

# ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication

Yufei Tao<sup>1</sup> Hekang Chen<sup>2</sup> Xiaokui Xiao<sup>3</sup> Shuigeng Zhou<sup>2</sup> Donghui Zhang<sup>4</sup>

<sup>1</sup>Chinese University of Hong Kong  
*taoyf@cse.cuhk.edu.hk*

<sup>2</sup>Fudan University  
*{hkchen, sgzhou}@fudan.edu.cn*

<sup>3</sup>Cornell University  
*xiaokui@cs.cornell.edu*

<sup>4</sup>Northeastern University  
*donghui@ccs.neu.edu*

## Abstract

Generalization is a well-known method for privacy preserving data publication. Despite its vast popularity, it has several drawbacks such as heavy information loss, difficulty of supporting marginal publication, and so on. To overcome these drawbacks, we develop ANGEL<sup>1</sup>, a new anonymization technique that is as effective as generalization in privacy protection, but is able to retain significantly more information in the microdata. ANGEL is applicable to any monotonic principles (e.g.,  $l$ -diversity,  $t$ -closeness, etc.), with its superiority (in correlation preservation) especially obvious when tight privacy control must be enforced. We show that ANGEL lends itself elegantly to the hard problem of marginal publication. In particular, unlike generalization that can release only restricted marginals, our technique can be easily used to publish any marginals with strong privacy guarantees.

**Keywords:** Privacy, generalization, ANGEL.

**To appear in *IEEE TKDE*.**

---

<sup>1</sup>The name reflects the fact that our approach captures two popular methods *ANatomy* [40] and *GENeralLization* [33, 34] as special cases.

# 1 Introduction

*Privacy preserving publication* has received considerable attention from the database community in the past few years. Specifically, let  $T$  be a table containing sensitive information. The objective is to release a modified version  $T^*$  of  $T$  such that  $T^*$  forbids adversaries from inferring the sensitive data of  $T$  confidently, but on the other hand, allows researchers to understand useful correlations in  $T$ . The table  $T$  is often called the *microdata*.

To illustrate, assume that a hospital wants to release the microdata of Table 1a. Here, *Disease* is sensitive, that is, the publication must prevent the disease of any patient from being discovered. Simply removing the names is insufficient due to the possibility of *linking attacks* [33, 35]. For example, consider an adversary that knows the age 21 and gender M of Alan. Given Table 1a (even without the names), s/he is still able to assert that the first tuple must belong to Alan, and thus find out his real disease *pneumonia*. As *Age* and *Sex* can be combined to recover a patient’s identity, they are referred to as *quasi-identify* (QI) attributes.

Name	Age	Sex	Disease
Alan	21	M	pneumonia
Bob	23	M	pneumonia
Carrie	38	F	bronchitis
Daisy	40	F	bronchitis
Eddy	41	M	pneumonia
Frank	43	M	pneumonia
Gloria	58	F	bronchitis
Helena	60	F	bronchitis

Age	Sex	Disease
[21,40]	*	pneumonia
[21,40]	*	pneumonia
[21,40]	*	bronchitis
[21,40]	*	bronchitis
[41,60]	*	pneumonia
[41,60]	*	pneumonia
[41,60]	*	bronchitis
[41,60]	*	bronchitis

(a) The microdata                      (b) 2-diverse generalization

Table 1: An example of generalization

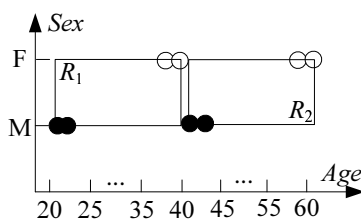


Figure 1: Regarding generalization as point-to-rectangle transformation

*Generalization* is a popular method of thwarting linking attacks. It works by replacing QI-values in the microdata with fuzzier forms. Table 1b is a generalized version of Table 1a. Notice that, for instance, the age 21 of the first tuple in Table 1a has been replaced with an interval [21, 40]

in Table 1b. Also observe that generalization creates *QI-groups*, each of which consists of tuples with identical (generalized) QI-values. For example, Table 1b has two QI-groups, including the first and last 4 tuples, respectively. To understand why generalization helps to prevent linking attacks, consider the same adversary aforementioned that knows Alan’s age and gender. Given Table 1b, s/he cannot tell exactly which of the first 4 tuples describes Alan. With a random guess, the adversary can correctly link Alan to *pneumonia* only with 50% probability.

It is often convenient to regard generalization as a point-to-rectangle transformation in the *QI-space*, which is a space formed by all the QI attributes. Figure 1 represents each tuple in Table 1a as a point, whose horizontal and vertical coordinates equal the tuple’s age and sex, respectively. A black (white) point indicates a tuple with sensitive value *pneumonia* (*bronchitis*). Rectangle  $R_1$  represents the first QI-group of Table 1b. The *Age*-extent of  $R_1$  is the *Age*-value [21, 40] of the QI-group, and its *Sex*-extent covers both F and M, corresponding to the wildcard ‘\*’ in the group. Similarly, rectangle  $R_2$  describes the second QI-group.

A microdata relation can be generalized in numerous ways. Various generalizations, however, may provide drastically different privacy protection. Hence, in practice, generalization needs to be guided by an *anonymization principle*, which is a criterion deciding whether a table has been adequately anonymized. Most notable principles include *k*-anonymity [33, 35], *l*-diversity [25], and *t*-closeness [24].

### 1.1 Motivation 1: Large Information Loss in Stringent Privacy Protection

Researchers keep observing the drawbacks of existing principles, and then developing new principles to give better privacy guarantees. For instance, *l*-diversity is proposed to overcome the defects of *k*-anonymity and yet, its own limitations led to *t*-closeness. Privacy, however, is a natural foe of utility. A privacy-safer principle reduces the number of selectable generalizations, thus decreasing the chance of finding a utility-friendly generalization.

Let us demonstrate the phenomenon by examining again the microdata in Table 1a. We want to obtain a 2-diverse [25] generalization, where at most half of the tuples in a QI-group can have the same sensitive value. For Table 1a (where only two sensitive values exist), this translates into a constraint that each QI-group must have as many tuples having *pneumonia* as *bronchitis*. It is easy

<i>Name</i>	<b>Age</b>	<b>Sex</b>	<b>Zip.</b>	<b>Disease</b>
<i>Alan</i>	21	M	10k	pneumonia
<i>Bob</i>	23	M	58k	flu
<i>Carrie</i>	58	F	12k	bronchitis
<i>Daisy</i>	60	F	60k	pneumonia
<i>Eddy</i>	70	M	78k	flu
<i>Frank</i>	72	M	80k	bronchitis

(a) Microdata

<b>Age</b>	<b>Sex</b>	<b>Zipcode</b>	<b>Disease</b>
[21,23]	M	[10k,58k]	pneumonia
[21,23]	M	[10k,58k]	flu
[58,60]	F	[12k,60k]	bronchitis
[58,60]	F	[12k,60k]	pneumonia
[70,72]	M	[78k,80k]	flu
[70,72]	M	[78k,80k]	bronchitis

(b) 2-diverse generalization

<b>Zipcode</b>	<b>Disease</b>
[10k,12k]	pneumonia
[10k,12k]	bronchitis
[58k,60k]	flu
[58k,60k]	pneumonia
[78k,80k]	flu
[78k,80k]	bronchitis

(c) A 2-diverse marginal

Table 2: Marginal publications by generalization

to verify that, the grouping of points in Figure 1 is the only generalization<sup>2</sup> under *strict global recoding* [23], where the rectangles of all QI-groups must be mutually disjoint. This choice loses considerable information: apparently no gender data is preserved, while each age is transformed into a long interval.

The previous observation implies a discouraging dilemma. As the community strives to enhance anonymity by seeking even safer anonymization principles, we risk further shrinking the already-limited pool of eligible generalizations, and hence, would eventually be unable to feed the public with useful scientific data.

## 1.2 Motivation 2: Marginal Publication

Typically, generalization loses less information when the number of QI attributes is smaller [1]. Therefore, besides a large table that covers all the QI attributes, the publisher may also release certain projections to enhance the public’s understanding on the underlying correlations. This approach is called *marginal publication*, and has been explored in [20, 37, 44].

For example, assume again that the microdata is Table 2a, and that the publisher has released Table 2b. After some time, a researcher requests refined correlations of *Zipcode* and *Disease*. To entertain the request, the publisher prepares a generalization of the marginal  $\{Zipcode, Disease\}$ , as shown in Table 2c. Clearly, it has more accurate Zipcodes (than Table 2b), and thus captures the correlations between *Zipcode* and *Disease* better.

Both Tables 2b and 2c are 2-diverse, but their simultaneous publication violates 2-diversity. Consider an adversary knowing Alan’s QI-values. From Table 2b, s/he is aware that Alan’s disease is in the set  $\{pneumonia, flu\}$ , whereas from Table 2c, s/he is sure that the disease also falls in

<sup>2</sup>Except the trivial and worse generalization that includes the whole table in a single QI-group.

$\{pneumonia, bronchitis\}$ . Hence, Alan must have contracted *pneumonia*. Unfortunately, when the set of marginals overlap with other in an arbitrarily complex manner, evaluating the privacy risk is NP-hard [20, 44]. This fact forces a publisher to release only those “easy” marginals for which privacy risk can be calculated efficiently.

The above problem lingers in all the existing solutions [20, 37, 44] to marginal publication. In particular, the authors of [44] explicitly acknowledge this, by explaining several cases where efficient assessment of privacy risk is impossible. The work of [20], on the other hand, is applicable only if all the marginals to be published form a *decomposable graph*. Finally, the method in [37] requires that, except the first marginal, no subsequent marginal released can have the sensitive attribute. For instance, after giving Table 2b away, the publisher immediately loses the option of releasing Table 2c (as it contains the attribute *Disease*). This is a severe drawback because the sensitive attribute is very important for analysis, and hence, is solicited in most marginal requests. Finally, note that the solutions of [20, 37, 44] are designed for  $k$ -anonymity and  $l$ -diversity. When a different principle is applied, those solutions are no longer applicable, and their adaptation requires considerable overhead.

### 1.3 Contributions

This paper develops ANGEL, a new anonymization technique that overcomes all the above problems. ANGEL is applicable to any monotonic anonymization principle (including  $k$ -anonymity  $l$ -diversity, and  $t$ -closeness, etc.). Compared to traditional generalization, it ensures the same privacy guarantee, but preserves significantly more information in the microdata. The superiority of ANGEL is especially obvious when stringent anonymity control is enforced. This is a highly desirable feature because, as mentioned in Section 1.1, the community continuously invents safer anonymization principles that fix the vulnerabilities of the previous ones.

