obtaining anatomized tables with enhanced utility.

In the sequel, we discuss other previous research relevant to anatomy. Section 2.1 summarizes existing results on generalization. Then, Section 2.2 reviews “data swapping”, which is a generic anonymization methodology that anatomy belongs to. Finally, Section 2.3 surveys the literature of statistical databases.

2.1 Generalization

Generalization has been extensively studied in the literature [2, 3, 7, 15, 17, 18, 20–23, 25–27, 31, 32, 35–37, 39, 41–43]. LeFevre et al. [21] present an interesting taxonomy to categorize alternative methods based on their “recoding schemes”, which impose different constraints on generalization. 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. Local recoding removes this requirement.

The category of global recoding can be further divided into *single-dimension recoding* and *multidimension recoding*. Specifically, a recoding is single dimensional, if the generalized forms of two arbitrary QI-groups on the same attribute are either disjoint or equivalent. When the condition is not satisfied, the recoding is multidimensional. For example, the generalization in Table 2 is multidimensional, because the *Sex*-value ‘M’ of the first QI-group overlaps with the value ‘*’ (including both ‘F’ and ‘M’) of the second QI-group.

Computing the optimal generalization is usually harder for recoding schemes with fewer constraints. Unfortunately, it is NP-hard to find the optimal solution, even for simple schemes and quality metrics [3, 22, 27]. Therefore, the existing algorithms rely on heuristics for pruning the search space, in order to discover reasonable generalization within a time limit. However, when the number d of QI attributes is large, any generalization necessarily loses considerable information in the microdata [2], 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.

A majority of the literature focuses on k -anonymous generalization. However, Machanavajjhala et al. [25]

observe that k -anonymity fails to secure anonymity 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 distribution of sensitive values in each QI-group. The observation leads to l -diversity (as will be formalized in Section 3), which guarantees stronger privacy control than k -anonymity.

This article is virtually orthogonal to all the above works. The proposed anatomy technique abandons the idea of generalization, but instead, falls in a classical anonymization framework, called data swapping, as reviewed in the next subsection. Anatomy remedies the defects of generalization. Specifically, nearly-optimal anatomized tables can be computed in polynomial time, and captures a significant amount of correlation for any dimensionality.

2.2 Data Swapping

Given a microtable T , *data swapping* produces an alternative table by interchanging the values (of the same attribute) among the tuples in T . Numerous solutions have been proposed following this methodology (see [14] for a survey). In general, the design of such a solution requires clarification of three issues:

- Determination of the pairs or groups of tuples in which swapping is performed.
- Attributes involved.
- (Algorithmic and conceptual) nature of swapping.

The simplest method, *random swapping* [9], distorts the sensitive values in T as follows. First, it randomly selects a set S of $\delta \cdot n$ tuples in T , where n is cardinality of T and $\delta \in [0, 1]$ is a parameter. Then, it takes the list of $\delta \cdot n$ sensitive values in S , randomly permutes the list, and assigns the i -th ($1 \leq i \leq \delta \cdot n$) value of the randomized list to the i -th tuple in S . To protect privacy, δ should be sufficiently large; otherwise, many tuples in T would remain unchanged after the the swapping and are thus vulnerable to privacy attacks. However, when δ is large, random swapping destroys the correlation between the sensitive attribute and the QI attributes, rendering the resulting table of limited use for analysis.

Rank swapping [29] alleviates the defects of random swapping, by demanding that swapping should be performed only between tuples with close sensitive values. For this purpose, the tuples of T are first sorted in ascending order of their sensitive values. Then, for each tuple t , the algorithm exchanges its sensitive value with another tuple t' , which is randomly selected among all the tuples whose positions (in the sorted list) differ from that of t by at most ε . Here, ε is a system parameter, controlling the degree of data distortion.

tuple ID	Age	Sex	Zipcode	Disease
1 (Alice)	15	F	10000	dyspepsia
2	20	M	10000	dyspepsia
3	30	F	20000	pneumonia
4	20	F	10000	pneumonia
5	30	M	20000	dyspepsia
6	20	F	20000	dyspepsia
7	30	M	10000	pneumonia

(a) The microtable T (b) The table T' after swapping

Table 4: Illustration of marginal-preserving swapping

Intuitively, the higher ε is, the table after swapping retains less information in T , but provides better privacy protection.

Neither random nor rank swapping offers solid guarantees regarding the preservation of statistics in T . A more rigorous solution is *Marginal-preserving swapping* (MPS) [34], which aims at deriving a table T' that captures all the λ -order marginals of T involving the sensitive attribute. Towards this goal, all the sensitive and QI attributes may be involved in swapping, as opposed to the previous two approaches that concern only the sensitive attribute. To illustrate the idea, let T and T' be Tables 4a and 4b, respectively. Observe that although the two tables do not share any common tuple, T' preserves all three 2-order marginals of T that include *Disease*. For example, the marginal $(Age, Disease)$ is the projection of T onto those attributes, with duplicates retained. $(Sex, Disease)$ and $(Zipcode, Disease)$ are the other two marginals preserved.

The importance of preserving a marginal lies in the fact that T' is simply as effective as T , for studies relevant only to the marginal. For instance, any query concerning *Age* and *Disease* has the same result on Tables 4a and b. In general, if T' preserves all the λ -order marginals (related to the sensitive attribute A^s), then every query on T involving A^s and any $\lambda' \leq \lambda - 1$ QI attributes can be precisely answered from T' . Unfortunately, even for the smallest $\lambda = 2$, the marginal-preserving T' does not always exist. For instance, no T' is possible, if T is the microdata in Table 1. Furthermore, even if T' does exist, its computation may be computationally expensive [14, 34] and practically intractable.

Motivated by this, Reiss [34] develops the *approximate MPS*, which produces a T' whose marginals are close, albeit not identical, to those of T . Strictly speaking, this algorithm falls out of the scope of traditional data swapping, since T' may no longer be transformed from T by swapping — in fact, some values in T' may not even exist in T . However, from a data user’s perspective, whether T' is swapping-transformable from T is not important, as long as T' reflects the statistical patterns of T . This observation has led to many distribution-based techniques (e.g., [16, 24]) with the objective of generating a T' according to the distribution of T . A T' thus obtained does not necessarily have marginals similar to T , but may preserve

other statistical properties of T , such as mean, standard deviation, quantiles, etc.

To the best of our knowledge, none of the previous data-swapping algorithms is designed with linking attacks in mind. As a result, they do not promise the prevention of such attacks. To understand the vulnerability of, for example, MPS, assume an adversary that knows the age 15 of Alice, and the existence of Alice in the microdata of Table 4a. Given that Table 4b preserves the marginal (*Age, Disease*), the adversary can precisely infer Alice’s disease, because (i) there must be one occurrence of $\{15, \text{dyspepsia}\}$ in the microtable, and (ii) no other record in the marginal contains age 15.

2.3 Statistical Databases

There exists a large body of research (see for example [1, 8, 11, 12, 30]) on the topic of *statistical databases* (StatDB), which is fundamentally different from, although relevant to, privacy-preserving publication (i.e., the focus of our article). Specifically, in a StatDB, the microdata is stored in a *statistical server*, which is responsible for answering OLAP-like (e.g., count) queries submitted by the public through, for example, a web-based interface. Even though the server returns only aggregate results (as opposed to individual tuples), an adversary may still be able to precisely infer the original microdata, by formulating queries in a clever way and auditing their answers. Thus, the objective of privacy control is to prevent such *malicious auditing*.

According to a recent taxonomy in [12], the previous StatDB solutions can be classified into two categories: *input perturbation* and *output perturbation*. Input perturbation replaces the microdata with a distorted version, where the distortion must fulfill two purposes. First, given a query, the server should be able to derive, from the distorted dataset, a good estimate of the real answer of executing the query on the original database. This estimate is returned to the user. Second, the precision of those estimates must be limited to such a level that does not allow an adversary to accurately reconstruct any individual tuple of the microdata. Output perturbation, on the other hand, does not perform any data alternation, but instead, modifies query results. Specifically, given a query, the server first obtains its exact answer on the microdata, adds a noise to the answer, and then, reports the resulting noisy answer to the user.

The two categories mentioned earlier provide different service qualities. Input perturbation is able to support an infinite number of queries, but at the cost of lower precision in their results. In particular, Dwork et al. [12] prove that, in order to achieve the so-called ϵ -*differential privacy*, a huge amount of distortion must be applied to the microdata, rendering large variances in the reported results. Output perturbation has the opposite characteristics. Namely, it achieves much lower query error, whereas the tradeoff is that only a finite number of queries can be supported in order to satisfy ϵ -differential privacy.

We note that ϵ -differential privacy is the state-of-the-art anonymization principle [12] in StatDB. It demands tighter privacy control than all the known anonymization principles (including l -diversity) for privacy-preserving publication. However, enforcing ϵ -differential privacy is a difficult issue such that the application of this principle is currently limited to output perturbation techniques [12].

3 Formalization of Anatomy

Let T be a microtable with d QI-attributes $A_1^{qi}, A_2^{qi}, \dots, A_d^{qi}$, and a sensitive attribute A^s . Each A_i^{qi} ($1 \leq i \leq d$) can be either categorical or numerical with a discrete domain, but A^s should be categorical, following the assumption of l -diversity [25]. Let n be the number of tuples in T . For any tuple $t \in T$, we use $t[i]$ ($1 \leq i \leq d$) to denote its A_i^{qi} value, and $t[d+1]$ to represent its A^s value. Clearly, t can be regarded as a point in a $(d+1)$ -dimensional space, denoted as Ω . In Section 3.1, we first clarify the relevant concepts and the privacy guarantee of anatomy. Then, Section 3.2 compares the degrees of privacy protection provided by anatomy and generalization. Finally, Section 3.3 elaborates why anatomy can be regarded as a form of data swapping.

3.1 Concepts and Privacy Guarantee

As with generalization, anatomy requires partitioning the microtable T . In particular, a *partition* of T includes 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, \dots, QI_m , where m is the number of subsets. We are interested only in l -diverse partitions:

Definition 1. (l -diversity² [25]) *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 .

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

Definition 2. (Anatomy) *Given an l -diverse partition P with m QI-groups, **anatomy produces a quasi-identifier table (QIT) and a sensitive table (ST) that satisfy the following conditions:***

²This is called $(\frac{1}{l-1}, 2)$ -diversity in [25], which also suggests several other instantiations of l -diversity providing various degrees of privacy protection.

Symbol	Description
T	the microtable
n	the number of tuples in T
d	the number of QI-attributes in T
A_i^{qi} ($1 \leq i \leq d$)	the i -th QI attribute in T
A^s	the sensitive attribute in T
λ	the domain size of A^s
Ω	the $(d + 1)$ -dimensional space formed by the attributes in T
$t[i]$ ($1 \leq i \leq d$)	the A_i^{qi} value of a tuple t
$t[d + 1]$	the A^s value of a tuple t
QI_j	the j -th QI-group in a partition of T
$c_j(v)$	the number of tuples in QI_j with a sensitive value v

Table 5: Frequently used symbols

1. *QIT* has schema $(A_1^{qi}, A_2^{qi}, \dots, A_d^{qi}, \text{Group-ID})$, where *Group-ID* has an integer domain. There exists a bijection from the microdata T to *QIT*. Specifically, let t be a tuple in T , and assume that the QI-group that includes t is the j -th ($1 \leq j \leq m$) one in P . Then, in the bijection, t is mapped to a tuple $(t[1], t[2], \dots, t[d], j)$ in *QIT*.
2. *ST* has schema $(\text{Group-ID}, A^s, \text{Count})$, where *Group-ID* and *Count* have integer domains. Furthermore, for every $j \in [1, m]$, *ST* has a tuple $(j, v, c_j(v))$ if and only if the j -th QI-group in P contains $c_j(v)$ tuples whose sensitive values equal v .

For instance, based on the 3-diverse partition suggested in Table 2, anatomy produces the QIT and ST in Tables 3a and 3b respectively. When there is no ambiguity, we refer to a pair of QIT and ST collectively as the *anatomized tables*. For convenience, Table 5 summarizes the notations that will be used frequently in this paper.

By incorporating l -diversity, anatomy provides rigorous privacy guarantee against linking attacks. To clarify this, let us consider an adversary who attempts to infer the sensitive value of a victim individual o , with the following prior knowledge:

- A1: the adversary has the QI values of o ;
- A2: the adversary knows that o is in the microdata.

Note that, other than the victim o , the adversary is not necessarily aware of the QI-values or presence of any other individual in the microdata. Furthermore, before consulting the published data, the adversary has no information about the sensitive attribute. We have the following theorem.

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. All proofs can be found in the appendix. □

3.2 Comparison with Generalization

In this section, we will acknowledge the advantages of generalization over anatomy, by showing that the former may actually provide stronger privacy protection in certain scenarios. Our acknowledgement aims at providing an accurate evaluation of the pros and cons of alternative techniques. It, however, does not weaken the importance of anatomy, which is a useful option for data anonymization, since it always secures adequate anonymity control (Theorem 1) and has its own significant merits (in preserving data correlation, as explained in the next section).

Let us revisit the two assumptions A1 and A2 in Section 3.1. In practice, usually both assumptions are satisfied in a linking attack. For example, in [36], the author identifies the medical record of an ex-governor of Massachusetts from the GIC dataset, by (i) utilizing the governor’s QI values obtained from public sources, and (ii) believing that the governor was involved in the dataset. If both A1 and A2 are true, anatomy provides as much privacy control as generalization: limiting breach probability to $1/l$.

Now, consider the case where A1 holds, but A2 does not. For an adversary prying into the privacy of Alice, her/his chance of success has a Bayes form:

$$Pr_{A2}(Alice^{q^i}) \cdot Pr_{breach}(Alice^s|A2), \quad (2)$$

where $Pr_{A2}(Alice^{q^i})$ is the probability of Alice being in the microdata, and $Pr_{breach}(Alice^s|A2)$ the likelihood for the adversary to correctly guess the disease of Alice on condition that Alice appears in the microdata. As mentioned earlier, anatomy and generalization give the same $Pr_{breach}(Alice^s|A2)$, which is simply the breach probability when both A1 and A2 are valid.

To compute $Pr_{A2}(Alice^{q^i})$, an adversary typically needs to consult another external database [41], which relates QI values to concrete identities for all the persons in the microdata, perhaps together with some other “extraneous people”. An example of such an external source is a voter registration list, partially demonstrated in Table 6, where the record of Emily is italicized to indicate that she is extraneous, with respect to the microdata of Table 1. In this scenario, generalization and anatomy make a difference. Specifically, judging from (the QI values of tuples 5-7 in) the generalized Table 2, the adversary thinks that each person

Name	Age	Sex	Zipcode
...
Adam	61	M	10001
<i>Emily</i>	<i>65</i>	<i>F</i>	<i>60000</i>
Alice	70	F	60000
Bella	70	F	60000
...

Table 6: The voter registration list (publicly accessible)

shown in Table 6 could be in the microdata with equal likelihood, and hence, calculates $Pr_{A2}(Alice^{qi})$ as 3/4. On the other hand, given the QIT in Table 3a, the adversary concludes $Pr_{A2}(Alice^{qi}) = 1$ (here s/he can figure out that Emily is extraneous). Thus, generalization achieves stronger privacy protection.

There are, however, two important points to note. First, since anatomy ensures $Pr_{breach}(Alice^s|A2) \leq 1/l$, it still guarantees an upper bound $1/l$ for Formula 2. Having such a bound is a crucial property of an anonymization solution, since it allows a publisher to provide a firm promise on the maximum likelihood of privacy breach. Second, although generalization may result in a lower breach probability, leveraging the advantage is difficult in computing generalized data. This is because the publisher cannot predict or control the external database to be utilized by an adversary, and therefore, must conservatively assume that the adversary knows exactly whether her/his victim o is in the microdata (i.e., by making assumption A2).

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

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

where x is a vector representing a possible set of QI values of Alice. $Pr_{A1}(x)$ equals the probability that x captures Alice’s real QI values, whereas Pr_{A2} and Pr_{breach} follow the same semantics as in Formula 2, but on condition that x is real. The comparison between anatomy and generalization is analogous to the previous case where A1 is true and A2 is not.

Machanavajjhala et al. [25] show that generalization can be used to anonymize a microtable with multiple sensitive attributes. This seems to be another advantage over anatomy, which, as in our article, is limited to a single sensitive attribute. Furthermore, besides linking attacks, generalization also effectively prevents *presence inferences* [31], which attempt to figure out whether an individual is in the microdata. Anatomy is designed to thwart linking attacks with provably good guarantees, but does not ensure strong protection against presence inferences. In other words, anatomy should be employed only in applications where presence inferences are not considered a threat. Fortunately, there are many such applications. For instance, suppose that the government publishes census data on citizens’ tax payments. The presence of an individ-

ual in the microdata implies only an innocuous fact: s/he is working. As another example, consider that a hospital releases its patient records to a medical institution. In this case, revealing an individual’s presence indicates that s/he has been treated in the hospital before. Such “revelation” is unlikely to be harmful, since everybody goes to see a doctor periodically anyway. It is worth pointing out that presence inferences cannot be eliminated by any technique that releases QI-values directly — the presence of a person is inevitably disclosed, as long as any of her/his QI-values is unique in the external database employed by the adversary. For example, data-swapping solutions are also vulnerable to such inferences.

3.3 Relations to Data Swapping

Anatomy fits in the generic framework of data swapping. We illustrate this by explaining how anatomy settles the three design issues of a data swapping solution, stated at the beginning of Section 2.2. First, the tuples, among which swaps are performed, are those in the same QI-group. Second, only the sensitive attribute is involved in swapping. Third, within a QI-group QI , swapping has an *implicit* and *permutating* nature. That is, unlike any of the previous swapping algorithms (see Section 2.2), anatomy does not explicitly combine a tuple with any (swapped) sensitive value; instead, any permutation of the sensitive values in QI determines a possible scheme of assigning the values to the tuples in QI . From the perspective of an adversary or a data user, each scheme has a chance to be the real one in the original microdata.

Anatomy positions itself as the first data-swapping solution that is specifically designed to thwart linking attacks, and achieves rigorous privacy control (through the enforcement of l -diversity) demanded by the modern community of data anonymization. Furthermore, anatomy is “statistically-solid”, since it offers theoretical guarantees on the degree of correlations retained from the microdata. Specifically, as shown in the next three sections, anatomy nearly minimizes the error of reconstructing the original microdata, up to an extremely small approximation ratio. Recall that, as reviewed in Section 2.2, MPS ensures capturing all marginals of up to a certain order, whereas random- and rank-swapping do not have any non-trivial guarantee. Therefore, in terms of statistical-solidity, anatomy and MPS outperform random- and rank-swapping.

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 the accuracy of the modeling.

4.1 Intuition

Let us first examine the correlation between *Age* and *Disease* in the microdata of Table 1. The two attributes define a 2D space $\Omega_{A,D}$. Every tuple in the table can be mapped to a point in $\Omega_{A,D}$. For example, tuple 1, denoted as t_1 , corresponds to point $(t_1[A], t_1[D])$, where $t_1[A]$ is the age 21 of t_1 , and $t_1[D]$ is disease ‘dyspepsia’. We can model t_1 using a probability density function (pdf) $\mathcal{G}_1 : \Omega_{A,D} \rightarrow [0, 1]$. Specifically:

$$\mathcal{G}_1(x) = \begin{cases} 1 & \text{if } x = (t_1[A], t_1[D]) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where x is a 2D random variable in $\Omega_{A,D}$. Figure 1a demonstrates the pdf.

Assume that a researcher wants to re-construct an approximate pdf $\tilde{\mathcal{G}}_1^g$ 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 equal probability $1/40$, but $t_1[D]$ must be dyspepsia. Hence,

$$\tilde{\mathcal{G}}_1^g(x) = \begin{cases} 1/40 & \text{if } x[A] \in [21, 60] \text{ and } x[D] = \text{dyspepsia} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

which is illustrated in Figure 1b.

Instead, suppose that the researcher re-constructs a pdf $\tilde{\mathcal{G}}_1^a$ from the QIT and ST in Tables 3a and 3b. This time, s/he knows that $t_1[A]$ must be 21 (since the age is published directly in QIT), but $t_1[D]$ can be one of the four diseases in the first QI-group with equivalent likelihood, according to ST. Therefore,

$$\tilde{\mathcal{G}}_1^a(x) = \begin{cases} 1/4 & \text{if } x = (21, \text{dyspepsia}) \text{ or } x = (21, \text{gastritis}) \text{ or} \\ & x = (21, \text{flu}) \text{ or } x = (21, \text{pneumonia}) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

as shown in Figure 1c. Obviously, the pdf approximated from the anatomized tables is more accurate than that (Figure 1b) from the generalized table. Although we focused on t_1 , in the same way, it is easy to verify that the anatomized tables permit better re-construction of all tuples in Table 1.

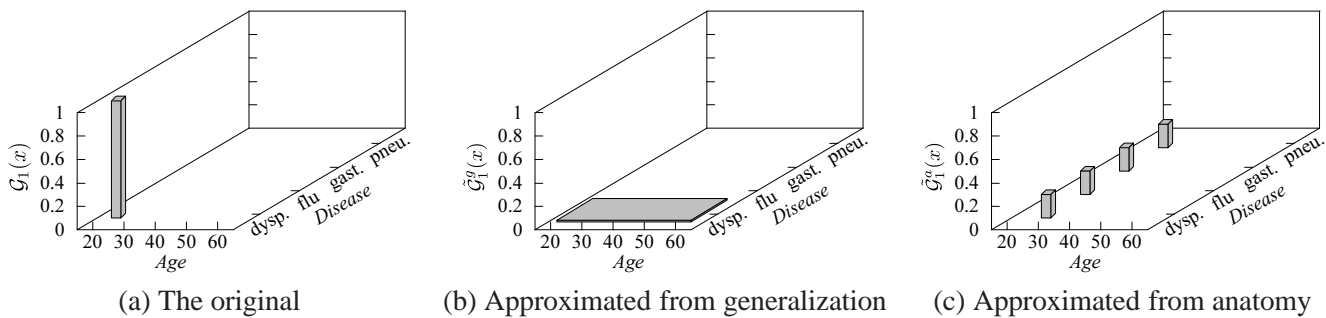


Figure 1: The original/re-constructed pdf of tuple 1 in Table 1

4.2 General Results and Quality Metric

As defined in Section 3, each tuple t in the microtable T can be regarded as a point in a $(d + 1)$ -dimensional space Ω (including all the QI- and sensitive dimensions). Next, we generalize discussion in Section 4.1 to Ω . We model t as a pdf $\mathcal{G}_t(x) : \Omega \rightarrow [0, 1]$:

$$\mathcal{G}_t(x) = \begin{cases} 1 & \text{if } x = t \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where x is a random variable in Ω . 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^* be the generalization of t . The i -th ($1 \leq i \leq d$) QI value $t^*[i]$ of t can be regarded as an interval³ enclosing $t[i]$. Denote the length of $t^*[i]$ as $L(t^*[i])$, equal to the number of different values in $t^*[i]$. Then, the reconstructed pdf $\tilde{\mathcal{G}}_t^g(x)$ of t is

$$\tilde{\mathcal{G}}_t^g(x) = \begin{cases} \frac{1}{\prod_{i=1}^d L(t^*[i])} & \text{if } x[i] \in t^*[i] \text{ for all } i \in [1, d] \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Next we discuss anatomized tables. Assume that QI is the QI-group containing t (in the underlying l -diverse partition). Let $v_1, v_2, \dots, v_\lambda$ be all the distinct A^s values in QI (e.g., for QI-group 1 in Table 3a, $\lambda = 4$, whereas for QI-group 2, $\lambda = 3$). Denote $c(v_h)$ ($1 \leq h \leq \lambda$) as the *Count* value in ST corresponding to v_h .

³If $A_i^{q_i}$ is categorical, we consider that there is a total ordering on $A_i^{q_i}$. Such an ordering naturally exists, under the common assumption that there is a generalization hierarchy over $A_i^{q_i}$. Specifically, it simply enumerates all the leaves of the hierarchy from left to right.

The reconstructed pdf $\tilde{\mathcal{G}}_t^a(x)$ of t is

$$\tilde{\mathcal{G}}_t^a(x) = \begin{cases} c(v_1)/|QI| & \text{if } x = (t[1], \dots, t[d], v_1) \\ \dots & \dots \\ c(v_\lambda)/|QI| & \text{if } x = (t[1], \dots, t[d], v_\lambda) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $|QI|$ is the number of tuples in QI .

Notice that $\tilde{\mathcal{G}}_t^a(x)$ is greater than 0, only when x lies at one of the λ points in Ω , as described in the if-conditions of Equation 8. That is, $\tilde{\mathcal{G}}_t^a(x)$ consists of λ “spikes” at these points ($\lambda = 4$ in Figure 1c). On the other hand, in practice, $\tilde{\mathcal{G}}_t^g(x)$ 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 (c.f. Figure 1b), thus deviating significantly from the actual $\mathcal{G}_t(x)$.

Given an approximate pdf $\tilde{\mathcal{G}}_t$ (Equation 7 or 8), we quantify its *approximation error* from the actual \mathcal{G}_t (Equation 6) as

$$E_t = \sum_{x \in \Omega} |\tilde{\mathcal{G}}_t(x) - \mathcal{G}_t(x)|^p, \quad (9)$$

where p is an integer. Taking into account all tuples $t \in T$, a good publication method should minimize the following *re-construction error* (RCE):

$$RCE = \sum_{\forall t \in T} E_t. \quad (10)$$

The formulation of E_t in Equation 9 is based on the L_p -norm⁴ In Section 5.2, we will discuss alternative information loss metrics.

5 A Nearly-Optimal Anatomizing Algorithm

We propose an efficient algorithm for computing anatomized tables that (almost) minimize RCE (Equation 10). Our algorithm runs in $O(n \log \lambda + n \cdot l)$ time, n is the cardinality of T , and λ the domain size of the sensitive attribute.