Another crucial feature of ANGEL is that it lends itself very nicely to marginal publication. It easily supports the publication of *any* set of marginals, thus settling a problem known to be very difficult with generalization. Furthermore, ANGEL supports all monotonic principles in exactly the same manner. As a result, no adaptation effort is necessary when a publisher decides to adopt a different principle. This is a significant advantage over the previous solutions to marginal publi-

cation (which are “hard-wired” to specific principles).

The rest of the paper is organized as follows. Section 2 provides a generic model to capture generalization and anonymization principles. Based on the model, Section 3 elaborates the details of ANGEL and proves its privacy guarantees, while Section 4 extends the results to marginal publication. Section 5 explains how to leverage the publication ANGEL for data analysis. Section 6 presents an experimental evaluation of the proposed technique. Section 7 reviews the previous work related to ours. Section 8 concludes the paper with directions for future work.

## 2 E-M Generalization Modeling

Let  $T$  be a microdata table, which contains  $d$  quasi-identifier (QI) attributes and a sensitive attribute (SA)  $X$ . The core of generalization is to

- divide the tuples of  $T$  into a set  $E$  of disjoint *equivalence classes* (EC), and then
- transform the QI-values of the tuples in each EC to the same format.

In Table 1b, for instance, each QI-group corresponds to an EC. Following the previous work [7, 15, 17, 22, 23, 21, 24, 37], we assume that the (generalized) QI-values of all ECs obey the *strict global recoding* [22]. Namely, there cannot be two ECs whose rectangles in the QI-space overlap each other (review the point-to-rectangle transformation illustrated in Figure 1).

**Definition 1** ( $k$ -anonymity [33, 35]).  *$E$  satisfies  $k$ -anonymity if every EC in  $E$  contains at least  $k$  tuples.*

$k$ -anonymity thwarts the so-called *presence attacks*, where an adversary obtains the precise QI-values of an individual, and wants to find out whether this individual exists in the microdata. However,  $k$ -anonymity alone provides weak protection against linking attacks [25]. Hence, advanced anonymization principles aim at constraining the distribution of the sensitive values in each EC. Next, we give a generic modeling of such principles. Let  $x_1, x_2, \dots, x_m$  denote all the values in the domain of the sensitive attribute  $X$ , where  $m$  is the domain size of  $X$ . Then:

**Definition 2** (SA-distribution). *Given a multi-set  $S$  of sensitive values, the **SA-distribution** in  $S$  is characterized by a pdf  $f : X \rightarrow [0, 1]$ , where  $f(x_i)$  equals the number of occurrences of  $x_i$  in  $S$*

divided by  $|S|$  ( $1 \leq i \leq m$ ). Whenever convenient, we may regard  $f$  as an  $m$ -dimensional vector  $\{f(x_1), f(x_2), \dots, f(x_m)\}$ .

Generalization maps tuples in the same EC to an identical SA-distribution. This is the key to privacy preservation, namely, a tuple's sensitive value is concealed into a distribution, which is shared by all the tuples in the same EC. Formally, if we use  $F$  to denote the entire family of all possible SA-distributions, generalization essentially defines a mapping

$$M : E \rightarrow F \tag{1}$$

which associates each EC  $C \in E$  with a SA-distribution  $f \in F$ , written as  $f = M(C)$ . In a linking attack,  $f$  corresponds to an adversary's understanding about an individual's sensitive value, after realizing that the individual's record falls in  $C$ .

Motivated by this, we introduce the following succinct definition of generalization.

**Definition 3** (E-M Generalization Modeling). A **generalization** of a microdata table  $T$  can be adequately represented as a pair of  $E$  and  $M$ , denoted as  $(E, M)$ .

**Example 1.** Let us elaborate the E-M modeling of Table 1b. Here,  $E$  includes two ECs  $C_1 = \{\text{Alan, Bob, Carrie, Daisy}\}$  and  $C_2 = \{\text{Eddy, Frank, Gloria, Helena}\}$ .  $C_1$  is mapped to a SA-distribution  $f_1$  where  $f_1(\text{pneumonia}) = f_1(\text{bronchitis}) = 0.5$ . Similarly,  $C_2$  is mapped to a SA-distribution  $f_2$  that happens to be equivalent to  $f_1$ .  $M$  thus includes two mappings:  $f_1 = M(C_1)$  and  $f_2 = M(C_2)$ . ■

Now we are ready to define anonymization principles.

**Definition 4** (Anonymization Principle). An **anonymization principle** is a constraint on a SA-distribution. A generalization  $(E, M)$  satisfies the principle if the SA-distribution of every EC in  $E$  satisfies the constraint.

For example, the constraint imposed by  $l$ -diversity is that<sup>3</sup>, in each EC  $C \in E$ , the frequency of the most frequent sensitive value must be at most  $1/l$ .  $t$ -closeness [24], on the other hand, forbids

---

<sup>3</sup>Precisely speaking,  $l$ -diversity requires each EC to contain at least  $l$  well-represented sensitive values. The meaning of well-represented can be interpreted in various ways [25], leading to different instantiations of  $l$ -diversity.

the SA-distribution of any EC to deviate significantly from the SA-distribution of the whole table. Formally, let  $f_0$  represent the SA-distribution of the entire microdata  $T$ . Then, an EC  $C$  qualifies  $t$ -closeness, if  $EMD(f, f_0) \leq t$ , where  $f$  is the SA-distribution of  $C$ , and function  $EMD(f, f_0)$  is the *earth mover distance* [24] between distributions  $f$  and  $f_0$ . Table 1b is actually a perfect generalization by this principle, as it fulfills 0-closeness, noticing that the SA-distribution  $f$  of each EC is exactly  $f_0$ , i.e.,  $EMD(f, f_0) = 0$ .

In this paper, we are interested in monotonic principles:

**Definition 5** (Monotonicity). *An anonymization principle is **monotonic** if the following is true: given any two multi-sets of sensitive values  $S_1$  and  $S_2$  whose SA-distributions obey the principle, the SA-distribution of the union  $S_1 \cup S_2$  also obeys the principle.*

Monotonicity is an important property that permits a top-down pruning strategy in computing a generalization [7, 15, 23, 21, 24, 25], which is the key to achieving low computation cost. Both  $l$ -diversity and  $t$ -closeness are monotonic [24, 25].

Our discussion focuses on generalizations  $(E, M)$  where

- (i)  $E$  is  $k$ -anonymous, and
- (ii)  $(E, M)$  satisfies the adopted anonymization principle.

By achieving these objectives, we provide sufficient protection against both presence and linking attacks. In particular, the parameter  $k$  in (i) controls the degree of presence-attack protection, whereas the principle in (ii) can be any monotonic principle against linking attacks.

**Discussion.** Any method for privacy preservation must be designed with a target adversary in mind. Specifically, there should be a clear assumption on the background knowledge of the adversary. As long as the assumption holds, the method must have a solid guarantee on how much the adversary can learn from an anonymized dataset. The assumption and the guarantee constitute what defines a privacy principle. For example,  $l$ -diversity assumes that an adversary knows only the QI-values, and guarantees that no individual's sensitive value will be revealed with probability higher than  $1/l$ . Tackling stronger background knowledge is the chief motivation of designing new principles.

For instance, *recursive*  $(c, l)$ -diversity [25] can guard against more powerful adversaries than the basic  $l$ -diversity. Even if (besides having QI-values) an adversary can exclude  $l - 2$  values (in the domain of the sensitive attribute) from belonging to an individual, recursive  $(c, l)$ -diversity still guarantees that the adversary can succeed in pinpointing the individual’s real value with probability at most  $c/(c+1)$ . A crucial feature of ANGEL (presented in the next section) is that it is applicable to all principles that are monotonic (hence, it applies to recursive  $(c, l)$ -diversity as well) without compromising their power of privacy protection at all.

### 3 The ANGEL Technique

We first provide an overview of ANGEL in Section 3.1, and then formalize it in Section 3.2. Section 3.3 establishes its privacy guarantees, and Section 3.4 clarifies its relevance to traditional generalization, and its anonymization algorithm. Finally, Section 3.5 explains the differences between ANGEL and *anatomy* [40].

#### 3.1 Overview

Suppose that we want to publish the microdata of Table 1a, conforming to 2-diversity. ANGEL first divides the table into *batches*:

Batch 1: {Alan, Carrie}, Batch 2: {Bob, Daisy},  
Batch 3: {Eddy, Gloria}, Batch 4: {Frank, Helena}

Observe that each batch obeys 2-diversity: it contains one *pneumonia*- and one *bronchitis*-tuple. ANGEL creates a *batch table* (BT), as in Table 3a, summarizing the *Disease*-statistics of each batch. For example, the first row of Table 3a states that exactly one tuple in Batch 1 carries *pneumonia*. Then, ANGEL creates another partitioning of Table 1, this time into *buckets* (which do *not* have to be 2-diverse):

Bucket 1: {Alan, Bob}, Bucket 2: {Carrie, Daisy},  
Bucket 3: {Eddy, Frank}, Bucket 4: {Gloria, Helena}

Finally, ANGEL generalizes the tuples of each bucket into the same form, producing a *generalized table* (GT). Table 3b demonstrates the GT. Note that GT does not include the *Disease* attribute, but stores, for each tuple of the microdata, the ID of the batch containing it. For instance, the first tuple of Table 3b has a *Batch-ID* 1, because its owner Alan belongs to Batch 1. Tables 3a and 3b are the final relations released by ANGEL.

### 3.2 Formalization

Let  $T$  be the microdata, and  $\mathcal{P}$  be the objective anonymization principle (e.g., 2-diversity in Section 3.1). We start by formalizing the notions of batch and bucket.

**Definition 6** (Batch). A **batch partitioning** of the microdata  $T$  consists of **batches**  $B_1, B_2, \dots, B_b$  such that

- each batch is a set of tuples in  $T$ ;
- $\cup_{i=1}^b B_i = T$  and, for any  $i \neq j$ ,  $B_i \cap B_j = \emptyset$ ;
- the SA-distribution in each batch  $B_i$  ( $1 \leq i \leq b$ ) satisfies principle  $\mathcal{P}$ .

We refer to the subscript  $i$  of batch  $B_i$  as the *batch-ID* of  $B_i$ .

**Definition 7** (Bucket). A **bucket partitioning** consists of **buckets**  $C_1, C_2, \dots, C_e$  such that

- each bucket is a set of tuples in  $T$ ;
- each bucket contains at least  $k$  tuples, where  $k$  is a parameter controlling the degree of protection against presence attacks;
- $\cup_{i=1}^e C_i = T$  and, for any  $i \neq j$ ,  $C_i \cap C_j = \emptyset$ .