⁴Strictly speaking, the L_p distance between the actual and approximate pdfs equals $(\sum_{\forall t \in T} E_t)^{1/p}$. Nevertheless, minimization of $\sum_{\forall t \in T} E_t$ also minimizes $(\sum_{\forall t \in T} E_t)^{1/p}$. Hence, for notational simplicity, we formulate RCE as in Equation 10.

5.1 The Anatomizing Algorithm

Figure 2 presents the algorithm *Anatomize* which, given a microtable T and a parameter l , obtains a pair of QIT and ST. *Anatomize* first computes an l -diverse partition of T (Lines 1-12), and then, produces QIT and ST (Lines 13-18) from the partition. The objective of *Anatomize* is to create *only* QI-groups with at least l tuples, all of which possess different sensitive values. As shown in the next subsection, this strategy leads to small reconstruction error RCE . Towards that objective, *Anatomize* produces one QI-group at a time, using tuples whose sensitive values are the most frequent, among the “remaining tuples” that have not been added to any QI-group yet. This approach ensures adequate diversity of sensitive values in the remaining tuples, so that we can always grab enough tuples with different sensitive values to form the next QI-group.

Specifically, *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 and 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.*

We use the term *residue tuple* to refer to a tuple remaining in a bucket, at the end of the group-creation phase. Property 1 implies 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). 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 2) always includes at least one QI-group.*

Correctness. Since Lines 13-19 of Figure 2 essentially implement Definition 2, *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.

Algorithm Anatomize (T, l)

Input: a microtable T , and a value l ; Output: anatomized tables QIT and ST

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
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 $(t[1], \dots, t[d], j)$ 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 tuple $(j, v, c_j(v))$ into ST
19. return QIT and ST

Figure 2: The anatomizing algorithm

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.*

Example 1. Let us illustrate *Anatomize* by applying it to the microdata in Table 1, assuming $l = 3$. First, the algorithm hashes all the tuples into 5 buckets by their diseases: $B_{\text{pneu}} = \{2, 6\}$ (i.e., containing tuples 2 and 6 carrying “pneumonia”), $B_{\text{gast}} = \{4, 7\}$, $B_{\text{dysp}} = \{1\}$, $B_{\text{flu}} = \{3\}$, $B_{\text{bron}} = \{5\}$. Now, the group-creation phase begins. To obtain the first QI-group QI_1 , *Anatomize* collects the set S of three largest buckets B_{pneu} , B_{gast} , B_{flu} . Note that the inclusion of B_{flu} is a random choice — B_{dysp} , B_{flu} , B_{gast} all have the same size, and therefore, *Anatomize* arbitrarily picks one of them into S . From each bucket in S , a tuple is randomly selected into QI_1 . Suppose that the selection results in $QI_1 = \{2, 4, 3\}$, after which B_{pneu} changes to $\{6\}$, B_{gast} to $\{7\}$, and B_{flu} to \emptyset . Similarly, to spawn the second QI-group QI_2 , *Anatomize* refreshes S to include three buckets currently having the largest sizes. Let S be $\{B_{\text{bron}}, B_{\text{pneu}}, B_{\text{gast}}\}$. Taking a tuple from each bucket respectively, $QI_2 = \{5, 6, 7\}$, after which those three buckets become empty. As there are fewer than $l - 1$ non-empty buckets (actually, there is only one B_{dysp}), the algorithm enters the residue-assignment step.

B_{dysp} has a residue tuple 1. *Anatomize* decides $S' = \{QI_1, QI_2\}$, since neither QI_1 nor QI_2 contains any tuple with disease “dyspepsia”. Then, tuple 1 is added to a random QI-group in S' . Assuming QI_1 is chosen, it is thus updated to $\{1, 2, 3, 4\}$, and the residue-assignment phase terminates. The final QI_1 and QI_2 constitute a 3-diverse partition of the microtable, which produces the QIT and ST in Tables 3a and 3b, by Definition 2.

5.2 Analysis

In this section, we analyze the efficiency and effectiveness of *Anatomize* (Figure 2). We first prove that *Anatomize* has linear time complexity, as shown in the following Theorem.

Theorem 2. *Anatomize terminates in $O(n(\log \lambda + l))$ time, where λ is the domain size of the sensitive attribute.*

In practice, $\log \lambda$ and l are significantly lower than n , and may be even regarded as constants. In that case, the running time of *Anatomize* is linear to n . Next, we establish the lower bound of the RCE achievable by any anatomized tables as follows.

Lemma 1. *For any pair of QIT and ST, their RCE under the L_p norm is at least $n(1 - 1/l)^p + n(l - 1)(1/l)^p$.*

Based on Lemma 1, we prove that the QIT and ST generated by *Anatomize* is nearly optimal in terms of information loss, as shown in Theorem 3.

Theorem 3. *If n is a multiple of l , under the L_p norm, the QIT and ST computed by *Anatomize* achieve the lower bound of RCE in Lemma 1. Otherwise, their RCE is higher than the lower bound by a factor at most $1 + \frac{2l-1}{n} \cdot \phi$, where*

$$\phi = \frac{(1 - \frac{1}{2l-1})^p + (2l - 2)(\frac{1}{2l-1})^p}{(1 - \frac{1}{l})^p + (l - 1)(\frac{1}{l})^p} - 1. \quad (11)$$

It is worth mentioning that the ϕ in Equation 11 is independent of n . For a large microtable, the RCE of the anatomized tables is extremely close to the lower bound in Lemma 1. As an example, when $p = 2$, the approximation ratio $1 + \frac{2l-1}{n} \cdot \phi = 1 + \frac{1}{n}$.

The result in Theorem 3 holds, when RCE is gauged by an L_p norm. It is natural to wonder whether *Anatomize* can minimize information loss, when RCE is defined according to other metrics. In the following, we provide a positive answer, with respect to two common definitions: *Kullback-Leibler (KL) divergence* [19] and the *discernability* metric [7].

KL-divergence. Let \mathcal{G}_t (in Equation 4) denote the actual pdf of a tuple t in the microtable T , and $\tilde{\mathcal{G}}_t^a$ (Equation 8) the pdf approximated from a pair of QIT and ST. The KL-divergence between \mathcal{G}_t and $\tilde{\mathcal{G}}_t^a$ equals (all logs have base 10):

$$E_t = \sum_{x \in \Omega} \mathcal{G}_t(x) \cdot \log \mathcal{G}_t(x) - \sum_{x \in \Omega} \mathcal{G}_t(x) \cdot \log \tilde{\mathcal{G}}_t^a(x). \quad (12)$$

Next, we derive a lowerbound of the RCE (Equation 10) of any anatomized tables, when E_t is calculated with Equation 12.

Lemma 2. *For any pair of QIT and ST, under the KL-divergence metric, their RCE is at least $n \log l$.*

We now show that the tables produced by *Anatomize* incur small information loss, under the KL-divergence metric.

Theorem 4. *If n is a multiple of l , under the KL-divergence metric, the QIT and ST computed by *Anatomize* achieve the lower bound of RCE in Lemma 2. Otherwise, their RCE is higher than the lower bound by a factor at most $1 + \frac{(2l-1) \cdot \log(2-1/l)}{n \cdot \log l}$.*

Again, the approximation ratio approaches 1 given a practical n . For instance, for $l = 10$, $1 + \frac{(2l-1) \cdot \log(2-1/l)}{n \cdot \log l} \approx 1 + \frac{5.3}{n}$.

Discernability. The discernability metric is defined on the number of tuples in each QI-group. Specifically, given a QI-group QI , the discernability metric charges, on each tuple $t \in QI$, a penalty

$$E_t = |QI|. \quad (13)$$

Consider the RCE calculated with Equation 10, but using the above E_t . The following lemma shows a lower bound of RCE in this case.

Lemma 3. *For any pair of QIT and ST, under the discernability metric, their RCE is at least $(n+r) \cdot l + r$, where $r = n \bmod l$.*

Based on Lemma 3, we prove the following theorem.

Theorem 5. *If n is a multiple of l , under the discernability metric, the QIT and ST computed by *Anatomize* achieve the lower bound of RCE in Lemma 3. Otherwise, their RCE is higher than the lower bound by a factor at most $1 + \frac{l^2 - 3l + 2}{n \cdot l + l^2 - 1}$.*

Summarizing Theorems 3, 4, and 5, the anatomized tables output by *Anatomize* achieve a reconstruction error worse than the minimum by a factor of $1 + O(\frac{1}{n})$, under all the metrics of information loss discussed.

tuple ID	Age	Sex	Zipcode	Disease
1 (Bob)	21	M	10001	gastritis
2	27	M	13000	dyspepsia
3	35	M	60000	flu
4	60	M	12000	pneumonia
5	61	M	10001	bronchitis
6	70	F	60000	pneumonia
7 (Alice)	70	F	60000	gastritis

Table 7: A possible microdata instance of the anatomy in Table 3

6 Query Processing with Anatomized Tables

This section explains how to analyze the characteristics of the microdata using anatomized tables. Our discussion primarily concentrates on count queries of the form:

```
SELECT COUNT(*) FROM Unknown-Microdata
WHERE  $pred^{qi}$  AND  $pred^s$ 
```

Here, $pred^{qi}$ and $pred^s$ are arbitrary predicates on the QI and sensitive attributes, respectively. We focus on count queries, because they are fundamental to numerous mining operations such as decision tree learning [28], association rule mining [4], etc. In Section 6.1, we will show that the output of anatomy can be interpreted as a probabilistic relation [10]. The interpretation motivates the formulation of several forms of probabilistic answers. Section 6.2 derives the concrete formulae for computing those answers. Finally, Section 6.3 discusses the processing of other types of queries.

6.1 Probabilistic Query Results

Different microtables can lead to the same QIT and ST. Consider the microdata in Table 7; if we place tuples 1-4 and 5-7 into two QI-groups respectively, the resulting QIT and ST are identical to Tables 3a and 3b. In other words, given the QIT and ST, it is impossible to tell whether the microdata is Table 1 or 7 — both of them are “possible microdata instances”.

Formally, let $P = \{QI_1, QI_2, \dots, QI_m\}$ be the partition of a microtable T that underlies a pair of QIT and ST. For each QI_j ($1 \leq j \leq m$), let us perform a random permutation of the sensitive values. Specifically, we first randomize the $|QI_j|$ sensitive values in QI_j , and then assign the i -th ($1 \leq i \leq |QI_j|$) sensitive value in the randomized list to the i -th tuple of QI . After permuting all QI-groups independently, the final tuples constitute a *possible microdata instance*.

Let M be the set of all possible microdata instances $T_1, T_2, \dots, T_{|M|}$. From the perspective of a researcher,

any instance in M has an equal probability to be the actual microtable. Therefore, s/he can model the microdata as a *probabilistic relation* \mathcal{T} such that

$$Pr\{\mathcal{T} = T_i\} = 1/|M|, \quad (14)$$

for any $i \in [1, |M|]$ ⁵.

Consider a count query q whose actual answer on T is act . To derive a probabilistically-correct estimate est of act , conceptually a researcher needs to evaluate q over all the possible instances of \mathcal{T} . We use ϕ_i to denote the result of q on T_i ($1 \leq i \leq |M|$). Let φ be the set $\{\phi_1, \phi_2, \dots, \phi_{|M|}\}$. As all elements in φ are equally likely to be act , est obeys a pdf $\mathcal{Q} : \varphi \rightarrow [0, 1]$ of the form

$$\mathcal{Q}(x) = \frac{|\{i \mid \phi_i = x\}|}{|M|}.$$

(The numerator gives the number of elements in φ equal to x .) \mathcal{Q} permits a researcher to compute several types of results. The most common one is the expectation $\langle est \rangle$ of est :

$$\langle est \rangle = \sum_{x \in \varphi} (x \cdot \mathcal{Q}(x)). \quad (15)$$

Alternatively, s/he may calculate a lower bound est_{\downarrow} and an upper bound est_{\uparrow} of est :

$$est_{\downarrow} = \min\{x \mid \mathcal{Q}(x) > 0\} \quad (16)$$

$$est_{\uparrow} = \max\{x \mid \mathcal{Q}(x) > 0\}. \quad (17)$$

Note that both bounds are potentially reachable, since they are the results of q on some possible microdata instances of \mathcal{T} . Finally, another useful result is the probability that est falls in a range $[c_1, c_2]$:

$$\sum_{x \in \varphi \wedge x \in [c_1, c_2]} \mathcal{Q}(x). \quad (18)$$

The above equations remain at the conceptual level. Specifically, they cannot be implemented in practice, due to the assumption of having all the possible instances available. Next, we remove the assumption, and explain how to efficiently extract each type of results.

⁵ \mathcal{T} can also be regarded as a set of *possible tuples*, each of which exists in T with a probability. Furthermore, these tuples are not mutually independent, because their presence is subject to certain constraints, e.g., the total number of tuples appearing in T simultaneously equals $|T|$.

6.2 Result Computation

Again, let $P = \{QI_1, QI_2, \dots, QI_m\}$ be the partition of T that produces the released QIT and ST, and q be a count query. Although a researcher does not know the exact content of any QI_j ($1 \leq j \leq m$), s/he knows

- the size n_j of QI_j ;
- the number u_j of tuples in QI_j satisfying $pred^{qi}$ (obtainable from QIT);
- the number v_j of tuples in QI_j satisfying $pred^s$ (obtainable from ST).

Based on the above information, s/he aims at deriving the number est_j of tuples in QI_j qualifying q . Then, the estimated result est of q can be represented as

$$est = \sum_{j=1}^m est_j. \quad (19)$$

As with est , each est_j ($1 \leq j \leq m$) is a random variable, whose distribution is determined by the possible microdata instances. Furthermore, for any $j \neq j'$, est_j is independent of $est_{j'}$, reflecting the fact that the j -th and j' -th QI-groups are permuted independently in creating a possible microdata instance. The next lemma gives the pdf of est_j .

Lemma 4. For each $j \in [1, m]$, est_j obeys a pdf $\mathcal{Q}_j : \mathbb{Z} \rightarrow [0, 1]$ of the form

$$\mathcal{Q}_j(x) = \begin{cases} \binom{u_j}{x} \binom{n_j - u_j}{v_j - x} / \binom{n_j}{v_j} & \text{if } x \in [\max\{u_j + v_j - n_j, 0\}, \min\{u_j, v_j\}] \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

We proceed to elaborate the calculation of each type of results about est discussed in Section 6.1.

Corollary 1. $\langle est_j \rangle = u_j \cdot v_j / n_j$.

It follows that $\langle est \rangle = \sum_{j=1}^m \langle est_j \rangle = \sum_{j=1}^m \frac{u_j \cdot v_j}{n_j}$, which theoretically justifies the method used in Section 1.2 to answer query A. Namely, the result 1 we obtained is indeed the expected value of est .

Corollary 2. Let est_{j-} and est_{j+} be the lower and upper bounds of est_j , respectively. Then, $est_{j-} = \max\{u_j + v_j - n_j, 0\}$, and $est_{j+} = \min\{u_j, v_j\}$.

Therefore, $est_{-} = \sum_{j=1}^m \max\{u_j + v_j - n_j, 0\}$, and $est_{+} = \sum_{j=1}^m \min\{u_j, v_j\}$. Finally, evaluation of Formula 18 requires the concrete formula of $\mathcal{Q}(x)$. Combining Equation 19 with the fact that est_1, est_2, \dots ,

est_m are mutually independent, we have:

$$\mathcal{Q}(x) = (\mathcal{Q}_1 * \mathcal{Q}_2 * \dots * \mathcal{Q}_m)(x),$$

where $*$ denotes the convolution operator⁶.

It is natural to wonder: can the above results be used by the publisher in computing anatomized tables? The answer is yes, *if a set of representative queries is given*. To see this, first note that Lemma 4 and Corollaries 1 and 2 hold only for a single query. Utilizing them, the publisher may compute an anatomy that returns the most accurate answer for that query. Extending the idea, when the publisher is given a *workload* of queries, it can prepare an anatomy that minimizes the average error of all queries (the error of each query can be calculated from the earlier lemma and corollaries). In fact, this is exactly the rationale behind the interesting technique of *workload-aware anonymization* [20].

The drawback of workload-aware anonymization is that it is not effective for ad-hoc analysis that involves arbitrary queries significantly different from those in the workload. Intuitively, this is caused by the fact that the anatomy output by workload-aware anonymization retains data correlation only in a certain part of the data space, but may lose significant correlation in the remaining parts. To enable better ad-hoc analysis, the publisher needs to resort to a generic metric of information loss (such as the metrics discussed in Section 5.2), i.e., finding an anatomy that minimizes the metric, and thus, retains sufficient correlations in all parts of the data space.

6.3 General Query Processing

Our discussion so far has concentrated on counting. The underlying concepts, however, can be extended to support *any* query q using anatomized tables, regardless of the nature of q , i.e., whether it is an aggregate query, a selection query, a data mining operation, and so on. This owes to the modeling, described in Section 6.1, of anatomized tables as a probabilistic relation \mathcal{T} . Specifically, denote the set M of possible instances of \mathcal{T} as $\{T_1, T_2, \dots, T_{|M|}\}$. Let φ be $\{\phi_1, \phi_2, \dots, \phi_{|M|}\}$, where ϕ_i ($1 \leq i \leq |M|$) is the result of q on T_i . Depending on the details of q , each ϕ_i may be a simple value (e.g., for an aggregate query q), a set of tuples (for a selection query q), or even a set of mining results (e.g., association rules). The set φ constitutes the final result returned by anatomy, indicating that each element in φ has an equal chance of being the real answer (of applying q to the original microdata). Although answering queries on a probabilistic relation is #P-hard in the worse case [10], there exist queries that can be processed efficiently on anatomized tables,

⁶For any functions $f, g : \mathbb{Z} \rightarrow \mathbb{R}$, their convolution $(f * g)(x)$ is given by $\sum_{y=-\infty}^{\infty} f(y) \cdot g(x - y)$

i.e., those for which φ can be adequately captured by a set of closed formulae. Section 6.2 has illustrated this for count queries.

It is worth mentioning that *probabilistic databases* (PrDB) have received considerable attention from the database community (see [10] and the references therein). By modeling the anatomized tables as a probabilistic relation, we allow a researcher to leverage the literature of PrDB to assist her/his study. For example, [33] defines several novel types of top-k retrieval on a probabilistic relation, and propose efficient query processing algorithms. Their results can also be applied to anatomized tables.

7 Improving the Utility of Anatomy

Given a microtable T , *Anatomize* (Figure 2) does not return a unique output, due to two random factors. First, in the group-creation phase, an iteration yields a QI-group, by choosing a tuple arbitrarily from each of the l largest buckets. Second, in the residue-assignment step, the QI-group that incorporates a residue tuple is also randomly determined (among a set of qualified QI-groups). Nevertheless, all the potential outputs of *Anatomize* achieve an equivalent RCE.

RCE is a general-purpose metric for measuring the utility of anatomy. Specifically, if (i) the publisher is not sure about the nature of analysis to be performed on the released data, or (ii) the publication will be employed in multiple drastically-different forms of studies, minimizing RCE is a good choice, since RCE directly quantifies the error of rebuilding the original microdata. In some applications, however, data recipients may wish to obtain an anonymized dataset that is well suited for a particular type of queries. To satisfy such a request, the publisher may need to consider additional optimization goals in applying anonymization. In the sequel, we explain how to optimize anatomy for count queries. Section 7.1 reveals a crucial factor that affects the accuracy of counting. Then, Sections 7.2-7.4 elaborate the concrete computation algorithms.

7.1 MBRs of QI-Groups

We use Ω^{qi} to denote the d -dimensional *QI-space*, whose axes are the QI attributes of T . Each tuple $t \in T$ can be regarded as a point in Ω^{qi} with coordinates $t[1]$, $t[2]$, ..., and $t[d]$. For every QI-group QI_j in a partition of T , where $1 \leq j \leq m$ and m is the number of QI-groups, we define its *minimum bounding rectangle* (MBR), denoted as R_j , as the smallest axis-parallel rectangle enclosing all the points representing the tuples in QI_j .

The MBR sizes of QI-groups influence the quality of anatomy in counting analysis. Consider again query A

in Section 1.1. We create a rectangle Q in the QI-space, whose projections on the *Age (Zipcode)* dimension is $[1, 30]$ ($[10k, 20k]$). and its projection on the *Sex* axis encloses the entire domain. A tuple may satisfy query A, only if its point representation in Ω^{qi} is covered by Q . For a QI-group QI_j , let act_j be the actual number of tuples in QI_j qualifying the query, and est_j the estimate computed from anatomy. Depending on the topological relationship between Q and R_j , est_j is identical to act_j in two cases:

- R_j is disjoint with Q . When this happens, $act_j = 0$. From QIT, an analyst obtains $u_j = 0$, where u_j is the number of tuples in QI_j qualifying $pred^{qi}$. Hence, by Corollary 2, s/he calculates $est_{j-} = \max\{v_j - n_j, 0\} = 0$ and $est_{j+} = \min\{0, v_j\} = 0$, where n_j is the size of QI_j , and v_j the number of tuples in QI_j whose A^s values are a stomach disease (a predicate of query A).
- R_j is contained in Q . Here, act_j equals v_j . From QIT, the analyst understands that u_j is identical to $|QI_j|$. Thus, Corollary 2 gives $est_{j-} = \max\{n_j + v_j - n_j, 0\} = v_j$ and $est_{j+} = \min\{n_j, v_j\} = v_j$.

It follows that est_j may not be act_j , only if R_j partially intersects Q . To reduce the chance of partial intersection, QI-groups should have small MBRs.

7.2 Tuple Swapping

Our objective is to obtain a pair of QIT and ST that (i) have the same RCE as the anatomized tables output by *Anatomize*, but (ii) have smaller QI-group MBRs. Specifically, given a microtable T , we aim at minimizing the *anatomy perimeter*, equal to the sum of the perimeters of all R_j ($1 \leq j \leq m$).

The core of our technique is *tuple swapping*, which exchanges a pair of tuples in different QI-groups. A swap is *legal*, if

- after swapping, each QI-group is still l -diverse, and
- the swap does not affect RCE (Equation 10).

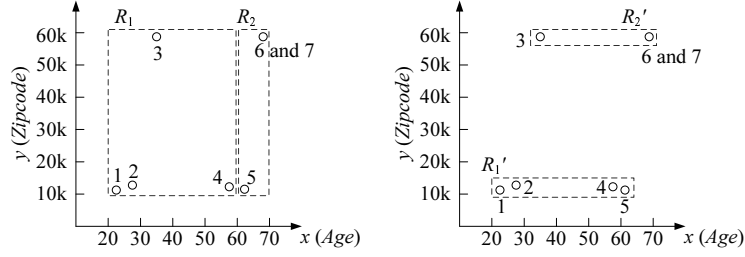
Consider, for example, the anatomy in Table 3. Let us exchange tuples 3 and 5 between QI-groups 1 and 2 respectively, which leads to the QIT and ST in Table 8. This swap is legal, as indicated by the following lemma.

Lemma 5. *Let QI_x (QI_y) be a QI-group in an l -diverse partition, such that all tuples in the group have distinct sensitive values. Two tuples $t_x \in QI_x$ and $t_y \in QI_y$ can be legally swapped if (i) $t_x.A^s = t_y.A^s$, or (ii) $t_x.A^s$ does not appear in QI_y , and $t_y.A^s$ does not appear in QI_x .*

tuple ID	Age	Sex	Zipcode	Group-ID
1	21	M	10001	1
2	27	M	13000	1
5	61	M	10001	1
4	60	M	12000	1
3	35	M	60000	2
6	70	F	60000	2
7	70	F	60000	2

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

(a) The quasi-identifier table (QIT) (b) The sensitive table (ST)
Table 8: Alternative anatomized tables for the microdata in Table 1



(a) Before swapping (Table 3) (b) After swapping (Table 8)
Figure 3: Changes of QI-group MBRs after swapping

A swap is *beneficial* if it reduces the anatomy perimeter. Figures 3a and 3b illustrate the QI-group MBRs respectively before and after the swap that produced Tables 8a and 8b (dimension *Sex* is omitted as it is not affected by the swap). Clearly, the anatomy perimeter drops after swapping.

Figure 4 formally describes our algorithm of performing a single swap. The algorithm invokes a procedure *Choose-QI-Groups*, whose details are clarified in the next subsection, to identify two appropriate QI-groups. Then, two tuples are exchanged between these groups, provided that the swap is both legal and beneficial.

7.3 Selection of QI-Groups for Swapping

This section elaborates *Choose-QI-Groups* in Figure 4. The procedure selects two QI-groups QI_x and QI_y , such that swapping tuples between them may lead to large decrease, denoted as $\Delta(QI_x, QI_y)$, in the anatomy perimeter. Let R_x (R'_x) be the MBR of QI_x before (after) the swapping. Use $PERI(R_x)$ and $PERI(R'_x)$ to represent the perimeters of R_x and R'_x , respectively. Similar notations are also adopted for QI_y . Since swapping does not influence the other QI-groups, $\Delta(QI_x, QI_y)$ equals

$$(PERI(R_x) + PERI(R_y)) - (PERI(R'_x) + PERI(R'_y)). \quad (21)$$

A naive implementation of *Choose-QI-Groups* is to simply examine all (QI_x, QI_y) pairs, compute the highest possible $\Delta(QI_x, QI_y)$ for each pair, and then choose the one that maximizes $\Delta(QI_x, QI_y)$. This approach, however, incurs expensive overhead. We provide a faster method to heuristically decide “good”

Algorithm Swap (QIT, ST)

Input: anatomized tables QIT and ST

Output: modified anatomized tables with the same RCE but a lower anatomy perimeter

1. $\{QI_x, QI_y\} = \mathbf{Choose-QI-Groups}(QIT, ST)$ /* see Section 7.3 */
2. if there exist tuples $t_x \in QI_x$ and $t_y \in QI_y$ such that:
 - they satisfy one of the two conditions in Lemma 5 and
 - swapping them decreases the anatomy perimeter
3. swap t_x and t_y
4. return the resulting anatomized tables

Figure 4: Algorithm for tuple swapping

QI_x and QI_y that may trigger significant decrease in the anatomy perimeter. As a tradeoff, the *Swap* algorithm in Figure 4 does not necessarily perform the best swap, due to the suboptimal nature of the adopted heuristics. Our method is based on an interesting observation:

Lemma 6. *After swapping a pair of tuples between QI_x and QI_y , we have:*

$$PERI(R'_x) + PERI(R'_y) \geq \min PERI(R_x, R_y) \quad (22)$$

where

$$\min PERI(R_x, R_y) = 2 \cdot \sum_{i=1}^d \left(\left| R_x[i_-] - R_y[i_-] \right| + \left| R_x[i_+] - R_y[i_+] \right| \right). \quad (23)$$

Here, $R_x[i_-]$ ($R_x[i_+]$) is the coordinate of the left (right) end point of the projection of R_x on the i -th dimension, and $R_y[i_-]$ ($R_y[i_+]$) has the same meaning with respect to R_y .