In ANGEL, any pair of bucket and batch partitionings determine an anonymized publication, called *angelization*:

**Definition 8** (Angelization). Given a batch partitioning  $\{B_1, B_2, \dots, B_b\}$  and a bucket partitioning  $\{C_1, C_2, \dots, C_e\}$  of the microdata  $T$ , an **angelization** of  $T$  is a pair of a **batch table** (BT) and a **generalized table** (GT), such that

- *BT has three columns:  $\{Batch-ID, X, Count\}$ , where  $X$  is the sensitive attribute of  $T$ . For every batch  $B_i$  ( $1 \leq i \leq m$ ) and every sensitive value  $x \in X$  that appears in  $B_i$ ,  $BT$  has a row  $(i, x, y)$ , where  $y$  is the number of occurrences of  $x$  in  $B_i$ .*
- *GT has all the QI-attributes of  $T$ , together with an extra column  $Batch-ID$ . Every tuple  $t \in T$  defines a row in  $GT$ , which stores the generalized QI-values of  $t$ , and the ID of the batch containing  $t$ . All tuples in the same bucket  $C_i$  ( $1 \leq i \leq e$ ) have equivalent generalized QI-values.*

ANGEL publishes  $BT$  and  $GT$ . For example, Table 3a (3b) is the  $BT$  ( $GT$ ) resulting from angelizing the microdata Table 1a, when  $\mathcal{P}$  is 2-diversity. The batch (bucket) partitioning, which leads to the  $BT$  and  $GT$ , contains the batches (buckets) as labeled in Table 3a (Table 3b).

### 3.3 Privacy Guarantees

Before analyzing the privacy guarantee of ANGEL, we fit it into the E-M modeling of generalization in Definition 3. For this purpose, we need to elaborate the semantics of  $E$  and  $M$  here. The meaning of  $E$  is obvious: the set of buckets. Specifically, each bucket  $C$  is an EC, as all the tuples in  $C$  possess the same generalized QI-values in  $GT$ .

Now, given any EC (i.e., a bucket)  $C \in E$ , we reveal its associated SA-distribution  $f = M(C)$ . Let  $I$  be the set of rows in  $GT$  created from  $C$ . For each batch-ID  $i \in [1, b]$ , use  $n(I, i)$  to denote the number of rows in  $I$  whose attribute  $Batch-ID$  equals  $i$ . Thus,  $f$  is given by

$$f = \frac{n(I, 1)}{|I|} f_1 + \frac{n(I, 2)}{|I|} f_2 + \dots + \frac{n(I, b)}{|I|} f_b \quad (2)$$

where  $f_i$  ( $1 \leq i \leq b$ ) is the SA-distribution of batch  $B_i$ , and all of  $f, f_1, \dots, f_b$  should be understood as  $m$ -dimensional vectors (see Definition 2) with  $m$  being the domain size of the sensitive attribute.

In other words, *the SA-distribution of each EC in an angelization is the synthesization of the SA-distributions of several batches*. As will be clear shortly, this is a salient characteristic of angelization that distinguishes it from traditional generalization. The bipartite graph in Figure 2 demonstrates the synthesization relationships between buckets and batches for the angelization in Table 3. For example, the SA-distribution of bucket 1 is synthesized from the SA-distributions of batches 1 and 2.

Batch -ID	Disease	Count
Batch 1	1 pneumonia	1
	1 bronchitis	1
Batch 2	2 pneumonia	1
	2 bronchitis	1
Batch 3	3 pneumonia	1
	3 bronchitis	1
Batch 4	4 pneumonia	1
	4 bronchitis	1

Age	Sex	Batch -ID
[21,23]	M	1
[21,23]	M	2
[38,40]	F	1
[38,40]	F	2
[41,43]	M	3
[41,43]	M	4
[58,60]	F	3
[58,60]	F	4

(a) The batch table (BT)      (b) The generalized table (GT)

Table 3: ANGEL publication

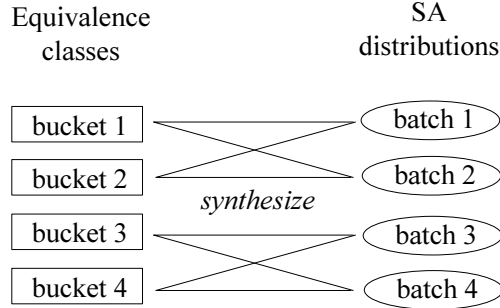


Figure 2: Synthesization graph of the angelization in Table 2

As with conventional generalization, in angelization, the SA-distribution  $f$  associated with an EC  $C$  is also an adversary’s understanding about an individual’s sensitive value, after realizing that the individual is in  $C$ . This is illustrated in the next example.

**Example 2.** Assume that, given the BT and GT in Table 3, an adversary wants to derive Alan’s disease, knowing his age 21 and sex M. By these QI-values, the adversary learns from GT (Table 3b) that Alan must be in Bucket 1, containing the first two rows of GT. Let  $I$  be the set of those two rows.  $n(I, 1) = 1$ , since a tuple in  $I$  has 1 as its *Batch-ID*. Similarly,  $n(I, 2) = 1$ ,  $n(I, 3) = 0$  and  $n(I, 4) = 0$ .

Based on the above information, the adversary assumes that Alan may fall in Batch 1 with a probability  $n(I, 1)/|I| = 1/2$ , and in Batch 2 with likelihood  $n(I, 2)/|I| = 1/2$ . As a result, her/his probabilistic modeling  $f$  of Alan’s disease should be a weighted sum of the SA-distributions  $f_1, f_2$  in Batches 1 and 2, respectively. Here, as indicated in BT (Table 3a),  $f_1$  is a vector  $\{0.5, 0.5\}$ , where the first and second values correspond to  $f_1(pneumonia)$  and  $f_1(bronchitis)$ , respectively. Likewise,  $f_2$  is also  $\{0.5, 0.5\}$ . Hence, by Equation 2,  $f$  evaluates to  $\{0.5, 0.5\}$  as well, implying

that, after the linking attack, the adversary believes that Alan contracted *pneumonia* or *bronchitis* with equal chance. ■

The privacy guarantee of ANGEL relies on:

**Lemma 1.** *If principle  $\mathcal{P}$  is monotonic,  $f = M(C)$ , as calculated by Equation 2, must satisfy  $\mathcal{P}$ .*

*Proof.* This lemma is a special case of Theorem 1 to be established in Section 4.2, which provides a much more general result. □

As Lemma 1 holds for any  $C \in E$ , the BT and GT output by angelization determine a generalization  $(E, M)$  that satisfies  $\mathcal{P}$ . Furthermore, by Definition 7, each EC has size at least  $k$ , namely,  $E$  is  $k$ -anonymous. Hence, angelization qualifies both goals given at the end of Section 2.

Since angelization has been captured by the E-M modeling (Definition 3), in the sequel, we often denote an angelization by  $(E, M)$ , which will be used interchangeably with the standard BT-GT definition. To avoid confusion, we adopt the term *simple generalization* to refer to an anonymized table obtained by conventional generalization. For example, Table 1b is a simple generalization of Table 1a. Remember that a simple generalization also has two equivalent representations: it can be indicated as a relation like Table 1b, or as a pair  $(E, M)$ .

### 3.4 Relevance to Simple Generalization

Angelization captures simple generalization as a special case: given any simple generalization  $T^* = (E, M)$  of the microdata  $T$ , we can always construct a corresponding angelization  $(E', M')$  such that  $E = E'$  and  $M = M'$ . Towards this purpose, it suffices to create a batch partitioning and a bucket partitioning that coincide with each other: every batch (bucket) corresponds to a QI-group in  $T^*$ . These two partitionings define the angelization  $(E', M')$  mentioned earlier.

To illustrate, consider the simple generalization in Table 1b (with respect to the microdata of Table 1a). To build its corresponding angelization, we create a batch (bucket) partitioning, where each batch (bucket) is a QI-group in Table 1b:

Batch (Bucket) 1: {Alan, Bob, Carrie, Daisy}

Batch -ID	Disease	Count
1	pneumonia	2
1	bronchitis	2
2	pneumonia	2
2	bronchitis	2

Age	Sex	Batch -ID
[21,40]	*	1
[21,40]	*	1
[21,40]	*	1
[21,40]	*	1
[41,60]	*	2
[41,60]	*	2
[41,60]	*	2
[41,60]	*	2

(a) The batch table (BT)
(b) The generalized table (GT)

Table 4: ANGEL version of Table 1b

Batch (Bucket) 2: {Eddy, Frank, Gloria, Helena}

They lead to the angelization in Table 4.

The above transformation is *one-way*, namely, *an angelization does not necessarily determine a generalization that can be equated to any simple generalization*. This can be best understood by recalling how the SA-distribution of an EC  $C$  is constructed in each case. As shown in Figure 2, in angelization, the SA-distribution of  $C$  (i.e., a bucket) may need to be synthesized from multiple SA-distributions. This never happens in simple generalization where, as demonstrated in Example 1, the SA-distribution of  $C$  (i.e., a QI-group) is always determined by a unique SA-distribution: that in  $C$  itself.

Next, we give a method of finding an angelization of the microdata  $T$ . The method leverages two (arbitrary) algorithms  $AL_{\mathcal{P}}$  and  $AL_k$  that compute a simple generalization conforming to principle  $\mathcal{P}$  and  $k$ -anonymity, respectively. Specifically, there are three steps:

- First, we run  $AL_{\mathcal{P}}$  to get a simple generalization  $T_{\mathcal{P}}^*$  of  $T$ . The set of QI-groups in  $T_{\mathcal{P}}^*$  is taken as a batch partitioning.
- Similarly, we execute  $AL_k$  to acquire a simple generalization  $T_k^*$  of  $T$ , whose QI-groups constitute a bucket partitioning.
- Finally, discarding  $T_{\mathcal{P}}^*$  and  $T_k^*$ , we derive an angelization from the batch and bucket partitionings according to Definition 8.

For example, let  $\mathcal{P}$  be 2-diversity, and  $k = 2$ . Given the microdata of Table 1a,  $AL_{\mathcal{P}}$  ( $AL_k$ ) would

<i>Name</i>	<b>Age</b>	<b>Sex</b>	<b>Disease</b>
<i>Alan</i>	[21,38]	*	pneumonia
<i>Carrie</i>	[21,38]	*	bronchitis
<i>Bob</i>	[23,40]	*	pneumonia
<i>Daisy</i>	[23,40]	*	bronchitis
<i>Eddy</i>	[41,58]	*	pneumonia
<i>Gloria</i>	[41,58]	*	bronchitis
<i>Frank</i>	[43,60]	*	pneumonia
<i>Helena</i>	[43,60]	*	bronchitis

(a)  $T_{\mathcal{P}}^*$

<i>Name</i>	<b>Age</b>	<b>Sex</b>	<b>Disease</b>
<i>Alan</i>	[21,23]	M	pneumonia
<i>Bob</i>	[21,23]	M	pneumonia
<i>Carrie</i>	[38,40]	F	bronchitis
<i>Daisy</i>	[38,40]	F	bronchitis
<i>Eddy</i>	[41,43]	M	pneumonia
<i>Frank</i>	[41,43]	M	pneumonia
<i>Gloria</i>	[58,60]	F	bronchitis
<i>Helena</i>	[58,60]	F	bronchitis

(b)  $T_k^*$

Table 5: Computing an angelization (microdata Table 1a)

output  $T_{\mathcal{P}}^* = \text{Table 5a}$  ( $T_k^* = \text{Table 5b}$ ), which is 2-diverse (2-anonymous), and determines the batch (bucket) partitioning given in Section 3.1. Recall that those batch and bucket partitionings lead to the angelization in Table 3. Note that the final generalized values in GT (Table 3b) are identical to those in  $T_k^*$  (Table 5b). Namely, the generalized values in  $T_{\mathcal{P}}^*$  (Table 5a) are not important at all, and discarded right away. Hence, if a certain recoding scheme needs to be enforced, it suffices to do so in  $T_k^*$ , i.e., *the QI-groups in  $T_{\mathcal{P}}^*$  can be formed freely, without being constrained by the recoding scheme at all.* For example, Table 5a does not conform to strict global recoding, as there is overlap in the Age-intervals of different QI-groups.

The fact that angelization captures simple generalization as a special case immediately implies that the former *never* loses more information than the latter. In the worst case, if the angelization computed as above turns out to retain less information than  $T_{\mathcal{P}}^*$  (a simple generalization under principle  $\mathcal{P}$ ), we can always create another angelization equivalent to  $T_{\mathcal{P}}^*$ . However, in practice, this is rarely necessary. The main observation is that  $k$ -anonymity is usually (much) easier to satisfy than  $\mathcal{P}$ . As a result, the  $k$ -anonymous generalization  $T_k^*$  of  $T$  preserves much more information than  $T_{\mathcal{P}}^*$ . Inheriting the QI-values of  $T_k^*$ , the GT of an angelization also preserves much more QI information than  $T_{\mathcal{P}}^*$ . Such a significant gain makes it rather unlikely for the overall angelization to lose more data than  $T_{\mathcal{P}}^*$ . Note that this is especially the case when  $T_{\mathcal{P}}^*$  is a stringent principle imposing strong privacy control (in which case there is even greater difference in the QI information retained by  $T_k^*$  and  $T_{\mathcal{P}}^*$ ).

Our angelization algorithm is applicable to any monotonic principle  $\mathcal{P}$ . Furthermore, it allows researchers to focus on studying algorithms that produce simple generalizations fulfilling  $\mathcal{P}$ . Once

Age	Sex	QI-Group
21	M	1
23	M	1
38	F	1
40	F	1
41	M	2
43	M	2
58	F	2
60	F	2

(a) QIT

QI-Group	Disease	Count
1	pneumonia	2
1	bronchitis	2
2	pneumonia	2
2	bronchitis	2

(b) ST

Table 6: Anatomy version of Table 1b

such an algorithm is available, we can immediately combine it with a  $k$ -anonymous algorithm to compute angelizations satisfying  $\mathcal{P}$ .

### 3.5 Comparison with Anatomy

Angelization should not be confused with *anatomy* [40], which looks similar only because it also publishes two tables. The strongest evidence is that anatomy goes hand-in-hand with simple generalization — every simple generalization can be mapped to an anatomy, and conversely, every anatomy can be mapped to a simple generalization. In contrast, as explained in Section 3.4, angelization is a superset of simple generalization, and hence, also a superset of anatomy.

The bijection between anatomy and simple generalization lies in the fact that both of them are based on a *single* partitioning of the microdata into QI-groups (anangelization, however, uses two partitionings). For example, corresponding to the simple generalization of Table 1b, anatomy produces a *QI table* (QIT) and a *sensitive table* (ST), illustrated in Table 6, as follows. First, it numbers the QI-groups in Table 1b: the first group has an ID 1, and the second 2. Then, for any tuple in Table 1b, anatomy inserts its *original* QI-values into QIT, together with the ID of the QI-group containing it. For instance, the first row of QIT (Table 6a) is obtained from the first tuple of Table 1b. ST, on the other hand, provides the statistics of the sensitive values in each QI-group. The first row of the ST (Table 6b), for example, says that two tuples in QI-group 1 carry *pneumonia*. This finishes the conversion from simple generalization to anatomy. Conversely, given an anatomy (i.e., a pair of QIT and ST), it is even easier to construct the simple generalization — just (i) take the set of QI-groups of the microdata as indicated in (the *QI-group* column of) QIT, and (ii) generalize each QI-group in an ordinary way.

Batch-ID	Disease	Count
1	pneumonia	1
1	flu	1
2	bronchitis	1
2	pneumonia	1
3	flu	1
3	bronchitis	1

(a) BT

Age	Sex	Zipcode	Batch-ID
[21,23]	M	[10k,58k]	1
[21,23]	M	[10k,58k]	1
[58,60]	F	[12k,60k]	2
[58,60]	F	[12k,60k]	2
[70,72]	M	[78k,80k]	3
[70,72]	M	[78k,80k]	3

(b) The first GT

Zipcode	Batch-ID
[10k,12k]	1
[10k,12k]	2
[58k,60k]	1
[58k,60k]	2
[78k,80k]	3
[78k,80k]	3

(c) The second GT

Table 7: Marginal publications by angelization (microdata Table 2a)

It is worth noting that, anatomy releases all the QI-values directly. This is yet another difference between anatomy and angelization. Disclosing the precise QI-values is not permitted in those applications where presence attacks (mentioned in Section 2) are a concern. Finally, when  $k$  equals 1, angelization actually gracefully degrades into anatomy. In other words, if an application allows direct publication of QI-values (e.g., those applications where anatomy is applicable), angelization can also be employed (by setting  $k$  to 1).

## 4 Marginal Publication

Our discussion so far focuses on releasing only a single anonymized version, including all the attributes of the microdata. In the sequel, we deal with marginal publication, which has been introduced in Section 1. Section 4.1 first elaborates how angelization can be utilized for this purpose. Then, Section 4.2 analyzes its privacy guarantees.

### 4.1 Deployment of Angelization

Following the notations of the previous section, let  $T$  be a microdata table with a sensitive attribute  $X$ , and  $\mathcal{P}$  the anonymization principle selected by the publisher. Next, we will explain how to leverage ANGEL to publish any marginals of  $T$ , while ensuring the guarantee of  $\mathcal{P}$ .

Without loss of generality, suppose that we need to release  $g$  marginals, denoted as  $G_1, G_2, \dots, G_g$ , respectively. Each  $G_i$  ( $1 \leq i \leq g$ ) is a set of attributes in  $T$ . We assume that  $X$  appears in all of  $G_1, G_2, \dots, G_g$ . Note that this assumption is reasonable, because it is trivial to publish a marginal  $G$  that does not contain  $X$  – we can simply release a  $k$ -anonymous generalization of (the projection of  $T$  onto)  $G$ , without worrying about privacy breach [20].

ANGEL accomplishes the task by releasing one BT and  $g$  GTs. Intuitively, the BT is shared by all marginals, and yet, every marginal has a GT of its own. Furthermore, the  $g + 1$  tables are obtained through  $g + 1$  simple generalizations, including a  $k$ -anonymous generalization (for obtaining the BT) and  $g$  generalizations conforming to  $\mathcal{P}$  (one for each GT). Formally, let  $AL_k$  be any algorithm for computing a  $k$ -anonymous generalization, and  $AL_{\mathcal{P}}$  any algorithm for computing a generalization under  $\mathcal{P}$ . ANGEL employs the following procedures for marginal publication:

1. Run  $AL_{\mathcal{P}}$  on  $T$  to obtain a simple generalization  $T_{\mathcal{P}}^*$ . Decide a batch partitioning of  $T$  according to the equivalence classes in  $T_{\mathcal{P}}^*$ . Denote the batch partitioning as  $E_{\mathcal{P}}$ . Note that  $E_{\mathcal{P}}$  is a set, where each element is a batch (which can also be regarded as a set of tuples in  $T$ ).
2. For each marginal  $G_i$  ( $1 \leq i \leq g$ ), run  $AL_k$  on  $\Pi_{G_i}(T)$  (i.e., the projection of  $T$  onto  $G_i$ ) to obtain a simple generalization  $T_{ki}^*$ . Decide a bucket partitioning of  $T$  according to the equivalence classes in  $T_{ki}^*$ . Denote the bucket partitioning as  $E_i$ , which is a set where each element is a bucket (also a set of tuples).
3. Having  $E_{\mathcal{P}}$  and  $E_1, \dots, E_g$ , we are ready to produce the BT and GTs according to Definition 8. Specifically, the batch partitioning  $E(\cdot)$  alone uniquely decides a BT as follows. For every batch  $B$  in  $E(\cdot)$  and every sensitive value  $x$  appearing in  $B$ , we create a row  $(i, x, y)$  in the BT, where  $i$  is the batch-ID of  $B$ , and  $y$  the number of occurrences of  $x$  in  $B$ .

After this, we generate a GT for each marginal  $G_i$  ( $1 \leq i \leq g$ ), using  $E_{\mathcal{P}}$  and the bucket partitioning  $E_i$ . Recall that  $E_i$  comes from a  $k$ -anonymous generalization  $T_{ki}^*$  of  $\Pi_{G_i}(T)$  (in other words,  $T_{ki}^*$  does not involve any attribute outside  $G_i$ ). For every tuple  $t$  of  $T$ , we insert a row in the GT containing the generalized values of  $t$  in  $T_{ki}^*$ , together with the ID of the batch in  $E_{\mathcal{P}}$  containing  $t$ . Apparently, such a GT has no attribute outside  $G_i$ .

**Example 3.** To demonstrate the above procedures, let us assume that the microdata table  $T$  is Table 2a, and we need to publish two marginals:

$$G_1 = \{Age, Sex, Zipcode, Disease\}$$

$$G_2 = \{Zipcode, Disease\}.$$

Let the privacy principle  $\mathcal{P}$  be 2-diversity, and the parameter  $k$  of ANGEL be 2. To apply ANGEL, we first choose two algorithms  $AL_{\mathcal{P}}$  and  $AL_k$  for computing 2-diverse and 2-anonymous (simple) generalizations, respectively (e.g., both  $AL_{\mathcal{P}}$  and  $AL_k$  can be the *Mondrian* algorithm in [23]). Given these algorithms, ANGEL computes a BT and 2 GTs as follows.