We set QI_x to a random QI-group whose perimeter is among the top 10% of all the QI-groups. This choice is intuitive — swapping tuples between QI-groups with small MBRs is not likely to lower the anatomy perimeter considerably. Next, we decide QI_y by “farthest neighbor” search. Specifically, for every QI-group QI different from QI_x , we define its “distance” from QI_x as:

$$\max \Delta(QI_x, QI) = (PERI(R_x) + PERI(R)) - \min PERI(R_x, R) \quad (24)$$

where R is the MBR of QI , and $\min PERI(R_x, R)$ given in Equation 23. QI_y is chosen as the QI that maximizes $\max \Delta(QI_x, QI)$, i.e., having the largest distance to QI_x .

7.4 An Improved Anatomizing Algorithm

Summarizing the previous findings, Figure 5 presents *Anatomize**, which outputs a pair of QIT and ST that possess small RCE and anatomy perimeter. Compared with *Anatomize* in Figure 2, *Anatomize** incorporates

Algorithm Anatomize* (T, l)

Input: a microtable T , and a value l ; Output: anatomized tables QIT and ST

Lines 1-5 are the same as Lines 1-5 in Figure 2

7. $QI_{gcnt} = QI_{gcnt} \cup \{t\}$
8. for $i = 2$ to l
9. $t =$ the tuple in the i -th bucket of S whose insertion into QI_{gcnt} causes the least increase in the perimeter of the MBR of QI_{gcnt}
10. $QI_{gcnt} = QI_{gcnt} \cup \{t\}$

Lines 11-13 are the same as Lines 11-13 in Figure 2

14. $QI =$ the existing QI-group in S' that incurs the least increase in the perimeter of its MBR, after including t
15. assign t to QI

Lines 16-21 are the same as Lines 11-13 in Figure 2

22. while time permits
23. $(QIT, ST) = \mathbf{Swap}(QIT, ST)$ /* invoke the algorithm of Figure 4*/
24. return QIT and ST

Figure 5: An enhanced anatomizing algorithm

several heuristics to construct QI-groups with small MBRs, as well as an additional *postprocessing step* which employs tuple swapping to reduce the anatomy perimeter.

As with *Anatomize*, *Anatomize** first executes a group-creation step, followed by a residue-assignment step. Recall that, in each iteration of the group-creation step, *Anatomize* forms a new QI-group by randomly extracting tuples from the l largest buckets. In *Anatomize**, tuple extraction is performed in a more deterministic manner by Lines 6-10 in Figure 5. Specifically, we first remove an arbitrary tuple t from the first bucket in S (the set of the l largest buckets), and insert t in the new QI-group QI_{gcnt} (Lines 6-7). Then, from all the tuples in the second bucket in S , we add to QI_{gcnt} the tuple t whose insertion causes the least perimeter increase of QI_{gcnt} . Next, the same process (Lines 8-10) is carried out for the third, fourth, ... l -th buckets in S , at which point QI_{gcnt} is complete with l tuples.

In the residue-assignment phase, *Anatomize** places each residue tuple t in an existing QI-group. For this purpose, (as with *Anatomize*) we first collect the set S' of QI-groups, which do not contain any tuple having the same sensitive value as t (Line 13 of Figure 5). Then, for each QI-group in S' , we calculate the increase of its perimeter, if t is included in the group. Then, t is assigned to the QI-group demanding the smallest increase (Lines 14-15). After all the residue tuples have been assigned, Lines 16-21 populate QIT and ST following Definition 2.

The postprocessing step (Lines 22-23) of *Anatomize** is executed in iterations, each of which swaps a pair of tuples between two QI-groups. This step is “time responsive”, because it keeps running until a time limit specified by the publisher. *Anatomize** does not guarantee returning the optimal anatomized tables with the

minimum anatomy perimeter. However, the longer it is allowed to run, the lower the anatomy perimeter becomes.

Example 2. We demonstrate the algorithm on the microdata in Table 1, setting l to 3. As with *Anatomize* in Example 1, *Anatomize** initializes 5 buckets: $B_{\text{pneu}} = \{2, 6\}$, $B_{\text{gast}} = \{4, 7\}$, $B_{\text{dysp}} = \{1\}$, $B_{\text{flu}} = \{3\}$, $B_{\text{bron}} = \{5\}$. Then, it proceeds to produce the first QI-group QI_1 . For this purpose, (still same as *Anatomize*) *Anatomize** fetches $S = \{B_{\text{pneu}}, B_{\text{gast}}, B_{\text{flu}}\}$. Now, it randomly picks a tuple, say tuple 2, from B_{pneu} , and adds it to QI_1 . Next, it chooses a tuple in B_{pneu} for inclusion in QI_1 . Unlike *Anatomize*, however, the choice is not random, but aims at minimizing the perimeter of QI_1 . Specifically, between the two tuples in B_{pneu} , tuple 4 is selected, because the MBR of tuples 2 and 4 is much smaller than that of tuples 2 and 6. *Anatomize** continues by inspecting B_{flu} . As B_{flu} has only one tuple, it is added to QI_1 , which becomes $\{2, 3, 4\}$. Accordingly, $B_{\text{pneu}} = \{6\}$, $B_{\text{gast}} = \{7\}$, $B_{\text{flu}} = \emptyset$. Computation of the second QI-group QI_2 is similar. Assuming that this time $S = \{B_{\text{bron}}, B_{\text{pneu}}, B_{\text{gast}}\}$, we have $QI_2 = \{5, 6, 7\}$. The group-creation phase is completed.

Given the residue tuple 1 in B_{dysp} , *Anatomize** determines $S' = \{QI_1, QI_2\}$, following the same reasoning as *Anatomize*. Next, the algorithm decides a QI-group in S' that tuple 1 should be assigned to. Different from *Anatomize*, the decision is deterministic: QI_1 is chosen because, compared to QI_2 , it incurs less perimeter increase after incorporating tuple 1. After updating QI_1 to $\{1, 2, 3, 4\}$, the residue-assignment step terminates. Based on QI_1 and QI_2 , *Anatomize** produces the QIT and ST in Table 3.

*Anatomize** enters the postprocessing stage. It finds out that tuples 3 and 5 (in QI_1 and QI_2 respectively) constitute a legal and beneficial swap. After exchanging them, the QIT and ST become Tables 8a and 8b. Suppose that the time limit set by the published has been exhausted. *Anatomize** returns Tables 8a and 8b as the final result.

Time Complexity. The analysis of the running time of *Anatomize** is similar to that of *Anatomize*, presented in the proof of Theorem 2. There are two major differences. First, each while-loop (Lines 4-10) here consumes $O(n + l \log \lambda)$ time. This change is caused by the fact that, decision of the tuple t at Line 9 requires scanning a bucket, such that l buckets must be scanned to obtain a QI_{gcnt} , which incurs $O(n)$ time. Since there are $O(n/l)$ loops, the group-creation step costs totally $O(n^2/l + n \log \lambda)$ overhead, which is also the time complexity before the postprocessing step.

Second, we also need to discuss the overhead of postprocessing, which repetitively executes the *Swap* algorithm in Figure 4, until the time limit is exhausted. Specifically, *Swap* includes two parts: *Choose-QI-Groups*

(Line 1 in Figure 4) and swapping between QI_x, QI_y (Lines 2-3). Recall that, *Choose-QI-Groups* first decides QI_x as a random one of the QI-groups whose perimeters are among the top 10% of all the QI-groups. As there are $O(n/l)$ QI-groups, this can be done in $O(n/l)$ time, provided that we have a binary tree on the perimeters of all QI-groups. Then, QI_y is determined as the QI-group QI that maximizes Equation 24. Evaluating the equation costs $O(1)$ overhead (treating d as a constant). Thus, QI_y can be found by examining all QI-groups in $O(n/l)$ time. It follows that *Choose-QI-Groups* demands $O(n/l)$ overhead. Tuple swapping between QI_x and QI_y can be accomplished in $O(l^3)$ time. To understand this, notice that both QI_x and QI_y have $O(l)$ tuples, which result in $O(l^2)$ pairs of tuples that may be swapped. For each pair, checking whether swapping is legal and beneficial can be easily achieved in $O(l)$ cost. After a swap has been done, $O(\log(n/l))$ cost is necessary to update the binary tree on the perimeters of the QI-groups. Hence, the overall complexity of *Swap* is $O(n/l + l^3)$.

8 Experiments

This section empirically evaluates the effectiveness and efficiency of anatomy. Section 8.1 first clarifies the settings of our experiments. Then, Section 8.2 tunes the postprocessing time for *Anatomize**. Next, Section 8.3 compares anatomy to other anonymization solutions in their effective for data analysis. Finally, Section 8.4 investigates their computation overhead.

8.1 Experiment Setup

We employ a real-world dataset CENSUS, downloadable at <http://www.ipums.org>, which contains the personal information of 500k American adults. The dataset has 9 discrete attributes summarized in Table 9, whose domains are normalized to a unit range $[0, 1]$. From CENSUS, we create two sets of microtables, 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 9 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 . Furthermore, to study the impact of cardinality, we generate multiple versions of each OCC- d (SAL- d) with various cardinalities n , by randomly sampling n tuples from the “full” OCC- d (SCC- d) with 500k tuples.

We compare anatomy against generalization, randomized response (RR) [38], and approximate marginal-preserving swapping (aMPS) [34]. For generalization, we utilize the state-of-the-art algorithm in [22], which

Attribute	Number of distinct values	Generalization method (inapplicable to anatomy)
<i>Age</i>	78	Free interval
<i>Gender</i>	2	Taxonomy tree (2)
<i>Education</i>	17	Free interval
<i>Marital</i>	6	Taxonomy tree (3)
<i>Race</i>	9	Taxonomy tree (2)
<i>Work-class</i>	8	Taxonomy tree (4)
<i>Country</i>	83	Taxonomy tree (3)
<i>Occupation</i>	50	NA (sensitive)
<i>Salary-class</i>	50	NA (sensitive)

Table 9: Summary of attributes

adopts multi-dimension recoding. Recall that each generalized QI value is an interval. The last column of Table 9 describes how these intervals are formed. 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, means that a generalized value corresponds to a node in a taxonomy with height x .

RR is a popular input perturbation technique in statistical databases (also called *random perturbation* in [5, 13]). For each tuple t in the microdata, RR computes its published version t' as follows. First, t' and t have identical QI-values. Second, the sensitive value of t' equals that of t with probability p , or is randomly generated in the domain of the sensitive attribute with probability $1 - p$, where p is a parameter called the *retention probability* (the greater p is, the less information loss). In our experiments, the retention probability p is set to the highest value that promises the degree of protection required by l -diversity⁷. aMPS is chosen as a representative of the existing data-swapping approaches. However, unlike the solutions described earlier, aMPS does not ensure the privacy control demanded by l -diversity (nor does any of the previous data-swapping methods, as mentioned in Section 2.2).

For OCC- d microtables, l distributes from 2 to 12, whereas l ranges from 2 to 20 for SAL- d datasets. Note that, regardless of d , 12 (20) is already the largest possible l for OCC- d (SAL- d). This is because $1/l$ must be at least the percentage of tuples having the most frequent sensitive value (the percentage equals 4.84% and 7.85% for OCC- d and SAL- d , respectively). Otherwise, no l -diverse partition exists.

The utility of an anonymization technique is evaluated in its accuracy of answering count queries of the form:

⁷RR offers the so-called ρ_1 -to- ρ_2 *guarantee* [13]. Such a guarantee demands that, if an adversary can correctly guess the sensitive value of a victim with only probability ρ_1 before examining an anonymized dataset, her/his probability of doing so should be bounded by ρ_2 after the examination. Under our assumption of an adversary’s prior knowledge stated in Section 3.1, ρ_1 equals $1/\lambda$, with λ being the domain size of the sensitive attribute, whereas ρ_2 needs to be $1/l$, for fulfilling the requirement of l -diversity.

Parameter	Value
l	2, 5, 10 , 12, 15, 20
cardinality n	100k , 200k, 300k, 400k, 500k
number of QI-attributes d	3, 4, 5 , 6, 7
query dimensionality qd	1, 2, ..., d
expected selectivity s	0.25%, 0.5%, 1% , 2%, 4%

Table 10: Parameters and tested values

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

Specifically, a query concerns qd random QI-attributes $A_1^{qi}, \dots, A_{qd}^{qi}$, 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 “ $A \in [x, y]$ ”. Here, $[x, y]$ is an interval randomly generated in the domain of A , and its length $y - x$ equals

$$|A| \cdot s^{1/(qd+1)} \quad (25)$$

where $|A|$ is the number of distinct values in the domain of A (see Table 9), and s a query parameter called *expected query selectivity*. A higher s leads to longer query ranges, and hence, larger results.

A *workload* consists of 10000 queries (with the same qd and s). The effectiveness of a method is measured as its *average relative error* in answering a workload. Specifically, for each query, the relative error equals $|act - est|/act$, where act is the actual result from the microdata, and est the estimate derived from the underlying anonymization approach. Specifically, for anatomy, est is calculated by Equation 15. For generalization, est is obtained in the manner exemplified in Section 1.1. For RR, computation of est follows the analysis in [5], whereas for aMPS, est is simply the query result on the anonymized dataset.

Table 10 summarizes the parameters of our experimentation, as well as their values examined. The values in bold are the defaults. Unless specifically stated, each parameter is set to its default value in the following experiments. All the reported results are obtained on a computer with a 3.4 GHz Pentium IV CPU and one-gigabyte memory.

8.2 Tuning the Postprocessing Time of *Anatomize**

Remember that *Anatomize** involves a postprocessing step, which iteratively performs tuple swapping to reduce the anatomy perimeter, until the permitted amount of time has been exhausted. The experiments

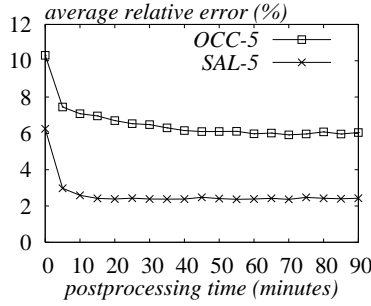


Figure 6: Query accuracy vs. postprocessing time of *Anatomize** ($l = 10, d = 5, qd = 5, n = 100k, s = 1\%$)

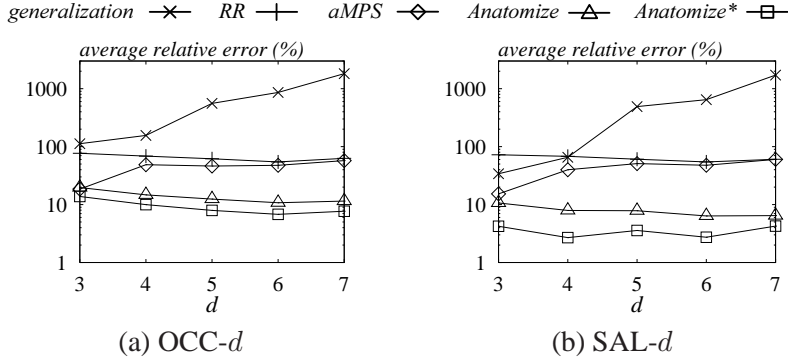


Figure 7: Query accuracy vs. number d of QI-attributes ($l = 10, n = 100k, qd = d, s = 1\%$)

in this section aim at deciding an appropriate time limit. For this purpose, given a microtable, we allow *Anatomize** to run continuously. Every 5 minutes after the postprocessing phase has started, we measure the accuracy of using the current anatomized tables to answer a workload with the default parameter values ($qd = d, s = 1\%$). Figure 6 plots the query error as a function of the elapsed time, for the microtables OCC-5 and SAL-5 with cardinality 100k. In both cases, the error drops drastically during the first 10 minutes of postprocessing, after which the rate of decreasing becomes considerably slower. This observation proves the effectiveness of the *Choose-QI-Groups* procedure in Section 7.3, i.e., it is able to identify two QI-groups whose tuple exchanging leads to large reduction of the anatomy perimeter. In the subsequent experiments, we set the time limit to 10 minutes.

8.3 Effectiveness of Data Analysis

This section compares alternative solutions on their accuracy in count analysis. For anatomy, we evaluate the performance of both *Anatomize* and *Anatomize**, proposed in Sections 5 and 7, respectively. The first set of experiments examines the effects of d . Figure 7a (7b) plots the error of all methods as a function of d , for datasets OCC- d (SAL- d). Apparently, generalization incurs by far the largest error, and its accuracy decays severely as the dimensionality increases, confirming the analysis of [2]. RR also entails huge error. aMPS

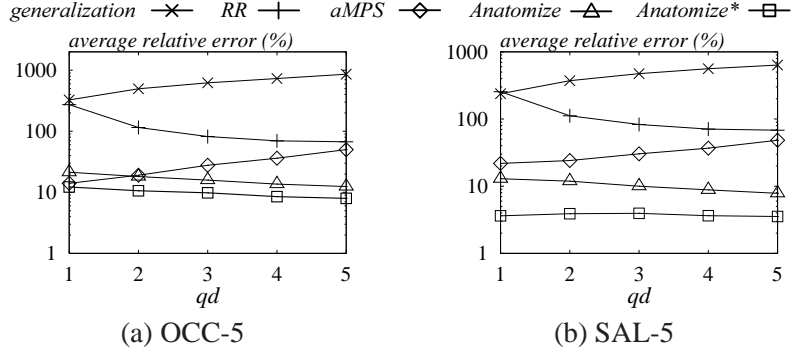


Figure 8: Query accuracy vs. query dimensionality qd ($l = 10$, $n = 100k$, $s = 1\%$)

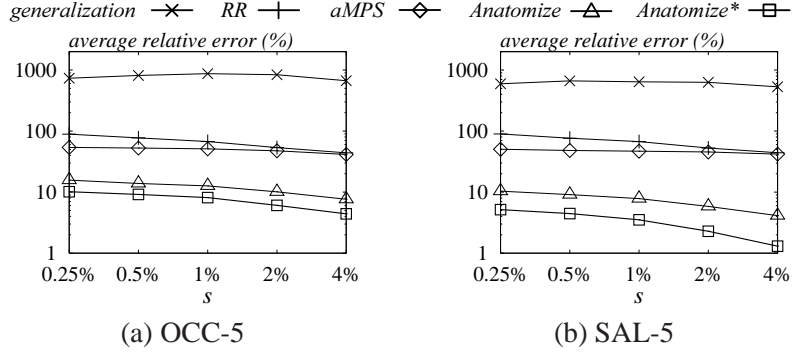


Figure 9: Query accuracy vs. expected query selectivity s ($l = 10$, $n = 100k$, $qd = d$)

works well for small d , but considerably deteriorates at high dimensionalities. This is expected because, as d grows, the total number of marginals escalates exponentially, and hence, the quality of marginal preservation drops significantly. Both *Anatomize* and *Anatomize** demonstrate excellent performance regardless of the dimensionality, and outperform the other competitors by an order of magnitude for $d \geq 5$.

Next, we examine alternative techniques on queries involving different number qd of QI-attributes. Figure 8 demonstrate the results for OCC-5 and SAL-5. Again, the error of anatomy is consistently small, whereas generalization and RR are again the worst techniques. aMPS achieves good accuracy only when qd is small, due to the fact that marginals of low orders are better preserved (compared to those of high orders). Figure 9 present the error as a function of query selectivity s . The relative superiority of the five approaches remains the same as before, except that their performance generally improves as s becomes larger.

Figure 10 examines the accuracy as l varies. As expected, the error of anatomy, generalization, and RR increases with l , because more data distortion is needed in order to enforce stronger privacy control. The performance of aMPS is unaffected, since it is totally independent from l (recall that aMPS does not guarantee the privacy protection required by l -diversity). Finally, Figure 11 presents the results under different dataset cardinalities n . The effectiveness of each method remains fairly stable at all cardinalities, except

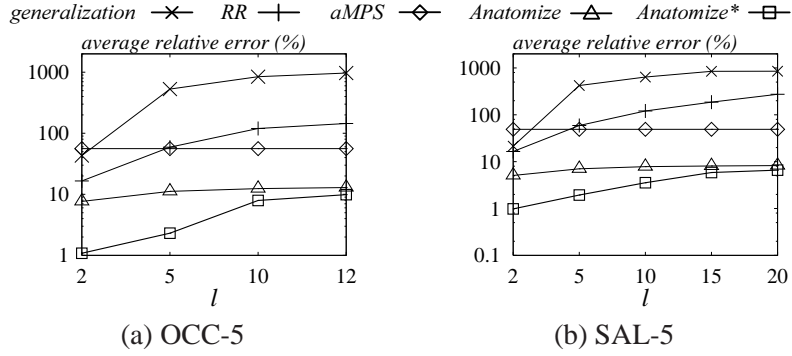


Figure 10: Query accuracy vs. l ($n = 100k$, $d = 5$, $qd = 5$, $s = 1\%$)

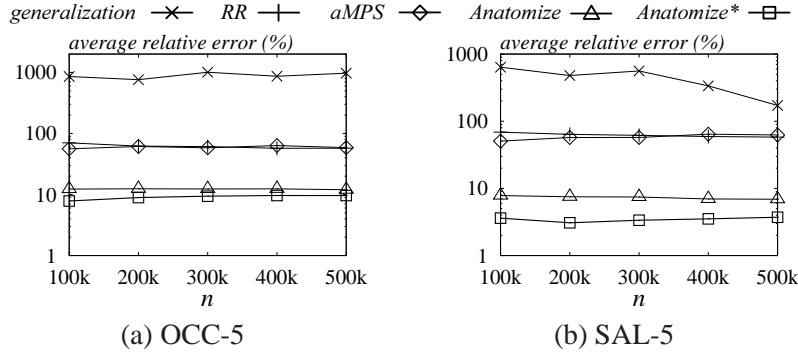


Figure 11: Query accuracy vs. dataset cardinality n ($l = 10$, $d = 5$, $qd = 5$, $s = 1\%$)

that, in Figure 11b, the error of generalization decreases to around 200% at the largest n .

In summary, we have shown that anatomy allows very accurate counting analysis, and its error is lower than the other solutions by a factor up to an order of magnitude. Furthermore, its performance is not influenced by data and query dimensionalities. Generalization and RR are ineffective for counting analysis, because they incur more than 100% error in most cases. aMPS has acceptable performance, only when the dataset or query dimensionality is small.

8.4 Computation Cost

Having demonstrated the effectiveness of anatomy for data analysis, we proceed to evaluate its efficiency. Figure 12a (12b) compares the time of anonymization required by *Anatomize*, *Anatomize**, generalization, RR and aMPS, on the OCC-5 (SAL-5) microtables with different cardinalities n . For *Anatomize**, we include only the cost of its group-creation and residue-assignment phases, since its postprocessing time is a user parameter. RR and aMPS are the most efficient algorithms, but their efficiency does not justify the huge error they entail in query processing. *Anatomize** is the slowest, because it incorporates expensive heuristics to reduce the anatomy perimeter. Nevertheless, the extra overhead pays off because, as shown in the previous

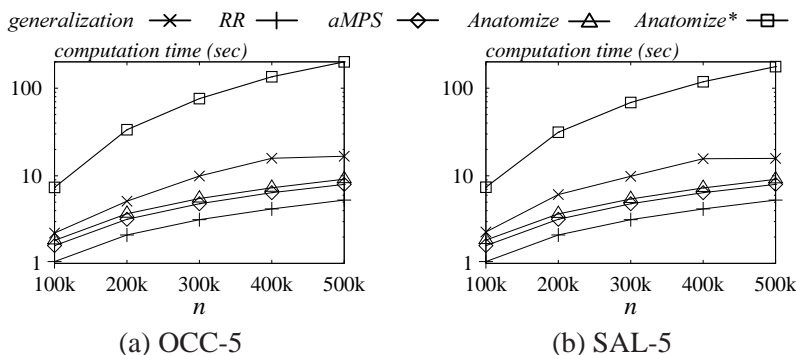


Figure 12: CPU cost vs. dataset cardinality n ($l = 10$, $d = 5$)

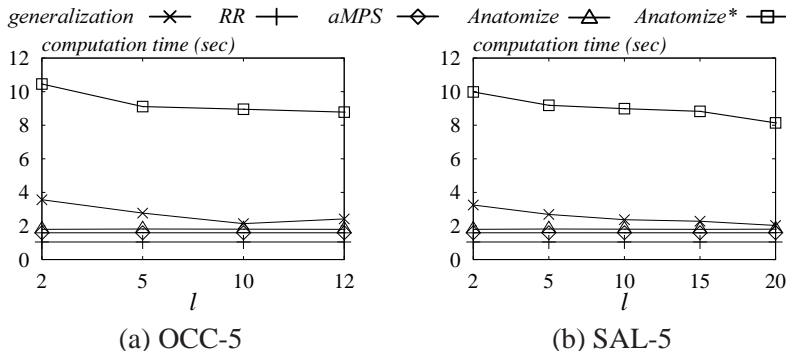


Figure 13: Computation cost vs. l ($n = 300k$, $d = 5$)

subsection, *Anatomize** permits the most accurate counting. In any case, all algorithms terminate within 4 minutes, even on the largest dataset. Figure 13 plots the computation time as a function of l , confirming the relative efficiency of alternative solutions observed in Figure 12. In both Figures 12 and 13, the behavior of *Anatomize* and *Anatomize** is consistent with the complexity analysis in Sections 5.2 and 7.4, respectively.