First, it applies  $AL_{\mathcal{P}}$  to obtain a 2-diverse generalization  $T_{\mathcal{P}}^*$  of  $T$ , and derives a batch partitioning  $E_{\mathcal{P}}$  from  $T_{\mathcal{P}}^*$ . For simplicity, let us assume that  $E_{\mathcal{P}}$  has these 3 batches:

$$B_1 = \{\text{Alan, Bob}\}, B_2 = \{\text{Carrie, Daisy}\}, B_3 = \{\text{Eddy, Frank}\}.$$

As a second step, ANGEL determines two bucket partitionings  $E_1$  and  $E_2$  for marginals  $G_1$  and  $G_2$ , respectively. Specifically,  $E_1$  comes from a 2-anonymous generalization of  $T$  (returned by algorithm  $AL_k$ ). For our example, assume that  $E_1$  includes 3 buckets:

$$C_1 = \{\text{Alan, Bob}\}, C_2 = \{\text{Carrie, Daisy}\}, C_3 = \{\text{Eddy, Frank}\}.$$

Similarly, to obtain  $E_2$ , ANGEL takes the projection of  $T$  onto  $G_2$ , invokes  $AL_k$  to find a 2-anonymous generalization of the projection, and spawns  $E_2$  from the generalization. For illustration, assume that  $E_2$  has these buckets:

$$C'_1 = \{\text{Alan, Carrie}\}, C'_2 = \{\text{Bob, Daisy}\}, C'_3 = \{\text{Eddy, Frank}\}.$$

In the last step, ANGEL determines the contents of a BT and two GTs from  $E_{\mathcal{P}}$ ,  $E_1$ , and  $E_2$ . Table 7a gives the BT, which is decided from  $E_{\mathcal{P}}$  alone. For example, the first row of Table 7 implies that only a single tuple in the batch  $B_1$  of  $E_{\mathcal{P}}$  carries *pneumonia*. Table 7b shows the GT released for marginal  $G_1$ , which is computed from  $E_{\mathcal{P}}$  and  $E_1$ . For instance, the first two tuples of Table 7b correspond to the bucket  $C_1$  of  $E_1$ . Both tuples have value 1 as their *Batch-ID* because the two individuals in  $C_1$  both appear in the bucket  $B_1$  of  $E_{\mathcal{P}}$ . Table 7c presents the GT released for marginal  $G_2$ , which is derived in the same fashion from  $E_{\mathcal{P}}$  and  $E_2$ . Note, however, that this GT does not have attributes *Sex* and *Zipcode*, as they are absent from  $G_2$ . ■

The next subsection proves that releasing the  $g + 1$  tables output by the above strategy *always* automatically fulfills the principle  $\mathcal{P}$ . Hence, unlike the solutions in [20], our approach does not require any post-processing step to assess the privacy risk.

## 4.2 Privacy Guarantees

Next, we will analyze the quality of privacy protection of ANGEL. Towards this purpose, Section 4.2.1 first elaborates a framework for quantifying the risk of publishing marginals with simple generalization. Then, Section 4.2.2 employs the framework to discuss the guarantee of ANGEL.

### 4.2.1 Simple Generalization

The  $d$  QI-attributes of the microdata  $T$  together define a  $d$ -dimensional *QI-space*. On the other hand, let us use the term *universe* to refer to the  $(d + 1)$ -dimensional space that involves all the columns of  $T$ . Namely, the universe has one more dimension (i.e., the sensitive attribute  $X$ ) than the QI-space. For simplicity, we assume that all dimensions are discrete. Our analysis can be extended to the continuous case with straightforward adaptation.

The distribution of the microdata  $T$  can be fully described by a  $(d + 1)$ -dimensional pdf  $D$ . Specifically, we can represent a point  $p$  in the universe as a pair of  $Q$  and  $x$ , where  $Q$  is a point in the QI-space (having  $d$  QI values) and  $x$  is a sensitive value. Then,  $D$  is represented as

$$D(Q, x) = |(Q, x)|/|T| \quad (3)$$

where  $|(Q, x)|$  gives the number of tuples in  $T$  whose QI-values are  $Q$  and their sensitive values are  $x$ . Specially, if there is no such tuple,  $|(Q, x)|$  equals 0. Obviously, no publication should allow the public to obtain  $D(., .)$  exactly; otherwise, the entire  $T$  is disclosed directly. Instead, the objective of publication is to permit users to build a close approximation  $D^*(., .)$  of  $D(., .)$ .

Without loss of generality, assume that we need to release (the anonymized versions of)  $g$  marginals of  $T$ . Let  $T_1^*, T_2^*, \dots, T_g^*$  be the generalized tables for the  $g$  marginals, respectively. In the sequel, we will first explain how to construct the distribution  $D^*(., .)$  from  $T_1^*, \dots, T_g^*$ , and then discuss the privacy revealed by  $D^*(., .)$ .

**Calculation of  $D^*$ .** Let  $O$  be the set of individuals in  $T$ , whose QI- and sensitive values are unknown to us. Rebuilding  $T$  is equivalent to conjecturing, for each person  $o \in O$ , her/his QI-values  $o.Q$  and sensitive value  $o.X$ . The only clues available are:

- $o.Q$  is a point in the QI-space.

- The tuple of  $o$  has been generalized to a tuple in each of  $T_1^*, T_2^*, \dots, T_g^*$ , respectively.

We take a *random-world approach* to rebuild  $T$ . The idea is as follows. Since we have no further information about  $o.Q$ , it is reasonable to postulate that  $o.Q$  may be any point  $Q$  in the QI-space with an equal probability. Likewise, for each  $i \in [1, g]$ , it is fair to assume that  $o$  has been generalized to any tuple of  $T_i^*$  with the same chance. Of course, our conjectures must be consistent with the available clues. Specifically, let  $t_i^*$  be the tuple that we think is associated with  $o$  in  $T_i^*$ . Then, two conditions must hold. First,  $Q$  is indeed covered by the (generalized) QI-values of  $t_i^*$ . Second, all of  $t_1^*, t_2^*, \dots, t_g^*$  must have an identical sensitive value (because the sensitive value  $o.X$  of  $o$  never changes in any generalization). Finally, obviously we must prevent any two individuals from being associated with the same tuple in any  $T_i^*$ .

Once the QI- and sensitive values of everybody in  $O$  have been determined as above, we have found a table that could have been the real  $T$ . Such a table is called a *world*. More formally, a world is defined by  $g + 1$  functions  $q(\cdot), h_1(\cdot), \dots, h_g(\cdot)$ . Function  $q(\cdot)$  returns the point in the QI-space of an individual  $o \in O$ , namely,  $o.Q = q(o)$ . Function  $h_i(\cdot)$  ( $1 \leq i \leq g$ ), on the other hand, maps each  $o \in O$  to a distinct integer  $h_i(o)$  in  $[1, |T_i^*|]$ , implying the association of  $o$  with the  $h_i(o)$ -th tuple in  $T_i^*$ . Denote this tuple as  $t_i^*$ . Then, the QI-values of each  $t_i^*$  ( $1 \leq i \leq g$ ) must cover  $q(o)$ , and all the  $t_1^*, \dots, t_g^*$  must carry the same sensitive value.

**Example 4.** To illustrate the concept of worlds, assume that the microdata  $T$  is Table 2a, and we have released two marginals  $T_1^*$  and  $T_2^*$  as Tables 2b and 2c, respectively. Here,  $O$  equals  $\{Alan, Bob, Carrie, Daisy, Eddy, Frank\}$ . Consider the following function  $q(\cdot)$ , which maps each individual to a point in the QI-space:

$$\begin{aligned} q(Alan) &= (21, M, 10k), q(Bob) = (23, M, 58k), q(Carrie) = (58, F, 12k), \\ q(Daisy) &= (60, F, 60k), q(Eddy) = (70, M, 78k), q(Frank) = (72, M, 80k). \end{aligned}$$

Also, consider the function  $h_1(\cdot)$  below:

$$\begin{aligned} h_1(Alan) &= 1, h_1(Bob) = 2, h_1(Carrie) = 3 \\ h_1(Daisy) &= 4, h_1(Eddy) = 6, h_1(Frank) = 5. \end{aligned}$$

<i>Name</i>	<b>Age</b>	<b>Sex</b>	<b>Zip.</b>	<b>Disease</b>
<i>Alan</i>	21	M	10k	pneumonia
<i>Bob</i>	23	M	58k	flu
<i>Carrie</i>	58	F	12k	bronchitis
<i>Daisy</i>	60	F	60k	pneumonia
<i>Eddy</i>	70	M	78k	bronchitis
<i>Frank</i>	72	M	80k	flu

Table 8: A possible world built from the marginal publication in Table 2

For instance,  $h_1(\textit{Alan}) = 1$  indicates that we think Alan has been generalized to the first tuple in  $T_1^*$ . Finally, consider function  $h_2(\cdot)$  as:

$$h_2(\textit{Alan}) = 1, h_2(\textit{Bob}) = 3, h_2(\textit{Carrie}) = 2$$

$$h_2(\textit{Daisy}) = 4, h_2(\textit{Eddy}) = 6, h_2(\textit{Frank}) = 5.$$

There is no inconsistency among  $q(\cdot)$ ,  $h_1(\cdot)$  and  $h_2(\cdot)$ . For example, the  $h_1(\textit{Alan})$ -st tuple of Table 2b indeed has the same sensitive value as the  $h_2(\textit{Alan})$ -st tuple of Table 2c. Furthermore, the QI-values of both tuples indeed cover  $q(\textit{Alan})$ .

The functions  $q(\cdot)$ ,  $h_1(\cdot)$ , and  $h_2(\cdot)$  define a world, which is a table (see Table 8) that could have been a possible instance of the microdata. Note that the instance differs from the real microdata in the diseases of Eddy and Frank. ■

Remember that our goal is to build an approximate version  $D^*(\cdot, \cdot)$  of the actual pdf  $D(\cdot, \cdot)$  of the microdata  $T$ . For every point  $(Q, x)$  in the universe, we must decide a value for  $D^*(Q, x)$ . Assuming all worlds are equally likely, we have

$$D^*(Q, x) = \alpha(Q, x) / \alpha_{all}, \quad (4)$$

where  $\alpha(Q, x)$  is the number of worlds that associate an arbitrary individual  $o \in O$  with QI-values  $Q$  and sensitive value  $x$ , and  $\alpha_{all}$  is the total number of worlds.