9 Conclusions

The existing anonymization techniques are inadequate because they have at least one of the following defects: they (i) do not provide sufficient privacy protection against linking attacks, or (ii) lose considerable information in the microdata, and thus, prohibit effective data analysis. This article develops the anatomy technique that remedies both defects. Specifically, anatomy ensures strong privacy control as demanded by l -diversity, and meanwhile, retains a significant amount of data correlation. Extensive experiments confirm that anatomy permits highly accurate statistical studies.

This work also initiates several directions for future investigation. For example, in this article, we focused on the case where there is a single sensitive attribute; extending our technique to multiple sensitive attributes is an interesting topic. Another direction concerns “workload-aware anatomy”, where the publisher is given

a sample workload of queries [20], and the objective is to compute anatomized tables that minimize the error of those queries. Finally, it would be useful to study how anatomy can be utilized for discovering complex patterns in the microdata, perhaps through minimization of specialized metrics for quantifying information loss (e.g., the classification metric [17]).

References

- [1] N. R. Adam and J. C. Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 21(4):515–556, 1989.
- [2] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proc. of Very Large Data Bases (VLDB)*, pages 901–909, 2005.
- [3] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *Proc. of International Conference on Database Theory (ICDT)*, pages 246–258, 2005.
- [4] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proc. of ACM Management of Data (SIGMOD)*, pages 207–216, 1993.
- [5] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving olap. In *Proc. of ACM Management of Data (SIGMOD)*, pages 251–262, 2005.
- [6] G. Arfken and H. Weber. *Mathematical Methods for Physicists*. Academic Press, 4th edition, 1995.
- [7] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 217–228, 2005.
- [8] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 128–138, 2005.
- [9] T. Dalenius and S. P. Reiss. Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6(1):73–85, 1982.
- [10] N. N. Dalvi and D. Suciu. Management of probabilistic data: foundations and challenges. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 1–12, 2007.
- [11] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 202–210, 2003.
- [12] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, pages 265–284, 2006.
- [13] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 211–222, 2003.
- [14] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by dalenius and reiss. *Journal of Official Statistics*, 21(2):309–323, 2005.
- [15] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 205–216, 2005.

- [16] J. M. Gouweleew, P. Kooiman, L. C. R. Willenborg, and P.-P. de Wolf. Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14:463–478, 1998.
- [17] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Proc. of ACM Knowledge Discovery and Data Mining (SIGKDD)*, pages 279–288, 2002.
- [18] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *Proc. of ACM Management of Data (SIGMOD)*, pages 217–228, 2006.
- [19] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [20] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *Proc. of ACM Knowledge Discovery and Data Mining (SIGKDD)*, 2006.
- [21] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k -anonymity. In *Proc. of ACM Management of Data (SIGMOD)*, pages 49–60, 2005.
- [22] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k -anonymity. In *Proc. of International Conference on Data Engineering (ICDE)*, 2006.
- [23] N. Li, T. Li, and S. Venkatasubramanian. t -closeness: Privacy beyond k -anonymity and l -diversity. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 106–115, 2007.
- [24] C. K. Liew, U. J. Choi, and C. J. Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems (TODS)*, 10(3):395–411, 1985.
- [25] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian. l -diversity: Privacy beyond k -anonymity. In *Proc. of International Conference on Data Engineering (ICDE)*, 2006.
- [26] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy-preserving data publishing. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 126–135, 2007.
- [27] A. Meyerson and R. Williams. On the complexity of optimal k -anonymity. In *Proc. of ACM Symposium on Principles of Database Systems (PODS)*, pages 223–228, 2004.
- [28] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1st edition, 1997.
- [29] R. A. Moore. Controlled data-wapping techniques for masking public use microdata sets. *Statistical Research Division Report Series*, pages RR96–04, 1996.
- [30] S. U. Nabar, B. Marthi, K. Kenthapadi, N. Mishra, and R. Motwani. Towards robustness in query auditing. In *Proc. of Very Large Data Bases (VLDB)*, pages 151–162, 2006.
- [31] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *Proc. of ACM Management of Data (SIGMOD)*, pages 665–676, 2007.
- [32] H. Park and K. Shim. Approximate algorithms for k -anonymity. In *Proc. of ACM Management of Data (SIGMOD)*, pages 67–78, 2007.
- [33] C. Ré, N. N. Dalvi, and D. Suciu. Efficient top- k query evaluation on probabilistic data. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 886–895, 2007.

- [34] S. P. Reiss. Practical data-swapping: The first steps. *ACM Transactions on Database Systems (TODS)*, 9(1):20–37, 1984.
- [35] P. Samarati. Protecting respondents’ identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6):1010–1027, 2001.
- [36] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5):557–570, 2002.
- [37] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *Proc. of International Conference on Management of Data (ICDM)*, pages 249–256, 2004.
- [38] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.
- [39] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *Proc. of ACM Knowledge Discovery and Data Mining (SIGKDD)*, pages 754–759, 2006.
- [40] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proc. of Very Large Data Bases (VLDB)*, pages 139–150, 2006.
- [41] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proc. of ACM Management of Data (SIGMOD)*, pages 229 – 240, 2006.
- [42] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In *Proc. of ACM Knowledge Discovery and Data Mining (SIGKDD)*, pages 785–790, 2006.
- [43] C. Yao, X. S. Wang, and S. Jajodia. Checking for k-anonymity violation by views. In *Proc. of Very Large Data Bases (VLDB)*, pages 910–921, 2005.

Appendix

Proof of Theorem 1. Consider an arbitrary individual o in the microtable T . Let v_o be the sensitive value of o . Without loss of generality, assume that f tuples in T , denoted as t_1, t_2, \dots, t_f , have the same QI-values as o . From the adversary’s perspective, each of these f tuples has $1/f$ probability of belonging to o . Let p_i be the probability that the adversary conjectures the sensitive value of t_i to be v_o , subject to the condition that s/he takes t_i as the tuple owned by o . Then, the adversary correctly infers the sensitive value of o with an overall probability $\frac{1}{f} \sum_{i=1}^f p_i$. Without loss of generality, assume that t_1 belongs to a QI-group QI . The adversary may obtain the Group-ID j of QI from the QIT, which, however, does not contain any A^s data. Consequently, the adversary can only conjecture that, $t_1[d + 1]$ has equal chance to be any of the multi-set of sensitive values (summarized in the ST) contained in QI . Therefore, from the adversary’s perspective, $p_1 = c(v_o)/|QI| \leq 1/l$, where $c(v_o)$ is the number of tuples in QI with a sensitive value v_o . By symmetry, we have $p_i \leq 1/l$ for any $i \in [1, f]$. This leads to $\frac{1}{f} \sum_{i=1}^f p_i \leq 1/l$. \square

Proof of Lemma 1. Assume that the QIT and ST are produced by a partition with m QI-groups QI_1, \dots, QI_m . For each $j \in [1, m]$, we use α_j to denote the average E_t (Formula 9) for all tuples $t \in QI_j$.

Then, RCE can be rewritten as $RCE = \sum_{j=1}^m (|QI_j| \cdot \alpha_j)$. The rest of the proof will show that $\alpha_j \geq (1 - 1/l)^p + (l - 1)(1/l)^p$, for all $j \in [1, m]$. As a result, the above equation leads to

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

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

Consider any $j \in [1, m]$. Without loss of generality, assume that QI_j contains λ distinct A^s values v_1, \dots, v_λ , and there exist $c(v_h)$ tuples in QI_j with sensitive value v_h ($1 \leq h \leq \lambda$). Consider an arbitrary tuple $t \in QI_j$ with A^s value v_h (for some $h \in [1, \lambda]$). The actual pdf \mathcal{G}_t and its approximation $\tilde{\mathcal{G}}_t^a$ are given in Equations 6 and 8, respectively. Thus, by Equation 9, we have

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

Taking into account all tuples in QI_j , we obtain

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

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

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

(the second constraint is due to Definition 1).

Let us ignore the second constraint ($c(v_h) \leq |QI_j|/l$) temporarily. Then, minimization of α subject to the first constraint is a standard problem tackled by the *Lagrange multiplier method* [6]. Application of the method results in $\alpha \geq (1 - 1/\lambda)^p + (\lambda - 1)(1/\lambda)^p$, where the equality holds only when $c(v_1) = \dots = c(v_h) = |QI_j|/l$. Now, we take into account the second constraint, which leads to $\sum_{h=1}^{\lambda} c(v_h) \leq \lambda \cdot |QI_j|/l$. The left side of the inequality equals $|QI_j|$. Hence, the inequality implies $\lambda \geq l$. Therefore, $\alpha \geq (1 - 1/\lambda)^p + (\lambda - 1)(1/\lambda)^p \geq (1 - 1/l)^p + (l - 1)(1/l)^p$. \square

Proof of Property 1. An l -diverse partition exists, if and only if T satisfies an *eligibility condition*⁸ [25]: at most n/l tuples have the same A^s value, where n is the number of tuples in T . We will show that Property 1 always holds under this condition.

Assume on the contrary that, after the group-creation phase, the largest bucket B has $x \geq 2$ tuples. Let y be the total number of iterations. As precisely l tuples are removed in each iteration, after all iterations, the number of tuples in the non-empty buckets equals $n - y \cdot l$. There are at most $l - 1$ such buckets; thus $n - y \cdot l \leq x \cdot (l - 1)$, leading to

$$x + y = (x \cdot l + y \cdot l)/l > (x \cdot (l - 1) + y \cdot l)/l \geq n/l.$$

The rest of the proof will establish a fact Z : B is among the l most sizable buckets before each iteration. It implies that a tuple is deleted from B in each iteration, indicating that B originally has $x + y > n/l$ tuples. This violates the eligibility condition, and hence, $x \geq 2$ cannot be true.

⁸If 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 $1/l$.

In fact, Z is not difficult to observe. Since the l -th largest bucket after the last iteration includes 0 tuple, the l -th largest bucket before the iteration contains exactly 1 tuple. As B has 2 tuples even after the iteration, it is definitely one of the l largest buckets before the last iteration. In the same way, it is easy to prove that B is also among the l largest, before all iterations. \square

Proof of Property 2. Assume, on the contrary, that S' is empty when processing tuple t (at Line 11 in Figure 2). 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. \square

Proof of Property 3. 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 holds. \square

Proof of Theorem 2. The hashing at Line 2 in Figure 2 takes $O(n)$ time. In the group-creation phase, we keep a binary tree on the sizes of the λ buckets. Using this tree, the set S at Line 5 can be decided in $O(l)$ time. Lines 6-8 require $O(l)$ cost. After these lines, $O(l \log \lambda)$ cost is needed in order to update bucket counters in the binary tree. Therefore, each while-loop (Lines 4-8) consumes $O(l \log \lambda + l)$ cost. Since there are totally $O(n/l)$ loops, the total complexity of the group-creation step is $O(n \log \lambda + n)$.

For each residue tuple t , the set S' at Line 11 can be collected by scanning all the existing QI-groups once in $O(n)$ time. Line 12 incurs $O(1)$ overhead. Since there are at most $l - 1$ residue tuples, the residue-assignment phase completes in $O(l \cdot n)$ time. Finally, populating the QIT and ST in Lines 13-18 can obviously be accomplished in $O(l \cdot n)$ time, thus establishing Theorem 2. \square

Proof of Theorem 3. Let $r = n \bmod l$. We distinguish two cases.

Case 1 ($r = 0$): *Anatomize* stops immediately after the group-creation phase. Hence, each QI-group contains exactly l tuples with distinct A^s values. Combining Equations 6, 8, and 9, we have, for each tuple $t \in T$, $E_t = (1 - 1/l)^p + (l - 1)(1/l)^p$. By Equation 10, $RCE = n(1 - 1/l)^p + n(l - 1)(1/l)^p$.

Case 2 ($r \neq 0$): Let us examine the moment when the group-creation phase finishes. At this point, $n - r$ (a multiple of l) tuples have been placed in QI-groups. The sum of E_t of the tuples t already in QI-groups equals $(n - r)(1 - 1/l)^p + (n - r)(l - 1)(1/l)^p$.

Next, each residue tuple is assigned to an existing QI-group in turn. Suppose, without loss of generality, that the i -th residue tuple t_i is assigned to a QI-group QI with β_i tuples. Notice that the β_i tuples have distinct A^s values, all of which are different from $t_i.A^s$. Before the assignment of t_i , the RCE of QI (i.e., the sum of E_t for all tuples $t \in QI$) equals $f(\beta_i)$, where function $f(\cdot)$ is defined as:

$$f(x) = x \cdot (1 - 1/x)^p + x \cdot (x - 1)(1/x)^p.$$

After t_i is included in QI , the new RCE of QI becomes $f(\beta_i + 1)$. As a result, the overall RCE (of the entire anatomy) increases by $\delta(\beta_i)$, where function $\delta(\cdot)$ is formulated as

$$\delta(x) = f(x + 1) - f(x).$$

Hence, after assigning all the residue tuples, the final RCE equals

$$RCE = (n-r)(1-1/l)^p + (n-r)(l-1)(1/l)^p + \sum_{i=1}^r \delta(\beta_i)$$

Let $f'(x)$ and $f''(x)$ be the first and second order derivatives of $f(x)$, respectively. Then,

$$f''(x) = \frac{p-1}{x^{p+1}} \left[p \cdot ((x-1)^{p-2} - 1) + (p-2) \cdot x \right].$$

Notice that the two integers p and x are at least 1 and 2, respectively (in particular, $x \geq 2$ is due to the fact that β_i is at least 2 for a meaningful $l \geq 2$). Hence, $f''(x)$ is always non-negative. This means that $f'(x)$ is non-decreasing as x increases. Since $\delta(x) = \int_x^{x+1} f'(x)dx$, $\delta(x)$ is also non-decreasing as x becomes larger.

Before the i -th residue tuple is assigned, the largest existing QI-group has at most $l+i-1$ tuples. Hence, β_i is at most $l+i-1$, leading to $\delta(\beta_i) \leq f(l+i) - f(l+i-1)$. As a result, from Equation 26, we obtain

$$\begin{aligned} RCE &\leq n \cdot (1-1/l)^p + n \cdot (l-1)/l^p + \sum_{i=1}^r [f(l+i) - f(l+i-1)] \\ &= n \cdot (1-1/l)^p + n \cdot (l-1)/l^p + f(l+r) - f(l) \\ (\text{By } r \leq l-1) &\leq n \cdot (1-1/l)^p + n \cdot (l-1)/l^p + f(2l-1) - f(l) \end{aligned}$$

Dividing both sides of the above inequality by the lower bound $n(1-1/l)^p + n(l-1)/l^p$ in Lemma 1, the right hand side can be simplified to $1 + \frac{2l-1}{n} \cdot \phi$, where ϕ is given in Equation 11. \square

Proof of Lemma 2. Consider an arbitrary QI-group QI which contains λ distinct A^s values v_1, \dots, v_λ . Let t be any tuple in QI , with a sensitive value v_h , for some $h \in [1, \lambda]$. Plugging Equations 6 and 8 into Equation 12, we obtain

$$E_t = \log(|QI|/c(v_h)).$$

Since QI is l -diverse, by Inequality 1, $c(v_h) \leq |QI|/l$, leading to $E_t \geq \log l$. Therefore, $RCE = \sum_{t \in T} E_t \geq n \log l$. \square

Proof of Theorem 4. Introducing $r = n \bmod l$, we examine two cases.

Case 1 ($r = 0$): *Anatomize* terminates by returning a partition with n/l QI-groups, each of which has l tuples with distinct sensitive values. By Equations 6, 8, and 12, we have $E_t = \log l$ for each tuple $t \in T$, which leads to $RCE = n \log l$.

Case 2 ($r \neq 0$): When the group-creation step terminates, *Anatomize* has constructed $(n-r)/l$ QI-groups. Their total RCE equals $(n-r) \cdot \log l$. Assume that, in the residue-assignment step, the i -th residue tuple t_i is assigned to a QI-group QI with β_i tuples. Before and after the assignment, the RCE of QI is $\beta_i \log \beta_i$ and $(\beta_i + 1) \log(\beta_i + 1)$, respectively. Hence, the total RCE of all the QI-groups increases by $\delta(\beta_i)$, where function $\delta(\cdot)$ is formulated as:

$$\delta(x) = (x+1) \cdot \log(x+1) - x \cdot \log x.$$

The derivative of $\delta(x)$ equals $\log(x+1) - \log x > 0$; therefore, $\delta(x)$ monotonically increases with x .

Before the i -th residue tuple is assigned, $\beta_i \leq l + i - 1$. As a result, at the end of *Anatomize*, we have

$$\begin{aligned} RCE &= (n - r) \cdot \log l + \sum_{i=1}^r \delta(\beta_i) \\ (\text{By } r \leq l - 1) &\leq n \cdot \log l + (2l - 1) \cdot \log(2l - 1) - l \cdot \log l. \end{aligned}$$

The right hand side of the above inequality is greater than the lower bound $n \log l$ in Lemma 2, by a factor of $1 + \frac{(2l-1) \cdot \log(2l-1)}{n \cdot \log l}$. \square

Proof of Lemma 3. Consider the partition P of T that produces the QIT and ST. Let m_1 be the number of QI-groups in P with exactly l tuples. Then, for any of the $m_1 \cdot l$ tuples t in those groups, $E_t = l$, while for any tuple that is not, $E_t \geq l + 1$. If $m_1 \leq (n - r)/l - r$,

$$RCE \geq m_1 \cdot l^2 + (n - m_1 \cdot l) \cdot (l + 1) \geq (n + r) \cdot l + r.$$

The rest of the proof focuses on the case when $m_1 > (n - r)/l - r$. Let S be a set containing $(n - r)/l - r$ QI-groups, which are randomly selected from the m_1 QI-groups with size l . The sum of E_t for all the tuples $t \in S$ equals $((n - r)/l - r) \cdot l^2 = (n - r) \cdot l - r \cdot l^2$. In the sequel, we will show that the sum of E_t for the tuples $t \in T - S$ is at least $r \cdot (l + 1)^2$. Hence, the total RCE is lower bounded by $(n - r) \cdot l - r \cdot l^2 + r \cdot (l + 1)^2 = (n + r) \cdot l + r$, which completes the proof.

The number of tuples in $T - S$ equals $n - ((n - r)/l - r) \cdot l = r \cdot (l + 1)$. Suppose that the tuples in $T - S$ is contained in m_2 QI-groups, whose sizes are x_1, x_2, \dots, x_{m_2} , respectively. The sum of the E_t of all the tuples $t \in T - S$ is $\sum_{j=1}^{m_2} x_j^2$. We employ the Lagrange multiplier method to derive the minimum of $\sum_{j=1}^{m_2} x_j^2$, subject to $\sum_{j=1}^{m_2} x_j = r \cdot (l + 1)$, and $x_j \geq l$ for all $1 \leq j \leq m_2$. This results in $\sum_{j=1}^{m_2} x_j^2 \geq r \cdot (l + 1)^2$. \square

Proof of Theorem 5. Let $r = n \bmod l$. We differentiate two cases.

Case 1 ($r = 0$): *Anatomize* completes execution right after the group-creation phase, so that each QI-group has l tuples. By Equations 13 and 10, we have $RCE = n \cdot l$.

Case 2 ($r \neq 0$): The group-creation phase terminates after constructing $(n - r)/l$ QI-groups (each with l tuples), whose total RCE is $(n - r) \cdot l$. Recall that, in the residue-assignment step, each of the r residue tuples is assigned to one of these QI-groups. Later, we will show a fact Z : after the i -th residue tuple is processed, the overall RCE of the QI-groups increases by at most $(l + i)^2 - (l + i - 1)^2$. Hence, the final RCE satisfies:

$$RCE \leq (n - r) \cdot l + \sum_{i=1}^r ((l + i)^2 - (l + i - 1)^2) = (n + r) \cdot l + r^2.$$

Let us divide the two sides of the inequality by the the lower bound $(n + r) \cdot l + r$ in Lemma 3, which yields

$$\frac{RCE}{(n + r) \cdot l + r} \leq 1 + \frac{r^2 - r}{(n + r) \cdot l + r}.$$

The right hand side of the above inequality monotonically increases with r (its derivative is always positive, when n, l , and r are at least 1). Since $r \leq l - 1$, the right hand side takes its maximum $1 + \frac{l^2 - 3l + 2}{n \cdot l + l^2 - 1}$ at $r = l - 1$.

It remains to establish fact Z . Without loss of generality, assume that the i -th residue tuple t_i is allocated to a QI-group QI with β_i tuples. By Equation 13, before and after the allocation, the RCE of QI equals β_i^2

and $(\beta_i + 1)^2$, respectively. Therefore, the overall RCE increases by $(\beta_i + 1)^2 - \beta_i^2$. Before t_i is processed, β_i is at most $l + i - 1$. Hence, $(\beta_i + 1)^2 - \beta_i^2 \leq (l + i)^2 - (l + i - 1)^2$. \square

Proof of Lemma 4. Let V the multi-set of sensitive values in QI_j . Use M to denote the set of possible microdata instances defined by the QIT and ST. Recall that each instance is obtained by independently permuting the sensitive values of each QI-group in P . Let us divide M into disjoint subsets S_1, \dots, S_w , such that all instances in a subset differ only in the j -th QI-group QI_j . Note that S_1, \dots, S_w have the same cardinality, which equals to the total number of permutations of the elements in V .

For any $h \in [1, w]$, let us focus on the set $S_h(x)$ of instances in S_h ($1 \leq h \leq w$), where the j -th QI-group contains precisely x tuples satisfying q . We will show that

$$\frac{|S_h(x)|}{|S_h|} = \begin{cases} \binom{u_j}{x} \binom{n_j - u_j}{v_j - x} / \binom{n_j}{v_j} & \text{if } x \in [\max\{u_j + v_j - n_j, 0\}, \min\{u_j, v_j\}] \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

Then, Equation 20 results from $\mathcal{Q}_j(x) = (\sum_{h=1}^w |S_h(x)|) / (\sum_{h=1}^w |S_h|)$.

Use U to denote the set of tuples in QI_j satisfying $pred^{qi}$. Also, let V_1 be the set of values in V fulfilling $pred(A^s)$, and $V_2 = V - V_1$ (V_1 and V_2 are both multi-sets). Apparently, $|U| = u_j$ and $|V_1| = v_j$. Furthermore, we introduce c_1 (c_2) to represent the number of permutations of the elements in V_1 (V_2), i.e., $c_1 = v_j!$ and $c_2 = (n_j - v_j)!$. It follows that

$$|S_h| = n_j! = \binom{n_j}{v_j} \cdot c_1 \cdot c_2. \quad (27)$$

The rest of the proof will solve $|S_h(x)|$ into a function of n_j , u_j , v_j , c_1 , and c_2 . The function, together with Equation 27, will establish Equation 26.

Each instance in $S_h(x)$ must have been generated by a “feasible permutation” of QI_j satisfying two properties: (i) exactly x tuples in U obtain sensitive values in V_1 after the permutation; (ii) the remaining $u_j - x$ tuples in U acquire sensitive values in V_2 . No such permutation exists, namely $|S_h(x)| = 0$, if $x > \min\{u_j, v_j\}$ (in which case property (i) never holds), or $x < \max\{u_j + v_j - n_j, 0\}$ (in which case property (ii) is always false). For any $x \in [\max\{u_j + v_j - n_j, 0\}, \min\{u_j, v_j\}]$, the number of feasible permutations, also the value of $|S_h(x)|$, equals $\binom{u_j}{x} \binom{n_j - u_j}{v_j - x} \cdot c_1 \cdot c_2$. \square

Proof of Corollary 1. Let X be set of integers in $[\max\{u_j + v_j - n_j, 0\}, \min\{u_j, v_j\}]$. By Lemma 4,

$$\begin{aligned} \langle est_j \rangle &= \sum_{x \in X} (x \cdot \mathcal{Q}_j(x)) = \sum_{x \in X} \frac{x \cdot \binom{u_j}{x} \binom{n_j - u_j}{v_j - x}}{\binom{n_j}{v_j}} = \sum_{x \in X} \frac{x \cdot \frac{u_j}{x} \cdot \binom{u_j - 1}{x - 1} \binom{n_j - u_j}{v_j - x}}{\frac{n_j}{v_j} \cdot \binom{n_j - 1}{v_j - 1}} \\ &= \frac{u_j \cdot v_j}{n_j} \cdot \frac{\sum_{x \in X} \binom{u_j - 1}{x - 1} \binom{n_j - u_j}{v_j - x}}{\binom{n_j - 1}{v_j - 1}} = u_j \cdot v_j / n_j. \quad \square \end{aligned}$$

Proof of Corollary 2. The corollary directly follows Lemma 4, and Equations 16, 17, and 19. \square

Proof of Lemma 5. Since QI_x and QI_y belong to an l -diverse partition, we have $|QI_x| \geq l$ and $|QI_y| \geq l$. When either of the two conditions is satisfied, after the swap, all tuples in QI_x (QI_y) will still have distinct

sensitive values; therefore, both QI_x and QI_y will remain l -diverse. The fact that RCE is not influenced by the swap can be easily verified, according to the concrete definition of E_t (Equation 9, 12, or 13) under each metric of information loss. \square

Proof of Lemma 6. Consider the projection of R_x on the i -th dimension ($1 \leq i \leq d$). Let t_{x+} (t_{x-}) be a tuple in QI_x whose i -th coordinate equals $R_x[i+]$ ($R_x[i-]$). Similarly, we introduce notations t_{y+} and t_{y-} in the same manner with respect to R_y . After the swapping, each of t_{x+} , t_{x-} , t_{y+} , and t_{y-} may or may not appear in a QI-group different from the one it belongs to before the swapping. Hence, there are 16 possibilities; it turns out that all of them satisfy Inequality 22. Since the discussion of all possibilities is analogous, next we will prove the lemma for only one possibility: t_{x+} , t_{x-} , t_{y+} and t_{y-} appear in QI_x , QI , QI_x , and QI , respectively. In fact, since R'_x covers t_{x+} and t_{y+} , its projection on the i -th dimension must be at least $|R_x[i+] - R_y[i+]|$. Similarly, the projection of R'_y on the i -th dimension is at least $|R_x[i-] - R_y[i-]|$. Therefore, Inequality 22 is correct. \square