**Privacy Risk.** Consider a linking attack where an adversary has the exact QI-values  $v.Q = Q$  of a victim  $v$ , and wants to derive a SA-distribution  $f(\cdot)$  for the sensitive value  $v.X$  of  $v$ . By studying  $T_1, \dots, T_g$ , s/he obtains a distribution  $D^*(\cdot, \cdot)$  as shown in Equation 4. Note that  $D^*(Q, x)$  is the joint probability that a person's QI-values are  $Q$ , and her/his sensitive value is  $x$ . On the other

hand,  $f(x)$  is the probability for a person to have sensitive value  $X$ , *given* that her/his QI-values are  $Q$ . Formally:

$$\begin{aligned} f(x) &= Pr[v.X = x | v.Q = Q] = \frac{Pr[v.x = x, v.Q = Q]}{\sum_{\forall x' \in X} Pr[v.x = x', v.Q = Q]} \\ &= \frac{D^*(v.Q, x)}{\sum_{\forall x' \in X} D^*(v.Q, x')}. \end{aligned} \quad (5)$$

Therefore, a marginal publication qualifies a principle  $\mathcal{P}$ , if and only if the  $f(\cdot)$  (calculated as above) of every person  $v$  fulfills  $\mathcal{P}$ .

It is worth mentioning that our derivation actually provides an alternative interpretation of the “maximum likelihood probability” in [20]. Our result, however, does not contradict the NP-hardness proved in [20]. In fact, the NP-hardness suggests that Equation 5 cannot be always computed in polynomial time, which excludes simple generalization as a practical approach for marginal publication.

#### 4.2.2 Angelization

Interestingly, we can leverage the above theory to analyze the privacy guarantee of ANGEL in marginal publication. To understand the intuition, consider the BT in Table 7a, and the two GTs in Tables 7b, 7c, respectively. Let us temporarily ignore the BT, and regard Tables 7b and 7c as two simple generalizations, and *Batch-ID* as the sensitive attribute. Thus, given the QI-values  $v.Q$  of any victim  $v$ , an adversary uses Equation 5 to compute a “SA”-distribution  $f'(\cdot)$ : for every integer  $i \in [1, 3]$  (where 3 is the number of batches; see Table 7a),  $f'(i)$  gives the probability that  $v$  belongs to batch  $B_i$ . Now, bring back the BT, and treat *Disease* as the real sensitive attribute. The adversary consults the BT and obtains the SA-distribution  $f_i(\cdot)$  in each  $B_i$  ( $1 \leq i \leq 3$ ). The final SA-distribution  $f(\cdot)$  of  $v$  equals  $\sum_{i=1}^3 (f'(i) \cdot f_i)$ , where  $f_i$  ( $1 \leq i \leq 3$ ) is the  $m$ -dimensional vector format of  $f_i(\cdot)$ , and  $m$  is the domain size of *Disease*.

Extending the previous discussion to the general case is straightforward. A linking attack derives the SA-distribution  $f(\cdot)$  of a victim  $v$  in two steps. In the first step, the adversary examines all the GTs, and applies Equation 5 to derive  $f'(\cdot)$ , which captures the probability that  $v$  falls in each batch  $B_i$  ( $1 \leq i \leq b$ , where  $b$  is the number of batches). Then, in the second step, the adversary

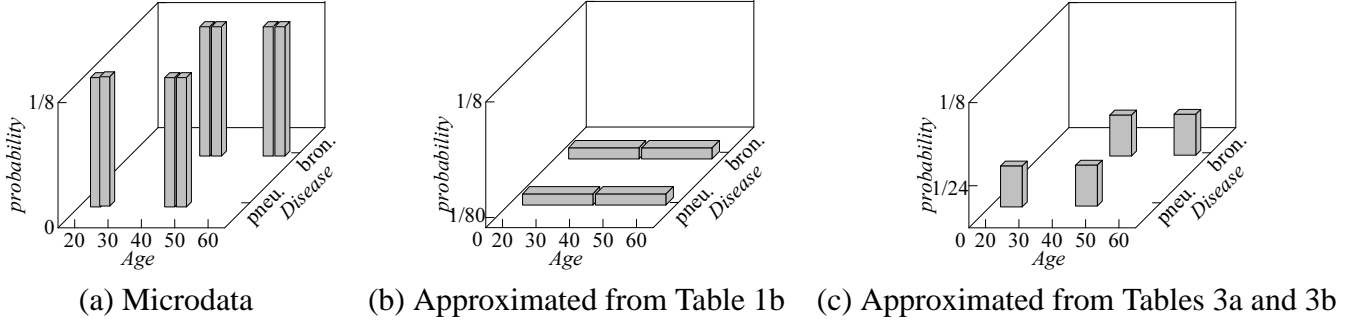


Figure 3: The original and approximated data distributions

computes

$$f = c_1 \cdot f_1 + c_2 \cdot f_2 + \dots + c_b \cdot f_b \quad (6)$$

where  $c_i$  equals  $f'(i)$ , and  $f_i$  is the  $m$ -dimensional vector format of the SA-distribution  $f_i(\cdot)$  in batch  $B_i$  ( $1 \leq i \leq b$ ). Remember that  $f_i(\cdot)$  can be obtained from BT.

As mentioned earlier, the exact computation of  $f'(i)$  in the above equation may not be settled in polynomial time. However, it does not matter — *the final  $f$  always qualifies the underlying anonymization principle  $\mathcal{P}$ , regardless of the values of  $f'(i)$* . We establish this guarantee with Lemma 2 and Theorem 1.

**Lemma 2.** *For any  $i \in [1, b]$ ,  $f'(i)$  is a rational value.*

*Proof.* It suffices to prove that Equation 5 returns only rational values. This is true because, for any  $Q$  and  $x$ ,  $D^*(Q, x)$  is rational — it is the ratio of integers  $\alpha(Q, x)$  and  $\alpha_{all}$  as shown in Equation 4.

□

**Theorem 1.** *If  $c_1, \dots, c_b$  are rational and  $\mathcal{P}$  is monotonic, the SA-distribution  $f(\cdot)$ , as calculated in Equation 6, always satisfies  $\mathcal{P}$ .*

*Proof.* We aim at constructing a set  $S$  of multi-sets of sensitive values such that (i) the SA-distribution of every multi-set qualifies  $\mathcal{P}$ , and (ii) the union of all multi-sets has exactly the SA-distribution  $f(\cdot)$ . Once such an  $S$  is found, the correctness of the theorem follows the monotonicity of  $\mathcal{P}$ .

Since  $c_i$  ( $1 \leq i \leq b$ ) is rational, it can be represented as  $\alpha_i/\beta_i$ , where  $\alpha_i$  and  $\beta_i$  are integers. Let  $S_i$  be the multi-set of sensitive values in  $B_i$ . Use  $c$  to denote the least common multiple of  $|S_1|, |S_2|, \dots, |S_b|$ , and  $\beta$  the least common multiple of  $\beta_1, \beta_2, \dots, \beta_b$ . We build an  $S$  as follows. Initially,  $S$  is empty. Then, for each  $S_i$  ( $1 \leq i \leq b$ ), we add  $\alpha_i \cdot (\beta/\beta_i) \cdot (c/|S_i|)$  copies of it to  $S$ . Since, by Definition 6, the SA-distribution  $f_i(\cdot)$  of every batch  $B_i$  ( $1 \leq i \leq b$ ) satisfies  $\mathcal{P}$ , the resulting  $S$  fulfills the conditions (i) and (ii) stated at the beginning of the proof.  $\square$

## 5 Data Analysis

Anonymized data is useless, unless it allows researchers to understand the distribution in the original microdata  $T$ . In fact, every generalization governed by the E-M modeling of Definition 3 allows an analyst to reconstruct an approximate version of the original microdata distribution. In the sequel, we explain the reconstruction, and demonstrate the superiority of angelization.

**Data Reconstruction.** Again, let  $T$  be a microdata table with  $d$  QI-attributes and a sensitive attribute  $X$ . We will continue using the concepts of *QI-space* and *universe* defined in Section 4.2. Recall that the distribution of  $T$  can be captured by a  $(d + 1)$ -dimensional pdf  $D$ .

Under the modeling of Definition 3, any generalization  $(E, M)$  decides a pdf  $D^*$  (which approximates  $D$ ) with a common mathematical form. The derivation leverages the property mentioned in Section 1 that the generalized QI-values of an EC  $C \in E$  can be regarded as a rectangle in the QI-space. We use  $R(C)$  to denote the rectangle, and  $|R(C)|$  to represent the number of points in the QI-space covered by  $R(C)$ .

To compute  $D^*(Q, x)$ , let  $C$  be the EC in  $E$  such that  $R(C)$  covers point  $Q$ . Under strict global-recoding, there is at most one such  $C$ . In case  $C$  does not exist,  $D^*(Q, x) = 0$ ; otherwise:

$$D^*(Q, x) = (|C|/|T|) \cdot (1/|R(C)|) \cdot f(x) \quad (7)$$

where  $f = M(C)$  is the SA-distribution associated with  $C$ .

The intuition of the formula is reflected in the following process of rebuilding a random tuple  $t$  in  $T$  from  $(E, M)$ . The probability of  $t$  belonging to a particular EC  $C$  equals  $|C|/|T|$ . Next, under the condition that  $t$  falls in  $C$ , we proceed to conjecture its QI-values, namely a point  $t.Q$  in the QI-

space, and its sensitive value  $t.X$ . Given only the information in  $(E, M)$ , we accept the modeling that  $t.Q$  coincides with a point  $Q$  in  $R(C)$  with probability  $1/|R(C)|$ , and independently,  $t.X$  takes a sensitive value  $x$  with likelihood  $f(x)$ . Thus, there is a probability of  $(|C|/|T|) \cdot (1/|R(C)|) \cdot f(x)$  for the process to yield a reconstructed tuple with  $t.Q = Q$  and  $t.X = x$ .

The previous discussion focuses on distribution reconstruction concerning the entire universe. In practice, researchers may wish to study the correlation between a subset of the QI-dimensions and the sensitive attribute  $X$ . The above analysis can be easily adapted to support this case. Equations 3 and 7 still apply, but with a slightly different interpretation of  $Q$  and  $R(C)$ . Specifically,  $Q$  should be understood as a point in the subspace of the QI-space that involves only the QI-attributes of interest;  $R(C)$  becomes the rectangle of  $C$  in this subspace.

**Example 5.** Figure 3a presents the distribution of the microdata in Table 1a, regarding the QI-attribute *Age* and the sensitive attribute *Disease*. Figure 3b shows the distribution approximated from the simple generalization Table 1b. Figure 3c demonstrates the distribution reconstructed from Tables 3a and 3b released by ANGEL. ■

**Probabilistic Counting.** A counting query has the form: *select count(\*) where  $\sigma$* , where  $\sigma$  can be an arbitrary predicate. Such queries play a significant role in data analysis. Particularly, in addition to being a useful stand-alone operation for OLAP applications, they are also the building brick of complex data mining tasks (e.g., association rule mining, decision tree construction, etc.).

Equation 7 provides a simple approach to derive a probabilistic answer for a counting query. Specifically, we first identify all the QI-attributes relevant to  $\sigma$ , which form a subspace of the QI-space. Then, we scrutinize the combination of every point  $Q$  in that subspace and every sensitive value  $x$ , and collect the set  $S$  of all the pairs  $(Q, x)$  that qualify  $\sigma$ . Then, the query answer equals

$$\sum_{\forall(Q,x) \in S} D^*(Q, x) \quad (8)$$

where  $D^*$  is given in Equation 7.

**Example 6.** We explain Formula 8 using the query: *select count(\*) where  $Age \in [35, 45]$  and  $Disease = pneumonia$* , which has an answer of 2 on the microdata Table 1a. The set  $S$  in Formula 8

contains 11 pairs:  $(35, pneumonia)$ ,  $(36, pneumonia)$ , ...,  $(45, pneumonia)$ . If Table 1b is employed to process the query, Formula 8 evaluates to 0.14, based on the distribution in Figure 3b. On the other hand, if we deploy Tables 3a and 3b (published by ANGEL), Formula 8 yields 2 by the distribution in Figure 3b, which coincides with the actual result. ■

**Marginal Publicaiton.** Given a set of anonymized marginals output by angelization (as described in Section 4.1), data analysis directly follows the above discussion, after the analyst has decided the GT to use. (This GT, together with the BT, leads to an approximate distribution in the subspace involving only the QI-attributes of the GT and the sensitive attribute.) A reasonable choice is to deploy the GT that contains the least QI-attributes irrelevant to the goal of analysis. For example, in Example 3, if the objective is to study the correlation between *Zipcode* and *Disease*, Table 7b should be employed, whereas Table 7c ought to be selected, if any other QI-attribute is included in the study.

## 6 Experiments

This section experimentally evaluates the effectiveness of angelization. We use a real dataset OCC downloadable at <http://ipums.org>, which contains 300k tuples each describing the personal information of an American. OCC contains 6 attributes *Marital-status*, *Work-class*, *Education*, *Gender*, *Age*, and *Occupation*, whose domain sizes are 6, 10, 17, 2, 78, and 50, respectively. *Occupation* is the sensitive attribute, whereas the others are QI-attributes.

We employ the Mondrian algorithm [23] to implement angelization, i.e.,  $AL_{\mathcal{P}} = AL_k = \text{Mondrian}$ , where  $AL_{\mathcal{P}}$  ( $AL_k$ ) is the algorithm for obtaining a batch (bucket) partitioning, as explained in Section 3.4. On each QI-attribute, the generalized value of a bucket equals the minimum bounding interval of the values of the tuples in the bucket. In the sequel,  $k$  is fixed to 10.

**Data Reconstruction Error.** The first experiment studies the accuracy of the data distribution reconstructed by angelization (in the way described in Section 5). Note that the distribution involves all the QI- and sensitive attributes; hence, the accuracy indicates how well angelization captures the correlation in the original microdata. We measure the reconstruction error as the KL-divergence between the reconstructed distribution and the original distribution. KL-divergence is

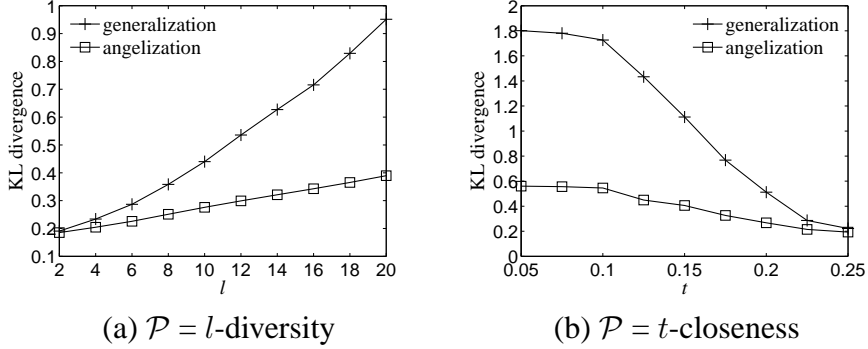


Figure 4: Data reconstruction error

a popular metric [20, 25] for quantifying the discrepancy of two distributions. For comparison, we also gauge the error of the distribution rebuilt from conventional generalization (also obtained with Mondrian).

Figure 4a (4b) demonstrates the KL-divergence (in the entire 6-dimensional universe) as a function of  $l$  ( $t$ ), when  $l$ -diversity ( $t$ -closeness) is the underlying anonymization principle. Angelization incurs significantly lower error than generalization, and their gap increases, when more stringent privacy protection is enforced (i.e., with a larger  $l$  or a smaller  $t$ ). This is expected, because tighter privacy control leaves conventional generalization with fewer choices (of permissible generalizations), whereas the influence on angelization is much smaller, as its generalization does not need to obey  $\mathcal{P}$  (recall that, as explained in Section 3.4, the generalized values in the GT of angelization are obtained from a  $k$ -anonymous simple generalization).

**Utility for Probabilistic Counting.** Next, we study the utility of angelized data in concrete analytical operations. For this purpose, following the practice of [40], we use probabilistic counting as the representative operation. Specifically, each query has the form: SELECT COUNT(\*) FROM OCC WHERE  $A_1 \in S_1$  AND  $A_2 \in S_2$  AND ... AND  $A_6 \in S_6$ . Here,  $A_i$  ( $1 \leq i \leq 6$ ) refers to the  $i$ -th attribute of OCC, and  $S_i$  includes a set of values in the domain of  $A_i$ . Each query has a parameter  $s \in [0, 1]$ , called *volume*, which decides the cardinality of  $S_i$  as  $\lceil |A_i| \cdot s^{1/6} \rceil$ . If  $A_i$  is numerical (i.e., *Age* and *Education*),  $S_i$  is a continuous interval; otherwise,  $S_i$  contains  $|S_i|$  random values in the domain of  $A_i$ . A workload contains 1000 queries with the same  $s$ . For both angelization and generalization, the estimated answer is given by Equation 8. For each technique, we measure its

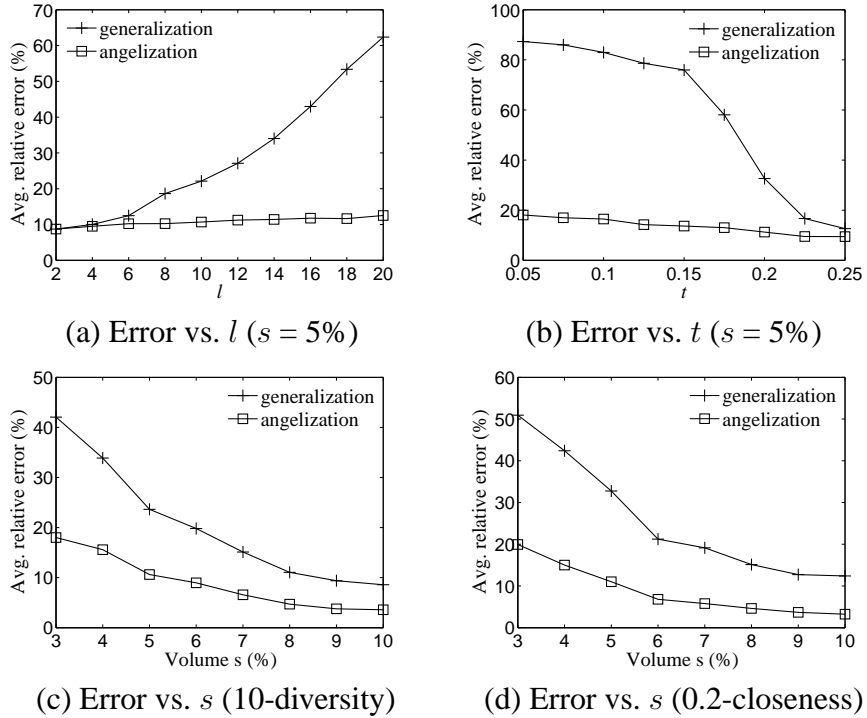


Figure 5: Accuracy in probabilistic counting

average relative error<sup>4</sup> in processing all the queries in a workload.

Fixing the volume  $s$  to 5%, Figure 5a (5b) plots the average error as a function of  $l$  ( $t$ ), for  $\mathcal{P} = l$ -diversity ( $t$ -closeness). Angelization achieves fairly high accuracy in all cases, whereas the error of generalization increases sharply with  $l$  and  $t$ . These results confirm the phenomena in Figure 4. In Figure 5c (5d), we use 10-diversity (0.2-closeness) as the anonymization principle, and inspect the query error as the volume varies from 3% to 10%. Both methods benefit from a higher volume, since in general probabilistic counting is easier when query results are larger. Angelization, again, consistently outperforms its competitor.

**Marginal Publication.** The last set of experiments demonstrates the advantages of marginal publication with respect to releasing only a single table. Towards this purpose, we examine four marginals  $G_1, G_2, \dots, G_4$ , whose dimensionalities are 2, 3, ..., 5, respectively. Specifically,  $G_1$  includes attributes *Marital-status* and *Occupation*.  $G_2$  contains another attribute *Work-class*,  $G_3$  yet another *Education*, and  $G_4$  still another *Gender*. Given an anonymization principle  $\mathcal{P}$ , we prepare

<sup>4</sup>Let  $act$  and  $est$  be the actual and estimated answers respectively. Then the relative error equals  $|est - act|/act$ .

Marginal	$G_1$	$G_2$	$G_3$	$G_4$
Error using GT	0.25	2.90	6.18	8.96
Error using $GT_i$	0.03	1.13	3.71	6.06

(a) KL-divergence unit  $10^{-2}$ ,  $\mathcal{P} = 10$ -diversity

Marginal	$G_1$	$G_2$	$G_3$	$G_4$
Error using GT	0.30	3.20	6.79	9.31
Error using $GT_i$	0.08	1.38	4.20	6.22

(b) KL-divergence unit  $10^{-2}$ ,  $\mathcal{P} = 0.2$ -closeness

Table 9: Reconstruction error of marginals

a marginal publication including a BT and 5 group tables:  $GT, GT_1, GT_2, \dots, GT_4$ , where  $GT$  is for all the QI attributes, and  $GT_i$  for  $G_i$  ( $1 \leq i \leq 4$ ). Now, for each  $G_i$  ( $1 \leq i \leq 4$ ), we compare the errors of the distributions (in the subspace decided by  $G_i$ ) reconstructed from (i) BT and  $GT$ , and (ii) BT and  $GT_i$ . A larger difference between the two errors indicates more significant benefits from publishing  $GT_i$ .

Table 9a (9b) illustrates the comparison results when  $\mathcal{P}$  is 10-diversity (0.2-closeness). In all cases, releasing marginals always reduces reconstruction error. Furthermore, the improvement becomes more obvious when a marginal has a lower dimensionality. This is expected, because generalization in a low-dimensional subspace incurs much smaller information loss, compared to generalization in the original universe.

## 7 Related Work

This section surveys the previous work on privacy preserving publication. We will first discuss the existing anonymization principles, and then review the known generalization algorithms. Finally, we will briefly cover alternative anonymization methodologies and other areas related to data privacy.

**Anonymization principles.** Privacy protection must take into account the knowledge of adversaries. A common assumption is that an adversary has the precise QI-values of all individuals in the microdata. Indeed, these values can be obtained, for example, by knowing a person or consulting an external source such as a voter registration list [34].

Under this assumption,  $k$ -anonymity,  $l$ -diversity, and  $t$ -closeness (elaborated in Section 2) aim at

preventing the accurate inference of individuals’ sensitive values. Many other principles share this objective.  $(\alpha, k)$ -anonymity [39] combines the previous two principles: each QI-group must have size  $k$  and at most  $\alpha$  percent of its tuples can have the same sensitive value.  $m$ -invariance [42] is a stricter version of  $l$ -diversity, by dictating each group to have exactly  $m$  tuples with different sensitive values. The *personalized approach* [41] allows each individual to specify her/his own degree of privacy preservation. The above principles deal with categorical sensitive attributes, whereas  $(k, e)$ -anonymity [45] supports numerical ones.  $(k, e)$ -anonymity demands that each QI-group should have size at least  $k$ , and the largest and smallest sensitive values in a group must differ by at least  $e$ .

$\delta$ -presence [29] assumes the same background knowledge as the earlier principles, but ensures a different type of privacy. It prevents an adversary from knowing whether an individual has a record in the microdata (i.e., a presence attack mentioned in Section 2).  $(c, k)$ -safety [26] tackles stronger background knowledge. In addition to individuals’ QI values, an adversary may have several pieces of *implicational knowledge*: “if person  $o_1$  has sensitive value  $v_1$ , then another person  $o_2$  has sensitive value  $v_2$ ”.  $(c, k)$ -safety guarantees that, even if an adversary has  $k$  pieces of such knowledge, no individual’s sensitive value can be disclosed with probability higher than  $c$ . Achieving a similar purpose, the *skyline privacy* [11] guards against an extra type of knowledge. Namely, an adversary may have already known the sensitive values of some individuals before inspecting the published contents.

**Generalization algorithms.** Numerous heuristic algorithms have been developed to compute generalization with small information loss. These algorithms are general, since they can be applied to many of the anonymization principles reviewed earlier. Specifically, a genetic algorithm is developed in [18], and the branch-and-bound paradigm is employed on a set-enumeration tree in [7]. Top-down and bottom-up algorithms are presented in [15, 43]. *Incognito* [22] borrows ideas from frequent item set mining, while *Mondrian* [23] takes a partitioning approach reminiscent of kd-trees. In [16], space filling curves are leveraged to facilitate generalization, and the work of [17] draws an analogy between spatial indexing and generalization. The above approaches minimize a generic metric of information loss, whereas a *workload-aware* method [21] uses a representative workload supplied by users. *Sequential publication* is addressed in [37]. As shown in [38],

the previous algorithms may suffer from *minimality attacks*, which can be avoided by introducing some randomization.

The algorithms mentioned earlier work well on practical datasets, but do not have attractive asymptotical performance in the worst case. This motivates studies on the theoretical aspects of the anonymization problem. Interestingly, all the known theoretical results focus on  $k$ -anonymity. Meyerson and Williams [27] are the first to prove the NP-hardness of optimal  $k$ -anonymous generalization, and give an  $O(k \log k)$ -approximation algorithm. Aggarwal et al. [4] reduce the approximation ratio to  $O(k)$ , which is further improved to  $O(\log k)$  by Park and Shim [30]. Unlike these solutions whose approximation ratios are functions of  $k$ , Du et al. [12] present a method having a ratio  $O(d)$ , where  $d$  is the number of attributes in the QID. Aggarwal et al. [3] develop constant approximation algorithms.

**Further Research on Privacy.** So far we have focused on generalization, while anonymized publication can also be achieved by other methodologies. Xiao and Tao [40] advocate *anatomy* that has been explained in Section 3.5. Aggarwal and Yu [2] design the *condensation method*, which releases only selected statistics about each QI-group. Rastogi et al. [31] employ *perturbation*. Finally, besides data publication, anonymity issues arise in many other environments. Some examples include anonymized surveying [6, 14], statistical databases [10, 13, 28], cryptographic computing [19, 32, 36], access control [5, 8, 9], and so on.

## 8 Conclusions

This paper proposes angelization as a new anonymization technique for privacy preserving publication, which is applicable to any monotonic anonymization principle. Angelization subsumes traditional generalization as a special case. It produces an anonymized relation that achieves the same privacy guarantee (as conventional generalization), but permits a much more accurate reconstruction of the original data distribution. Furthermore, angelization offers a simple and rigorous solution to anonymized marginal publication, which was previously a difficult issue with conventional generalization.

This work also initiates several directions for future work. Recall that ANGEL utilizes the exist-

ing algorithms of simple generalization to perform angelization. It would be interesting to study whether it is possible to obtain better angelization directly, without resorting to simple generalization. Furthermore, in this paper, we have considered only static microdata that do not need to be updated. In practice, there may be a need to publish another version of the microdata after it has received sufficient insertions and/or deletions [42]. Extending our technique to such a republication scenario is an exciting topic. Finally, it would be challenging to investigate how to employ the distribution reconstructed from angelization (in the way described Section 5) to perform advanced data mining such as decision tree classification, association rule mining, etc.

## Acknowledgements

Yufei Tao was partially supported by GRF grants CUHK 4173/08, CUHK 4161/07, and CUHK 1202/06 from HKRGC. Shuigeng Zhou was supported by National Natural Science Foundation of China under grant No. 60873070 and Shanghai Leading Academic Discipline Project No. B114. Donghui Zhang was partially supported by NSF CAREER Award IIS-0347600.

## References

- [1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.
- [2] C. C. Aggarwal and P. S. Yu. A condensation approach to privacy preserving data mining. In *EDBT*, pages 183–199, 2004.
- [3] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *PODS*, pages 153–162, 2006.
- [4] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, pages 246–258, 2005.
- [5] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. Hippocratic databases. In *VLDB*, pages 143–154, 2002.
- [6] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *SIGMOD*, pages 439–450, 2000.
- [7] R. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE*, pages 217–228, 2005.
- [8] E. Bertino, C. Bettini, E. Ferrari, and P. Samarati. An access control model supporting periodicity constraints and temporal reasoning. *TODS*, 23(3):231–285, 1998.
- [9] E. Bertino and E. Ferrari. Secure and selective dissemination of xml documents. *ACM Trans. Inf. Syst. Secur.*, 5(3):290–331, 2002.
- [10] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, pages 128–138, 2005.

- [11] B.-C. Chen, R. Ramakrishnan, and K. LeFevre. Privacy skyline: Privacy with multidimensional adversarial knowledge. In *VLDB*, pages 770–781, 2007.
- [12] Y. Du, T. Xia, Y. Tao, D. Zhang, and F. Zhu. On multidimensional  $k$ -anonymity with local recoding generalization. In *ICDE*, pages 1422–1424, 2007.
- [13] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.
- [14] A. V. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222, 2003.
- [15] B. C. M. Fung, K. Wang, and P. S. Yu. Top-down specialization for information and privacy preservation. In *ICDE*, pages 205–216, 2005.
- [16] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis. Fast data anonymization with low information loss. In *VLDB*, pages 758–769, 2007.
- [17] T. Iwuchukwu and J. F. Naughton.  $K$ -anonymization as spatial indexing: Toward scalable and incremental anonymization. In *VLDB*, pages 746–757, 2007.
- [18] V. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD*, pages 279–288, 2002.
- [19] W. Jiang and C. Clifton. A secure distributed framework for achieving  $k$ -anonymity. *The VLDB Journal*, 15(4):316–333, 2006.
- [20] D. Kifer and J. Gehrke. Injecting utility into anonymized datasets. In *SIGMOD*, pages 217–228, 2006.
- [21] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. In *SIGKDD*, 2006.
- [22] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain  $k$ -anonymity. In *SIGMOD*, pages 49–60, 2005.
- [23] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional  $k$ -anonymity. In *ICDE*, pages 277–286, 2006.
- [24] N. Li, T. Li, and S. Venkatasubramanian.  $t$ -closeness: Privacy beyond  $k$ -anonymity and  $l$ -diversity. In *ICDE*, pages 106–115, 2007.
- [25] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkatasubramanian.  $l$ -diversity: Privacy beyond  $k$ -anonymity. In *ICDE*, page 24, 2006.
- [26] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. Worst-case background knowledge in privacy. In *ICDE*, 2007.
- [27] A. Meyerson and R. Williams. On the complexity of optimal  $k$ -anonymity. In *PODS*, pages 223–228, 2004.
- [28] S. U. Nabar, B. Marthi, K. Kenthapadi, N. Mishra, and R. Motwani. Towards robustness in query auditing. In *VLDB*, pages 151–162, 2006.
- [29] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, pages 665–676, 2007.
- [30] H. Park and K. Shim. Approximate algorithms for  $k$ -anonymity. In *SIGMOD*, pages 67–78, 2007.

- [31] V. Rastogi, S. Hong, and D. Suciu. The boundary between privacy and utility in data publishing. In *VLDB*, pages 531–542, 2007.
- [32] J. Rothe. Some facets of complexity theory and cryptography: A five-lecture tutorial. *ACM Computing Surveys*, 34(4):504–549, 2002.
- [33] P. Samarati. Protecting respondents’ identities in microdata release. *TKDE*, 13(6):1010–1027, 2001.
- [34] L. Sweeney. Achieving  $k$ -anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):571–588, 2002.
- [35] L. Sweeney.  $k$ -anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems*, 10(5):557–570, 2002.
- [36] J. Vaidya and C. Clifton. Privacy-preserving  $k$ -means clustering over vertically partitioned data. In *SIGKDD*, pages 206–215, 2003.
- [37] K. Wang and B. C. M. Fung. Anonymizing sequential releases. In *SIGKDD*, pages 414–423, 2006.
- [38] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, pages 543–554, 2007.
- [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 *SIGKDD*, pages 754–759, 2006.
- [40] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *VLDB*, pages 139–150, 2006.
- [41] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, pages 229–240, 2006.
- [42] X. Xiao and Y. Tao.  $m$ -invariance: towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, pages 689–700, 2007.
- [43] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In *SIGKDD*, pages 785–790, 2006.
- [44] C. Yao, X. S. Wang, and S. Jajodia. Checking for  $k$ -anonymity violation by views. In *VLDB*, pages 910–921, 2005.
- [45] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *ICDE*, pages 116–125, 2007.